Influence of Conformational Flexibility on Complexation-Induced Changes in Chemical Shift in a Neocarzinostatin Protein−Ligand Complex

Marina Cioffi,[†] Christopher A. Hunter,*[,†] and Martin J. Packer[‡]

Department of Chemistry, Centre for Chemical Biology, Krebs Institute for Biomolecular Science, University of Sheffield, Sheffield S3 7HF, U.K., and AstraZeneca, Alderley Park, Cheshire, SK10 4TG, U.K.

Received January 25, 2008

In this paper is described an analysis of the effects of protein flexibility on the observed CIS values and the impact on the accuracy of 3D structures determined using a ^{1}H NMR CIS approach. The effects of protein conformational mobility have been investigated by using a set of different protein structures as starting points for the calculation: the unbound X-ray crystal structure, the unbound NMR solution structure, and the bound NMR solution structure of the protein. The results indicated that loop movement does have a significant impact on the quality of the structure generated by the CIS structure determination methodology. The implementation of methods to treat loop flexibility within our protocol, however, did not improve the results for calculations based on the unbound protein frame.

Introduction

Determining the interactions that occur when small molecules (ligands) bind to receptors and identifying the three-dimensional structures of the complexes are important issues for understanding and modulating protein function. NMR spectroscopy is often the method of choice for determining the structures of protein complexes. NMR measurements such as NOEs, *J*-coupling values, and residual dipolar couplings provide structural information about the orientation of ligands within the protein active site.[1] Complexation-induced changes in chemical shift (CIS[a]) provide a simple approach to mapping the intermolecular interface of a complex and locate the binding pocket in a protein.[2] The SAR by NMR method that makes use of CIS mapping has become an important tool in structure-based drug design.[3,4] CIS mapping can be used in a qualitative way for identifying the binding site or, quantitatively, providing more detailed information on the structure of the complex.[5–8]

The elucidation of high-resolution structures of complexes using classical NMR methods is far from routine. It requires complete assignment of all of the signals due to the ligand, the protein backbone, and the side chains and observation of a sufficient number of intermolecular NOEs to accurately locate the position and orientation of the ligand in the binding pocket. Integration of NMR chemical shift information as additional distance constraints in the structure refinement process can therefore be useful.[9,10] CIS data are relatively straightforward to collect and have been used to investigate host−guest complexes, protein folding, and protein complexes.[11–14] Reliable prediction tools are available to estimate the CIS expected for a particular 3D structure, and comparison with the experimental data allows an assessment of the accuracy of that structure.[5,15,16] Thus, CIS-based scoring functions have been developed using the difference between predicted and experimental CIS data to rank predicted protein−ligand binding modes or to introduce ambiguous restraints to restrict the location of the ligand during sampling.[17,18]

Although the accuracy of the structure determination and prediction methods for protein−ligand complexes has improved enormously, one of the main challenges is dealing with molecular flexibility and conformational changes. Protein−ligand recognition is a dynamic event in which both protein and ligand can change conformation in order to minimize the total free energy change on association.[19] In many docking methods, the ligand is treated as flexible but the protein conformation is restricted and, for practical reasons, it is often assumed to be rigid. This approximation is due to the combinatorial explosion on the size of the search space when both ligand and protein flexibility are considered.[20] Erickson et al.[21] attempted to explore the effect of protein flexibility on the accuracy of a docking calculation and found that the protein conformation is important to accurately dock ligands. They showed that the accuracy of the structure is correlated with the degree of protein movement allowed in the active site,[22,23] and this problem will be exacerbated in algorithms that make use of intermolecular distance restraints between specific atoms to describe the information contained in experimental CIS data.[12]

Many methods have been developed that allow partial protein flexibility to address this problem.[24,25] Protein flexibility can be treated implicitly by allowing interpenetration of molecules or explicitly by exploring all possible conformations, allowing

* To whom correspondence should be addressed. Phone: +44 114 2229476. Fax: +44 114 2229346. E-mail: C.Hunter@shef.ac.uk.

† University of Sheffield.

‡ AstraZeneca.

[a] Abbreviations: CIS, complexation-induced chemical shift changes; BB NH, backbone amide protons; $RMSD_{BB}$, root mean squared difference of the backbone heavy atoms; $RMSD_{LIG}$, root mean squared difference of the ligand heavy atoms; *n*-$RMSD_{CIS}$, normalized root mean squared difference between the calculated and experimental CIS; 44 NMR vdw set, set of data represented by the 44 NMR solution structures of the complex calculated by NMR for which CIS have been predicted; 44 NMR *K* set, set of data represented by the 44 NMR solution structures of the complex calculated by NMR for which CIS have been adjusting on the basis of the binding scaling factor optimization; 44 NMR CIS set, set of data represented by the 44 NMR solution structures of the complex calculated by NMR for which a geometry and binding scaling factor optimization has been carried out; X-ray vdw set, set of poses generated, with no optimization, using as starting protein frame the unbound X-ray protein; X-ray *K* set, set of poses generated using as starting protein frame the unbound X-ray protein and carry on a binding constant optimization; X-ray CIS set, set of poses generated using as starting protein frame the unbound X-ray protein and carry on a binding constant and geometry optimization; bound NMR CIS set, set of poses generated using as starting protein frame a bound NMR protein frame and carry on a binding constant and geometry optimization; unbound NMR CIS set, set of poses generated using as starting protein frame an unbound NMR protein frame and carry on a binding constant and geometry optimization.

