Docking of Flexible Molecules Using Multiscale Ligand Representations

Meir Glick, Guy H. Grant, and W. Graham Richards*

Department of Chemistry, Central Chemistry Laboratory, University of Oxford, South Parks Road, Oxford, OX1 3QH, United Kingdom

Received January 23, 2002

Structural genomics will yield an immense number of protein three-dimensional structures in the near future. Automated theoretical methodologies are needed to exploit this information and are likely to play a pivotal role in drug discovery. Here, we present a fully automated, efficient docking methodology that does not require any a priori knowledge about the location of the binding site or function of the protein. The method relies on a multiscale concept where we deal with a hierarchy of models generated for the potential ligand. The models are created using the k-means clustering algorithm. The method was tested on seven protein−ligand complexes. In the largest complex, human immunodeficiency virus reverse transcriptase/nevirapin, the root mean square deviation value when comparing our results to the crystal structure was 0.29 Å. We demonstrate on an additional 25 protein−ligand complexes that the methodology may be applicable to high throughput docking. This work reveals three striking results. First, a ligand can be docked using a very small number of feature points. Second, when using a multiscale concept, the number of conformers that require to be generated can be significantly reduced. Third, fully flexible ligands can be treated as a small set of rigid k-means clusters.

Introduction

The genome projects will reveal a plethora of new targets for drug discovery.[1] In tandem, structural genomics will provide the three-dimensional (3D) structures of many target proteins.[2] Exploiting this information may soon become the bottleneck in drug discovery. The identification of the binding site and its function are not necessarily straightforward, as was recently shown in the case of the anthrax toxin.[3] Indeed, there is now a shift of emphasis from genome mapping to the determination of genome function, hence functional genomics. If the binding site has no known ligand, such as in the case of orphan receptors,[4] the problem is even more challenging. As the number of target sites grows, there is a need for a virtual screening method without any a priori knowledge about the location of the binding site or its function.

Docking algorithms are essential in rational drug design. They differ in the size of the search space, the binding energy function, and the search strategy. Algorithms such as DOCK,[5−7] which is based on a sphere-matching procedure, or FLEXX,[8] which is an incremental construction method, assume that the binding site is known and limit the search to its boundaries. Heuristic methods such as Monte Carlo simulated annealing,[9−11] genetic algorithms,[12−16] or the Multiple Copy Simultaneous Search (MCSS)[17] can cover a larger translational space by minimizing the drug candidate−protein interaction energy. By starting the simulation near the binding site, the performance of such algorithms is reasonable. These methods, however, do not explore the conformational space exhaustively. Utilizing them without any a priori knowledge about the binding site might be too computationally expensive and skew the accuracy of the prediction. Exhaustive, discrete search methods such as GRID[18,19] cover consistently the entire search space, however, the trade off in computing time limits the method to small functional groups.

Many virtual screening methods take into account ligand flexibility. The geometry and the estimated binding free energy of protein−ligand complexes are calculated for each conformer. This strategy consumes vast amounts of computing time and disk space during high throughput docking (HTD) of millions or even billions of molecules. Indeed, in our screen saver project, which includes 3.5 billion molecules (www.chem.ox-.ac.uk/curecancer.html), the computational demands are massive (more than 1.5 million distributed computers).

Recently, we showed that by employing the k-means clustering algorithm it is possible to identify the binding sites on proteins and the orientation of rigid ligands in the binding site without any a priori knowledge.[20] Here, we significantly extend the methodology into full flexible docking.

Materials and Methods

First, we systematically created all of the conformers for the ligand. The problem of adequate coverage of the conformational space in a reasonable computing time was also addressed (vide infra). Only the torsion angles were modified, not the bond lengths or angles. Then, the conformers were ranked energetically. For this, the all-atom Consistent Valence Force Field (CVFF)[21] model was employed. The energy of a conformer was computed by eq 1 with the nonbonding 12-6 Lennard−Jones and electrostatic energy terms, where $A_{i,j}$ is the repulsion parameter for the two (i, j) atoms, $B_{i,j}$ is their attractive polarizability parameter, q_i is the partial charge, $r_{i,j}$ is the distance between atoms, and ϵ is the dielectric constant; V_n is the torsional potential barrier height for a torsion angle Φ, n being the multiplicity and γ the phase factor. If the energy term exceeded a given threshold, the conformer was ignored and was not clustered or docked. At the end of

* To whom correspondence should be addressed. Tel: (01865)-275908. Fax: (01865)275905. E-mail: graham.richards@chem.ox.ac.uk.

10.1021/jm020830i CCC: $22.00
Published on Web 09/06/2002

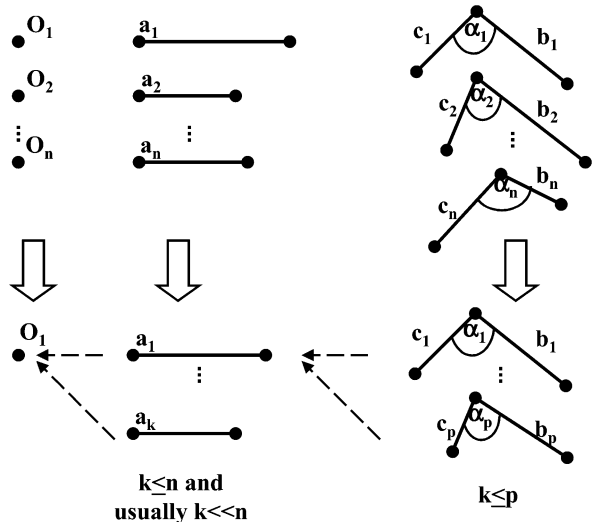

Figure 1. Clustering results (k-means) for an increasing number of feature points (from 1 to 3) for *n* conformers. In a single point representation, all of the conformers' representations (O_1, O_2...O_n) are identical and can be purged into a single point. The two point cluster is defined as a line (a_1, a_2...a_n) and the three point cluster as two lines and an angle (b_1, b_2...b_n), (c_1, c_2...c_n), (α_1, α_2...α_n). All of these clusters can still be purged into a much smaller number of *n*, either *k* for the two points or *p* for the three points.

