JCTC Journal of Chemical Theory and Computation

Prediction of the Structure of Complexes Comprised of Proteins and Glycosaminoglycans Using Docking Simulation and Cluster Analysis

Tsubasa Takaoka,[†] Kenichi Mori,[†] Noriaki Okimoto,[‡] Saburo Neya,[†] and Tyuji Hoshino*[,†,§]

Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan, Bioinformatics Group, GSC, RIKEN, Yokohama, Kanagawa 230-0046, Japan, and PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 332-0012, Japan

Received February 5, 2007

Abstract: A typical docking simulation provides information on the structure of ligand−receptor complexes and their binding affinity in terms of a docking energy. We have developed a potent method combining a docking simulation with cluster analysis to extract adequate docking structures from the many possible output structures of the simulation. First, we tried to predict the structure of basic fibroblast growth factor (bFGF) bound to heparin, using the docking simulation program AutoDock 3.0. Two X-ray crystal structures had already been obtained for bFGF. One was a complex of the protein and heparin, a kind of glycosaminoglycan, and the other, only the protein itself, hereafter called a simplex. We docked a heparin molecule onto the protein simplex and generated many trial structures for the bFGF−heparin complex. The structures of those docked complexes were optimized through energy minimization by AMBER8. Although neither the docking energy calculated by AMBER8 nor that calculated by AutoDock 3.0 could be used satisfactorily by themselves to select a proper heparin-binding complex from the output structures, the majority of the structures generated by AutoDock 3.0 were fairly close to each other in atom geometry, and the averaged geometry over these structures was also close to that of the crystal. Hence, we utilized only the atom geometry for evaluation and carried out cluster analysis with the collection of geometries. This procedure enabled selection of a structure considerably close to the crystal's. We applied this approach to two other heparin-binding proteins: antithrombin and annexin V. Two crystal structures, a complex and a simplex, had been elucidated for these proteins as well as for bFGF. Our trials gave an exact prediction of the heparin-binding structures of these proteins, showing the approach in this study is effective in studying the docking of ligands that have a variety of docking conformations due to the presence of multiple rotatable bonds and charged chemical groups.

Introduction

In silico screening is a powerful and indispensable computational tool in drug discovery and development because it enables analysis of the intermolecular interactions between proteins (receptors) and chemical compounds (ligands) and prediction of their interaction energy. A key process in in silico screening is modeling of complexes of lead compounds and target proteins. The structure for the target protein bound to a chemical compound is usually not available even if structural information on the unliganded protein has been disclosed by X-ray crystal analysis or nuclear magnetic resonance (NMR). In this case, the complex must be modeled from the individual structures of the unliganded protein

* Corresponding author e-mail: hoshino@faculty.chiba-u.jp.
[†] Chiba University.
[‡] RIKEN.
[§] PRESTO, Japan Science and Technology Agency.

10.1021/ct700029q CCC: $37.00
Published on Web 10/10/2007

simplex and the chemical compound. A docking simulation is a computational technique for enabling such modeling. In the docking simulation, a small molecule, such as a peptide or compound, is to be bound to a macromolecule, such as a protein or enzyme. Numerous conformations of the small molecule and the corresponding energies when bound to the macromolecule are calculated in one simulation. A lower binding energy indicates a higher probability of formation of a complex. The calculated binding energy, however, is poorly correlated with the closeness of the structure of the complex to that of the crystal, and it is difficult to select an optimal structure from the numerous ligand conformations generated by the docking simulation. This problem is particularly serious in the docking of glycosaminoglycans (GAGs) and GAG-binding proteins because the intended ligands, glycosaminoglycans, have many conformational variations.

The aim of this study is to establish a procedure for finding an adequate conformation of heparins bound to a target protein. Most of the currently available software programs for docking simulations are based on the assumption that a ligand molecule is held inside the binding pocket of the target protein, a pocket often composed of many hydrophobic amino acid residues. Paul and Rognan attempted to reproduce 100 crystal structures of ligand−protein complexes using several kinds of docking software programs.[1] They evaluated the ability of the software programs to correctly predict the ligand-binding structures and found that the rates for correct prediction were 39% for DOCK,[2] 51% for FlexX,[3] and 56% for GOLD 3.0.[4] In addition, they incorporated cluster analysis into the three docking programs and succeeded in improving the accuracy of the docking output. Success in the docking of glycosaminoglycans (GAGs) to proteins, however, has not been reported yet. The aim of this work is to provide a promising approach for the docking of GAGs to proteins. GAG−protein docking is very challenging because of the highly flexible nature of the GAG chain, high charge-density of the GAG binding site, and weak surface complementarity at the GAG−protein interface.