10.1021/jm800075r CCC: $40.75
Published on Web 07/15/2008

side chain and backbone flexibility. However, few degrees of freedom of the protein can be included explicitly without increasing the complexity of the conformational search too much. The degrees of freedom are usually restricted to rotations around side chain single bonds, since they are the most flexible part of the protein. For example, in the program HADDOCK, flexibility is introduced as a refinement of structures generated by a rigid body docking procedure. Flexibility is allowed first along the side chains at the interface, then for the backbone and side chains of both interacting molecules.[17] The correct selection of torsional degrees of freedom to include in a calculation requires a considerable amount of a priori knowledge of alternative binding modes for a given receptor. This knowledge usually is the result of the availability of experimental structures obtained under different conditions and/or using different ligands. If multiple experimental structures are not available, some information can be obtained from simulation methods such as Monte Carlo (MC) or molecular dynamics (MD) that model explicitly all degrees of freedom of the system including the solvent. MD simulations have provided detailed information on the fluctuations and conformational changes of proteins. MD simulations enable exploration of the conformational energy landscape accessible to a protein and have proved useful for assessing protein flexibility and dynamics on a nanosecond time scale.[26–28]

The alternative approaches use a more implicit flexibility that is much simpler to implement in docking applications. Soft receptors can be easily generated by lowering the energy penalty for the overlap of a ligand atom with an atom of the receptor structure. By reduction of the van der Waals contributions to the total energy score, the receptor is made softer, and ligands that are too large may still fit into the binding site. The reason for this approach is that the receptor structure has some inherent flexibility, which can be adapted to slightly differently shaped ligands by resorting to small variations in the orientation of binding site side chains and backbone positions. Although the use of soft receptors presents a number of advantages, such as ease of implementation and computational speed, it also makes use of conformational and energetic assumptions that are difficult to justify.[16] Implicit treatment of flexibility can also be achieved by performing rigid body docking of an ensemble of conformations. Since proteins in solution do not exist in a single minimum energy static conformation but are in fact constantly jumping between low energy conformational substates, the best description for a protein structure is that of a conformational set of slightly different protein structures coexisting in a low energy region of the potential energy surface. A representative subset of the typical conformational ensemble of a given receptor is currently obtained experimentally from X-ray crystallography or NMR or generated via computational methods such as MC or MD simulations. Multiple structures account for the full flexibility of the protein without an exponential increase in computational cost that would come from including all the degrees of freedom of the protein. On the other hand, as just a small fraction of the conformational space of the receptor is represented, the method used to obtain the multiple receptor structures has a significant influence on the results.[29]

In a previous work,[30] we described a new method for determining the structures of protein−ligand complexes from experimentally determined backbone amide proton CIS values. Experimental CIS values were first mapped onto the VDW surface of the X-ray crystal structure of the unbound protein to identify the ligand binding site. A wide range of different possible orientations of the ligand in the protein binding site (poses) was then produced, and each of them then was refined using the experimental CIS data: optimization involved calculation of the expected CIS value for each proton in each structure and minimization of the difference between these calculated values and the experimental ones as a function of the position and orientation of the ligand relative to the binding site. The final CIS optimized structure matched the structure determined experimentally using conventional NOE methods with an RMSD of less than 1 Å. However, in our calculation, the unbound protein structure was treated as a rigid body, and the differences observed between the NOE and CIS structures appear to be connected to the movements of the walls of the binding site on complexation. This system therefore provides an ideal opportunity to test methods for the treatment of protein flexibility, since the free and bound conformations of the protein are known. This paper describes an analysis of the effects of protein flexibility on the observed CIS values and the impact on the accuracy of 3D structures determined using the CIS approach. The prospects of implementing methods to treat protein flexibility within our protocol are also explored.

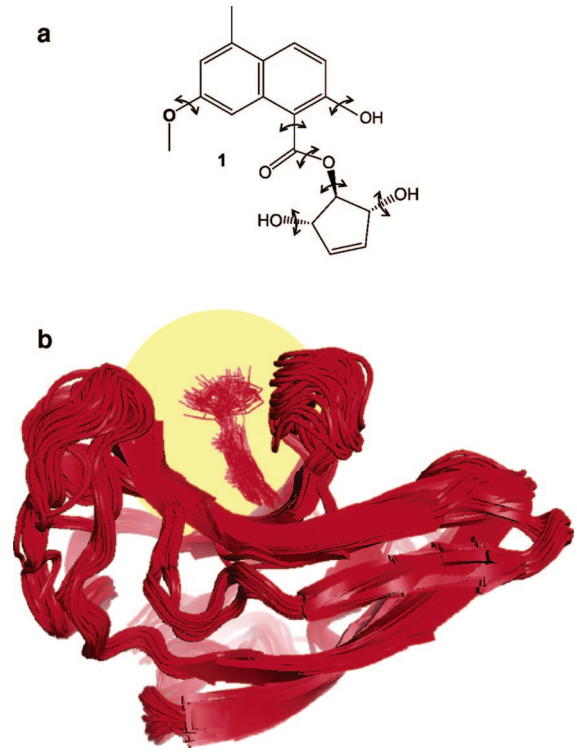

Figure 1. (a) Structure of the ligand. Torsion angles that were allowed to vary during the structure determination process are indicated. Conformational flexibility in the five-membered ring was not considered in the calculations. (b) Overlay of the 44 solution structures of the complex NCS/lig.1 obtained by NMR experiments using NOE restraints for structure refinement (PDB entry code 1J5I). The protein binding site is highlighted in yellow.

Results and Discussion

Approach. The system used is the complex formed by the chromoprotein antitumor antibiotic neocarzinostatin (NCS) and a synthetic chromophore (lig.1, Figure 1a), for which the 3D structure of the complex in solution has been determined by conventional NMR methods using NOE restraints for structure refinement.[31] The NMR structure of the complex is represented by 44 similar structures that provide some indication of the flexibility of the complex and provide excellent experimentally

derived data to test the impact of small conformational changes on the robustness of the CIS structure determination method. The coordinates of the protein frames used in the following calculations were obtained from the Protein Data Bank, and the structure of the ligand was created with the XED 6.1.0 software[32] using standard bond lengths and angles and energy-minimized. The method we have developed has three key stages: (a) definition of the receptor binding site using the backbone amide CIS values; (b) generation of a set of ligand conformations and orientations for introduction into the receptor binding site (poses); (c) optimization of each pose based on comparison of the experimental and calculated CIS values for the amide backbone protons.

