Fully Automated Molecular Mechanics Based Induced Fit Protein−Ligand Docking Method

Jürgen Koska,† Velin Z. Spassov,† Allister J. Maynard,* Lisa Yan, Nic Austin, Paul K. Flook, and C. M. Venkatachalam

Accelrys Inc., 10188 Telesis Court, San Diego, California 92121

Received March 7, 2008

We describe a method for docking a ligand into a protein receptor while allowing flexibility of the protein binding site. The method employs a multistep procedure that begins with the generation of protein and ligand conformations. An initial placement of the ligand is then performed by computing binding site hotspots. This initial placement is followed by a protein side-chain refinement stage that models protein flexibility. The final step of the process is an energy minimization of the ligand pose in the presence of the rigid receptor. Thus the algorithm models flexibility of the protein at two stages, before and after ligand placement. We validated this method by performing docking and cross docking studies of eight protein systems for which crystal structures were available for at least two bound ligands. The resulting rmsd values of the 21 docked protein−ligand complexes showed values of 2 Å or less for all but one of the systems examined. The method has two critical benefits for high throughput virtual screening studies. First, no user intervention is required in the docking once the initial binding site selection has been made in the protein. Second, the initial protein conformation generation needs to be performed only once for a given binding region. Also, the method may be customized in various ways depending on the particular scenario in which dockings are being performed. Each of the individual steps of the method is fully independent making it straightforward to explore different variants of the high level workflow to further improve accuracy and performance.

INTRODUCTION

The ability to dock ligands to protein binding sites using computational methods continues to receive considerable attention.[1−4] In particular, prediction of the correct binding modes of different ligands into a wide variety of receptor sites has remained the principal objective of the many ligand docking methods available today. It is generally accepted that for a docking procedure to work reliably, the following factors need to be considered at some level: (a) the general location of the binding site; (b) conformational flexibility of the ligand; (c) positional and orientational sampling of the ligand; (d) selection of a force field that can adequately describe the energy of interaction of the ligand with the receptor; (e) scoring of the ligand poses to prioritize the docked ligand poses; (f) explicit solvent effects; and (g) protein flexibility.

Docking methods vary widely in how they address these factors, and many different approaches have been developed to solve different aspects of the problem.[5−27] The various methods differ in two areas in particular: energy and scoring functions employed to prune and prioritize ligand poses and the method employed for ligand placement and orientational sampling. While scoring functions differ in the type of descriptors and the number of terms they employ, computation of the energy itself is often performed using grid-based methods[5,12] for computational efficiency.

Optimal ligand placement is sought using one of the following methods: systematic/stochastic search of the binding region, shape-based placement,[6,7,10,11,20,28−30] or feature-based alignment.[31,32] In shape-based methods the algorithm seeks ligand poses that correspond to the shape of the receptor cavity. In feature-based docking, the binding site of a protein is analyzed obtaining a description of the optimal locations of ligand features, such as hydrogen bond donors, acceptors, and nonpolar regions. The method then considers only docking poses that overlay the ligand features onto the desired positions in the binding site. Feature-based methods have an advantage over general search methods in that only the relevant part of the binding cavity is considered. However, there is no consensus on the best available ligand-docking method which is a reflection of the complexity of the problem, although, several scoring functions have been proposed to prioritize docked poses of various ligands docked into various proteins.[20]

One of the factors influencing the accuracy of docking that has received increasing emphasis is protein flexibility.[21,33−35] There is increasing experimental evidence to indicate that when different ligands are bound to the same protein, significant conformational changes occur in that protein. This is especially true with the conformations of side chains in the vicinity of the ligand. The majority of older docking methods assume a rigid protein, and consequently their accuracy is limited for many systems.

An example of the importance of protein flexibility in docking is demonstrated by the human estrogen receptor ligand-ligand binding domain when bound with different ligands. Table 1 shows the difference in the χ_1 and χ_2 angles

* Corresponding author phone: (858)799-5000; fax: (858)799-5100; e-mail: amaynard@accelrys.com.

† These authors contributed equally to this work.

10.1021/ci800081s CCC: $40.75
Published on Web 09/25/2008

Table 1. Difference in Torsion Angles χ_1 and χ_2 between Estrogen Receptor Structures 1err and 3ert

residue	$\Delta\chi_1$	$\Delta\chi_2$
LEU345	−8.6	19.4
LEU346	−3.1	13.6
THR347	−17.2	0.0
ARG352	−10.2	−12.1
LEU354	3.1	−12.5
GLU385	−15.6	−14.0
ILE389	−14.2	−11.2
TRP393	6.1	−18.2
ARG394	−15.0	−23.3
MET421	−107.7	47.3
HIS524	128.5	−31.3
LEU525	−69.3	44.5
LEU536	120.2	−94.4
TYR537	179.6	82.0
ASP538	−14.0	−55.0
LEU539	32.4	−10.6

of selected residues of two estrogen receptor crystal structures, PDB codes, 1err (complexed with raloxifene) and 3ert (complexed with 4-hydroxytamoxifen). The corresponding structures are shown in Figure 1. Seven residues undergo a significant ($\Delta\chi > 30°$) side-chain conformational change depending on whether they are docked with raloxifene or 4-hydroxytamoxifen.

The need for accuracy in docking methods is particularly acute in the case of virtual high throughput screening where thousands of ligands may be docked to a given protein. Virtual high throughput screening is often employed to develop hit rate plots (for an example see refs36−38). Clearly, ignoring protein flexibility in protein systems like the estrogen receptor could lead to misleading results when employing a wide variety of ligands.

In this paper, we propose a novel ligand docking approach that takes into account flexibility of the ligand and selected protein side chains. The method uses a previously described algorithm[39] to predict different protein side-chain conformations corresponding to a docked ligand. We present results using this method and demonstrate the power and flexibility of our approach.