The interaction of GAGs with proteins plays a significant role in the regulation of many physiological processes, such as homeostasis, growth factor activity, anticoagulation, cell adhesion, and enzyme regulation.[5−8] For example, heparin is now used as a coagulator in surgery. However, little is known about the mechanism of the interaction of GAGs with proteins. Since it is difficult to crystallize a GAG−protein complex, few crystal structures of GAG−protein complexes have so far been obtained. Consequently, modeling software for GAG−protein complexes would be a useful tool for analysis of the interaction of GAGs with proteins.

AutoDock 3.0, a docking program provided by Garrett M. Morris, can explore an extensive conformational space.[9] Morris et al. demonstrated the accuracy of AutoDock 3.0 using seven protein−ligand complexes whose tertiary structures and binding constants were known. They classified the protein−ligand complexes into three groups. The first group contained complexes that have small and rigid ligands. This group was utilized as the simplest docking test case. The second group contained moderately flexible ligands, providing a typical test set of intermediate difficulty. The third group contained ligands having many rotatable bonds and diverse chemical characteristics, the most difficult test cases. They compared the performances of the Monte Carlo simulated annealing algorithm (SA) used in earlier versions of AutoDock,[10,11] a genetic algorithm (GA),[12] and a Lamarckian genetic algorithm (LGA) newly employed in AutoDock 3.0.[9] In their study, there was little difference in computational results among the three methods for the first test group. In the test cases with the intermediate and highest levels of difficulty, a structure close to the crystal one was rarely generated by using the SA or GA method. On the other hand, many structures generated by using the LGA method were very close to that of the crystal, even for the third test group.

Goodford and his co-worker reported success in prediction of the binding for two GAGs—monosaccharide and disaccharide—using the docking programs GRID (version 15, Molecular Discovery Ltd.),[13−16] AutoDock 2.4,[17] and DOCK.[2] However, the closeness in structure of the predicted complex to the crystal's was not sufficient when they carried out the docking for hexasaccharide.[18] The results of our preliminary trial employing AutoDock 3.0 for docking of GAGs to proteins were also unsatisfactory. This failure is essentially due to the fact that GAGs have a large electric charge and many rotatable bonds.

In this paper we propose a method for reliable modeling of GAG−protein complexes using a docking simulation and cluster analysis together. We have executed a procedure comprised of the generation of ligand binding conformations by AutoDock 3.0, energy minimization with AMBER8,[19] and cluster analysis for selecting a reasonable ligand structure. We focused on basic fibroblast growth factor (bFGF),[20,21] antithrombin,[22,23] and annexin V,[24,25] whose structures had already been experimentally determined for both the simplex and heparin-bound complex. The effectiveness of our methodology for predicting the GAG−protein structure was evaluated by assessing the similarity between the experimentally determined crystal structures and the computationally derived structures.

Method

Docking Simulation by AutoDock 3.0. The following Brookhaven database entries were used for the docking simulations: (A) bFGF, 1bfc[20] and 1bfg;[21] (B) antithrombin III (ATIII), the L-chain of 1e03[22] and 1e04;[23] (C) and (D) annexin V, 1a8a[24] and 1g5n.[25] All of these test cases ((A)−(D)) satisfy the condition that tertiary structures are available both for the GAG-bound conformation and the unbound one. All of them are heparin-binding proteins. That is, the GAG-bound crystal in each case is a complex of heparin and protein.

Annexin V, a calcium-binding protein, has two heparin-binding sites. These two distinct GAG-binding sites are positioned on the protein surfaces opposite to each other, so annexin V provides two test cases: (C) and (D). One site, (C), holds Ca^{2+} ions that influence the interaction with heparin. This site is formed by two calcium-binding loops, I_{AB} and I_{DE}, termed from the numbering of domains and

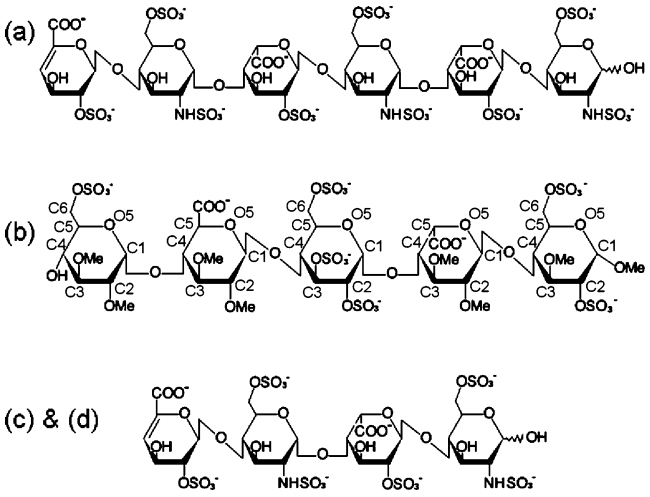

Figure 1. Chemical structures of heparins used in the docking simulation and the subsequent cluster analysis. RMSDs are measured with respect to the atoms composing the sugar ring, which are usually labeled as C1, C2, C3, C4, C5, C6, and O5 as shown in (b).