The structure determination protocol is summarized as follows. Experimental CIS values of the backbone amide protons (BB NH) were first mapped onto the VDW surface of the protein to create a ligand *j*-surface representing possible locations for the center of the ligand in the binding pocket. The ligand structure was then docked into the center of this surface, and an ensemble of starting orientations of the ligand in the protein binding site (poses) was produced purely on the basis of shape complementarity, allowing us to sample a wide range of different possible conformations. Each of these poses was used as an independent starting point for structure optimization calculations using the experimental CIS data. Optimization involved the calculation of the CIS of each BB NH proton for every change of the ligand orientation in the protein binding site using a semiempirical function that is described in detail elsewhere.[14] A genetic algorithm (GA) was used to minimize the difference between the calculated and the experimental CIS values, as a function of the position and orientation of the ligand relative to the binding site and its internal torsion angles. In addition, an association constant scaling factor (K) was included as a variable to scale the experimental CIS values to allow for ambiguity in the extent of protein binding in the NMR experiment.

Analysis of the 44 NMR Solution Structures of the Complex. We first examine in detail the properties of the solution structure of the complex determined by NMR experiments that represents our target and best possible outcome (Figure 1b, PDB code 1J5I). NMR structure determination does not give a single structure for the complex. Generally, between 100 and 1000 molecular mechanics structures that satisfy a range of experimentally determined distance restraints are produced, and then all the structures with low energies are superimposed. The root-mean-squared difference of the heavy atoms of the protein backbone ($RMSD_{BB}$) is a measure of the degree of convergence and resolution of the structure. Usually, $RMSD_{BB}$ values are interpreted as an indication of the mobility of the protein, but they are also dependent on the availability of sufficient experimental constraints to fully define the structure. Figure 2 shows the range of $RMSD_{BB}$ values obtained by comparing each of the 44 NMR structures of the NCS·lig.1 complex with all of the others, along with the root-mean-squared difference of the heavy atoms of the ligand ($RMSD_{LIG}$). The protein structures differ by 1.0−1.5 Å mainly through movement of the loops that flank the binding site (Figure 1b), and variation in the position of ligand is slightly larger due to mobility of the ester side chain. The range of values of $RMSD_{LIG}$ provides us with a measure of the certainty of the experimental structure and indicates that any CIS-based structure that we calculate that falls within 2 Å of any of the 44 NOE-based structures would represent a match with the experimental structure that is as good as the experimental structure itself.

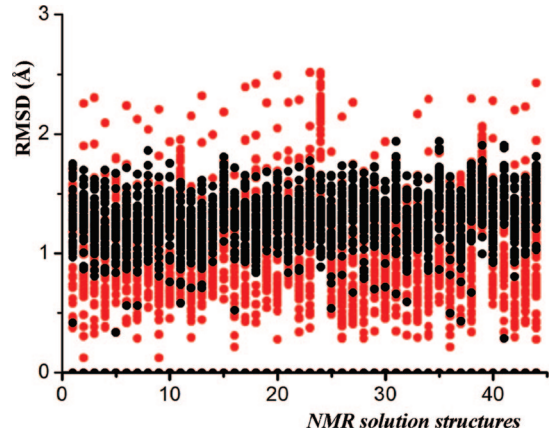

Figure 2. Range of $RMSD_{BB}$ (black circles) and $RMSD_{LIG}$ (red circles) values for the 44 NMR solution structures of the complex NCS/lig.1 obtained by pairwise comparison of all of the structures. The horizontal axis represents the identity of the structure used as the reference point for the comparison with each of the 44 structures.

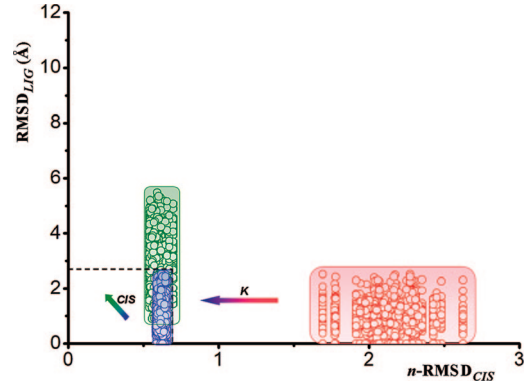

Figure 3. Comparison of $RMSD_{LIG}$ with n-$RMSD_{CIS}$ for the 44 NMR solution structures of the complex with no optimization (red), K optimization (blue), and CIS and K optimization (green). The black dotted box represents the target region corresponding to agreement with the NMR solution structure.

The potential impact of the structural variations that are observed in the experimental structure on the CIS values can be estimated using the CIS calculation method implemented in our structure determination software. This in turn will allow us to assess the influence of protein flexibility on the accuracy of the CIS-based structure determination protocol. Predicted CIS values were calculated for each of the 44 NMR solution structures (poses) and compared with the corresponding experimental data using the root-mean-squared difference in ppm ($RMSD_{CIS}$). A plot of $RMSD_{LIG}$ versus $RMSD_{CIS}$ provides a straightforward graphical method to evaluate the relationship between the 3D structure of the complex and the corresponding CIS values. If the CIS method is a useful structure determination tool, then $RMSD_{LIG}$ and $RMSD_{CIS}$ should be correlated so that they both tend to zero as the quality of the structure improves. Figure 3 shows the results for the 44 NOE-based structures (red box in Figure 3). The value of $RMSD_{CIS}$ is very high for all of these structures because the protein is only 40% bound in the experiments. The association constant scaling factor (K) used in our calculations allows for experimental uncertainty in the fraction bound, and optimization of this parameter without changing the geometry of the complexes resulted in good agreement between the calculated and experimental CIS values (blue box in Figure 3). In other words, the pattern of CIS values calculated from the experimental NOE-based structures agrees

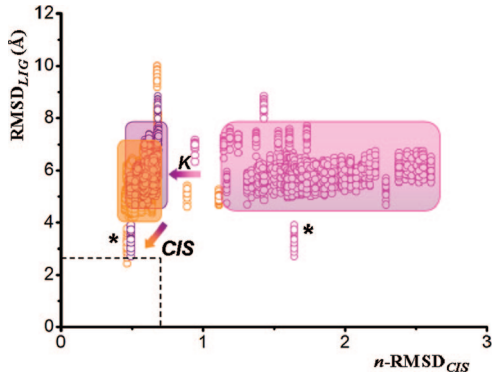

Figure 4. Comparison of n-RMSD$_{CIS}$ with RMSD$_{LIG}$ for the set of 100 structures generated using the unbound X-ray protein frame with no optimization (magenta), K optimization (violet), and CIS and K optimization (orange). The black dotted box represents the target region corresponding to agreement with the NMR solution structure.