METHODS

1. Preparation of Protein and Ligand Models. The following eight protein systems were selected for validation: thymidine kinase, human estrogen receptor, cyclin-dependent kinase 2 (CDK2), cyclooxygenase 2 (COX2), neuraminidase, thermolysin, HIV-1 RT, and factor Xa. A summary of the protein−ligand systems used in the study is shown in Table 2, including the PDB entry code and the resolution for each structure. These were chosen because they are systems where protein side-chain conformations are known to vary when different ligands are bound. For each protein system, two protein−ligand complexes were chosen for which crystal structures were available. Both native and cross docking experiments were performed. Cross docking was also performed with a rigid protein ignoring protein flexibility to gauge the importance of incorporating protein flexibility.

Protein structures were retrieved from Protein Data Bank (www.pdb.org). From each PDB file, the protein chain A was extracted for the study, water molecules were removed, and metal ions were retained. The 'Protein Health' tool in Discovery Studio (Accelrys, Inc.) was employed to identify problem areas in the protein structures, including missing atoms, atoms with alternate conformations, incorrectly named atoms, and incorrect bond orders. The protein "Clean" tool in Discovery Studio was used to correct minor problems: missing side-chain atoms which were added in an extended conformation. The first conformation is retained for atoms with alternate coordinates, while other conformers were removed. Hydrogen atoms were added to the protein structure if they were missing in the original PDB file. Standard protonation states of the titratable residues were adopted to produce the charged acidic and basic side chains as well as N- and C-termini. The ligand was extracted, bond orders and formal charges on each atom were manually inspected and corrected for correct chemistry, and hydrogen atoms were added. Protonation and tautomerization states were assigned according to published data.

Although we have only considered chains A, we recognize that it would be interesting to investigate the differences between the various chains for each protein system. However, we have not attempted to do that in this study. The input files prepared for this study are available from the authors upon request.

2. Low Energy Protein Side-Chain Generation - ChiFlex. ChiFlex is a CHARMm based molecular mechanics method that is used to generate a set of low energy side-chain conformations for a protein structure. It is a generalized version of the ChiRotor algorithm.[39] While ChiRotor returns a single protein conformation corresponding to the lowest energy structure, ChiFlex generates an ensemble of low energy protein conformations with varied side-chain conformations. A summary of the algorithm is provided in Figure 2.

The following procedure was employed to identify side chains to be varied using ChiRotor and ChiFlex: With the ligand in X-ray pose in place, all receptor residues within 3.5 Å from any of the ligand atoms were selected. Gly, Ala, Val, and Pro were unselected. If only backbone atoms of the selected residues are within 3.5 Å from the ligand, then those residues were also unselected. The computing time for ChiFlex increases with the number of side chains. In general, we attempted to keep the number of residues to less than 10. In some cases to help us achieve this, side chains with less than 3 atoms within 3.5 Å were also unselected. The same residues were selected for all structures of a given protein.

3. Ligand Conformation Generation. The second stage of the algorithm is to provide a set of low energy ligand conformations to be docked into the protein. For this purpose we used the CatConf component in Discovery Studio (Accelrys, Inc.), a well validated program widely used for generating diverse sets of low energy conformations.[40−44] In the current workflow of this method we used CatConf in the 'BEST' mode to generate up to 255 ligand conformations and used these for subsequent docking.

4. Hotspot Identification. The next step of the method involves identification of protein ligand interactions that will be used to guide the initial ligand placement. The LibDock[31] program was used for this purpose to compute the locations in the binding site where ligand polar and nonpolar groups may be placed to obtain favorable interactions between the ligand and protein. These are referred to as hotspots. The

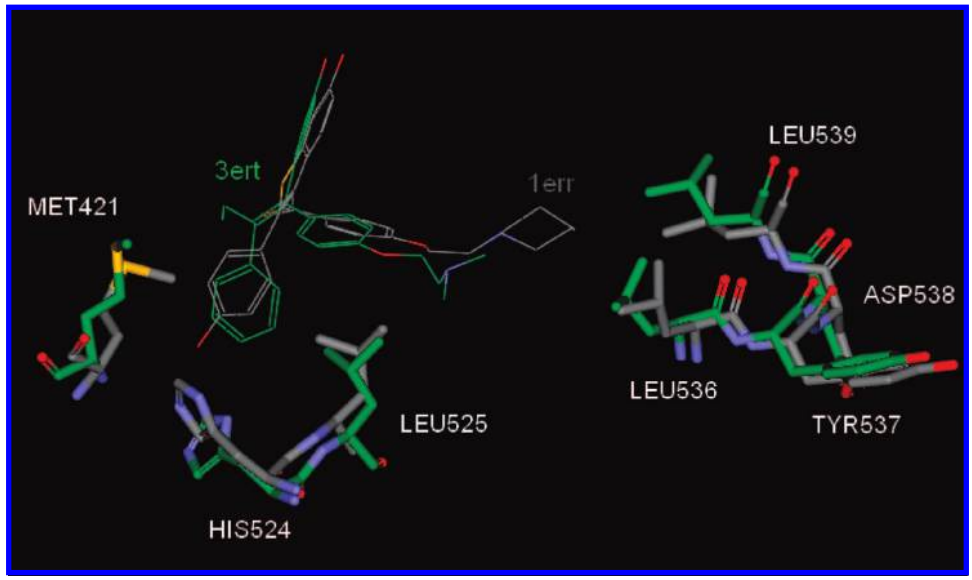

Figure 1. Superimposed estrogen binding site structures containing the ligands for 3ert (4-hydroxytamoxifen) and 1err (raloxifene). Only seven residues close to the ligand structures are shown.

Table 2. Protein Systems with PDB Entry Code and Resolution (Å)

protein system	PDB entry	resolution
thymidine kinase	1kim	2.14
	1ki4	2.34
estrogen	1err	2.60
	3ert	1.90
	1sj0	1.90
CDK2	1aq1	2.00
	1dm2	2.10
COX2	1cx2	3.00
	3pgh	2.50
neuraminidase	1nsc	1.70
	1a4q	1.90
HIV-1 RT	1rev	2.60
	1s1x	2.80
	1fk9	2.5
	1rth	2.2
factor Xa	1ksn	2.1
	1xka	2.3
thermolysin	1kr6	1.8
	1kjo	1.6

method works by first constructing a grid in the binding site. The binding site itself is identified by specifying a sphere of given radius located in the active site. The center of the sphere is placed at the geometric center of the ligand in the X-ray structure. In the protein systems considered, a radius of 12 Å covered the binding site being targeted. At each grid point a score is computed by placing a probe of a given type (namely polar or nonpolar) at the grid point and evaluating the interactions of the probe with the receptor. The list of hotspots is derived by clustering and retaining the top scoring hotspots. This is performed for both polar and nonpolar probes. For each protein conformation generated by ChiFlex, a list of hotspots is calculated.