helices of annexin V. No sulfate groups of heparin directly interact with the Ca^{2+} ions. Instead, the sulfate oxygen atoms make hydrogen bonds with the backbone nitrogen atoms in the I_{AB} calcium-binding loop. Additional hydrogen bonds are formed with water molecules coordinated to the Ca^{2+} ions in both loops and with the side chain of a serine residue in the I_{DE} loop. In contrast, the second heparin-binding site (D) is located on the concave surface of the protein and is not associated with calcium binding.[25] We carried out docking simulations, targeting these two heparin-binding sites separately.

The protocol for the docking simulations is as follows. First, the initial heparin structure was deduced from each crystal structure of the complex. To detach the heparin from the heparin-binding site, only the coordinate of the heparin was translated outward by 8−10 Å. The translated heparin was assumed to be a ligand for the docking simulation (Figure 1). The unliganded protein simplex was regarded as a receptor, which is called a macromolecule in AutoDock 3.0. Docking simulations were carried out using standard AutoDock 3.0 parameters, with 256 runs, the maximum value for AutoDock 3.0, performed for each protein. Grids with a spacing of 0.375 Å were generated around the geometric center of the original ligand position, so that the grid dimensions were (A) 82 × 54 × 54; (B) 100 × 80 × 80, (C) 100 × 80 × 80, and (D) 100 × 80 × 80. The rotatable bonds of the glycosidic link of the heparin were set rigid. If they are flexible or partially rigid, the heparins often have an unrealistic structure. Other rotatable bonds, for example, hydroxyl groups (−OH), sulfate groups ($-OSO_3$), and methyl groups ($-CH_3$), were set flexible. A genetic algorithm and local search procedure were employed. The calculation of internal electrostatic energy in a docking run was activated because heparin has a large negative charge. In cases (C) and (D), the charges of calcium ions were set to −2.0.

Minimization with AMBER8. The output file of AutoDock 3.0 contains the binding ligand structures and their binding energies but not information on the protein structure,

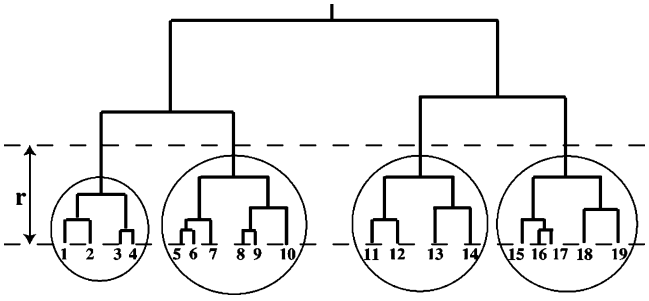

Figure 2. Example of a family tree in the cluster analysis. The RMSD between any two structures in a cluster is less than *r*. The structures, with sequential numbers in a circle, belong to the same cluster.

because a receptor is regarded as a rigid body in AutoDock 3.0. To obtain the structures of the ligand−protein complex, it is necessary to combine the protein structure with the ligand structures that are extracted from the output file. Hence, in each test case, 256 structures generated by AutoDock 3.0 for the ligand were saved as PDB format files. The protein coordinate was added to each file to prepare 256 protein−heparin complexes. Energy minimization was executed for the complexes using the AMBER8 sander module[19] with the parm99 all-atom force field for proteins[26] and with the glycam04 parameters for heparin.[27,28] Since parameters for heparin with sulfate groups are not provided in glycam04, they were prepared by ourselves. The parameters for the bonded terms were assigned in accordance with the parm99 force field. In order to determine atom charges for sulfate groups, the structures of glucosamine and iduronic acid extracted from the respective crystal structure were optimized at the HF/6-31g(d, p) level using the Gaussian03 program.[29] The charges of the atoms of these glycans were calculated by the two-stage RESP method[30] using the electrostatic potential computed at the rb3lyp/cc-pvtz level and with an ether solvation condition in a manner similar to that used in a previous study.[31,32] This procedure is the same as that used for the development of ff03.[33] To relax the strain in the complexes, the complexes were energetically minimized for 5000 steps by the generalized Born method.[34]

After energy minimization of the docking complexes, the calculated structures for each complex were superimposed on the crystal structure with respect to the main chain atoms of protein, and the coordinates of the heparins were saved. Simultaneously, the similarities between the docking structures of heparin and the crystal structure were measured by examining the root-mean-square deviations (RMSDs) of atom coordinates for the C1, C2, C3, C4, C5, C6, and O5 atoms. The VMD package[35] was used for superposition, RMSD measurement, and visualization.