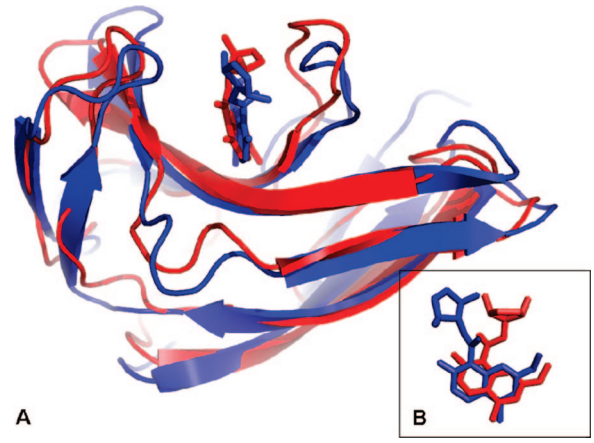

Figure 5. (A) Overlay of the optimized CIS structure (blue) and the NMR solution structure (red) of the complex NCS/lig.1. (B) Closer view of the orientation of the ligand in the two structures.

with the experimental CIS values, and so structure optimization based on CIS values has potential for locating these structures.

To test whether the geometries of the 44 NOE-based structures could be further optimized, these structures were used as starting points for a full structure optimization based on the experimental CIS values. The green box in Figure 3 shows the results of simultaneous geometry and binding constant optimization. The values of RMSD$_{CIS}$ are reduced somewhat as expected, but the spread of RMSD$_{LIG}$ values is dramatically increased. Thus, it appears that while structure determination based on CIS values provides a family of conformations that encompass the experimental structure, CIS values alone are not sufficient to accurately define the structure of the complex to the same degree as NOE constraints. The results that fall within the black box in Figure 3 represent optimized structures that agree with the 44 experimental structures as well as these structures agreeing with each other, and a significant number of structures do fall inside this region. However, a substantial number of structures also lie outside this region. The optimization procedure explores a relatively large conformational space, and there is clearly a range of structures that all satisfy the CIS constraints equally well. The black box in Figure 3 represents our target "perfect" result for ab initio structure determination using the CIS method. The results shown in green indicate the best that we can expect to obtain; i.e., in the population of structures generated by the CIS structure determination method, it will be possible to find a large number that match the experimental structure, but there will also be a significant population that is different.

Structure Determination Using the X-ray Crystal Structure of the Protein. Ab initio structure determination of the complex was first attempted using the X-ray crystal structure of the unbound protein (PDB entry code 1NCO[33]) as the protein input frame. By use of random starting points, a set of 100 poses was generated purely on the basis of shape complementarity. This set of poses was then optimized as a function of the binding constant parameter, and finally a full geometry optimization was carried out on the basis of the CIS values. For each step of this procedure, each of the 100 poses was compared with each of the 44 NOE-based experimental structures, and the results are analyzed using the RMSD$_{CIS}$ versus RMSD$_{LIG}$ plot in Figure 4. The binding constant optimization has a large effect on RMSD$_{CIS}$ as before because the protein is only 40% bound in the experiment. However, the important observation is that during the CIS-based structure optimization stage of the calculation, the results move toward the origin of the plot and the black dotted box that defines the limits of the accuracy of the experimental data (Figure 4). This demonstrates that structure optimization based on CIS values improves the quality of the 3D structure that is obtained. Nevertheless, none of the results fall inside the black box that represents good agreement with the experimental structure. The results are displayed in boxes that represent similar structures and indicate a level of convergence, but there are some outliers due to the range of conformational space explored. One set of structures, highlighted with asterisks in Figure 4, shows significantly better agreement with the experimental NOE-based structures. However, this simply reflects the element of chance involved in the random generation of starting points and is related to the conformation of the ester side chain discussed below.

Figure 5 shows an overlay of one of the top ranked poses with the corresponding NMR solution structure of the complex. The position of the ligand is almost the same in the two complexes, and the main difference, which leads to the relatively high value of RMSD$_{LIG}$ (2.42 Å), lies in the orientation of the flexible ester side chain. Comparison of the two protein frames shows some movement of the loops flanking the binding site. Since in our calculation, the protein structure is treated as a rigid body, the differences found in the orientation of the ligand may be due to movements of the walls of the binding site that alter the shape of the pocket.

Influence of Flexible Protein Loops. To test the effects of loop movement on the structure determination process, the calculations described above were repeated using both the NMR solution structure of the unbound protein (PDB code 1J5H) and a bound protein frame from the NMR solution structure of the complex. The differences in the structures of these protein frames are highlighted in Figure 6. By use of random starting points, a set of 100 poses was optimized as described above using the experimental CIS values. The RMSD$_{CIS}$ versus RMSD$_{LIG}$ plot in Figure 7 shows the results. Both of the NMR solution structures give results that are an improvement on those obtained using the X-ray crystal structure of the protein. Many of the structures now fall inside the black box that represents agreement with experiment. This demonstrates that the conformation of the protein and loop movement have a significant effect on the quality of the structure obtained from the CIS method. The best agreement is found for the solution structure of the bound protein, as might be expected. The structures obtained using the bound protein frame (green box in Figure 7) fall in exactly the same region of the RMSD$_{CIS}$ versus RMSD$_{LIG}$ plot as the structures obtained by optimizing the NMR

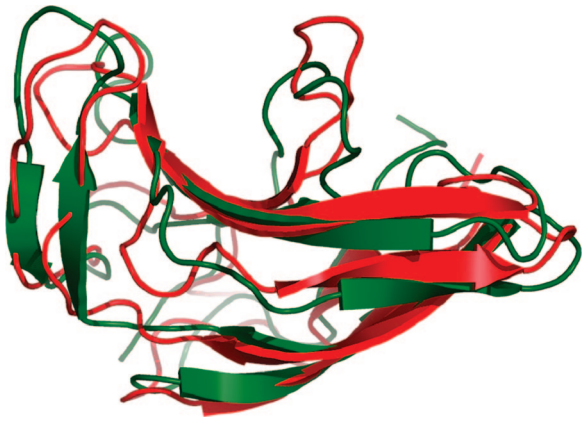

Figure 6. Overlay of the backbone of the unbound (green) and bound (red) NMR solution structures.

