Available online at www.sciencedirect.com
ScienceDirect

Journal of Molecular Graphics and Modelling 26 (2007) 656–666

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: Adding flexibility to the search for ligand kin

Andrew C. Good *

Bristol-Myers Squibb, PO Box 5100, Wallingford, CT 06492, USA

Accepted 28 March 2007
Available online 4 April 2007

Abstract

With readily available CPU power and copious disk storage, it is now possible to undertake rapid comparison of 3D properties derived from explicit ligand overlay experiments. With this in mind, shape software tools originally devised in the 1990s are revisited, modified and applied to the problem of ligand database shape comparison. The utility of Connolly surface data is highlighted using the program MAKESITE, which leverages surface normal data to a create ligand shape cast. This cast is applied directly within DOCK, allowing the program to be used unmodified as a shape searching tool. In addition, DOCK has undergone multiple modifications to create a dedicated ligand shape comparison tool KIN. Scoring has been altered to incorporate the original incarnation of Gaussian function derived shape description based on STO-3G atomic electron density. In addition, a tabu-like search refinement has been added to increase search speed by removing redundant starting orientations produced during clique matching. The ability to use exclusion regions, again based on Gaussian shape overlap, has also been integrated into the scoring function. The use of both DOCK with MAKESITE and KIN in database screening mode is illustrated using a published ligand shape virtual screening template. The advantages of using a clique-driven search paradigm are highlighted, including shape optimization within a pharmacophore constrained framework, and easy incorporation of additional scoring function modifications. The potential for further development of such methods is also discussed.

Keywords: Molecular similarity; Gaussian functions; DOCK; Shape; Pharmacophores

1. Introduction

Long standing efforts dating back to the early 1990s abound for the quantitative comparison of ligand shape. Due to limitations in CPU power, initial research often focused on rapidly comparable shape-based and pharmacophoric fingerprints and data prints [1–3]. Some techniques also coalesced multiple conformers for a given ligand into a single descriptor in order to conserve restricted hard disk storage [4]. While such techniques are still being developed [5–9], the lower levels of property approximation and easy visualization of comparisons based on explicit molecular alignment are beginning to gain more attention. In the early years of such calculations, continuous molecular properties were often approximated through the application of rectilinear grids or quantum chemical properties [10–12]. Innovative efforts in the development of methods fast enough to allow conformer database searching were pioneered by Moon and Howe [13] (atom distance mapping) and van Geerestein and co-workers [14,15] (gnomonic projection). These and other techniques [16–18] were generally restricted in their utility due to the hardware limitations of the last century, particularly in the context of explicit conformer generation and storage. The CAMD community was thus typically restricted to using pharmacophore mapping [19–24] for the majority of its explicit molecular alignment calculations.

Continuing improvements in CPU power and breakthroughs in disk storage have led to renewed focus in the area of molecular similarity calculations based on explicit alignment. Particular emphasis has been placed on the application of Gaussian function property approximations. The numerical foundations that underpin grid and distance based techniques

* Tel.: +1 203 677 6761; fax: +1 203 677 7702.
E-mail address: andrew.good@bms.com.

1093-3263/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2007.03.016

such as those alluded to above lead to noisy similarity hyper surfaces, rendering alignment optimization prone to premature convergence. In contrast, properties approximated by rapidly calculable Gaussian functions whose similarities are evaluated using product-based numerator terms (e.g. Tanimoto index) can be optimized analytically. The Richard's group pioneered this use of Gaussians, devising functions for both molecular electrostatic potential (MEP) (inverse distance approximation) [25] and shape (STO-3G derived atomic electron density) [26]. Willett and co-workers [27,28] applied the MEP Gaussian approach using genetic algorithms, MEP field graphs and bit climbers to allow the rapid comparison of MEPs for virtual screening. Grant et al. [29] extended the application of Gaussians to shape using an elegant hard sphere model approximation. This methodology has been made commercially available in the ROCS program [30], a rapid shape alignment tool that also permits atom coloring, allowing Gaussian interactions to be weighted based on atom type. The technique has been shown to have the potential to enrich actives at the top of its hit lists [31,32], providing a degree of validation for the approach.

With renewed interest in such approaches coming from both from the general and internal CAMD community, the author decided to revisit the issue of ligand-based shape alignment. One of the major issues to overcome when undertaking such searches is the ability to sample search space rapidly enough to determine the global function minimum. ROCS achieves this by mapping target molecules to four possible ‘‘inertial’’ starting points derived from Gaussian shape moments [33]. While this forms an economic solution, it also ties the optimization more tightly into the shape of the targets. For shape only searches this is fine, but it can be argued that mapping key pharmacophore elements is at least as (and often more) important to activity. Gaussian ‘‘coloring’’ by binding interaction type provides some access to this information [34]. Such a whole molecule measure has difficulty deconvoluting key binding elements from other extraneous functionality, however. Further, undertaking a separate pharmacophore filter, while useful, loses the reference frame with respect to ligand shape, thus limiting filter resolution. Given the long history and undoubted utility of pharmacophore-based mapping [23,24], the ability to measure shape within such a framework appears particularly attractive. With these points in mind and the aim of incorporating a scoring function derived from more than just shape, a clique driven approach to initial superposition has been pursued.

The clique searching engine within DOCK is a powerful yet under utilized feature of the program that has found extensive use within The CAMD department at BMS [35]. DOCK allows each atom in a template to be explicitly assigned to a specific chemical type [36], with further categorization to critical points or regions also possible [37]. In addition the user can control exactly which chemical types are assigned to each search template point (or atom when running in ligand-shape mode—see below). Combining these two features allows exquisite user control not only over what chemical types exist in a given template but also which atoms or groups of atoms must be mapped. In addition, applying critical regions without chemical typing also permits constrained generic shape searching. DOCK is by default designed to handle ligand–protein interactions. A cursory analysis of the code (which comes with the binary distribution) suggested modifications could be undertaken with relative ease to handle ligand-based comparisons. In addition, the supplementary program MAKEPOINT [35] originally used in site point generation could be inverted to create a shape cast of a ligand rather than site points within a protein active site (MAKESITE). This would permit DOCK to handle ligand-based shape searching without code modifications, allowing shape comparison based on surface as opposed to volume properties. The DOCK code modifications, MAKESITE methodology and their application are detailed below.

2. Methods

2.1. Search test data

To illustrate the behavior of KIN and DOCK running with MAKESITE, the PART shape template used in a ROCS search for small molecule inhibitors of the antibacterial target ZipA-FtsZ was chosen [31]. The query is shown in Fig. 1, together with two hits obtained in the original study. This query was chosen to allow some sort of comparison with ROCS search results. It also has the advantage of being rigid in nature, allowing a relatively simple reproduction of shape via visual reconstruction based on the conformation depicted in the paper. The template used in these studies has been built in MAESTRO, with subsequent minimization in the MACRO-

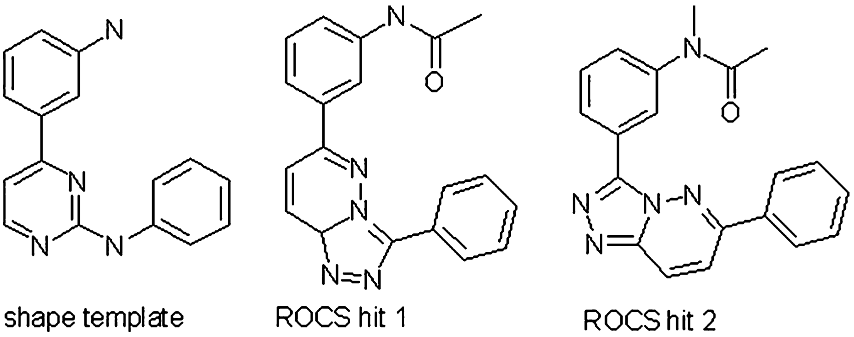

Fig. 1. Shape template and hits extracted from ROCS study [31].