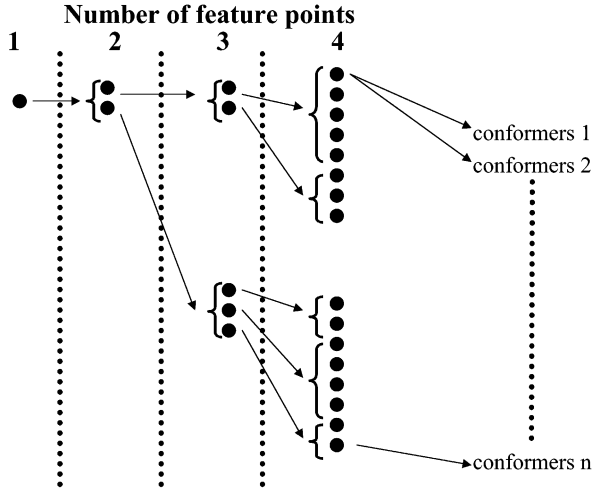

Figure 2. Purged clusters are sorted to form a search tree. The search begins from a single point purged cluster. Then it moves to the lower energy two point purged cluster, three point, and four. The search is done only on the relevant branches.

this stage, we retained the population of lowest energy conformers.

$$E_{\text{pot}} = \sum_{i<j}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\epsilon^* r_{i,j}}\right) + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\Phi - \gamma)] \quad (1)$$

To locate the binding site, we used a multiscale concept as described in detail previously,[20] where we dealt with a hierarchy of models generated for the potential ligand. Our search method is illustrated in Figures 1 and 2. We modeled each conformer at various scales by employing the *k*-means clustering algorithm.[20] Each set of conformers was assumed to be a single ligand. Figure 1 depicts *k*-means for an increasing number of feature points (from 1 to 3) on the HIV reverse transcriptase inhibitor, nevirapine.[22] Nevirapine has

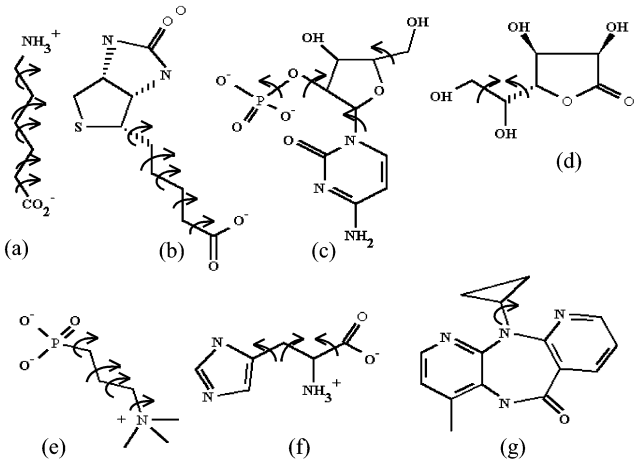

Figure 3. Seven ligands used as test cases. (a) ϵ-Amino capronic acid, (b) biotin, (c) cytidylic acid, (d) L-ascorbic acid, (e) phosphocholine, (f) L-histidine, and (g) nevirapin. The rotations are shown with arrows. Methyl moieties are not rotated.

one rotatable bond (we ignore the torsion of the methyl moiety) as shown in Figure 3g. If we systematically create all of the conformers in 5° intervals and assume all that of the conformers are energetically acceptable, we end up with $n = 72$ conformers. In the first model, a single point, we ended up with *n* points (O_1, O_2...O_n), as shown in Figure 1. Clearly, all of these points are near identical and can be represented once instead of *n* times. The second model was two points separated by a certain distance (a_1, a_2...a_n) that was related to the dimensions of the "main axis" of the conformer. In the case of nevirapine, the two feature point representation has an average distance of 4.6 Å and minimal and maximal values of 4.3 and 4.9 Å, respectively. Docking all of the 72 clusters at this low level of detail was redundant. Instead, the $n = 72$ clusters were purged by a purge criterion to *k* clusters, where $k \leq n$, and in most cases (see Results), $k \ll n$. Instead of docking *n* clusters, we docked *k* clusters. A rather strict purge criterion of $\Delta a \leq 0.6$ Å yielded a single representation in the case of nevirapine, i.e., from $n = 72$ clusters, we simplified the problem to $k = 1$ cluster. Indeed, at this level of representation and purge criterion, the ligand was treated as one rigid cluster, although it was flexible. The third model was defined by two distances (b_1, b_2...b_n) and (c_1, c_2...c_n) separated by an angle (α_1, α_2...α_n). It was purged into *p* presentations where the purge criterion was two distances and one angle. The number of three point purged clusters was equal to or bigger than those for the two points representation, $k \leq p$. For the four feature point representation, the purge criterion was three distances, two angles, and one dihedral. The process can be repeated for an increasing number of feature points.

The algorithm built a hierarchy tree of purged clusters as shown in Figure 2. This figure describes a typical search tree for a flexible ligand. Each node contains one such purged cluster and a list of the ensemble of conformers that this cluster stands for. In this example, two nodes of the purged clusters for two feature points emerged from the purged representation of a single point. The algorithm docked both of them and proceeded with the node of lower energy. This process was repeated for the third point and so on. In this search, only a small number of the purged representations were docked. If we assume that the four point representation is the last one to be docked, the worst case in this example is one rigid docking of a one point, two of two points, three of three points, and five of four points.