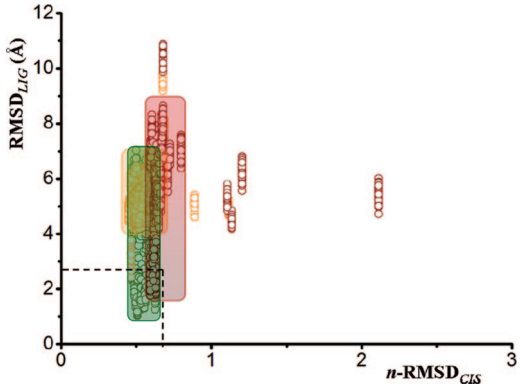

Figure 7. Comparison of n-RMSD$_{CIS}$ with RMSD$_{LIG}$ for the structures found after CIS and K optimization using the unbound X-ray crystal structure (orange), the unbound NMR solution structure (red), and a bound NMR solution structure (green) for the protein frame. The black dotted box represents the target region corresponding to agreement with the NMR solution structure.

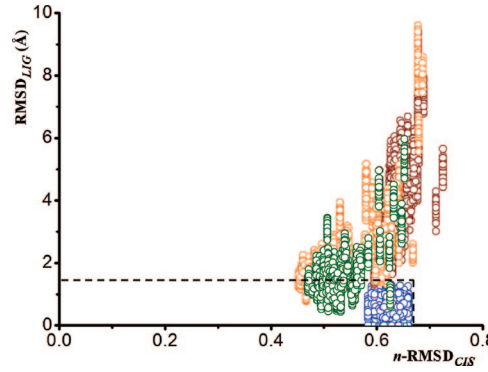

Figure 8. Comparison of n-RMSD$_{CIS}$ with RMSD$_{LIG}$ considering only the heavy atoms of the ligand−naphthoate moiety for the 44 NMR solution structures after K optimization (blue) and for the structures found after CIS and K optimization using the unbound X-ray crystal structure (orange), the unbound NMR solution structure (red), and a bound NMR solution structure (green) for the protein frame. The black dotted box represents the target region corresponding to agreement with the NMR solution structure.

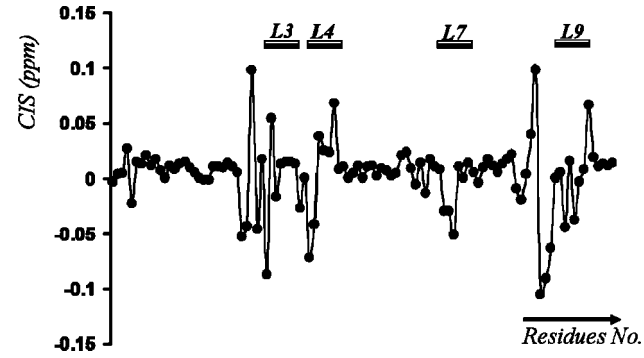

Figure 9. Experimental CIS values for the backbone amide protons. The backbone amide protons that are remote from the ligand have negligible CIS values, and the biggest variations are found for the residues that form flexible loops around the protein binding site.

structures of the complex (green box in Figure 3), indicating that the conformational search procedure works well and converges to the same population of structures from very different starting points.

Influence of Ligand Flexibility. The ester side chain of the ligand is torsionally flexible and carries little magnetic anisotropy, so its conformation is not well-defined by the experimental CIS data. We therefore reanalyzed the results of structure determination calculations considering only the naphthoate moiety of the ligand in the calculation of RMSD$_{LIG}$ (Figure 8). This significantly reduces the magnitude of RMSD$_{LIG}$ as expected, but more importantly, the calculated structures are now much more tightly clustered in the RMSD$_{LIG}$ dimension. The implication is that divergence in the orientation of the ligand side chain not only is responsible for most of the discrepancy between the CIS-based structures and the NOE-based structures but also accounts for the large range of RMSD$_{LIG}$ values observed in the Figures 3, 4, and 7. The apparently large range of different structures generated by the CIS structure determination simply reflects a lack of experimental data to define the conformation of the side chain, which is invisible to this technique. This plot also highlights the correlation between the quality of the structure obtained and the agreement between the calculated and experimental CIS data.

Refinement of Protein Loops. Flexibility of both the protein and the ligand clearly plays a role in the accuracy of the structure determined using this method. The approach described above treats ligand flexibility explicitly, but protein flexibility is a more challenging problem. If we can predict ab initio possible movements that the protein frame is able to undertake during binding, it might be possible to use the CIS data to discriminate between different possibilities, since the results above show that the CIS method is sensitive to changes in protein conformation. We therefore investigated the possibility of introducing degrees of freedom for the protein loops that form the binding site, since this is where the largest changes in chemical shift are observed (Figure 9) and where the largest changes in conformation are observed on binding (Figure 6).

There are four loops in the protein binding site: loop L3 (residues Thr39−Gln45), loop L4 (residues Trp48−Gly52), loop L7 (residues Thr77−Ser81), and loop L9 (residues Thr99−Ser107). Conformational searches were carried out for each of these loops using PRIME,[34,35] a loop prediction algorithm from Schrödinger Software. The X-ray crystal structure of the unbound protein was used as the starting point for the refinement of each loop individually, and 60 different structures were obtained: 20 different protein frames for loop L3, 10 structures for L4, 15 for L7, and 15 for L9. Figure 10 shows an overlay of the protein conformations generated by PRIME.