5. Ligand Placement. This step uses the LibDock[31] program to generate ligand poses consistent with hotspots identified for polar and nonpolar groups. Each ligand conformation is placed by rigidly aligning the ligand to the hotspots. Three ligand-atom hotspots are matched to three binding site hotspots taking into account interatom distances and the interhotspots distances. If the alignment results in significant overlap with the receptor, it is rejected. If the alignment is accepted, it is scored. Five top scoring poses are retained for each ligand.

6. Protein -Chain Reconstruction - ChiRotor. The next step, side-chain reconstruction at the binding site, is critical for achieving a successful docking. ChiRotor[39] is a CHARMm based program that performs refinement of specified protein side chains. When a ligand is included in the calculation, the CHARMm energy calculations and energy minimizations will also include interactions of the protein side-chain atoms with the ligand atoms. The algorithm starts with a 3D structure of a protein with or without a ligand present. Side-chain atoms of user-selected residues are removed, rebuilt, and then refined using a CHARMm-based energy minimization. A more detailed outline of the ChiRotor method is provided in Figure 3. Since the specified side-chain atoms are first deleted and rebuilt using this algorithm, the output of this method is independent of the initial conformations of the specified side chains.

7. Docking Refinement - CDOCKER. CDOCKER[45,46] is a CHARMm-based docking refinement application that employs high temperature molecular dynamics and energy minimizations to perform ligand pose sampling and refinement. This algorithm assumes a rigid protein and permits only the ligand to be flexible. Since our workflow already uses LibDock for initial docking placement, we omit the conformation and pose randomization phases of the CDOCKER script and instead reduce the final refinement stage to a simulated annealing phase followed by a minimization.

8. Workflow Implementation. We implemented each of the above six steps as Pipeline Pilot components. Pipeline Pilot is a data pipelining environment from Accelrys Inc. that enables rapid development and execution of custom pipelines. The software allows rapid integration of existing computational engines such as CHARMm and LibDock. As part of this study we explored different variants of the final workflow and were also able to examine different sampling schemes to assess the optimal number of protein conformations to use in the initial ChiFlex stage of the workflow.

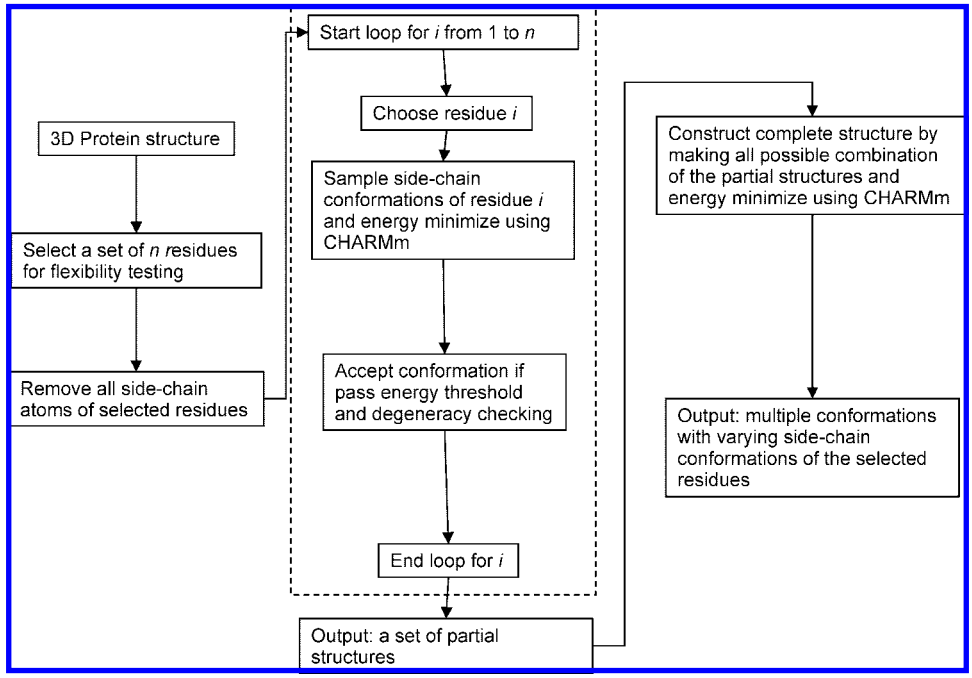

Figure 2. Summary of the ChiFlex algorithm.