The translation space was searched in a manner identical to that reported previously.[20] The only difference was the rapid tree search for each translation. The search began from the single point representation. At the end of this stage, the algorithm moved on the branch toward the two point and

Table 1. Protein Ligand Chosen as Test Cases

protein	PDB code	resolution (Å)	no. of residues in protein	ligand	no. of atoms in ligand[a]	no. of rotatable bonds	initial no. of translations	RMSD[b] between the lowest energy prediction and the crystal structure	energy gap[c] between the 1st and 5th low energy predictions
hydrolase (serine protease)	2pk4	2.25	80	ϵ-amino capronic acid	22	5	151 156	2.43	1.42
streptavidin	1stp	2.6	121	biotin	31	5	257 186	1.28	4.16
hydrolase (endoribonuclease)	1rob	1.6	124	cytidylic acid	33	4	270 396	0.90	7.01
isomerase	1xid	1.70	387	L-ascorbic acid	22	2	960 492	2.12	2.21
McPC-603	2mcp	3.1	442	phosphocholine	24	4	984 528	2.25	0.92
histidine binding protein	1hsl	1.89	476	L-histidine	20	3	1 236 235	1.92	3.62
HIV-reverse transcriptase	1vrt	2.2	926	nevirapine	34	1	2 666 664	0.29	0.79

[a] Including hydrogens. [b] RMSD is given in Ångstroms and was calculated between all nonhydrogen atoms of the ligands. [c] Given in kcal/mol.

docked the purged representation that evolved from this single point. The process was repeated for the next purged representation where the parent of this representation was the one with the most favorable interaction energy with the binding site. At the last stage, we employed a rapid, local optimization to refine the structure using the complete conformer with the binding site held fixed.

A grid-based method was employed for energy evaluation between the purged representation and the binding site[20] by precalculating ligand−protein pairwise interaction energies to form a lookup table. The energy was computed by eq 2 with the CVFF all-atom model nonbonding terms. Atom i belongs to a feature point k, and j is a protein grid point.

$$E_{\text{pot}} = \sum_{k}\sum_{i,j}\left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}} + \frac{q_i q_j}{\epsilon^* r_{kj}}\right) \qquad (2)$$

Results

To test the methodology, a number of protein−ligand complexes were downloaded from the Protein Data Bank (PDB).[23] The proteins that were chosen cover a range of sizes (80−926 residues), resolutions (1.6−3.1 Å), and fold families as shown in Table 1 and Figure 4. These proteins were hydrolase[24] (serine protease) (PDB entry 2pk4, panel a), streptavidin[25] (1stp, panel b), hydrolase[26] (endoribonuclease) (1rob, panel c), isomerase[27] (1xid, panel d), McPC-603[28] (2mcp, panel e), histidine binding protein[29] (1hsl, panel f), and HIV reverse transcriptase[22] (1vrt, panel g). All ligands were flexible, with the number of rotatable bonds ranging from 1 to 5 and the number of atoms varying from 24 to 34 as shown in Table 1 and Figure 3.

We assumed the harshest situation where both the binding sites and the bioactive conformations of the ligand were unknown. The ligands were deleted, and an attempt was then made to find the correct docking site and the bioactive conformation of the ligands in one simulation. To avoid bias toward the experimental binding mode, hydrogens were added to the target proteins without the presence of the bound ligand. Arginine, lysine, and the N terminus were assumed to be fully protonated, while the C terminus, aspartic, and glutamic acids were charged. All histidines were assumed to be monoprotonated on Nϵ2.

In all test cases, we placed the host protein in a box of dimensions 3 Å greater in each direction than the extent of the protein. We employed a molecular grid with a 0.7 Å resolution and a rotation angle of 5°. A distance-dependent dielectric constant of $\epsilon = 4r$ was used. We restricted the maximal representation of the ligands to four feature points. All conformers were systematically generated in 30° intervals. Questions over the justification of creating the conformers in 30° intervals are addressed below. If the number of conformers exceeded the threshold of 1 000 000, then 1 000 0000 conformers were sampled at random. The lowest energy population of conformers (up to a threshold of 12 kcal/mol) was clustered. The clusters were purged using the following thresholds: distance, <1.3 Å; angle, <30°; and dihedral, <30°. Conjugate gradients local optimization was then employed on the complex from the starting geometry of the ligand in its predicted position. The protein's atoms were held fixed, and the ligand was allowed to move until a convergence criterion of 0.01 kcal/Å had been achieved.

The search results are shown in Table 1 and Figures 4 and 5. The initial number of translations varied according to the size of the protein: from 151 156 for the smallest one, hydrolase (serine protease) (Figure 4a), up to 2 666 664 translations for HIV reverse transcriptase (Figure 4g). The root mean square deviation (RMSD) values when comparing our lowest energy prediction, i.e., the best-ranked solution (which may not necessarily be the solution with the lowest RMSD) to the crystal structure ranged from 0.29 Å (HIV reverse transcriptase/nevirapin, Figure 5g) to 2.43 Å (hydrolase/ϵ-amino capronic acid, Figure 5a) with an average value of 1.60 Å.

Figure 4 is a snapshot into the end of the third iteration. Even at this low level of detail, the calculation clearly converges to the binding site. In hydrolase (serine protease) (panel a), streptavidin (panel b), hydrolase (endoribonuclease) (panel c), and HIV reverse transcriptase (panel g), all of the remaining translations, even those with a low energetic score that were not yet evicted, fall into the binding site. The algorithm clearly discriminated between the binding site and the other binding pockets on the surface of the protein. In isomerase (panel d) and McPC-603 (panel e), the majority of translations, including those with the highest score (data not shown), fall into the binding site. At the end of the last iteration in isomerase, all of the five lowest energy solutions (besides the third solution) fall into the binding site, while in McPC-603 only the first and seventh lowest energy solutions fall into the binding