Each of the 60 protein structures was used as the starting protein frame for CIS-based determination of the structure of the complex. Figure 11 shows the results. We are clearly exploring a much larger conformational space, and some

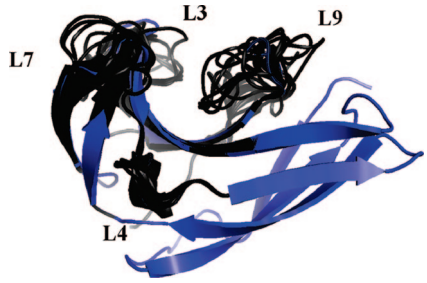

Figure 10. Overlay of the different binding site loop conformations of the NCS protein generated with the program PRIME (Schrödinger software) using the X-ray crystal structure of the unbound protein as the starting point. The loops that were refined are L3 (residues Thr39−Gln45), L4 (residues Trp48−Gly52), L7 (residues Thr77−Ser81), L9 (residues Thr99−Ser107).

structures do achieve a significantly improved $RMSD_{LIG}$ (less than 2 Å), but the results with and without loop refinement are qualitatively similar. No dramatic improvement is observed even for loops that move significantly on ligand binding. In order to sample protein conformations that are closer to the bound structure, it will clearly be necessary to sample more conformational space, but that carries the burden of increased computational cost.

Conclusion

In summary, we have explored the influence of conformational flexibility on the accuracy of a method for determining the structures of protein−ligand complexes using experimental CIS values. The ligand is allowed to sample a wide range of conformational space in the calculations, and ligand functional groups that have significant magnetic anisotropy are accurately located by the calculation, but groups that are spectroscopically silent are not. The result for the complex of NCS with lig.1 studied here is that the position and orientation of the ligand chromophore can be pinpointed with a high degree of accuracy in the protein binding pocket, but the aliphatic ester side chain samples a range of conformations. Although we do not allow movement of the protein during the structure calculation, the effects of protein conformational mobility were investigated by using a set of different protein structures as starting points for the calculation. The structure of the complex was reliably reproduced using the unbound X-ray crystal structure, the unbound NMR solution structure, and the bound NMR solution structure of the protein. However, the best agreement with experiment was obtained using the bound NMR protein frame, indicating that loop movement does have a significant impact on the quality of the structure generated by the CIS structure determination methodology. Attempts to treat loop flexibility using molecular mechanics to sample a range of loop conformations did not improve the results for calculations based on the unbound protein frame. In all of the calculations, there is a clear correlation of the agreement between the calculated and experimental CIS data with the quality of the structure obtained, indicating that the method has promise for applications in complex structure determination.

Experimental Section

The coordinates of the protein frames used in the calculations were obtained from the protein data bank (PDB entry code 1NCO[33] for the X-ray crystal structure of the unbound protein, PDB entry code 1J5H[31] for the NMR solution structure of the unbound protein, and PDB code 1J5I[31] for the NMR solution structure of the bound protein and complex). The structure of the ligand was created with the XED 6.1.0 software[32] using standard bond lengths and angles and energy-minimized. Our computational approach consists of a set of Perl and C^{++} scripts that implement the three main software packages used in the protocol and analyze the results. The three programs used are Jsurf to define the receptor binding site using the backbone amide CIS values, GOLD to generate a set of ligand conformations and orientations for introduction into the receptor binding site, and Shifty to optimize each pose based on comparison of the experimental and calculated CIS values for the amide backbone protons.

Analysis of the 44 NMR Solution Structures of the Complex. The analysis of the 44 NMR solution structure was carried out starting from the geometry of each of the 44 poses and using Shifty for the CIS-based structure optimization. First, without carrying out any optimization, the CIS values for each BB NH proton were calculated and compared to the experimental CIS values. Second, to allow for the incomplete saturation of the protein binding site in the experiment, the calculated CIS values were optimized by varying the association constant scaling factor (K) with no geometry optimization. The scaling factor was allowed to vary by up to a factor of 10. Third, the geometry of each of the 44 poses was fully optimized on the basis of the CIS values using Shifty.

Shifty. Optimization involved calculation of the CIS of each backbone NH proton using a semiempirical function[14] and minimization of the difference between the calculated and the experimental CIS values, using a genetic algorithm to vary the position and orientation of the ligand relative to the binding site as well as the ligand torsion angles. The conformational search in Shifty was divided into two steps, each with population sizes of 50 runs for 50 generations. The option create_offset_file was set to 0 so that the orientation of the ligand in the initial PDB file was retained, and the optimization was carried out using these geometries as starting points. In the first step, the intermolecular distance limit was set to 1.5 Å and the range of allowed rotations of one molecule relative to the other was set to ±10°. Intramolecular torsions were allowed to change within the full range of ±180°. In the second step, these parameters were reduced to 1 Å, 5°, and 90°, respectively. To reduce the conformational space, a steric clash penalty was added for distances of less than 2 Å for intermolecular clashes and for distances less than 1 Å for intramolecular clashes for non-hydrogen atoms. The association constant (K) was allowed to vary by a factor of 10. All of the optimized poses were then ranked by their fitness values. The fitness of a particular structure is defined using the normalized root-mean-squared difference between the experimental and calculated CIS values ($1/n\text{-}RMSD_{CIS} = \Delta\delta_{exp}/\Delta\delta_{cal}$, where $\Delta\delta_{cal}$ is the root-mean-squared difference between the calculated and experimental CIS values and $\Delta\delta_{exp}$ is the root mean square of the experimental CIS values).

Structure Optimization Procedure. The structures of the complex were determined using the procedure described below, the only difference between different calculations being the protein input frame used: X-ray crystal structure of the unbound protein, NMR solution structure of the unbound and bound protein, and the 60 different protein frames obtained after loop refinement.