Cluster Analysis. We carried out the hierarchical cluster analysis using the RMSDs on the C1, C2, C3, C4, C5, C6, and O5 atoms of the docking structures (Figure 2).

The 256 ligand structures are labeled s_1, s_2, ...,s_{256}. Initially, corresponding 256 clusters are designated C_1, C_2, ..., C_{256} with each cluster containing only a single structure. The group containing these 256 clusters is labeled A_{256}. The RMSD between s_i and s_j is represented as $d(s_i,s_j)$, and the distance between two clusters C_m and C_n is defined as

$$D(C_m,C_n) = \max_{si\in Cm, sj\in Cn} \mathrm{d}(s_i,s_j) \quad (1)$$

First, the distances of all pairs of C_m and C_n, $D(C_m, C_n)$, in A_{256} are measured. The pair that has the smallest distance is coupled and registered as a new cluster labeled C_{257}. As a result, the number of clusters becomes 255. The new group containing 255 clusters is labeled A_{255}. Next, the distances among the 255 clusters are measured, and the pair that has the smallest distance in A_{255} is coupled and registered as a new cluster. By iterating this procedure until A_1 is obtained, a family tree for the 256 structures is derived.

The cluster analysis suggests how the structures are distributed and where the generated ligand structures are concentrated. First, we examine all of the ligand structures, setting the distance r at 1.5 Å. Then an additional cluster analysis is carried out with the largest cluster in the first step set as a parent group and the distance r set at 1.2 Å. The structure closest in the RMSD sense to the average of the atom coordinates of all the ligand structures composing the largest cluster in the second cluster analysis is concluded to be our solution and is called a "representative model".

In our calculations of the number of hydrogen bonds, a combination of donor, hydrogen, and acceptor atoms is regarded as forming a hydrogen bond when the donor−acceptor distance is within 3.5 Å and the hydrogen-donor−acceptor angle is within 60°. This generous criterion for hydrogen bonding is applied so as not to miss even weak interactions and has been adopted in our previous studies to closely survey intermolecular interactions.[36−38]

Docking Simulation by GOLD 3.1. In order to examine the performance of our approach when other docking software is used, we have executed the same cluster analysis using GOLD 3.1. The protocol for the docking simulation is the same as that for AutoDock 3.0. The PDB files used in AutoDock 3.0 were converted into mol2 files by the BABEL[39] program in all the cases (A), (B), (C), and (D). For each test case, 256 runs were performed using standard GOLD 3.1 parameters. In order to obtain 256 docking ligand structures, the calculation was not terminated even if the top solutions in ranking were close to each other in RMSD. The binding site for the docking search was set to within 20 Å from the position of the grid center in AutoDock 3.0. The rotatable glycosidic bonds are rigid, while the other rotatable bonds are flexible.

Results

AutoDock 3.0 predicts the binding of small ligand molecules to receptors. In this study, 256 docking ligand structures were generated, and their docked energies were computed. In the docking results for bFGF, no significant correlation is observed between the RMSDs of the ligand structures, relative to the crystal structure, and their docked energies (Figure 3). For example, the lowest docked-energy structure shows a quite large RMSD value. The ligand is bound to the heparin-binding site in this structure, but the direction of the glycan chain is opposite to that of the crystal structure (Figure S1 in the Supporting Information).

In spite of poor correlation between the docked energy and the RMSDs from the crystal structure, a certain number of the ligand structures are fairly close to the crystal structure. Accordingly, we speculated that many of the structures generated by AutoDock 3.0 are distributed around the crystal structure and plotted the geometrical centers of the structures generated by AutoDock 3.0 to obtain scatter diagrams. The diagrams suggest that the majority of structures are concentrated in a specific area (Figure 4). It is reasonable to assume that the structure at the center of this area is considerably similar to the crystal. Hence, a representative model can be extracted from the docked ligand structures, and a reasonable structure for the GAG-bound complex is obtained without information on the crystal structure. The two-step cluster analysis is a promising method for extracting a good representative model because carrying out the cluster analysis twice is effective in removing structures that are localized to an area but not the major area.