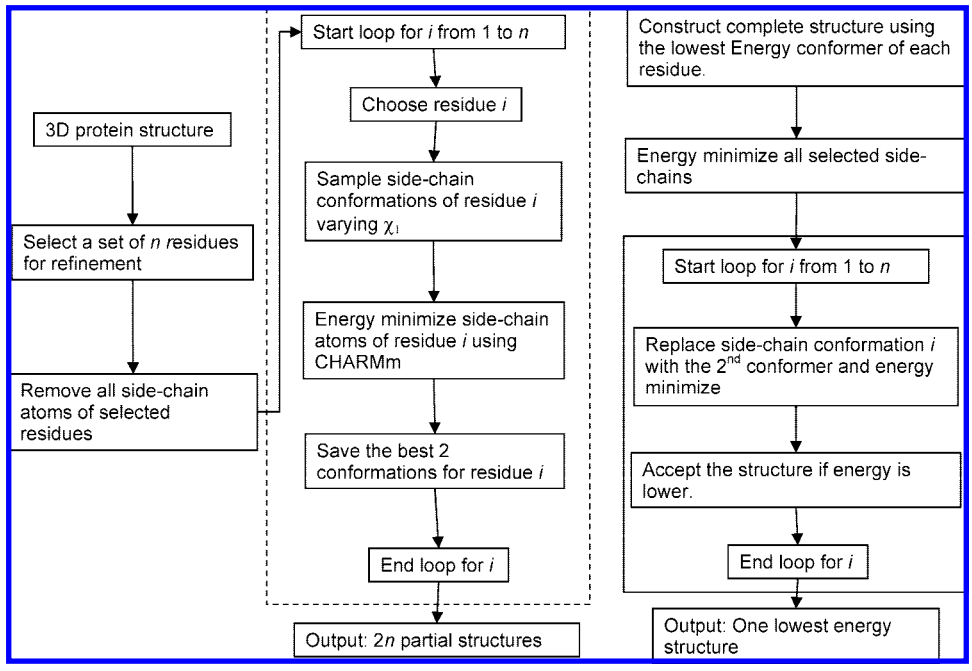

Figure 3. Summary of the ChiRotor algorithm.

By exploring different workflows in this way, we identified an optimal workflow for performing induced fit protein−ligand docking. This workflow is summarized in Figure 4. The ChiFlex component was employed to generate 20 diverse protein conformations. For each protein conformation, hotspots were computed. Independently, ligand conformations were generated by the CatConf component (up to 255). The LibDock component generated several docked poses for each protein conformation. However, only the top five ligand poses were retained. Therefore, for a given ligand up to 100 poses are generated. The ligand positional and orientational samplings are explored by using the hotspots and ligand placement methods in the docking steps 2 and 4. Even though the initial side-chain conformations are as determined by ChiFlex in step 1, these side-chain conformations may be altered in step 5 consistent with the initial ligand placement.

Due to its componentized design it is also straightforward to deploy a parallelized version of this method in a high performance computing environment (e.g., a Linux cluster). Figure 5 summarizes the parallelization strategy for this method.

9. Calculation of rmsd. For each protein system, we used two or more crystal structures corresponding to the same protein bound with two or more ligands. Referring to two structures as Aa and Bb where A and B are protein structures and a and b are corresponding ligand structures in those crystal structures, four dockings may be considered: two native dockings, Aa and Bb, and two cross dockings, Ab

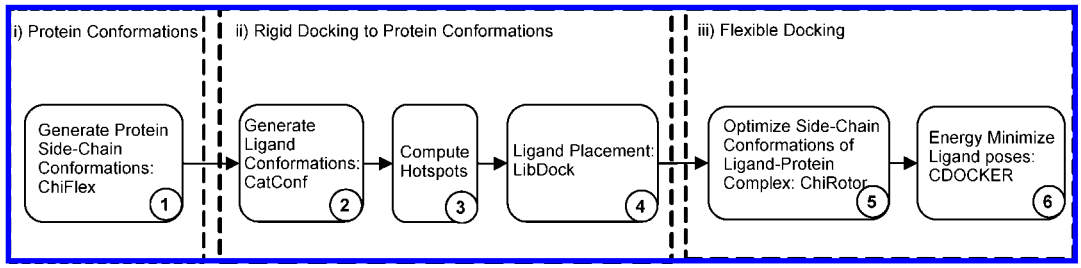

Figure 4. Summary of the flexible docking workflow. There are three distinct phases: (i) protein side-chain conformation side-chain generation, (ii) generation of ligand conformations and rigid ligand placement into each protein conformation, and (iii) induced fit protein flexibility and final pose refinement. Note that protein flexibility is introduced in the first and third phases.

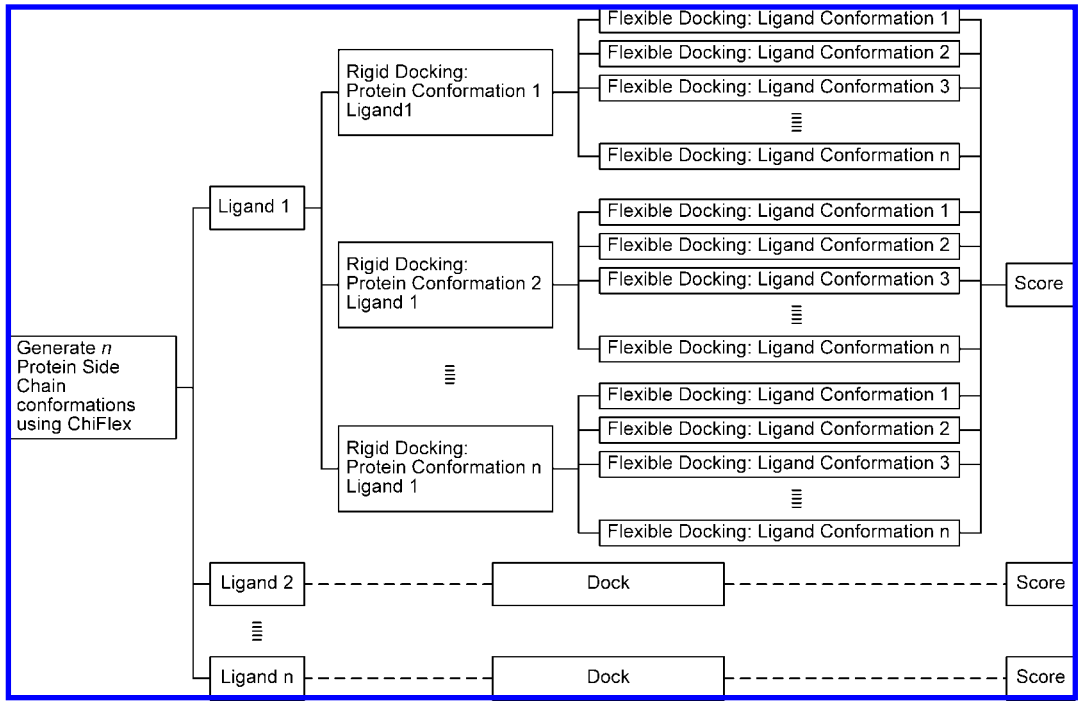

Figure 5. Parallelization strategy for a virtual screening study. In a typical scenario multiple ligands are docked into the same protein structure. As a result the first phase of protein side-chain conformation generation need only be performed once. Our study indicates that ten initial protein conformations are sufficient to recover an accurately docked ligand. For each protein ligand system the rigid docking phase is used to generate 20 initial ligand placements per conformation. The final flexible docking phase takes approximately five minutes per ligand conformation. Note that the third phase, where most of the time is spent, can be very effectively parallelized to minimize total run times.

and Ba. The quality of the native dockings is measured by computing the rms difference between computed ligand pose and the X-ray pose. In the cross dockings, a reference ligand pose is generated from the two crystal structures employed. For instance, to evaluate the cross docking Ab, the crystal structure Bb is overlaid to Aa by mapping the part of the protein backbone near the binding site of B to that of A. For this purpose, the backbone atoms of residues within 4 Å of the X-ray ligand in A were selected. The resulting overlaid pose of b is taken as the reference pose with respect to which ligand rms differences are computed.