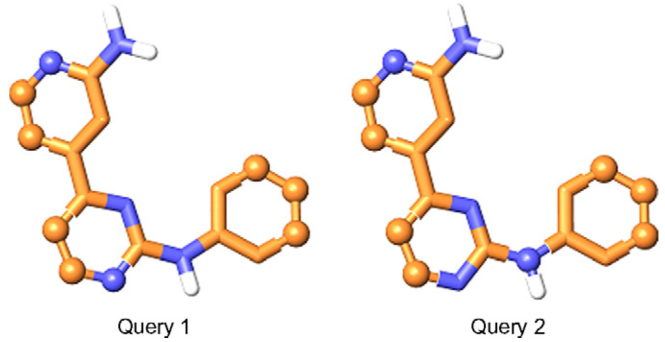

Fig. 2. DOCK search templates used in tests. Each cluster of atoms marked as spheres were defined as individual critical regions (one atom from each region has to be mapped, together with one of the remaining atoms in the template to form a matching clique). For the second search the critical atom donor NH was also constrained by chemistry to allow only donor atom mappings for a valid clique match. The remaining atoms were permitted to match any chemistry type except donor.

MODEL minimizer using the MMFFs force field [38]. The two ROCS hits shown in Fig. 1 were generated in the same fashion. This force field was chosen to maintain compatibility with the conformer database generated using OMEGA [39]. OMEGA was run in default mode on the 59864 ZINC database compounds of the 0.8 Tanimoto lead-like cluster set to generate said database [40,41] (1,840,485 conformers in total generated). OMEGA was also applied to the two ROCS hits to create a hit conformer data set comprising 56 conformers.

Two primary search variants were undertaken. In the first the search template was set up with three critical regions that map the edges of the template. Chemistry matching was turned off so that the resulting clique constraints were purely geometric in nature (Fig. 2). Based on DOCK critical region search criteria for the four node searches undertaken, one atom from each of the regions plus one of the remaining atoms must exhibit matching distances to the ligand for the cliques to match. In the second search a five node query was set, with three geometric regions and one chemically mapped atom (the donor) defined as critical (again one of the remaining atoms defines the element of the clique search criteria).

2.2. MAKESITE

MAKESITE finds it foundations in the Connolly surface program MS [42,43]. The software has been modified so that it forms a subroutine in other programs, taking mol2 files as input and outputting surface coordinate data together with their associated surface normals (unit vectors directed back toward the original probe position used to generate the surface point). Each surface point is used as an origin and the surface normal vector applied to position carbon atoms exactly one van der Waals (vdW) radius distant from the surface. An iodine atom is also positioned along each vector one vdW radius distant from the carbon, defining an approximation for protein bulk. The resulting atom ensemble creates a ligand shape cast that is output as a mol2 file (site.mol2). This mol2 file can be read directly into DOCKs grid program and used in the creation of a contact map. Fig. 3 shows the MAKESITE application in action both schematically and by example. The contact score used by DOCK is a simple atom counting routine that adds up the no. receptor atoms within a prescribed distance range defining potentially attractive interactions [37]. Replacing the receptor with the site.mol2 shape cast turns the contact score into a local surface count score that can be used directly within DOCK to undertake a ligand search. In addition the atoms in site.mol2 define the potential bounds of the receptor, allowing them to be used as bump constraints. Since the shape cast is in mol2 format, the atoms can also be easily edited to modify constraints/shape in the context of any existing SAR. For the MAKESITE shape cast used in these calculations a surface density of 100 points/Å^2 was used with a probe radius of 1 Å. The exclusion spheres used to constrain some of the KIN searches detailed below were defined using the structure shown in Fig. 3c. Surface density was reduced to 2.0 points/Å^2 was used with a probe radius of 1 Å, creating a total of 173 exclusion Gaussians.

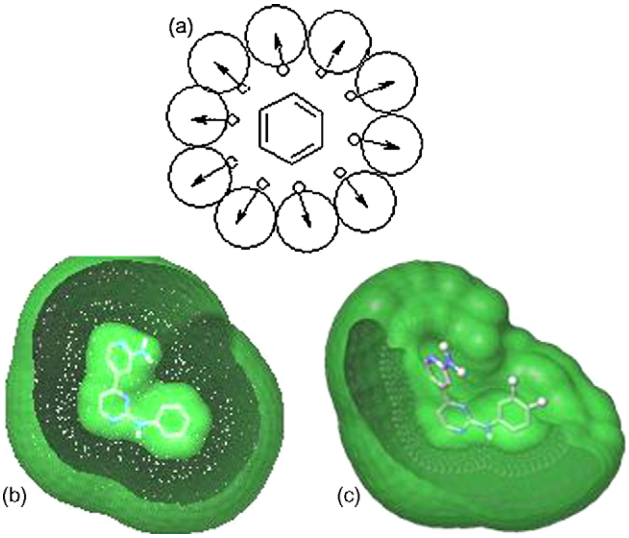

Fig. 3. (a) Schematic of how first atom shell is created using MAKESITE. Carbon atom is placed along the surface normal of each point, exactly one vdW radius from associated surface point. (b) Example surface atom ensemble and associated shape cast created for ROCS template. (c) Illustration of how the shape cast can be modified through atom deletion within the cast. Allows the user to remove cast atoms from regions with poorly defined shape SAR. In this case the additional carbon atoms shown as spheres had their atom names set to *open*. MAKESITE removes all surface points associated with these atoms from the output.

MAKESITE is available from the author on request for those users who already have a copy of the MS software.

2.3. DOCK

DOCK 4.0 code forms the basis of these studies. The default DOCK searching protocols utilizing MAKESITE shapes involve a slightly modified version that includes a function used to constrain critical atom matches. This is controlled through the addition of a parameter constrain_to_critical added to the dock.in file. This constraint is added as a weight in the fifth column of the sphere file used for template clique mapping. The penalty applied to the score for atom deviations is

Table 1
Donor definitions used in search extracted from in-house dock chem.defn file

Name	Donor definition
# generic	N. (H)
# imidazole-like tautomer	N.2 [H] (C.2 (N.pl3 (H)))
# pyrazole-like tautomer	N.2 [H] (C.2) (N.pl3 (H))
# conjugated anilinic	N.2 [H] (C.ar (C.ar (N.pl3 (H))))
# halogens/sulfur as donor	Cl (C.)
	Br (C.)
	I (C.)
	S.