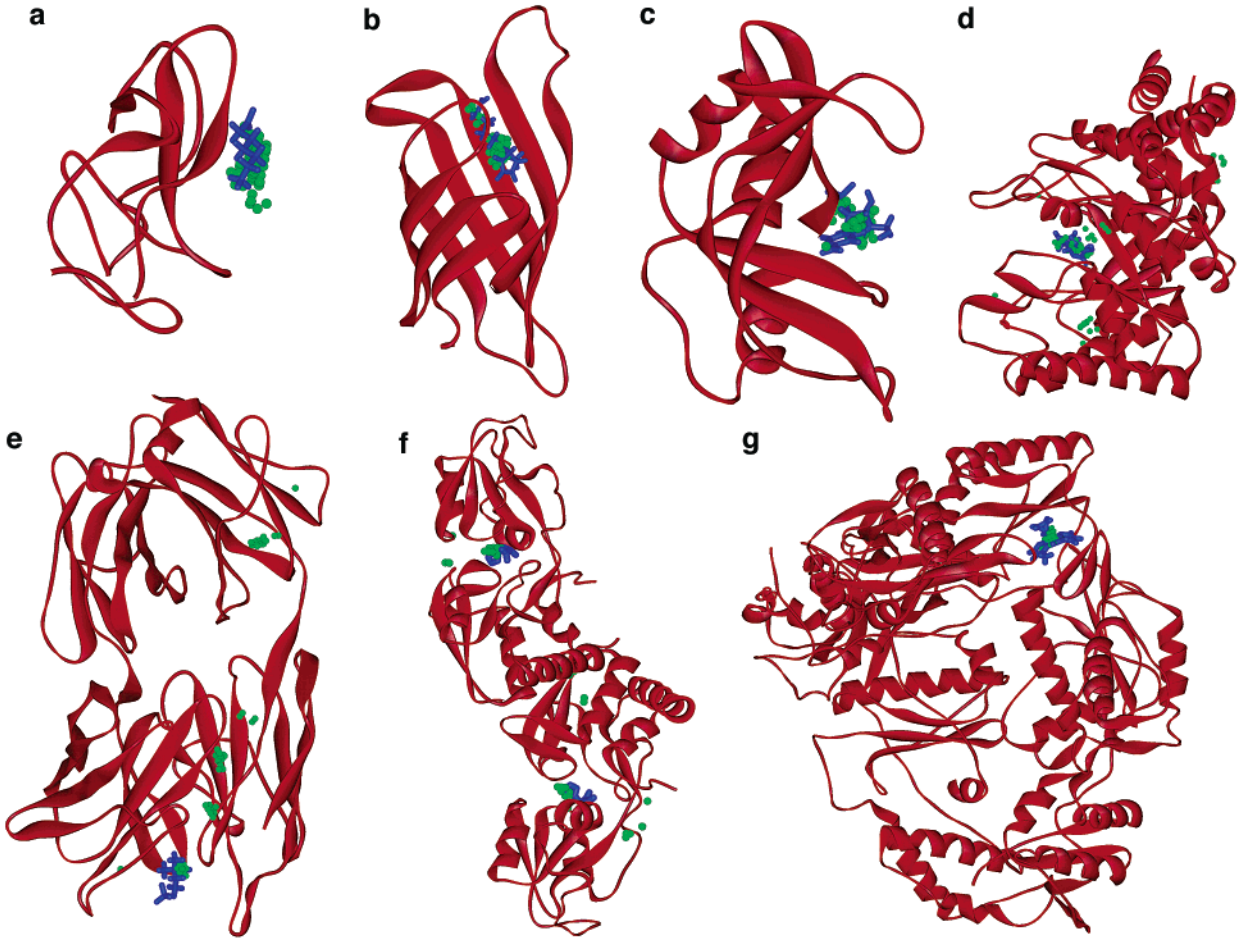

Figure 4. Seven proteins chosen as test cases. (a) Hydrolase (serine protease), (b) streptavidin, (c) hydrolase (endoribonuclease), (d) isomerase, (e) McPC-603, (f) histidine binding protein, and (g) HIV reverse transcriptase. The proteins are shown as red ribbons. The ligand in the crystal structure is shown in blue. The remaining possible translations of the ligand at the end of the third iteration are shown as green spheres.

site. The histidine binding protein/L-histidine complex (panel f) was deposited in the PDB as a dimer where L-histidine is bound to each of the monomers. We challenged the methodology and provided as an input the dimer and not a monomer. Both binding sites have been identified and ranked as the binding points with the lowest and equal energy, i.e., the first and second lowest energy solutions. The fourth and fifth solutions converged into the first binding pocket. The third solution did not converge to any of the pockets. This trend is suggested in Figure 4f, where at the end of the third iteration most of the feature points already converge onto the two binding sites.

The fact that from one up to four feature points are coarse representations of the ligand raises an intriguing question: Can we create conformers in larger dihedral angle intervals than 10° and still retain a reasonable coverage of all conformational space? Table 2 compares the computing time ratios needed to create conformers in 10° and 30° intervals. Biotin has $(360°/10°)^5 = 60\,466\,176$ conformers if we rotate in 10° intervals. If we change the 10° to 30°, the number of conformers will be $(360°/30°)^5 = 248\,832$. In other words, we gain $60\,466\,176/248\,832 = 243$ times speedup. In the seven ligands that we tested, the speedup varies from three for nevirapin up to 243 times for biotin and ϵ-amino capronic acid. We systematically created the conformers in 10° and 30° for the seven ligands, clustered all the conformers, and purged the clusters using the same criteria of distance, <1.3 Å; angle, <30°; and dihedral, <30° and compared populations within the lowest 12 kcal/mol. Table 2 shows the number of purged clusters needed to represent all of the conformers of the seven ligands. At one feature point, it does not matter how flexible the ligand is or by what increment we rotate the dihedral angles. In the case of two feature points, all ligands' conformations were still represented by one purged cluster, except biotin, which was represented by two. In other words, at this level of representation, all of the conformational space of the ligands was compressed into one rigid line, except biotin, which was branched into two families. When we increase the level of detail to three and four feature points, more purged clusters are needed to represent the available conformational space. In the four point representation, the purged clusters number ranges from one for nevirapin up to 22 for biotin. This number is extremely small as compared to the real number of conformers: 36 and 60 466 176, respectively (when creating the conformers in 10° intervals). When comparing the number of purged clusters when working in 10° and 30° intervals, in the case of one and two feature points, we got an identical number. This trend is preserved when moving to three feature points in six out of seven ligands. The only