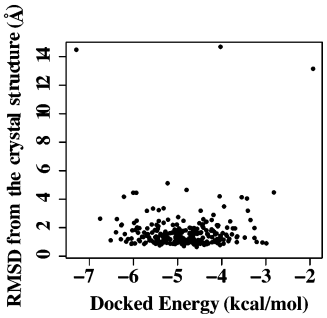

Figure 3. Comparison between the docked energies calculated by AutoDock 3.0 and the RMSDs from the crystal structure for the 256 structures for bFGF generated by AutoDock 3.0. Docked energy represents the stability of FGF−heparin complexes. No significant correlation is observed between RMSD and docked energy.

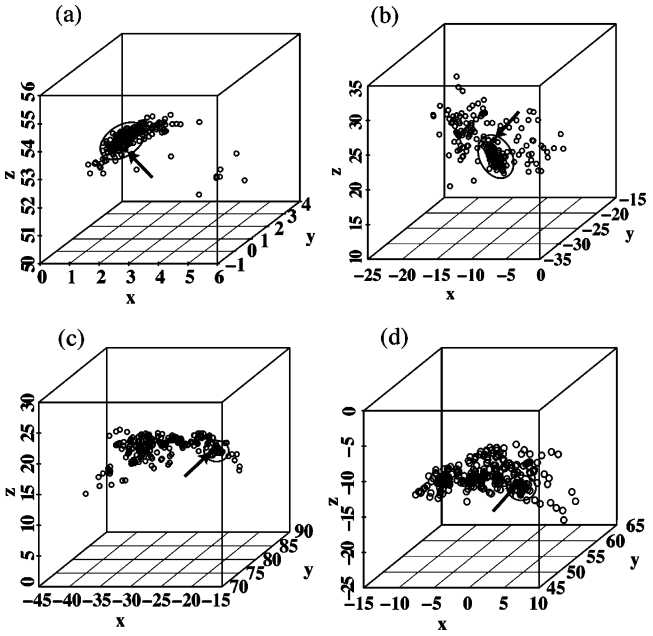

Figure 4. Scatter diagrams of geometrical centers of the structures generated by AutoDock 3.0. The units for the *x*, *y*, and *z* axes are Å. The densest concentration of structures is marked by an oval. The crystal structure is marked by an arrow. (a) bFGF, (b) antithrombin, (c) annexin V-Ca(+), (d) annexin V-Ca(−).

Table 1. Comparison of RMSDs in the Cluster Analysis

protein		*N*[a]	max RMSD[b] (Å)	average RMSD[c] (Å)
bFGF	all[d]	256	14.57	1.53
	first[e]	73	1.34	0.98
	second[f]	30	0.84	0.85
antithrombin	all	256	16.32	7.05
	first	38	1.32	2.63
	second	17	0.72	2.59
annexin V−Ca(+)[g]	all	256	16.54	8.77
	first	13	1.39	1.56
	second	7	1.09	1.28
annexin V−Ca(−)[g]	all	256	14.08	8.52
	first	16	0.68	0.62

[a] Number of structures in each cluster. [b] Largest RMSD measured from the averaged geometry over the structures in the cluster. [c] Average RMSD of all structures relative to the crystal structure. [d] A cluster containing all 256 structures. These structures are generated by AutoDock 3.0 and are minimized by AMBER8. [e] A cluster of the structures categorized into the largest group when performing a cluster analysis using "all" as a parent set. [f] A cluster of the structures categorized into the largest group when performing a cluster analysis using "first" as a parent set. [g] Ca(+) and Ca(−) indicate whether Ca^{2+} ions are present or not to interact with heparin.

Table 2. RMSDs between the Experimental Crystal Structure and the Model Selected by Cluster Analysis or the Model Selected from the Lowest Binding Energy of AutoDock 3.0

	cluster analysis		AutoDock 3.0	
protein	rmsd (Å)	rank[a]	rmsd (Å)	rank[a]
bFGF	0.66	3	14.25	255
antithrombin	2.45	7	6.95	150
annexin V−Ca(+)	0.83	4	10.89	175
annexin V−Ca(-)	0.57	6	8.05	105

[a] The rank of the model is the order among all 256 models, determined by closeness to the crystal structure.