RESULTS

ChiFlex. Validation of the methodologies used in the CatConf, CDOCKER, and LibDock steps in a docking context have already been reported elsewhere.[31,40−43,45,46] The ChiFlex program has not been published previously but may be thought of as a generalization of the ChiRotor algorithm. The role of ChiFlex is for generating reasonable alternative side-chain conformations for a given protein. While exploring various conformations of specified side chains, ChiFlex is also able to gauge the flexibility of the specified side chains. Highly flexible side chains result in a large number of low energy conformations within a given energy window from the lowest energy conformer. Figure 6 shows ChiFlex relative energy as a function of conformation number. A steep rise in energy corresponds to proteins with the least flexible side chains. A gradual rise in the energy is seen in the cases where ChiFlex is able to find larger numbers of side-chain conformations by various combinations of side-chain torsion angles corresponding to a low energy. One can also see from Figure 6 that there is a large difference in the number of conformations within an energy threshold between the two protein systems. On the other hand the differences within protein systems are small, at least for a low energy threshold. The difference in flexibility of side chains between the two systems is qualitatively captured well by this calculation.

ChiRotor. The function of ChiRotor is central to our workflow, and it is useful to describe the validation of

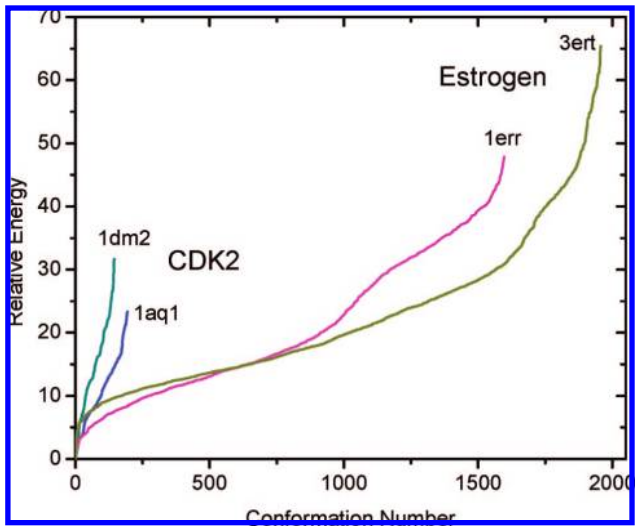

Figure 6. Relative ChiFlex energy vs conformation number in the absence of a ligand for estrogen and CDK2. Conformations have been sorted by ChiFlex energy and the lowest energy conformation is taken as zero. For the CDK2 systems (1aq1, 1dm2) there are fewer conformations and the steep increase in energy indicates a relative rigid receptor, whereas for estrogen (1err, 3ert) more conformations are found. Note that small differences within in each protein system exist due to differences in the backbone structure.

ChiRotor in the context of the flexible docking workflow in more detail. Since in the current implementation the protein is held rigid during the final ligand minimization step, it is important that the side chains are optimally oriented for the ligand. It is necessary to verify that ChiRotor builds optimal side-chain conformations. This may be performed in the presence as well as in the absence of a ligand in the binding site. We tested this in both cases. We give two examples here, one with and one without the ligand in the X-ray pose. In both cases ChiRotor was employed to rebuild residues in the binding site.

An example of a protein analyzed in the absence of a ligand is the CDK2. Using the 1aq1 crystal structure of CDK2, we used ChiRotor to rebuild several residues in the binding site. Fifteen residues in the vicinity of the inhibitor in that crystal structures were manually selected (see Table 3). Table 3 shows the rms differences between calculated and crystal structure residues. It may be seen that all selected residues are properly reconstructed except Leu83, Gln131, and Asn132. These three residues are at the extremities of the binding site. Figure 7 illustrates the binding-site residues with the bound ligand hidden. Based on these results, we conclude that this is an acceptable performance by ChiRotor on this system.

For 12 crystal structures of thymidine kinase, we have reconstructed side chains using ChiRotor in the presence of the ligand in the X-ray pose. For this validation we selected all residues within 4 Å of the crystal ligand. Table 4 shows the overall side-chain rms difference to crystal structures for these protein structures. All rms deviations are less than 2 Å. This demonstrates the reliability and robustness of the side-chain rebuilding of as many as about 15 residues in the vicinity of the ligand by ChiRotor in the presence of various ligands.

Accuracy of the Workflow. As described in the Methods section, eight protein systems were considered for flexible docking runs. For each ligand the lowest rms difference in the top 5 hits is reported. For most of the protein systems, the workflow performs well and returns a ligand pose close to that of the crystal structure (within 2 Å) with the top five hits based on CHARMm-based energy score. The rigid docking was performed by employing LibDock with one protein conformation retaining the top five poses for the rmsd calculation. Table 5 shows rms differences obtained with a rigid receptor obtained using LibDock. For these cases, the observed rms differences are large, demonstrating the importance of considering side-chain flexibility in these systems.