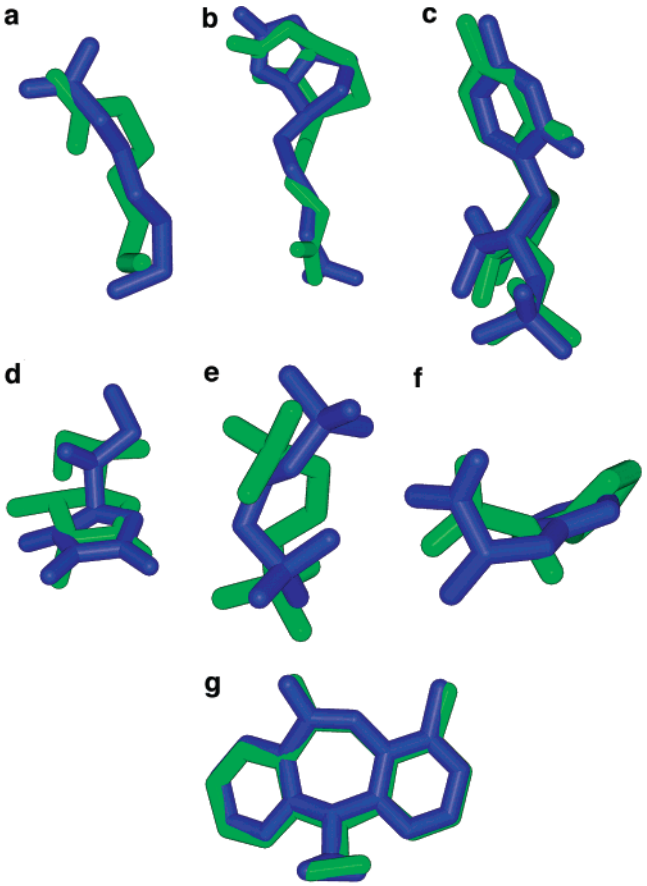

Figure 5. Comparison between our results (shown in green) and the conformation of the ligand at the crystal structure (shown in blue) for the seven ligands. (a) ϵ-Amino capronic acid, (b) biotin, (c) cytidylic acid, (d) L-ascorbic acid, (e) phosphocholine, (f) L-histidine, and (g) nevirapin.

exception, L-histidine, where there are two purged clusters in the 30° intervals as compared to three in the case of 10°. When we move to a four feature point representation, six out of seven ligands show two-thirds or more of the number of purged clusters when comparing the use of 10° and 30° increments.

We studied the applicability of the methodology to HTD where the binding site is normally known. We picked 25 structures at random from the Genetic Optimization for Ligand Docking (GOLD) data set[13] as shown in Table 3. We employed our algorithm with the same parameters as the above test cases with two differences: the conformers were created in 60° intervals and not 30°, and we limited the search to a sphere with a diameter of 28 Å around the binding site, where the centroid of the ligand was not allowed to move outside the boundaries of the sphere. In 18 test cases, the RMSD values when comparing our lowest energy prediction, i.e., the best-ranked solution to the crystal structure, ranged from 0.3 to 2.5 Å. In two test cases, we received RMSD values between 2.5 and 3.0 Å. In the remaining five complexes, the RMSD values were higher than 3.0 Å. The computing time (creation of the energy grid, generation of the conformers, and the docking calculations) on a single processor low-end PC (Pentium III 666Mz, 96Mb RAM laptop) was less than 1 min for 13 test cases, between 1−2 min for seven test cases, and more than 2 min for the remaining five test cases.

Ideally, one would like to accommodate protein flexibility during the docking process[30] instead of assuming a rigid protein structure. Regrettably, screening each conformer for a given ligand against each protein configuration is prohibitively slow to be considered as an efficient tool for in silico screening.[30] This raised the question as to whether the algorithm can provide a moderate docking prediction that can be a starting point for more elaborate calculations such as molecular dynamics or Monte Carlo. As a host protein, we utilized HIV reverse transcriptase[22] (PDB entry 1vrt), which is solved with nevirapine bound. We deleted nevirapine and tried to dock two other nonnucleoside inhibitors, 1051U91 and α-anilino phenyl acetamide (α-APA). It is hard to describe the results in RMSD terms since we utilized a different ligand than in the PDB file. In Figure 6a,b, we compare the bound conformation of nevarpine to our lowest energy prediction, i.e., the best-ranked solution of 1051U91 (panel a) and α-APA (panel b). Visual inspection of the results hints that the prediction is fairly accurate. We employed the program on the Nuclear Pregnane X Receptor-SR12813 complex[20,31] where we utilized the *apo* structure of the protein. Although SR12813 binds in three distinct orientations (Figure 6c shows only one of them), the translation of the our lowest energy predicted conformation was reasonable as shown by an overlap of the six-membered rings at the center of the molecule but the rotation did not resemble any of the three orientations. We also employed the algorithm on isomerase[32] (PDB entry 1xid). We removed the ligand, L-ascorbic acid, that appears in the crystal structure and tried to dock two other ligands, 1,5-dianhydrosorbitol and D-sorbitol. Similarly to the HIV reverse transcriptase test case, the lowest energy predicted conformation of 1,5-dianhydrosorbitol (Figure 6d) is fairly accurate but the one of D-sorbitol is inaccurate. Only the second-ranked solution (Figure 6e) is reasonable.