Add dock manual Ref. [38].

equal to the weight times the distance between mapped atoms (in Å). In this way pharmacophore constraints can be applied simultaneously with shape similarity measurement. The correction has been added as a correction to the *calc_score* routine using an additional constraint_id term added to the molecule structure. This is set in the *extract_clique* function. DOCK can be run without this term using MAKESITE, but optimization will then only take shape similarity into account, allowing a potentially significant shift in the Pharmacophore mapping during shape optimization. To highlight the utility of running DOCK in shape only mode, this function has not been used for the DOCK searches. The function has been maintained and used in KIN, however, and for these pharmacophore constrained searches a weight of 0.1 has been applied.

In addition to this small code modification, we use an extensively modified version of the chem.defn chemical definitions file [37]. This has been extended to incorporate a number of different types (general/donor/acceptor/donor_acceptor/acid/base/aromatic/hydrophobic). While the language for this file is simple [37], it can handle >90% of the definitions one might want to include. The primary limitation we have found is the inability to define ring systems, which can on occasion lead to some ambiguity. Sample *chem.defn* definitions are shown in Table 1 for the donor definitions used in these studies.

2.4. KIN

The following function modifications were made to the source code of DOCK 4.0 to create KIN.

make_receptor_grid: The DOCK sphere file stores normally reserved for site points can instead be used to store ligand atom positions, chemical typing, pharmacophore constraint weightings and critical region definitions. The file does not provide a vehicle to pull in other potentially useful data such as exclusion volumes, however. To this end *make_receptor_grid* has been altered to keep coordinates, with all code relating to grid creation removed. This allows template exclusion volume data and ligands element typing data to be accessed through a mol2 file read using the receptor_atom_file parameter.

calc_pairwise_contact: modified to control shape similarity calculation. All functions passing data from the parent calc_score through to calc_pairwise_contact modified to disable grid calculations and ensure coordinate data passed from make_receptor_grid. Similarly all calls to grid reads disabled in get_anchor_score.

calc_overlap: function created to calculate shape similarity based on STO-3G derived electron density. Atomic radii defined in DOCK vdw.defn used to determine atom type overlap combinations. Each atom-pair overlap integral was originally defined based on a three Gaussian function approximation to the electron density [26]. This results in each overlap integral consisting of six Gaussian terms. To increase calculation speed for each atom pair combination, all Gaussian overlap integrals contributing <10% to the total integral have been removed form the function (one Gaussian typically dominates the overall overlap term). The function is called using *calc_pairwise_contact*. To provide an additional speed up in the calculation, Gaussian overlap is only calculated for those atoms in vdW contact with each other based on their inter-atomic distance separation.

In the original work undertaken by the Richards group in the field of Gaussian functions, shape similarity calculations from STO-3G derived electron density were generally found to produce results very similar to those obtained using a grid-based hard-sphere model [26,44]. QSAR models constructed using the same functions suggested that the hard sphere model offered additional discrimination for small differences seen in the congeneric series studied [45]. These 3D QSAR models involved careful manual conformation selection and molecule superposition. In contrast database searching must rely on automated overlay and target conformation generation. Given the approximations inherent in automation and the general aim of a shape screen to provide lead scaffold hops, the softer STO-3G derived Gaussian functions were deemed suitable for incorporation. It should be noted, however, that any choice of Gaussian function could be built into the same clique search framework with modest effort. The original fortran source used to derive the *calc_overlap* function and its electrostatic potential brethren is available on request.

In addition to the standard overlap function, additional excluded volume Gaussian code has been included in this function. The ligand template data read in through the *make_receptor_grid* function includes charge data. The code has been set to interpret all atoms with a point charge >9.99 as an exclusion atom the size of carbon (taken from MAKESITE data in the examples given below). The *contact_clash_penalty* value has also been hijacked to weight the exclusion sphere values. The resulting exclusion overlap values direct optimization through subtraction from the similarity numerator term.

Main: The nature of DOCK clique searches is such that many of the initial orientations sampled exhibit a high degree of overlap. As such the application of a tabu-like search constraint [46] was deemed to be worthwhile to reduce the level of clique searching redundancy. This is of potentially major impact for the Gaussian searches given the robust nature of their optimization. To accomplish this, an array of MOLECULE structures has been added to main. Once the first clique match is oriented the resulting molecule data is copied into the array using the copy_molecule function. All subsequent clique match orientations are compared to this (and any subsequent matches that pass the test) molecule copy using the calc_rmsd function. Only those orientations with a resulting rmsd value greater than the user defined tabu value (entered using an input variable tabu_rms added to the list of get_parameter function calls) are added to the array and passed to the similarity optimization routine.

2.5. Key DOCK/KIN input parameters

For the contact score grid derived from the MAKESITE shape cast, grid spacing was set to 1 Å, cutoff distance to 4 Å and bump overlap to 0.85 × the vdW radii. For all KIN searches a clique distance tolerance value of 0.75 Å was applied unless otherwise specified. This value controls the maximum permitted error for any clique atom-pair distance. 0.75 Å equates to approximately half a bond length and has been found empirically to provide a good balance between speed and exhaustive search space coverage (this is a particularly important control variable with a general rule of thumb that each 0.5 Å increase in the distance tolerance value decreases search speed by approximately an order of magnitude). The minimum clique atom pair distance was set to 2.5 Å for all KIN searches. The same setting was used for the non-bump constrained searches used in DOCK. Because of the noisy similarity surface of constrained searches, this value was reduced to 2.0 Å to permit additional sampling. Where bump matches were used up to six bumps were permitted before scoring was cancelled, and a contact clash penalty of 100 per atom bump was applied to the similarity score. For minimizations, only one cycle was used for the DOCK searches with 100 iterations allowed and initial translations and rotations set to 0.5 Å and 0.25 rad, respectively. The precipitous nature of the similarity surface for MAKESITE-based searches reduces the minimizer to more of a local sampling technique. In contrast the KIN Gaussian functions allow for robust optimization, so the minimizer was set to run for up to 10 cycles with a maximum of 50 iterations each. Similarity convergence was set to 0.01 with cycle convergence and contact termination set to 0.1 Å for default simplexing, and 0.0025/0.01 Å respectively for tight convergence. Empirical analysis also determined that reducing the initial translation and rotations to 0.1 Å and 0.025 radians allowed for more rapid convergence in KIN. Minimum ligand heavy atom count for searching to initiate was set to 15 heavy atoms. All input, template and ROCS hit files are available on request.

3. Search test results

(1) Using query 1 (Fig. 2) shape tests were run against the ROCS hits shown in Fig. 1 using a variety of search conditions. This has been done both to illustrate the tabu filter and allow direct albeit limited comparison between DOCK, KIN and ROCS search results. The tabu search results comparisons are shown in Table 2. The top scoring KIN superposition for ROCS hit 2 is shown in Fig. 4 (similarity score = −0.853). For all searches the more negative the score the better it is in DOCK (and by extension KIN). As such similarity scores will be reported as negative values, with −1 being exact similarity for KIN. For DOCK scores are reported relative to the score found when mapping the template onto itself (−2550). Fig. 5 shows the top scoring overlays determined for both hits under a variety of conditions. For all these tests ROCS hits hit 1 produced very consistent results (optimal similarity = 0.87–0.88, RMSD of optimized ligand relative to starting orientation = 8.4–8.5 Å).