JSurf. In the first stage of the protocol, the location of the binding site was obtained using the program JSurf.[36] All points on the j-surface less than 2.5 Å from the protein backbone were removed, and the remaining points were averaged to give the coordinates of a single point: the center of the binding site. Experimental CIS values of the BB NH residues were first mapped into the protein VDW surface. The protein frame was colored depending on the intensity of the experimental CIS values; the biggest absolute changes were represented in red, moderate changes in yellow, and the lowest changes in blue. The largest CIS values are clustered in the natural cleft of the protein, and no significant perturbations were identified elsewhere on the protein surface. Afterward, spheres were constructed centered on each perturbed proton and then filled randomly with dots as an indication of the perturbation suffered by each BB NH proton and with a radius proportional to the intensity of the perturbation. A ligand j-surface consequently was

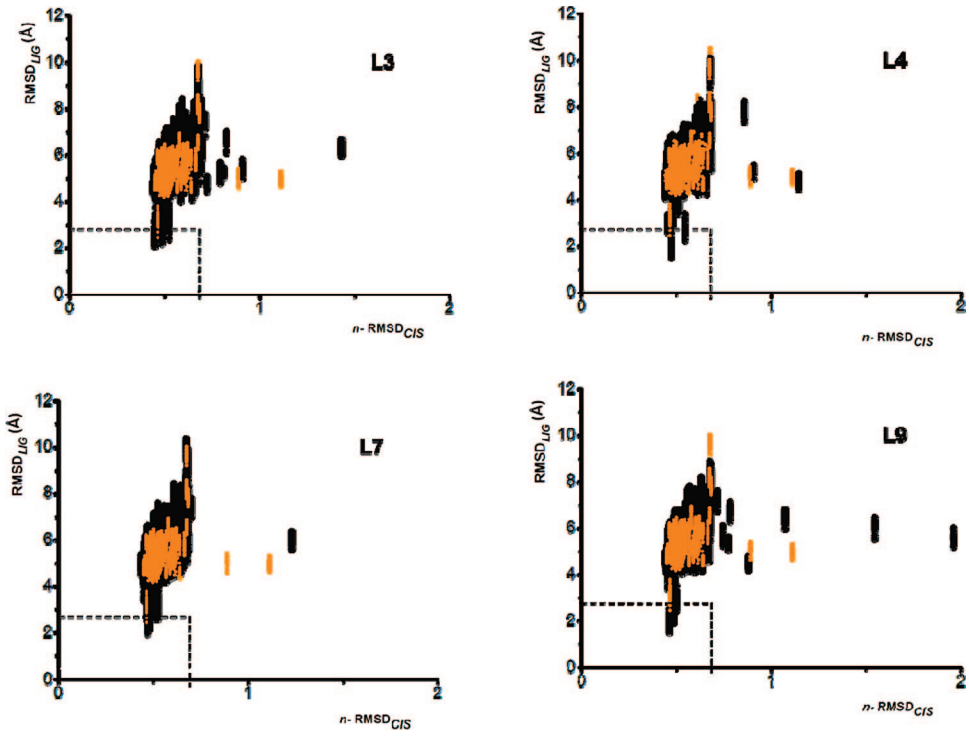

Figure 11. Comparison of n-RMSD$_{CIS}$ with RMSD$_{LIG}$ considering only the heavy atoms of the ligand for the structures found after CIS and K optimization using the unbound X-ray crystal structure for the protein frame (orange) and using the set of 60 protein structures generated by the loop refinement algorithm PRIME (black). The black dotted box represents the target region corresponding to agreement with the NMR solution structure.

created in which each dot represented a possible location for the center of the ligand. The *j*-surface represented a small area almost in the middle of the binding site. Next, all points in the *j*-surface less than 2.5 Å from the protein backbone were removed, and the remaining points were averaged to give the coordinates of a single point identifying the center of the binding site, obviating the need to search a large amount of redundant conformational space remote from the binding site.

GOLD. In the second stage, the generation of an ensemble of poses located in the binding site was carried out using the GOLD, version 2.2, software.[37] The center of the ligand was located at the center of the binding site determined using Jsurf, and a set of 100 conformations of the ligand for each set of data was generated. The GOLD scoring function was modified by setting the contribution of the hydrogen bond energy term (H_BOND_WT) to 0.001 in gold.parms. This allows us to use this software as a rapid method to generate a set of poses considering only the protein−ligand VDW energies and ligand intramolecular strain energy, i.e., based on shape complementary only. GOLD was used to generate structures using 10 runs with a population size of 100 for 100 000 generations. The "early termination" parameter used by GOLD in the default setting (the option that instructs the program to terminate runs as soon as a specified number of runs have given essentially the same answer) was switched off in our calculations in order to ensure that GOLD generated a diverse sample of structures.

Shifty. In the final stage, structure optimization was carried out using Shifty.[14] The 100 conformations generated by GOLD were used as independent starting points for structure optimization, which was carried out as described above. The calculation of the root-mean-squared difference of the heavy atoms of the whole ligand structure and of the ligand naphthoate moiety only (RMSD$_{LIG}$) was then carried out after overlaying the protein BB of the optimized poses with each of the 44 structures of the complex determined by conventional NMR methods using NOE restraints for structure refinement.[31]

Prime. Four loops were identified in the protein binding site: L3 (residues Thr39−Gln45), L4 (residues Trp48−Gly52), L7 (residues Thr77−Ser81), and L9 (residues Thr99−Ser107). These were refined using the commercial Prime package (Schrödinger, Inc.).[34,35] The loop prediction algorithm generates many loop conformation candidates following an ab initio procedure, clusters them to reduce redundancy, and selects only the representative candidates. Side chain optimization of the selected loops is then carried out, followed by energy minimization. Energy calculations used an all-atom model based on the OPLS-AA force field and the surface generalized Born implicit solvent model.[5] The lowest energy structures were selected as the output of the loop prediction algorithm. The loop refinement retrieved 60 different structures: in particular, 20 different protein frames for loop L3, 10 structures for L4, 15 for L7, and 15 for L9.

Acknowledgment. We thank Dr. G. Moyna for making available the source code of the JSURF program, Professor M. P. Williamson for some of the code used to develop Shifty, and Professor D. N. Woolfson for providing the experimental CIS data for the complex. We thank AstraZeneca for funding.