Discussion

This work presents a robust methodology for the docking of flexible molecules to proteins with unknown binding site or function. The approach confirms three assumptions. First, it is possible to discretize the search space for both the translation and rotation and the conformational flexibility of the ligand. Second, we can search this space at various levels of detail using a multiscale approach. Third, the result that emerges from this search is close enough to the optimal position that by employing a rapid local optimization we can get a reliable prediction. The algorithm was robust enough to find the binding site and the geometry of seven flexible ligands using only two basic assumptions. First, the protein 3D structure is known; second, we have a reasonable "cost function", in this case based on a molecular mechanics force field.

An advantage of the method is that even if there is more than one binding site, such as in the case of the histidine binding protein/L-histidine complex, one simulation is sufficient. It is impressive to see that both binding sites were identified and were ranked by identical energies (data not shown). Furthermore, the ligand itself was docked with a reasonable degree of accuracy.

Table 2. Number of Purged Clusters When Rotating in 10° and 30° Intervals

		when rotating in 10° intervals					when rotating in 30° intervals				
			no. of purged clusters for					no. of purged clusters for			
ligand	rotatable bonds	no. of conformers	1 point	2 points	3 points	4 points	conformers	1 point	2 points	3 points	4 points	speedup
ϵ-amino capronic acid	5	60 466 176	1	1	2	4	248 832	1	1	2	3	243
biotin	5	60 466 176	1	2	7	22	248 832	1	2	7	21	243
cytidylic acid	4	1 579 616	1	1	1	4	20 736	1	1	1	3	81
L-ascorbic acid	2	1296	1	1	2	6	144	1	1	2	4	9
phosphocholine	4	331 776	1	1	3	9	20 736	1	1	3	6	16
L-histidine	3	46 656	1	1	3	6	1728	1	1	2	3	27
nevirapine	1	36	1	1	1	1	12	1	1	1	1	3

Table 3. Applicability of Our Algorithm Towards HTD

PDB code	AMSD[a] between the lowest energy prediction and the crystal structure	computing time
1abe	2.5	0 min 27 sec
1acj	2.9	0 min 29 sec
1aco	1.3	0 min 47 sec
1aha	0.3	0 min 52 sec
1azm	0.9	0 min 59 sec
1cil	2.3	4 min 27 sec
1com	1.1	0 min 56 sec
1cps	2.5	5 min 28 sec
1die	1.9	0 min 47 sec
1hsl	2.1	1 min 14 sec
1imb	>3.0	1 min 47 sec
1lah	1.2	1 min 31 sec
1mdf	>3.0	9 min 36 sec
1nis	>3.0	1 min 7 sec
1phd	2.5	0 min 32 sec
1slt	2.4	2 min 5 sec
1tng	2.3	0 min 33 sec
1tni	2.7	3 min 56 sec
1tnl	1.6	0 min 56 sec
2r07	3.0	6 min 20 sec
4cts	2.2	1 min 30 sec
4fab	>3.0	1 min 8 sec
4ptb	>3.0	9 min 45 sec
6abp	2.2	0 min 27 sec
6rsa	1.5	1 min 13 sec

[a] RMSD is given in Ångstroms and was calculated between all nonhydrogen atoms.

There was no correlation between the size of the protein (reflected by the initial number of translations) and the accuracy of the prediction. Strikingly, in the largest complex (HIV reverse transcriptase/nevirapin), we obtained the best prediction while in the smallest one (hydrolase/ϵ-amino capronic acid), we had the highest RMSD value. Hence, the quality of the prediction (but not the complexity of the problem) is not correlated to the number of initial translations. There was no correlation between the size of the ligand and the quality of the prediction as well. Indeed, the ligands with more than 30 atoms exhibited lower RMSD values than the ligands that had fewer than 30 atoms. The resolution of the crystal structure was not correlated to the quality of the prediction.

There is a weak correlation between the flexibility of the ligand and the quality of the predictions. Nevirapin, which was the least flexible ligand, showed the lowest RMSD value, while ϵ-amino capronic acid, which has five rotatable bonds, showed the highest RMSD value. On the other hand, ligands such as cytidylic acid, with four rotatable bonds, showed a lower RMSD than ligands with two rotatable bonds such as L-ascorbic acid.

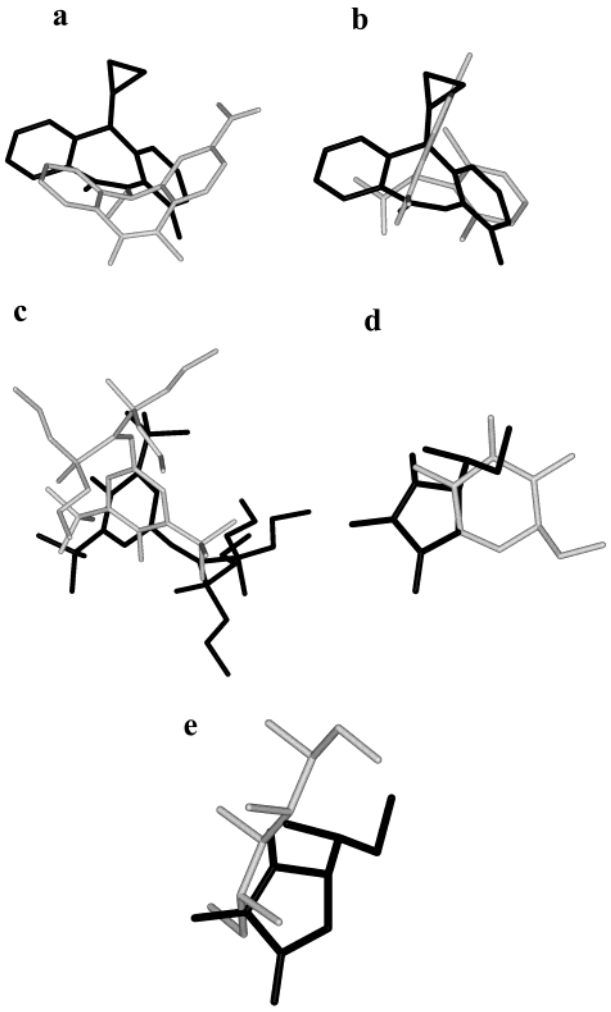

Figure 6. Comparison between our results for 1051U91 (shown in gray) and the conformation of the ligand at the crystal structure (a) nevarpine (shown in black), (b) α-anilino phenyl acetamide/nevarpine, (c) SR12813 docked to the *apo* protein/SR12813 in the bound form, (d) dianhydrosorbitol/L-ascorbic acid, and (e) D-sorbitol/L-ascorbic acid.