(2) For the second test, both queries 1 and 2 were run against the ZINC conformer database (described in Section 2.1)

Table 2
Results for a variety of search settings run for ROCS hit 2 conformers vs. query 1

Taboo Search Setting[a]	Search time (ligand/s)	No. of overlays with sim. <−0.8	ROCS hit 2 best[b]	ROCS hit 1 RMSD spread[c]
No tabu	0.2	20	−0.85 (5)	8.5–9.5
0.5	0.4	21	−0.85 (4)	9.0–9.5
1.0	1.0	20	−0.85 (4)	9.0–9.5
2.0	3.5	19	−0.85 (2)	9.2
3.0	7.1	19	−0.85 (1)	9.2
4.0	12.7	17	−0.84 (2)	9.1–9.4
5.0	18.1	17	−0.85 (1)	9.2
No tabu tight	0.1	21	−0.85 (5)	8.4–9.5
4.0 tight	6.3	20	−0.85 (5)	8.4–9.5
5 tight	10.2	19	−0.85 (3)	8.9–9.5
6.0 tight	18.7	21	−0.85 (4)	8.8–9.5[d]
7.0 tight	26.7	16	−0.84 (1)	8.5
6.0 tight no con	5.9	20	−0.85 (5)	8.4–9.5
6.0 tight exc 0.25	6.0	10	−0.82 (5)/–	8.6–9.5/–
6.0 tight exc 0.75	5.5	2	–	–

[a] Tight refers to use of tight simplex convergence criteria—see DOCK/KIN input info. The exc term refers to searches incorporating exclusion spheres (see Fig. 5). Associated number refers to the clash weight applied.

[b] Numbers in brackets refer to the no. of conformers found with this similarity.

[c] RMSD spread of conformers with best similarity value.relative to initial conformer position.

[d] Missing conformer (rmsd 8.4) present with similarity = −0.84.

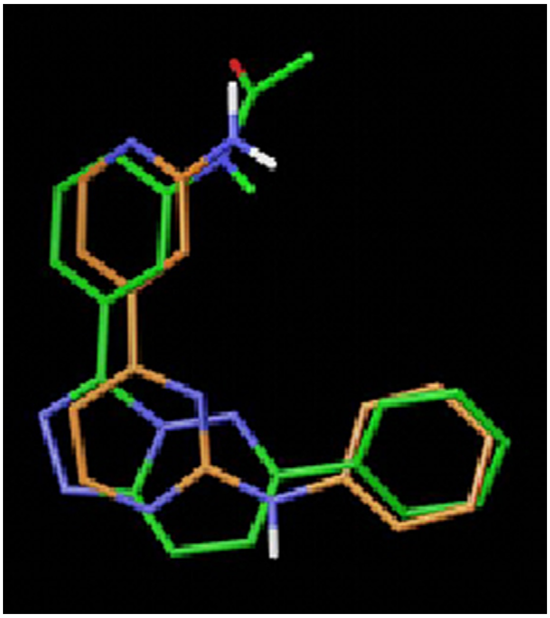

Fig. 4. Highest scoring overlay for ROCS hit 2. Superposition is visually very close to that seen in original ROCS paper [31].

under a variety of search conditions using DOCK and KIN. General search statistics for these tests are shown in Table 3. Top scoring hit ensembles for a variety of searches are shown in Fig. 6. Fig. 7 illustrates a selection of top scoring hits from these searches.

4. Discussion

Table 2 highlights the utility of using the tabu function in KIN, particularly in conjunction with tight simplex optimization criteria. The overlay count with scores less than −0.8 were designed provide insight into sampling behavior, while the best scores seen and the RMSD spread highlight convergence behavior. Using more standard simplex convergence settings, convergence performance begins to tail off somewhat with a tabu setting beyond 2.0 Å, with general performance falling rapidly at 3.0 Å and beyond. When tight convergence simplex conditions are applied, however, performance holds up until the tabu value shifts beyond 6.0 Å, allowing greater than two orders of magnitude improvement in speed relative to equivalent searches without a tabu filter. This highlights both the high level of redundancy present in the clique match superpositions and the robust nature of the Gaussian function optimizations. Based on these and other search tests a tabu setting of 6.0 Å with tight simplex convergence settings was chosen as the default for KIN searches involving shape only. This was reduced to 1.0 Å on application of pharmacophore distance constraints as the more constrained nature of the search already dramatically reduces the number of clique orientations tested.

Fig. 4 highlights the top hit for ROCS hit 2 from KIN. By visual inspection it a very similar solution to that shown in the original ROCS paper [31]. This is consistent with earlier comparisons between Gaussian shape and grid-based hard-sphere similarity evaluations for overlays of differing