As shown in the average RMSD of Table 1, the number of structures close to the crystal's contained in the 1st_cluster is larger than that of the whole group in every case. The max RMSD of the 1st_cluster in (A), (B), and (C) is larger than 1.3 Å. Accordingly a more concentrated area of docked structures was extracted with the 1st_cluster set as a parent group. The results of the second cluster analysis show that the number of structures close to the crystal's contained in the 2nd_cluster is larger than that in the 1st_cluster in every test case (Table 1).

The max RMSD of the 1st_cluster in (D) is very small (0.68 Å). The structures in this cluster are especially close to one another. Hence, further classification was not necessary for (D), and we did not carry out a second cluster analysis. Consequently, the final representative models in (A), (B), and (C) are the closest to the averaged geometry of all structures in the 2nd_cluster, while that of (D) is closest to the averaged geometry of all structures in the 1st_cluster. The RMSDs between the representative model and the crystal structure are shown in Table 2. The accuracy is moderate in (B) and good in (A), (C), and (D). The reason for this difference is that the structures generated by AutoDock 3.0 in (B) contain very few structures that are close to the crystal's. A comparison of the representative model and the crystal structure shows that our approach has a high level of accuracy, especially for (A), (C), and (D) (Figure 5).

The rank of the representative model among all 256 structures with respect to closeness to the crystal structure is shown in Table 2. The representative model selected from the cluster analysis has a rank within the top ten in every case. On the other hand, the model selected from the binding energy of AutoDock 3.0 is not good, with all of their ranks over 100. Hence, the energy analysis using AutoDock 3.0 does not discriminate an adequate docking structure for the binding of heparin from the generated structures. Although even the cluster analysis could not extract the closest model, i.e., rank 1, in this study, the ranks of the representative models demonstrate that this method is a useful approach for predicting an adequate heparin-binding structure.

Furthermore, when the first cluster analysis was carried out, there were some structures, "singular models" not comprising a cluster with any others, that provided useful information on GAG binding in docking simulations. The average RMSDs of the singular models in Table 3 are larger than those of the whole group in Table 1, implying they are very different from the crystal structure in cases (A)−(D). By inspecting what residues interact with the singular models, we found several residues positioned far from the heparin-binding site of the protein. Details of these residues are shown in Figures S2 and S3 in the Supporting Information.

Additionally, we examined the applicapability of our method to GOLD 3.1, which is one of the most widely used docking simulation tools. The entire procedure of generation of 256 structures with GOLD 3.1, energy minimization with AMBER8, and cluster analysis were performed in a manner similar to that for AutoDock 3.0. The energy minimization could not be completed for antithrombin because GOLD 3.1 generates many inadequate structures in which heparin is bound to the inside of antithrombin, and the forces on atoms are too large to execute molecular mechanical calculation in AMBER8. The clustering results are shown in Table S1 in the Supporting Information. The average RMSDs from the crystal structure of all 256 structures generated by GOLD 3.1 are equal or larger than those of the structures generated by AutoDock 3.0. Table S1 clearly indicates that the level of accuracy for predicting the heparin-binding structure was low compared with the results from AutoDock 3.0. A comparison of the models selected by the cluster analysis and those selected from the lowest binding energy from GOLD 3.1 shows that the selected models differ considerably from the crystal; i.e., their rankings among the 256 structures are not good with respect to closeness to the crystal structure. No notable improvement is observed except for the case of annexin V−Ca(+) (Table S2 in the Supporting Information). Furthermore, even the structure closest to the crystal's is inferior to the representative model selected by the cluster analysis combined with AutoDock 3.0 in every test case (Table S3 in the Supporting Information). Accordingly, GOLD 3.1 is deemed inappropriate for GAG−protein docking.