This implies that there are other factors that also affect the accuracy of the prediction. Three of them and their impact on the sensitivity of the results are discussed below.

First, the level of detail changes when discretizing the search space. Increasing this level, for example, by rotating the ligand in smaller increments or using a higher grid resolution than 0.7 Å is likely to improve the results. Out of these three parameters, the results are most sensitive to the angle increments in which the

ligand is rotated. The impact of the rotation is correlated to the distance from the center of rotation. In the case of long and large ligands, 5° rotations might be significant; consequently, the binder might miss potential binding pockets and clash with protein's atoms. On the other hand, small binders such as benzamidine may be rotated in larger increments thus saving substantial computing time. In this work, we showed that 5° rotations are a reasonable compromise; however, further work is needed in order to correlate the angle rotation to the dimensions of the binder. We employed a grid with a relatively low resolution (0.7 Å). Because the multiscale approach relies on fuzzy ligand representations (up to four feature points in this work), increasing the resolution to 0.4 Å is not a good trade between speed and accuracy; that is, the improvement of the RMSD values is not significant, and the computing time needed to created the grid significantly increases (data not shown).

Second, the detail level when using a multiscale approach can affect the results. The number of four feature points should be taken with an extreme caution, and further work is needed in order to parameterize and determine the optimal number of feature points. It is reasonable that large ligands may need more feature points. On the other hand, addition of feature points increases the ligand's detail level; hence, the angle intervals that the ligand should be rotated should be smaller, otherwise many translations that are close to potentially good locations will not be found due to clashes with the protein. The sensitivity of the purge criterion is also crucial to the success of the method. Employing a loose criterion will lead to a small number of purged clusters (and nodes on the search tree) and will significantly reduce the computing time. However, such representations might be too coarse and consequently skew the results. On the other hand, employing a strict purge criterion will lead to a large number of clusters. This detailed presentation of a plethora of conformers is indeed more accurate. However, it will increase the computational demands. At first glance, there seems no clear relation between the angle intervals by which the torsions are rotated when creating conformers and the purge criterion. However, if we employ a loose purge criterion, there is no point in rotating the torsions in very small intervals to generate an immense number of conformers since all, or the majority of them, will be purged.

Finally, more accurate energy functions than simple molecular mechanics nonbonding and torsion terms, which will consider other parameters such as conformational entropy loss due to the binding, are likely to improve the predictions. The atomic force field descriptor for a feature point comprises charge and Lennard−Jones parameters. It is straightforward to increase the sophistication of the model by adding more descriptors to be used in the interaction energy evaluation. Clearly, once the active site location is known, a more expensive but accurate calculation can be repeated. The aim of the work was to demonstrate that by employing a simple and general energy function, such as CVFF, a multiscale-based flexible docking methodology is viable.

The use of a multiscale approach enables one efficiently to break a problem down into a number of small steps. Dismantling a problem in this manner enables efficient distribution of computing time so that only the most fruitful areas are considered in any detail. In most of the test cases, when the location of the binding site was known, the entire computing time, which included the creation of energy grid, conformers, and docking the ligand was less than a minute on a low end machine. Indeed, the methodology saves a substantial amount of computing time in several ways. First, it evicts at a very early stage all of the unfavorable translations using a negligible amount of computing time. Second, it uses a very small number of feature points in order to dock the seven ligands. In this work, we clustered into four feature points seven ligands with numbers of atoms ranging from 24 to 34. Furthermore, most of the translations were already evicted in 1−3 feature points in a less expensive computing step.

Our methodology "rigidifies" the ligands. We are able to compress all of the conformers into a small number of rigid feature points. This allows us to simplify the flexibility problem. The most extreme example is nevirapin, which although flexible, was treated as rigid during the docking. In six out of seven ligands, we showed that in the two point representation the flexibility problem can be ignored.

In many cases, the target protein was not in its appropriate bioactive conformation either because it was in its *apo* form or the structure was solved when the protein was bound to a different ligand. Furthermore, the proteins' polar hydrogens, such as serine, threonine, and tyrosine hydroxyls, are likely to change their positions once binding takes place and misplaced hydrogens are likely to skew the results.[33] We received reasonable results however in all of the test cases; the hydrogens were added to the target protein without considering the ligand in the crystal structure. Therefore, the simplistic representation of the ligand that we utilized is advantageous in that sense. We previously docked rigid ligands into target proteins taken from complexes with other substrates.[20] Regrettably, when both the conformation of the binding site and the ligand are unknown, it is difficult reliably to dock and score flexible ligands with the current version of the program, although the algorithm provides a good starting position for more accurate calculations. We believe that simultaneous multiscale representations of both the ligand and the binding site (if it is known) are likely to improve the results. Multiple conformations for the binding site may be efficiently created by a stochastic method suggested by Glick et al.[34] or taken from NMR structures.

Unlike other docking methods, which normally rotate the ligand bonds about in 10−15° intervals, we employed here much larger increments of 30° and 60°. We have shown that such increments are justified when using a multiscale approach. In addition, we have shown a speedup that ranges from three to 243 times. In HTD, generating the conformers, storing them, and later evaluating them is one of the bottlenecks. The problem of highly flexible compounds with a prohibitively large number of conformers that cannot be evaluated exhaustively is a major one in chemoinformatics.[35] In the anchor-first procedure,[36] a rigid core is docked and the flexible parts are reattached incrementally. There is a

degree of bias in this method since the geometry of the conformer in the binding site depends on the placement of the anchor. This is a limitation if the position of a ligand is dictated by a small functional group such as ammonium or carboxylate or if the ligand is aliphatic. Other chemoinformatics algorithms such as Statistical Classification of Activities of Molecules for Pharmacophore Identification (SCAMPI)[37] sample confomations randomly. Here, we suggest a third option: creating the conformers in larger intervals and clustering them. The advantage of our strategy is in many cases, a uniform, nonrandom, and unbiased coverage of the whole conformational space. We believe that this software will provide a major opportunity in the area of rational drug discovery.