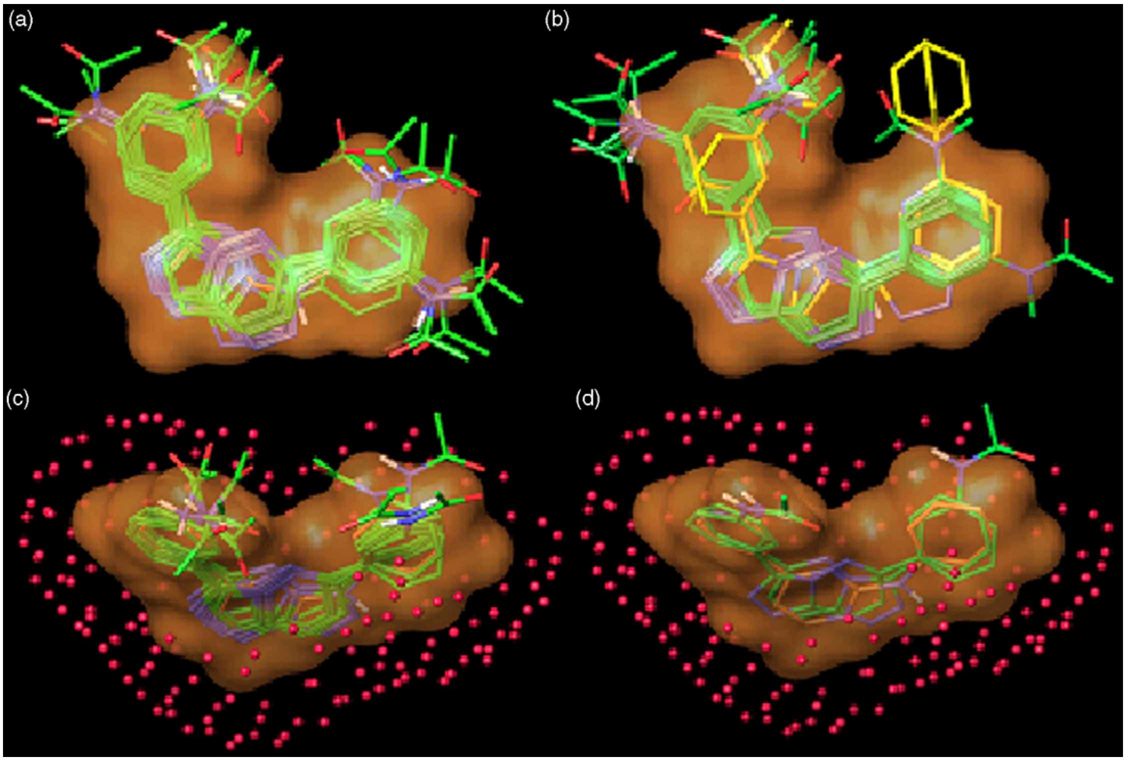

Fig. 5. Top scoring superposition results for the two ROCS hits using multiple search criteria. (a) The 20 hits (similarity range −0.8 to −0.88) typically found by KIN with no exclusion constraints set. (a) DOCK with MAKESITE shape cast score <−0.7 × self-test score. Results similar to KIN, with the exception of the structures highlighted yellow. (c) KIN with MAKESITE derived exclusion constraint atoms (shown in red), and exclusion overlap weighting set to 0.25. Overlays with a similarity <−0.8 have been retained. (d) As for (c) but with a clash overlap weighting of 0.75.

Table 3
Results for ZINC database searches using a variety of search conditions

Search settings[a]	Search time (ligand/s)	Hit count[b] (<−0.8/−0.85/−0.9, <−1750/−2000/−2250)
KIN tabu 6	28	76,184/13,525/821
KIN tabu 1 NH	284	6,874/1,227/85
KIN tabu 1 NH 1.0 tol.	74	12,940/2,064/102
KIN tabu 6 exc.	13	5,330/864/56
KIN tabu 1 NH exc. 1.0 tol	7	762/116/17
KIN tabu 6 pharm screen[c]	–	32,117/6,161/378
DOCK	22	15,002/730/21
DOCK bump	154	852/71/6
DOCK bump 1.5 tol.	10	1,700/128/9

[a] Query 1 used unless otherwise denoted. Use of tabu *n* refers tabu rmsd filter setting. NH denotes use of query 2 (inclusion of donor constraint). Application of exc. flags use of exclusion spheres (clash weight 0.75) in subsequent search to filter KIN tabu 6 results. NH exc. refers to similar procedure involving both exclusion spheres and query 2 constraints. Bump denotes used of bump filter during DOCK run (6 bumps with 100 point clash penalty). 1.0/1.5 tol. refers to increase of distance tolerance from default value of 0.75 Å.

[b] Hit count refers to the number of hits found at each given similarity cutoff. The larger numbers refer to DOCK searches. For comparison a superposition of the template onto itself produces a score of −2550. Note that the score convention for DOCK and by extension KIN is that the more negative a score the higher the similarity.

[c] Number of hits from KIN tabu 6 search also hitting the pharmacophore screen set in query 2 when screen run as separate filter rather than as direct constraint to shape search.

chemotypes [26,44]. Fig. 5 shows the ensemble of top hits for both DOCK and KIN. As one would expect, shape similarity is driven by ring matching, with the pendent amides able to map to varying positions with relatively small changes in similarity. Comparing the top left molecular ensemble of KIN with those of DOCK (top right), one can clearly see that DOCK is able to reach similar top scoring solutions. The solutions highlighted in yellow highlight the noisier nature of DOCK superposition, however, with the methyl amide of ROCS hits 2 able to substitute for a ring in a few cases. In this case the top scoring overlays were the same for both systems, however. The images at the foot of Fig. 4 highlight the impact of exclusion atoms (shown in red) on search behavior. By including an overlap function calculation filter based on vdW distance search times triple with incorporation of the exclusion criteria, a reasonable penalty to pay given the large number of atoms incorporated (173). With a clash penalty weight of 0.25 (bottom left), direct clashes of the amide with the exclusion regions are removed from the top scoring (<−0.8) solutions. When this is increased to 0.75 (bottom right), only solutions without direct amide clashes that exhibit a higher level of dihedral similarity relating the two distal aromatic moieties are retained (ROCS 1 overlays only). These results highlight the possibilities available to fine tune searches through incorporation of carefully weighted exclusion regions.

Table 3 highlights the results obtained running the ZINC database selection using a variety of search criteria. Fig. 6 illustrates examples of hit ensembles obtained from these searches, while Fig. 7 highlights some of the top scoring hits found. Running KIN with query 1 in shape only mode, 13,525 hits are found to have a similarity of −0.85 or greater. This tails off to 821 at −0.9, while expanding rapidly to 76,184 at a cutoff of −0.8. Fig. 4 illustrates that a high level of shape similarity still exists at −0.85. Even at −0.8 molecules can look quite similar, rendering the assignment of a similarity cutoff tricky at best. One way to improve the precision of such a cutoff is of course to add additional constraints to the search criteria. Use of query 2 to add the NH donor constraint to the clique search increases search speed to over 250 structures per second and reduces hit count by approximately 10-fold. Experience with the distance-based pharmacophore atom map penalty suggests that the term can produce convergence issues. This is illustrated by the additional query 2 run using a larger distance tolerance of 1.0 Å. Search speed is reduced to 74 ligands/s, with around twice as many hits being found at the lower similarity cutoffs. At the top similarity cutoff of −0.9 the difference tails off to around 15%. Modifying the pharmacophore constraints to become part of the Gaussian function would improve convergence here. It should be noted, however, that this will be most relevant when simple pharmacophores such as query 2 are applied, since there are still multiple starting point matches available. For stringent pharmacophores where very few matching orientations exist, this is less of an issue.

Applying the exclusion filter to the initial KIN results slows search speeds to a still respectable 13 ligands/s, while filtering hit counts by around 15-fold. When query 2 is also applied to this filtering process (with a 1.0 Å distance tolerance), search speed drops to 7 ligands/s (the pharmacophore distance penalty produces a significant slow down in optimization convergence), with hit count filtering increased to around a 100-fold. In contrast using the query 2 pharmacophore as a separate filter rather than as an integral clique constraint produces only a 2.5-fold decrease in count. These results highlight both the flexibility and filtering capabilities of including pharmacophore constraints and exclusion Gaussians as integral part of the shape search.

A closer look at Fig. 6 highlights the effects of the search criteria set. The top 500 hits for the KIN shape search (top left) clearly show good shape similarity with the template. It also the case, however, that molecules are breaking the surface of the template molecule in a wide variety of positions. The effect of the Gaussian exclusion atoms can be seen in the top 500 ensemble of the KIN search incorporating exclusion atoms (top right). Molecules no longer break through the lower surface of the template where the Gaussian exclusion atoms are concentrated. Comparing the top 100 of the KIN tabu 6 search (center left) with the top 100 of the KIN query 2 NH constrained search (center right), there is a clear increase in polar hydrogen atoms in the bottom region of the query around the NH donor (as expected). The generally higher polar hydrogen content of these top 100 hits further differentiates them from the shape only ensemble.