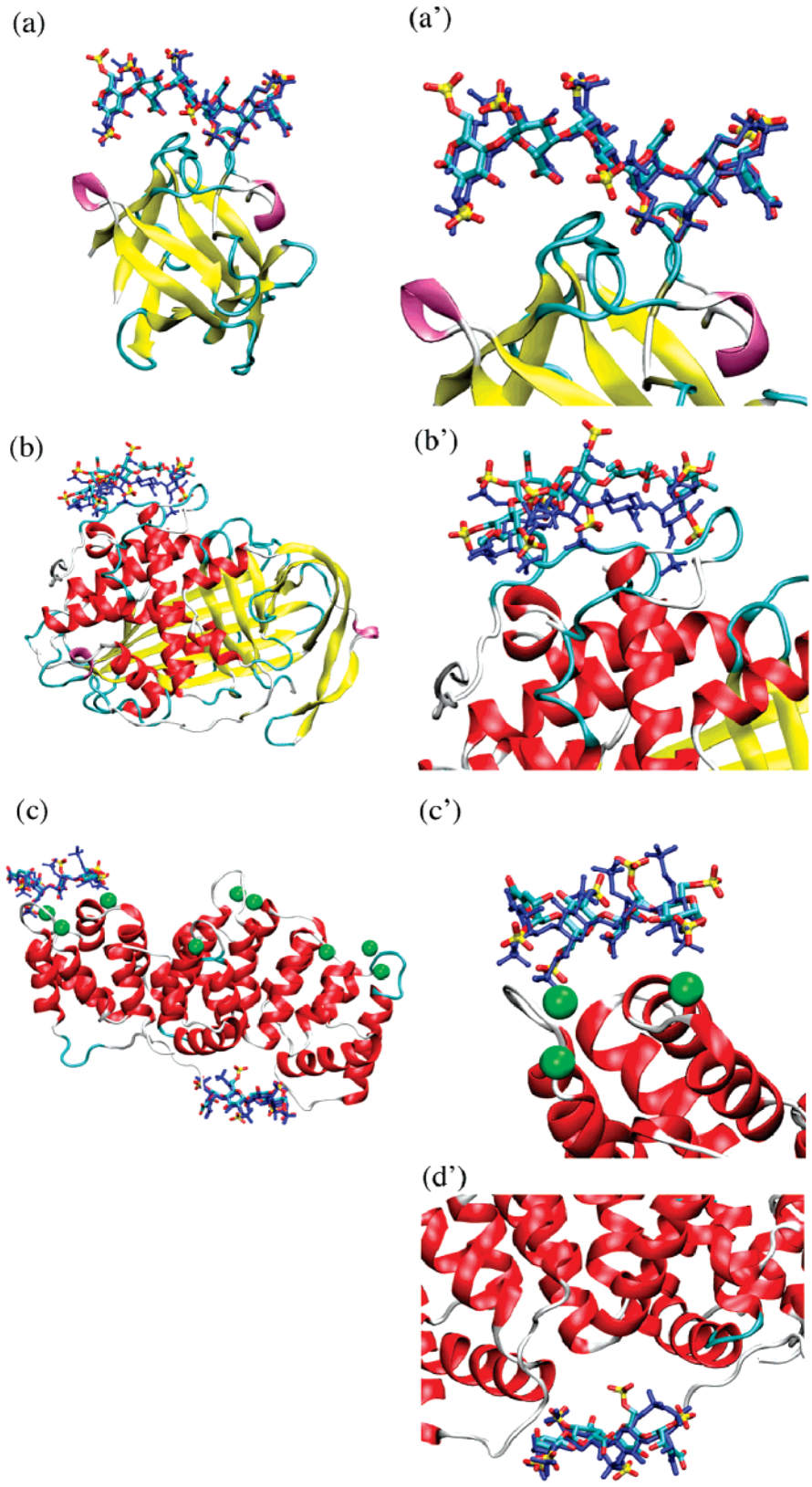

Figure 5. Comparison of the representative model to the crystal structure. Proteins and heparins are shown as cartoon and ball-and-stick representations. Green spheres represent Ca^{2+} ions. Heparins colored blue are crystal structures, and those colored cyan, yellow, and red are representative models. (a) bFGF, (a′) magnification of (a), (b) antithrombin, (b′) magnification of (b), (c) annexin V, (c′) magnification of Ca(+) area of (c), (d′) magnification of Ca(−) area of (c).

Discussion

Cluster analysis seems to be effective in predicting the structure of GAG−protein complexes. AutoDock 3.0 is able to generate many structures close to the crystal structure, and, as shown in the average RMSDs in Table 1, it is plausible that the largest cluster contains the structure most

Table 3. Average RMSD of All Singular Models Measured from the Crystal Structure

protein	average RMSD (Å)	number of singular models
bFGF	4.21	18
antithrombin	10.03	88
annexin V−Ca(+)	9.50	81
annexin V−Ca(-)	9.72	136

adequately reproducing the crystal structure. In this study, the representative model is determined by selecting the structure closest to the averaged atom geometry of the structures in the 2nd_cluster (exceptionally in the 1st_cluster for (D)). Although the top-ranked structure could not be extracted in each case, the average RMSD of the 2nd_cluster in Table 1 and the RMSD in Table 2 demonstrate that the representative model is acceptable and that its rank is satisfactory.