Table 3. Lowest Energy Side-Chain Conformations in CDK2 Predicted by ChiRotor in the Absence and Presence of the Staurosporine Ligand[a]

residue	rmsd (Å) presence of ligand	rmsd (Å) absence of ligand
ILE10	0.07	0.08
PHE80	0.26	0.21
LYS33	0.52	0.37
VAL64	0.20	0.16
PHE80	0.21	0.24
GLU81	0.18	0.70
PHE82	0.40	0.40
LEU83	1.44	1.44
HIS84	0.08	0.50
GLN85	0.42	0.46
ASP86	0.88	0.81
GLN131	2.66	2.58
ASN132	1.96	1.95
LEU134	0.46	0.38
ASP145	1.02	0.64

[a] The RMSD is for several side chains in the binding site with respect to crystal structure with the ligand.

Optimization for Virtual Screening Studies. Because of the additional conformational dimension in a ligand docking procedure that allows for protein flexibility, computational times can be potentially overwhelming compared to the time required for a rigid protein docking. The workflow described here overcomes this challenge in two ways.

First, the protein side-chain sampling strategy used to explore flexibility in the protein is extremely efficient. The computing times of ChiFlex and ChiRotor depend on the number of side-chain residues to be varied. On a PC with a 2 GHz processor, (Dell Precision M70 with Pentium processor with 2 GHz and 2GB of RAM) ChiFlex takes 4 min if three side chains are varied, 1 h if 6 side chains are to be varied, and several hours if more than 6 residues are considered for flexibility. Table 6 shows the side-chain residues selected for variation by ChiFlex in each protein system. These are large residues in the vicinity of the binding site. As one selects more side-chain residues to be flexible, ChiFlex generates larger numbers of protein conformations. Of course the ChiFlex algorithm needs to be executed only once for a protein with a specific binding region. Generated protein conformations may be reused for docking of several ligands such as in virtual high throughput screening work. ChiRotor is a fast algorithm; ChiRotor takes an average of four minutes per ligand pose to rebuild specified protein side chains (*cf.* Table 4).

The second area for applying this method in a high performance computing environment comes from its modular design. The bulk of the computational time is spent in the

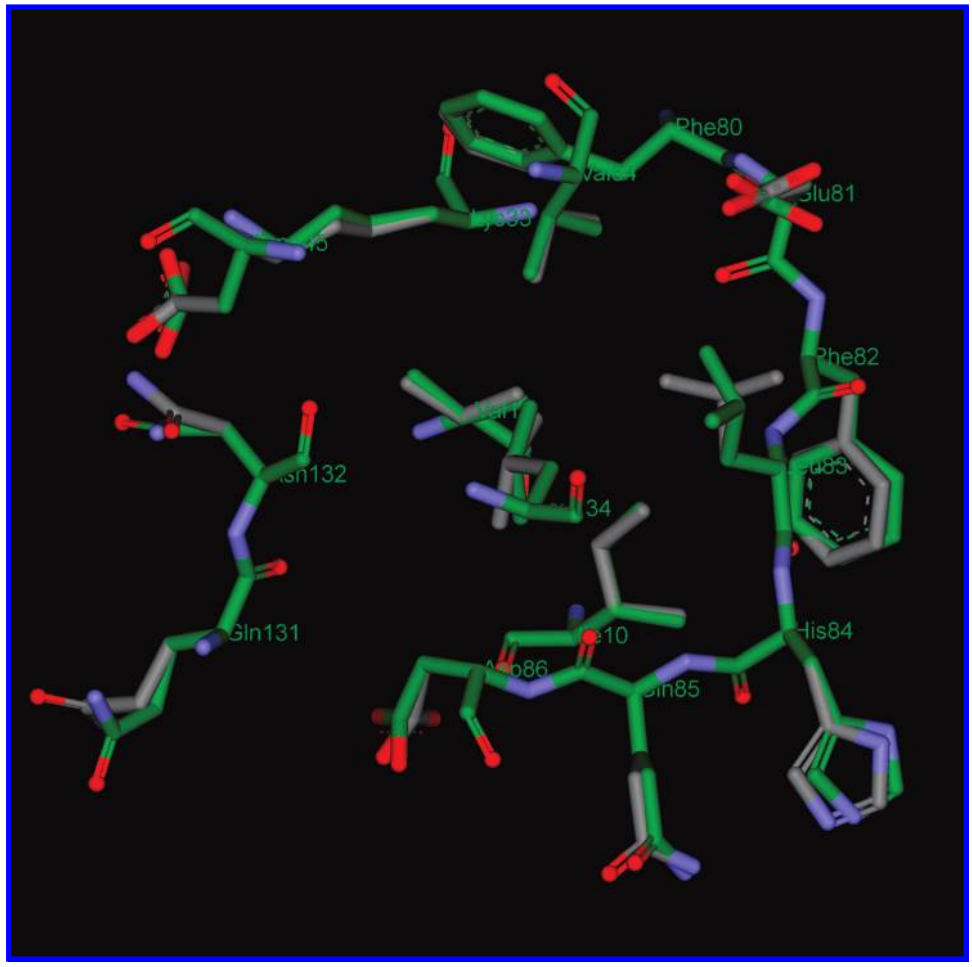

Figure 7. Calculated side-chain conformations and X-ray structure of CDK2 in the presence of the ligand (not shown). The side chains calculated with ChiRotor are shown in dark green.

Table 4. RMSD Values of Side-Chain Reconstructions of Thymidine Kinase Structures Using ChiRotor[a]

complex	rmsd (Å)	time (s)
1e2k	1.09	176
1e2m	1.85	216
1e2p	1.34	211
1ki2	1.90	215
1ki3	1.35	219
1ki4	1.16	322
1ki6	0.76	345
1ki7	0.79	379
1kim	0.74	270
1qhi	1.48	435
2ki5 -1	1.30	274
2ki5 -2	1.02	343

[a] Results indicate the overall side-chain RMS difference to crystal structures for these protein structures.

ChiRotor side-chain refinement step. However, since the computing time for a single ligand conformation is relatively short, this phase (including side-chain refinement and ligand pose minimization) can be very effectively parallelized. For example, on an eight node dual processor Linux cluster it is possible to reduce the actual run time of this phase to less than one hour.