An analysis of the DOCK results again highlights the utility and limitations of simple shape searches directly within DOCK. Running without bump checks DOCK finds over 15,000 hits above the score cutoff of −1750 (0.7 times the score found for the template overlaying onto itself). There is a significantly

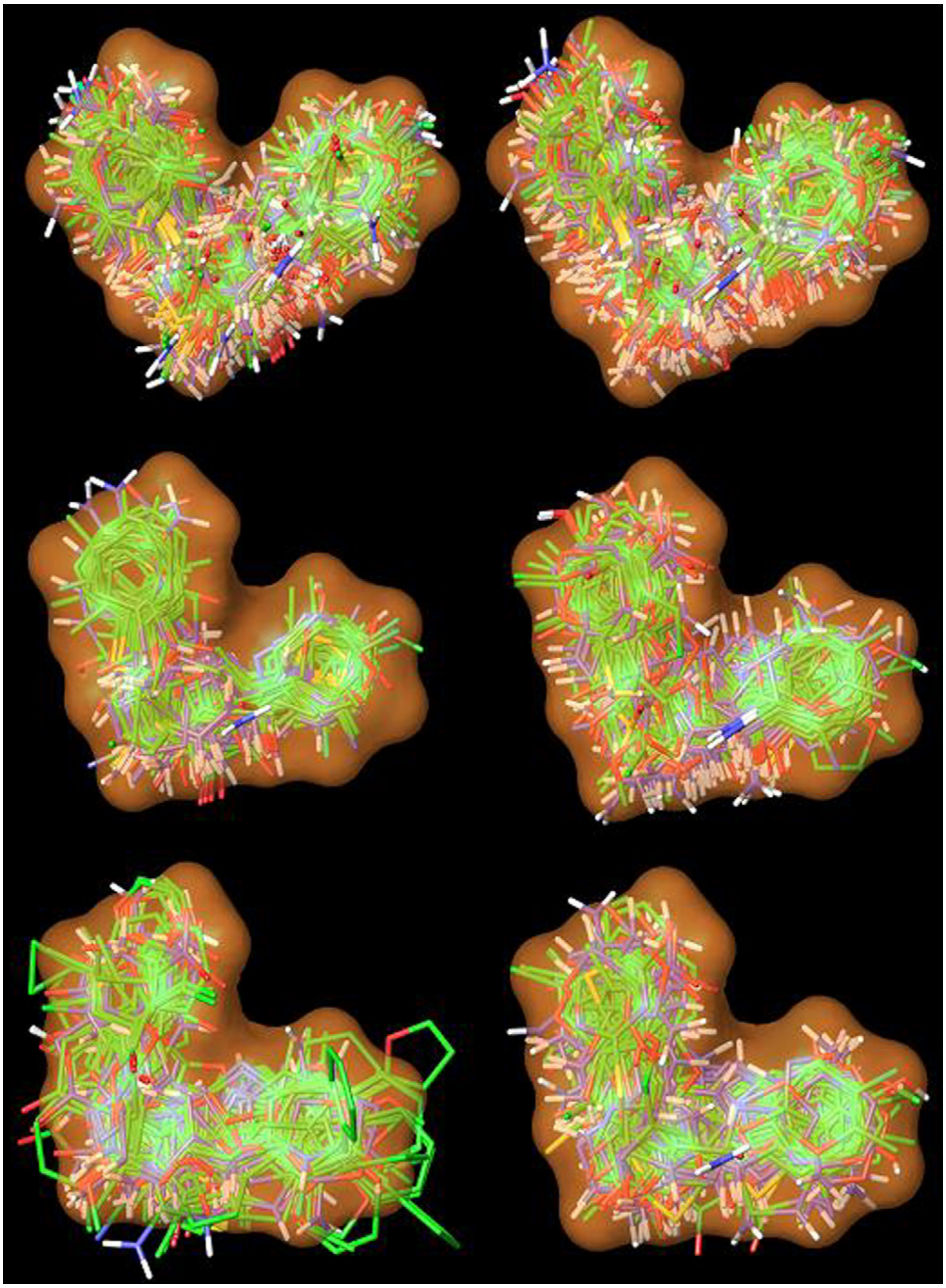

Fig. 6. Hit ensembles for a variety of the ZINC database searches. Query surface is marked in orange, hits shown with polar hydrogens only. Top left: top 500 from KIN tabu shape search. Top right: top 500 from KIN tabu with exclusion atom constraints (see Fig. 5), clash weight set to 0.75. Center left: top 100 from KIN tabu shape search. Center right: top 100 KIN hits using query 2 NH constraint. Bottom left: DOCK top 100. Bottom right: DOCK top 100 with bump check and contact penalty.

more rapid drop off at higher scores relative to KIN, however, with only ~0.14% of hits having a score <-2250 versus $>1\%$ showing a similarity score of <-0.9 in KIN. This further highlights the more sensitive nature of scoring derived from the MAKESITE derived shape cast. The lower ensembles shown in Fig. 6 highlight results obtained for two of the DOCK searches. The DOCK search ensemble shown without bump constraints (bottom left) highlights a limitation of running shape searches without modification in DOCK. Without a true similarity equation to penalize the non-overlapping sections of molecules. DOCK is prone to finding molecules with significantly non-overlapping portions as well as those with good overall match. This is illustrated by the molecules shown with significant functionality outside the template surface. The ensemble of DOCK bump filtered hits (bottom right) shows how incorporation of the bump filtering alleviates this issue. The results in Table 3 highlight how sampling can become a larger problem when this is done, however. Running at standard sampling only 852 hits pass the maximum score cutoff of -1750. Increasing the sampling by upping the distance tolerance to 1.5 Å more than doubles the hit count, but drops the search speed from 150 to 10 ligands/s. This is partly due to the fact that bumps are