Acknowledgment. This work was supported by the Wellcome Trust and partially by the National Foundation for Cancer Research. We acknowledge the support of the Royal Society through the provision of an equipment grant.

References

(1) Service, R. F. Structural genomics offers high-speed look at proteins. *Science* **2000**, *287*, 1954−1956.
(2) Sali, A. Target practice. *Nat. Struct. Biol.* **2001**, *8*, 482−484.
(3) Glick, M.; Grant, G. H.; Richards W. G. Pinpointing anthrax-toxin inhibitors. *Nat. Biotechnol.* **2002**, *20*, 118−119.
(4) Williams, S. C. Biotechnology match making: screening orphan ligands and receptors. *Curr. Opin. Biotechnol.* **2000**, *11*, 42−46.
(5) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505−524.
(6) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(7) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking. *Protein Eng.* **1993**, *6*, 723−732.
(8) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(9) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195−202.
(10) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. J. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1094.
(11) Kirkpatrick, S.; Gelatt, C. D., Jr; Vecchi, M. P. Optimization by simulated annealing. *Science* **1983**, *220*, 671−680.
(12) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *254*, 43−53.
(13) Jones, G.; Willet, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(14) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: Parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.
(15) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(16) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J. Goodsell, D. S. Automated docking to multiple target structures: Incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins: Struct., Funct., Genet.* **2002**, *46*, 34−40.
(17) Miranker, A.; Karplus, M. Functionality maps of binding sites: a multiple copy simultaneous search method. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 29−34.
(18) Wade, R. C.; Goodford, P. J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 140−147.
(19) Wade, R. C.; Goodford, P. J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 148−156.
(20) Glick, M.; Robinson, D. D.; Grant, G. H.; Richards, W. G. Identification of Ligand Binding Sites on Proteins Using a Multi-Scale Approach. *J. Am. Chem. Soc.* **2002**, *124*, 2337−2344.
(21) MSI, San Diego, CA.
(22) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Kirby, C. R. I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D. I.; Stammers, D. High-resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nat. Struct. Biol.* **1995**, *2*, 293−302.
(23) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(24) Wu, T. P.; Padmanabhan, K.; Tulinsky, A.; Mulichak, A. M. The refined structure of the epsilon-aminocaproic acid complex of human plasminogen kringle 4. *Biochemistry* **1991**, *30*, 10589−10594.
(25) Weber, P. C.; Ohlendorf, D. H.; Wendolski, J. J.; Salemme, F. R. Structural origins of high-affinity biotin binding to streptavidin. *Science* **1989**, *243*, 85−88.
(26) Lisgarten, J. N.; Gupta, V.; Maes, D.; Wyns, L.; Zegers, I.; Palmer, R. A.; Dealwis, C. G.; Aguilar, C. F.; Hemmings, A. M. Structure of the crystalline complex of cytidylic acid (2′-CMP) with ribonuclease at 1.6 Å resolution. Conservation of solvent sites in RNase-A high-resolution structures. *Acta Crystallogr.* **1993**, *D49*, 541−547.
(27) Carrell, H. L.; Hoier, H.; Glusker, J. P. Modes of binding substrates and their analogues to the enzyme D-xylose isomerase. *Acta Crystallogr.* **1994**, *D50*, 113−123.
(28) Padlan, E. A.; Cohen, G. H.; Davies, D. R. On the specificity of antibody 3-antigen interactions-phosphocholine binding to MCPC603 and the correlation of 3-dimensional structure and sequence data. *Ann. Immunol. Sect. C* **1985**, *136*, 271−276.
(29) Yao, N.; Trakhanov, S.; Quiocho, F. A. Refined 1.89-Å structure of the histidine-binding protein complexed with histidine and its relationship with many other active transport/chemosensory proteins. *Biochemistry* **1994**, *33*, 4769−4779.
(30) Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57*, 213−218.
(31) Watkins, R. E.; Wisely, G. B.; Moore, L. B.; Collins, J. L.; Lambert, M. H.; Williams, S. P.; Willson, T. M.; Kliewer, S. A.; Redinbo, M. R. *Science* **2001**, *292*, 2329−2333.
(32) Carrell, H. L.; Glusker, J. P.; Burger, V.; Manfre, F.; Tritsch, D.; Biellmann, J. F. X-ray analysis of D-xylose isomerase at 1.9 Å: native enzyme in complex with substrate and with a mechanism-designed inactivator. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 4440−4444.
(33) Glick, M.; Goldblum, A. A Novel Energy-Based Stochastic Method for Positioning Polar Protons in Protein Structures from X-rays. *Proteins: Struct., Funct., Genet.* **2000**, *38*, 273−287.
(34) Glick, M.; Rayan, A.; Goldblum, A. A stochastic algorithm for global optimization and for best populations: A test case of side chains in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 703−708.
(35) Bravi, G.; Gancia, E.; Zaliani, A.; Pegna, M. SONHICA Simple Optimized Non-Hierarchical Cluster Analysis: a new tool for analysis of molecular conformations. *J. Comput. Chem.* **1997**, *18*, 1295−1311.
(36) Leach, A. R.; Kuntz, I. D. Conformational Analysis of Flexible Ligands in Macromolecular Receptor Sites. *J. Comput. Chem.* **1992**, *13*, 730−748.
(37) Chen, X.; Rusinko, A., III; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.

JM020830I