CONCLUSION

The modeling of protein flexibility in ligand-docking studies has attracted increasing attention among the research community in recent years. In this work we have emphasized its importance for a set of well characterized systems and have presented a flexible docking workflow that predicts the bound ligand pose with high accuracy. The diversity of the target systems studied and general accuracy of the results emphasizes the robustness of the method.

It is useful to compare the method employed here with other docking methods that consider receptor flexibility. The term 'ensemble docking' has been used to refer to different types of docking: One where a ligand is docked to each of a set of alternate protein conformations and the other where a ligand is docked *simultaneously* to a collection of receptor structures. The former is often called ***sequential*** or ***serial*** docking, while some flavors of the latter are termed ***soft-docking***. In soft-docking, often a modified energy function is employed to account for receptor flexibility and thereby enabling a single docking of a ligand to collection of receptor structures.[47] The collection of receptor structures is usually obtained from NMR NOE studies as well as crystal structures of the same protein obtained with different ligands bound.[48,49] When compared to these methods, the method we have adopted in this paper more closely resembles a sequential or serial method with an important difference. In a sequential method one usually docks a ligand to several rigid snapshots of the receptor. In our method even though ChiFlex may be viewed as generating several rigid snapshots of the receptor, the ChiRotor step in the workflow makes modifications to

Table 5. Lowest RMS Differences to the X-ray Structure (Å) for the New Flexible Docking Algorithm (Flexible) and the Rigid Docking Algorithm LibDock (Rigid)

protein system	ligand	receptor	flexible lowest rmsd	rigid lowest rmsd
thymidine kinase	1kim	1ki4	1.2	1.1
	1ki4	1kim	1.2	2.7
estrogen	1err	3ert	1.2	5.7
	3ert	1err	1.0	5.2
	1sj0	1err	1.7	4.9
	1sj0	3ert	1.8	5.1
CDK2	1aq1	1dm2	0.9	5.7
	1dm2	1aq1	0.7	3.6
COX2	1cx2	3pgh	2.0	6.1
	3pgh	1cx2	1.9	5.1
neuraminidase	1nsc	1a4q	1.6	1.8
	1a4q	1nsc	1.7	4.2
HIV-1 RT	1rev	1rth	1.5	4.9
	1s1x	1rth	0.5	5.3
	1s1x	1rev	0.6	4.2
	1s1x	1fk9	0.5	3.7
	1fk9	1rev	0.4	4.8
	1fk9	1rth	0.6	5.5
	1fk9	1s1x	0.7	5.6
	1rth	1rev	1.5	3.7
	1rth	1s1x	1.1	4.3
	1rth	1fk9	1.3	6.8
factor Xa	1ksn	1xka	2.0	7.0
	1xka	1ksn	2.0	8.6
thermolysin	1kr6	1kjo	1.2	4.9
	1kjo	1kr6	3.0	5.2

Table 6. Selection of Residues for Generating a Set of Protein Conformations using ChiFlex

system	residues selected for ChiFlex
thymidine kinase	His58, Glu83, Ile100, Gln125, Arg163, Tyr172, Arg222, Glu225
estrogen receptor	Met421, His524, Leu525, Leu536, Tyr537, Asp538, Leu539
CDK2	Ile10, Phe80, Glu81, Leu83, Gln85, Asp86, Asn132, Leu134, Asp145
COX2	His90, Arg120, Tyr355, Arg513, Phe518
neuraminidase	Arg115, Asp148, Arg149, Glu274, Arg291, Tyr408
thermolysin	Asn112, Glu143, Tyr157, Glu166, His231
HIV-1 RT	Lys101, Lys103, Val179, Tyr188, Tyr318
factor Xa	Glu97, Tyr99, Arg143, Glu147, Phe174, Gln192, Trp215

the side chains when a ligand is in the binding site. Furthermore, that step is not just a side-chain refinement, but it is a side-chain rebuilding quite capable of exploring alternate energetically attractive side-chain conformations independent of the starting side-chain conformation from ChiFlex. Thus this method is an *induced*-fit model instead of a pre-existing equilibrium model.

Significantly, the methodology presented here requires minimal manual input other than identification of the protein binding site. All other parts of the workflow, including protein preparation, can be fully automated if desired. This aspect of the method is of particular value in high throughput screening studies where large numbers of ligands are being surveyed against a single protein system.

An important feature of this workflow is that the components on which the method is implemented are based on discrete, scientifically validated units. Together these provide a rich toolkit for the computational scientist. While this specific workflow is shown to perform well in the systems considered here, it should be noted that the component based approach allows for custom modification as demanded by specific application. For example, if more extensive exploration of ligand pose minima is desired, one can replace the final *in situ* energy minimization with a CDOCKER component that performs high temperature dynamics followed by annealing by merely changing the input parameters of CDOCKER.

Another variant of the workflow that could be explored is the initial generation of the protein conformations. In contrast to the automated conformation generation used in this study, specific protein conformations, such as those generated from a molecular dynamics simulation, NMR structures, or rotamer libraries, could be used to generate the initial snapshots of receptor structures instead of ChiFlex as employed here. These are certainly reasonable alternatives. However, with the exception of experimental structures, these methods have to deal with the combinatorial problem of sampling side-chain conformations. A difference between the ChiFlex method and a rotamer library based method will be in the number of side-chain conformations considered for energy optimization for a given side-chain residue. Both methods however have to deal with the combinatorial problem. We have employed a coarser side-chain search in ChiFlex than normally employed in a rotamer library based approach, since in the present method the snapshots are only initial conformations and the ChiRotor step modifies the side-chain conformations when the ligand is placed.

Another advantage of using ChiFlex for the conformation generation is that a consistent molecular mechanics based approach is adopted for treatment of the protein in all stages of the workflow. One example of this benefit concerns the treatment of cofactors. The only system for which we were unable to recover a pose within 2 Å of the X-ray structure was for one of the cross dockings in thermolysin, the 1kjo ligand into 1kr6 receptor (see Table 5). Such large RMSDs have also been reported by others for this system.[34] This is probably a reflection of poor parametrization of zinc in this environment, which may be resolved by deriving new parameters or using a higher level of theory in such cases.