Supporting Information Available: Structures of ligands that bind NCS, CIS values, parameters for structure optimization, n-RMSD$_{cis}$, and Shifty fitness. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lepre, C. A.; Moore, J. M.; Peng, J. W. Theory and applications of NMR-based screening in pharmaceutical research. *Chem. Rev.* **2004**, *104* (8), 3641–3675.
(2) Zuiderweg, E. R. P. *Mapping Protein-Protein Interactions in Solution by NMR Spectroscopy Biochemistry* **2002**, *41* (1), 1–7.
(3) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–1534.
(4) Schieborr, U.; Vogtherr, M.; Elshorst, B.; Betz, M.; Grimme, S.; Pescatore, B.; Langer, T.; Saxena, K.; Schwalbe, H. How much NMR data is required to determine a protein−ligand complex structure? *ChemBioChem* **2005**, *6* (10), 1891–1898.
(5) McCoy, M. A.; Wyss, D. F. Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR

chemical shift perturbations. *J. Am. Chem. Soc.* **2002**, *124* (39), 11758–11763.
(6) Medek, A.; Hajduk, P. J.; Mack, J.; Fesik, S. W. The use of differential chemical shifts for determining the binding site location and orientation of protein-bound ligands. *J. Am. Chem. Soc.* **2000**, *122* (6), 1241–1242.
(7) Wyss, D. F.; Arasappan, A.; Senior, M. M.; Wang, Y.-S.; Beyer, B. M.; Njoroge, F. G.; McCoy, M. A. Non-peptidic small-molecule inhibitors of the single-chain hepatitis C virus NS3 protease/NS4A cofactor complex discovered by structure-based NMR screening. *J. Med. Chem.* **2004**, *47* (10), 2486–2498.
(8) McCoy, M. A.; Wyss, D. F. Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. *J. Biomol. NMR* **2000**, *18* (3), 189–198.
(9) Polshakov, V. I.; Birdsall, B.; Feeney, J. Characterization of rates of ring-flipping in trimethoprim in its ternary complexes with *Lactobacillus casei* dihydrofolate reductase and coenzyme analogues. *Biochemistry* **1999**, *38* (48), 15962–15969.
(10) Polshakov, V. I.; Birdsall, B.; Frenkiel, T. A.; Gargaro, A. R.; Feeney, J. Structure and dynamics in solution of the complex of *Lactobacillus casei* dihydrofolate reductase with the new lipophilic antifolate drug trimetrexate. *Protein Sci.* **1999**, *8* (3), 467–481.
(11) Korukottu, J.; Bayrhuber, M.; Montaville, P.; Vijayan, V.; Jung, Y.-S.; Becker, S.; Zweckstetter, M. Fast high-resolution protein structure determination by using unassigned NMR data. *Angew. Chem., Int. Ed.* **2007**, *46* (7), 1176–1179.
(12) Hunter, C. A.; Packer, M. J.; Zonta, C. From structure to chemical shift and vice-versa. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *47* (1−2), 27–39.
(13) Spitaleri, A.; Hunter, C. A.; McCabe, J. F.; Packer, M. J.; Cockroft, S. L. A ^{1}H NMR study of crystal nucleation in solution. *CrystEngComm* **2004**, *6*, 489–493.
(14) Hunter, C. A.; Packer, M. J. Complexation-induced changes in ^{1}H NMR chemical shift for supramolecular structure determination. *Chem.—Eur. J.* **1999**, *5* (6), 1891–1897.
(15) Dobrodumov, A.; Gronenborn, A. M. Filtering and selection of structural models: combining docking and NMR. *Proteins: Struct., Funct., Genet.* **2003**, *53* (1), 18–32.
(16) Morelli, X. J.; Palma, P. N.; Guerlesquin, F.; Rigby, A. C. A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. *Protein Sci.* **2001**, *10* (10), 2131–2137.
(17) van Dijk, A. D. J.; Boelens, R.; Bonvin, A. M. J. J. Data-driven docking for the study of biomolecular complexes. *FEBS J.* **2005**, *272* (2), 293–312.
(18) Schwieters, C. D.; Kuszewski, J. J.; Tjandra, N.; Marius Clore, G. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **2003**, *160* (1), 65–73.
(19) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Freer, S. T.; Rose, P. W. Complexity and simplicity of ligand−macromolecule interactions: the energy landscape perspective. *Curr. Opin. Struct. Biol.* **2002**, *12* (2), 197–203.
(20) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein−small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16* (3), 151–166.
(21) 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* (1), 45–55.
(22) Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M. Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, *266* (2), 424–440.
(23) Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13* (6), 547–562.
(24) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47* (4), 409–443.
(25) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5* (4), 375–382.
(26) Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2002**, *12* (2), 190–196.
(27) Wong, C. F.; McCammon, J. A. Protein flexibility and computer-aided drug design. *Annu. Rev. Pharmacol. Toxicol.* **2003**, *43*, 31–45.
(28) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
(29) Carlson, H. A.; McCammon, J. A. Accommodating protein flexibility in computational drug design. *Mol. Pharmacol.* **2000**, *57* (2), 213–218.
(30) Cioffi, M.; Hunter, C. A.; Packer, M. J.; Spitaleri, A. Optimization of the structure of protein−ligand complexes using ^{1}H NMR chemical shifts. *J. Med. Chem.*, in press.
(31) Urbaniak, M. D.; Muskett, F. W.; Finucane, M. D.; Caddick, S.; Woolfson, D. N. Solution structure of a novel chromoprotein derived from apo-neocarzinostatin and a synthetic chromophore. *Biochemistry* **2002**, *41* (39), 11731–11739.
(32) Vinter, J. G. Extended electron distributions applied to the molecular mechanics of some intermolecular interactions. II. Organic complexes. *J. Comput.-Aided Mol. Des.* **1996**, *10* (5), 417–426.
(33) Kim, K. H.; Kwon, B. M.; Myers, A. G.; Rees, D. C. Crystal structure of neocarzinostatin, an antitumor protein-chromophore complex. *Science* **1993**, *262* (5136), 1042–1046.
(34) Schrödinger. http://www.schrodinger.com/.
(35) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004**, *55* (2), 351–367.
(36) McCoy, M. A.; Wyss, D. F. Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. *J. Am. Chem. Soc.* **2002**, *124*, 11758–11763.
(37) 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* (3), 727–748.

JM800075R