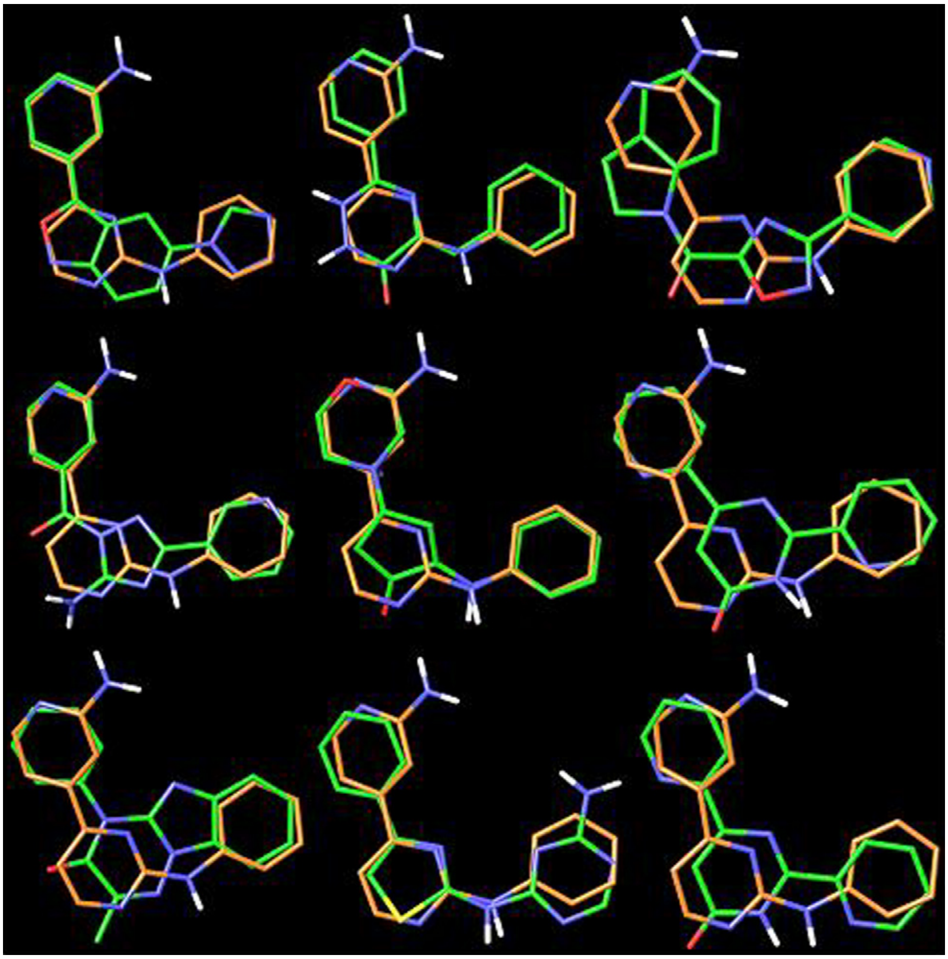

Fig. 7. Selected hits from the shape searches. The top 3 hits are from the KIN tabu 6 search, middle 3 from KIN query 2 NH constrained search and bottom 3 from the DOCK search with bump check (standard distance tolerance setting). Note that many of these hits are found in multiple searches (typically 25–30% of the top 500 overlap). Hit 6 (center right) and 9 (bottom right) show an example of this. Hit 9 looks essentially identical for both KIN and DOCK MAKESITE shape only searches. The shift in hit 6 is produced by the pharmacophore constraint used in the NH pharmacophore constrained searches.

penalized using a fixed penalty term (in this case 100), with the lack of gradient in the penalty further increasing optimization noise. Incorporating a correction for the target molecule and using a smoother clash penalty would help alleviate these shortcomings.

Fig. 7 illustrates how each of the techniques applied are able to find similarly shaped molecules from multiple chemotypes unrelated to the primary template, exactly as one would hope for a shape similarity search technique. Many of the top hits are found in multiple searches (typically 25–30% of the top 500 overlap for the searches compared). Hit 6 (center right) and 9 (bottom right) show an example of this. Hit 9 looks essentially identical for both KIN and DOCK MAKESITE shape only searches. The shift in hit 6 is produced by the pharmacophore constraint used in the NH pharmacophore constrained searches.

Overall the noise and sensitivity intrinsic to the DOCK searches lead to said screens missing some of the hits found in KIN. Nevertheless through application of the MAKESITE program DOCK can be massaged into a useful ligand shape docking tool with little or no modification. KIN by nature of its robust optimization makes for a more reliable search method. Search speeds for the two methods are similar (10–100 ligands/s range). This is certainly fast enough for database searching, but slower than ROCS. It should be noted, however, that this version of KIN is hamstrung by distance-based pharmacophore constraints and the simplex optimizer used by DOCK. DOCK requires this optimization technique for its default minimizations due to the fact that it uses grid-based scoring functions. KIN, by contrast can employ gradient minimization techniques by virtue of the fact that its scoring function is differentiable. McMahon and King showed that employing gradient methods (steepest descent), an order of magnitude speed up could be obtained for Gaussian functions similarity searches relative to the application of simplex-based optimization [47]. Such a speed up would render KIN search speeds similar to those obtained in ROCS (potentially faster when stringent pharmacophore constraints are employed). Other features that would further improve KIN include: addition of special chemical types to permit fragment R group mapping for de novo design, incorporation of pharmacophore constraint mapping into the Gaussian functions weighting, chemical (color) scoring and individual weightings for exclusion spheres, the addition of other Gaussian properties (e.g., MEPs) and template specific atom types to decouple color scoring from pharmacophore constraints. These features are the subject of ongoing research and will be reported on in a subsequent article.

5. Conclusions

Application of the MAKESITE program allows DOCK to be molded into a useful ligand shape similarity search tool. Additional modification to create the KIN program through incorporation of Gaussian-based shape functions allows the creation of an extremely flexible shape search tool. Use of a tabu-like filtering function permits respectable performance for shape only searches by removing redundant starting overlays from the shape optimization. The true utility of this approach becomes apparent when DOCK's highly flexible clique matching features of critical regions and chemical matching are used to drive pharmacophore constraints searches, however. By using such constraints in conjunction with shape similarity, search times are enhanced (speeds approaching 100 ligands/s are possible even when a simple pharmacophore is used) while at the same time improving search resolution. This can have significant advantages over a color force field similarity approach, since it is possible to force the presence of key binding groups directly within the ligand shape framework without reference to functionality redundant to binding. Further, by using a flexible superposition approach independent of the underlying scoring function, additional terms can easily be added to the scoring function with relative ease. In the version of KIN discussed here additional functionality has been incorporated in the form Gaussian exclusion functions. Their application allows the inclusion of additional shape constraints while still permitting relatively rapid search speed. Driving superposition using clique matching divorces the initial overlay from the underlying scoring function. Many other terms can thus easily be added to scoring to permit further search customization, allowing the creation of a highly flexible ligand-based similarity search tool.

Acknowledgements

I wish to express gratitude to Graham Richards. Enthusiastic without ever being heavy handed, studying as part of his research group was always a pleasure and an inspiration. Both MAKESITE and the original fortran code for calculating Gaussian STO-3G shape and electrostatic potential similarity are available from the author on request. Their creation was formulated under Graham's able tutelage. With the application of molecular similarity calculations finally maturing to a point where it rivals his original vision, It is a fitting time for the technology to find a wider audience.

References

[1] G.W. Bemis, I.D. Kuntz, A fast and efficient method for 2D and 3D molecular shape description, J. Comput. Aided Mol. Des. 66 (1992) 607–628.

[2] R. Nilakantan, N. Bauman, R.J. Venkataraghavan, New method for rapid characterization of molecular shapes: applications in drug design, Chem. Inf. Comput. Sci. 331 (1993) 79–85.

[3] R.P. Sheridan, M.D. Miller, D.J. Underwood, S.K. Kearsley, Chemical similarity using geometric atom pair descriptors, J. Chem. Inf. Comput. Sci. 361 (1996) 128–136.

[4] A.C. Good, T.J.A. Ewing, D.A. Gschwend, I.D. Kuntz, New molecular shape descriptors: application in database screening, J. Comput. Aided Mol. Des. 9 (1995) 1–10.

[5] J.S. Mason, D.L. Cheney, Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores, Pac. Symp. Biocomput. (1999) 456–467.

[6] M.J. McGregor, S.M. Muskal, Pharmacophore fingerprinting. 1. Application to QSAR and focused library design, J. Chem. Inf. Comput. Sci. 39 (1999) 569–574.