However, although we did not resolve this completely in this study, the use of a single force-field in our method allows us to pursue further inquiry into force field-related factors that may tend to favor the generated poses. Indeed, work on other CHARMm based force-fields than the one used in our study has emphasized treatment of specific cofactors and potentially represents an interesting way forward.[50]

In addition to exploring the integration of different steps into the workflow, one other avenue for future development of this method concerns explicit solvent effects, which have been ignored in the current study. This version of the method will need further work if water molecules are to be considered in the docking process. While conserved functional water molecules may be included during docking a single ligand, in the case of docking a variety of ligands, mobility of water molecules will have to be taken into account.

ACKNOWLEDGMENT

The authors thank and acknowledge valuable discussions with Dr. Frank Brown and Dr. Dipesh Risal.

REFERENCES AND NOTES

(1) Klebe, G. Virtual Screening: Scope and Limitations. In *Virtual Screening in Drug Discovery*; Alvarez, J., Shoichet, B., Eds.; Taylor & Francis: Boca Raton, FL, 2005; *Chapter 1*, pp 1−24.
(2) Perola, E.; Walters, W. P.; Charifson, P. S. An Analysis of Critical Factors affecting Docking and Scoring. In *Virtual Screening in Drug Discovery*; Alvarez, J., Shoichet, B., Eds.; Taylor & Francis: Boca Raton, FL, 2005; *Chapter 3*, pp 47−86.
(3) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput.-Aided Mol. Des.* **1997**, *18*, 1175–1189.
(4) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151–166.
(5) Sun, Y.; Ewing, T. J.; Skillman, A. G.; Kuntz, I. D. CombiDOCK: structure-based combinatorial docking and library design. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 597–604.
(6) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449–62.
(7) Wang, J.; Kollman, P. A.; Kuntz, I. D. Flexible ligand docking: a multistep strategy approach. *Proteins* **1999**, *36*, 1–19.
(8) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.
(9) Trosset, J. Y.; Scheraga, H. A. Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 8011–8015.
(10) Sudarsanam, S.; Virca, G. D.; March, C. J.; Srinivasan, S. An approach to computer-aided inhibitor design: application to cathepsin L. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 223–233.
(11) 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.
(12) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. Flexible ligand docking using a genetic algorithm. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 113–130.
(13) Pattabhiraman, N.; Levitt, M.; Ferrin, T. E.; Langridge, R. Computer Graphics in Real-time Docking with Energy Calculation and Minimization. *J. Comput. Chem.* **1985**, *6*, 432–436.
(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) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-Based Energy Evaluations. *J. Comput. Chem.* **1992**, *13*, 505–524.
(16) Makino, S.; Ewing, T. J.; Kuntz, I. D. DREAM++: flexible docking program for virtual combinatorial libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 513–532.
(17) Hahn, M. Three-Dimensional Shape-Based Searching of Conformationally Flexible Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 80–86.
(18) Luty, B. A.; Wasserman, Z. R.; Stouten, P. F. W.; Hodge, C. N. A Molecular Mechanics/Grid Method for Evaluation of Ligand-Receptor Interactions. *J. Comput. Chem.* **1995**, *16*, 454–464.
(19) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 435–451.
(20) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395–407.
(21) Jones, G.; Willett, 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.
(22) Gschwend, D. A.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 123–132.
(23) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1–5.
(24) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. *J. Med. Chem.* **1988**, *31*, 722–729.
(25) DesJarlais, R. L.; Sheridan, R. P.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **1986**, *29*, 2149–2153.
(26) DesJarlais, R. L.; Dixon, J. S. A shape- and chemistry-based docking method and its use in the design of HIV-1 protease inhibitors. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 231–242.
(27) Bacon, D. J.; Moult, J. Docking by least-squares fitting of molecular surface patterns. *J. Mol. Biol.* **1992**, *225*, 849–858.
(28) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573–591.
(29) Fradera, X.; Knegtel, R. M. A.; Mestres, J. Similarity-Driven Flexible Ligand Docking. *Proteins* **2000**, *40*, 623–636.
(30) Goldman, B. B.; Wipke, W. T. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). *Proteins* **2000**, *38*, 79–94.
(31) Diller, D. J.; Merz, K. M., Jr. High throughput docking for library design and library prioritization. *Proteins* **2001**, *43*, 113–124.
(32) Kirchhoff, P. D.; Brown, R.; Kahn, S.; Waldman, M.; Venkatachalam, C. M. Application of Structure-Based Focusing to the Estrogen Receptor. *J. Comput. Chem.* **2001**, *22*, 993–1003.
(33) Claussen, H.; Buning, C.; Rarey, M.; Lengauer, T. FlexE: efficient molecular docking considering protein structure variations. *J. Mol. Biol.* **2001**, *308*, 377–395.
(34) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49*, 534–553.
(35) Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discovery* **2003**, *2*, 527–541.
(36) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759–4767.
(37) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.
(38) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(39) Spassov, V. Z.; Yan, L.; Flook, P. K. The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: a side-chain prediction algorithm based on side-chain backbone interactions. *Protein Sci.* **2007**, *16*, 494–506.
(40) Dimitris, A. Conformational Sampling of Bioactive moleucles. *J. Chem. Inf. Model.* **2007**, *47*, 1067–1086.
(41) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection; CONAN. *J. Comput. Chem.* **2003**, *24*, 10–20.
(42) Smellie, A.; Kahn, S.; Teig, S. Analysis of Conformational Coverage. 1.Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285–294.
(43) Smellie, A.; Kahn, S.; Teig, S. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295–304.
(44) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422–430.
(45) Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCK-ER-A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549–1562.
(46) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47*, 45–55.
(47) Knegtel, R. M.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266*, 424–440.
(48) Huang, S. Y.; Zou, X. Efficient molecular docking of NMR structures: application to HIV-1 protease. *Protein Sci.* **2007**, *16*, 43–51.
(49) Huang, S. Y.; Zou, X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins* **2007**, *66*, 399–421.
(50) Stote, R. H.; Karplus, M. Zinc Binding in Proteins and Solution: A Simple but Accurate Nonbonded Representation. *Proteins* **1995**, *23*, 12–31.

CI800081S