[7] A.C. Good, S.-J. Cho, J.S. Mason, Descriptors you can count on? Normalized and filtered pharmacophore descriptors for virtual screening, J. Comput. Aided Mol. Des. 18 (2004) 523–527.

[8] S. Renner, G. Schneider, Fuzzy pharmacophore models from molecular alignments for correlation–vector-based virtual screening, J. Med. Chem. 47 (2004) 4653–4664.

[9] J.A. Haigh, B.T. Pickup, J.A. Grant, A. Nicholls, Small molecule shape-fingerprints, J. Chem. Inf. Model. 45 (2005) 673–680.

[10] A.C. Good, Application of 3D molecular similarity index calculations to QSAR studies, in: P.M. Dean (Ed.), Molecular Similarity in Drug Design, Blackie Academic and Professional, Glasgow, 1995, pp. 24–56.

[11] A.C. Good, W.G. Richards, Explicit calculation of 3D molecular similarity, Perspect. Drug Discov. Des. 9 (1998) 321–338.

[12] R. Carbo, E.A. Besalu, General survey of molecular quantum similarity, Theomchemistry 451 (1998) 11–23.

[13] J.B. Moon, W.J. Howe, 3D database searching and de novo construction methods in molecular design, Tetrahedron Comput. Methodol. 3 (1990) 697–711.

[14] V.J. van Geerestein, N.J. Perry, P.D.J. Grootenhuis, C.A.G. Haasnoot, 3D database searching on the basis of shape using the SPERM prototype method, Tetrahedron Comp. Methodol. 3 (1992) 595–613.

[15] N.C. Perry, V.J. van Geerestein, Database searching on the basis of three-dimensional molecular similarity using the SPERM program, J. Chem. Inf. Comput. Sci. 32 (1992) 607–616.

[16] M.J. Hahn, Three-dimensional shape-based searching of conformationally flexible compounds, Chem. Inf. Comput. Sci. 37 (1997) 80–86.

[17] S. Putta, C. Lemmen, P. Beroza, J. Greene, A novel shape-feature based approach to virtual library screening, J. Chem. Inf. Comput. Sci. 425 (2002) 1230–1240.

[18] S. Putta, J. Eksterowicz, C. Lemmen, R. Stanton, A novel subshape molecular descriptor, J. Chem. Inf. Comput. Sci. 435 (2003) 1623–1635.

[19] A.C. Good, J.S. Mason, Computational Screening of 3D Databases, Reviews in Computational Chemistry, vol. 7, VCH, New York, 1995, pp. 67–127.

[20] G.W.A. Milne, M.C. Nicklaus, S. Wang, Pharmacophores in drug design and discovery, SAR QSAR Environ. Res. 9 (1998) 23–38.

[21] W.A. Warr, P. Willett, The principles and practice of three-dimensional database searching, in: Des. Bioact. Mol., American Chemical Society, Washington, D.C., 1998, pp. 73–95.

[22] O.F. Guner (Ed.), Pharmacophore Perception, Development, and Use in Drug Design, International University Line, 2000, p. 390.

[23] A.C. Good, J.S. Mason, S.D. Pickett, Pharmacophore pattern application in virtual screening, library design and QSAR, in: H.-J. Böhm, G. Schneider (Eds.), Virtual Screening for Bioactive Molecules, vol. 10, Wiley, 2000, pp. 131–154.

[24] J.H. van Drie, Pharmacophore discovery—lessons learned, Curr. Pharm. Des. 9 (2003) 1649–1664.

[25] A.C. Good, E.E. Hodgkin, W.G. Richards, The utilisation of Gaussian functions for the rapid evaluation of molecular similarity, J. Chem. Inf. Comput. Sci. 32 (1992) 188–191.

[26] A.C. Good, W.G. Richards, Rapid evaluation of shape similarity using Gaussian functions, J. Chem. Inf. Comput. Sci. 33 (1993) 112–116.

[27] D.J. Wild, P. Willett, Similarity searching in files of three-dimensional chemical structures: flexible field-based searching of molecular electrostatic potentials, J. Chem. Inf. Comput. Sci. 36 (1996) 159–167.

[28] D.A. Thorner, D.J. Wild, P. Willett, P.M. Wright, Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm, J. Chem. Inf. Comput. Sci. 36 (1996) 900–908.

[29] J.A. Grant, M.A. Gallardo, B.T. Pickup, A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape, J. Comp. Chem. 17 (1996) 1653–1666.

[30] ROCS, developed and distributed by OpeneyeScientific Software Inc. www.eyesopen.com/products/applications/rocs.html (accessed February 2007).

[31] T.S. Rush III, A.J. Grant, L. Mosyak, A. Nicholls, A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction, J. Med. Chem. 485 (2005) 1489–1495.

[32] P.C.D. Hawkins, A.G. Skillman, A. Nicholls, Comparison of shape-matching and DOCKing as virtual screening tools, J. Med. Chem. 50 (2007) 74–82.

[33] http://www.eyesopen.com/docs/html/rocs/node5.html (accessed February 2007).

[34] http://www.eyesopen.com/docs/html/rocs/colorsect.html (accessed February 2007).

[35] A.C. Good, D.L. Cheney, D.F. Sitkoff, J.S. Tokarski, T.R. Stouch, D.A. Bassolino, S.R. Krystek, Y. Li, J.S. Mason, T.D. Perkins, Analysis and optimization of structure-based virtual screening protocols. 2. Examination of docked ligand orientation sampling methodology: mapping a pharmacophore for success, J. Mol. Graph. Model. 22 (2003) 31–40.

[36] B.K. Shoichet, I.D. Kuntz, Matching chemistry and shape in molecular docking, Prot. Eng. 6 (1993) 723–732.

[37] DOCK 4 Manual. http://dock.compbio.ucsf.edu/Old_Versions/dock4.0_manual.pdf (accessed January 2007).

[38] MAESTRO, developed and distributed by Schrödinger Inc. www.schrodinger.com.

[39] OMEGA, developed and distributed by Openeye Inc. www.eyesopen.com.

[40] J.J. Irwin, B.K. Shoichet, ZINC—a free database of commercially available compounds for virtual screening, J. Chem. Inf. Model. 45 (2005) 177–182.

[41] http://blaster.docking.org/zinc/ - file 1_t80.smi (abstracted 09/06).

[42] Quantum Chemistry Program Exchange program, 429. http://qcpe.chem.indiana.edu/.

[43] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.

[44] A.C. Good, The extension and application of molecular similarity calculations. D.Phil Thesis, Oxford University, 1993.

[45] A.C. Good, S.J. Peterson, W.G. Richards, QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods, J. Med. Chem. 36 (1993) 2929–2937.

[46] C.A. Baxter, C.W. Murray, D.E. Clark, D.R. Westhead, M.D. Eldridge, Flexible DOCKing using Tabu search and an empirical estimate of binding affinity, Proteins 33 (1998) 367–382.

[47] A.J. McMahon, P.M. King, Optimization of Carbo molecular similarity index using gradient methods, J. Comp. Chem. 18 (1997) 151–158.