Cluster analysis has already shown a substantial degree of success in predicting protein folding structures. Based on the supposition that there are a greater number of conformations surrounding the correct folding structure than the incorrect folding one, Shortle et al. performed cluster analysis for small proteins with the 1000 lowest-energy conformations produced by random structure generation and subsequent energy minimizaion.[40] They clearly suggested that the analysis can identify conformations considerably closer to the native structure than the conformation with the lowest energy. This finding is the basis for our trial prediction of heparin-binding structure by cluster analysis. Zagrovic and co-workers closely examined the average structure of the ensemble generated by molecular dynamics simulations for small polypeptides both in folded and unfolded states.[41] They found that none of the conformations of the unfolded state exhibited a nativelike structure but that the mean structure obtained by averaging over the entire set of unfolded conformations showed a nativelike geometry. This approach is quite useful because information on the native heparin structure is usually not available when performing binding predictions. Zagrovic et al. further suggested the advantage of evaluation with the distance-based RMSD and the preference of the average structure over the unfolded ensemble of small protein structures for predicting the native geometry.[42] Their reports provided good justification for our present trial.

The criteria for selecting the representative model from 256 structures should be determined carefully because the averaged atom geometry is greatly influenced by the selection of clusters. The structures in (C) and (D) are dispersed compared with those in (A) and (B) (Figure 4). As a result of the first cluster analysis in (C) and (D), the number of structures is less than 10 for all the clusters except the 1st_cluster, which contains structures fairly close to the crystal's (Table 1). This result suggests that proper selection of the 1st_cluster is very important for prediction of the structure of the GAG−protein complex. In the first clustering shown in Table 1, the distance r was set to 1.5 Å. To examine the dependency of the clustering results on the distance criteria, the first clustering was executed with r set at 2.0, 1.0, and then 0.5 Å (Table 4). For 0.5 Å, clusters became too small, and a cluster far from the crystal structure occasionally became the largest cluster. This caused a decrease in the level of accuracy of prediction, as seen for annexin V−Ca(+) in Table 4. For r equal to 2.0 Å, the largest cluster was likely to contain structures too different from the crystal structure, thus lowering the prediction accuracy. Judging from these findings, an r from 1.0 Å to 1.5 Å is most suitable for the first clustering.

In the geometry search step, AutoDock 3.0 computes the interaction energy between a receptor and a ligand by intermolecular van der Waals, hydrogen bond, and Coulomb potentials and evaluates the stability of the ligand after generating a large number of ligand binding conformations by the Lamarckian genetic algorithm. The scoring function for the geometry search is[43]

$$\Delta E = \sum\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{6}}\right) + \sum\left(\frac{q_i \times q_j}{\epsilon(r_{ij}) \times r_{ij}}\right) + \sum K_\phi(1 + \cos(n\phi) - \delta) + \Delta H_{\mathrm{vdW}}^{\mathrm{ligand}} + \Delta H_{\mathrm{eleq}}^{\mathrm{ligand}} + \Delta H_{\mathrm{hbond}}^{\mathrm{ligand}} \quad (2)$$

where A and B are the van der Waals parameters for atoms i and j, C and D are hydrogen bond parameters, r_{ij} is the interatomic distance between atom i and atom j, q is the Coulomb charge of each atom, and $\epsilon(r)$ is the distance-dependent dielectric function. The first three terms are the receptor−ligand interaction energy terms. The next four are the internal energies of the ligand—the torsion potential, van der Waals force, electrostatic force, and intramolecular hydrogen bonding, respectively. In the LGA, the structure with the highest ΔE is deleted, the structure with the lowest ΔE is always retained, and the other structures are merged using a crossover technique or random mutation technique.[9] The estimation of ΔE will highly influence the accuracy of our approach using cluster analysis because this energy dominates the structures generated in the docking simulation.

The most prominent difference between (B) and the other cases is in the functional groups of the different species of heparin. Uronic acids or glucosamine of natural heparin has a hydroxyl group at C3. Tetrasaccharide and hexasaccharide fragments of porcine mucosal heparin, the heparin of the crystal structure in (A), (C), and (D), were experimentally prepared by partial digestion with heparin-lyase I, followed by collection with strong anion exchange liquid chromatography.[21,25,44] On the other hand, the heparin of the crystal structure in (B) is a synthetic heparin analog.[45,46] This pentasaccharide contains O-alkyl ethers in place of hydroxyls that are usually sulfated at the early stage of the synthesis. As in the representative model (B) of Figure 5, carbon (red) and oxygen (cyan) atoms are seen at the C3 site of uronic acids or glucosamine. Consequently, the ligand in the docking simulation of (B) contains O-alkyl ethers instead of hydroxyls. AutoDock 3.0 does not consider any nonpolar atoms of a ligand. Accordingly, protein atoms directly interact with the carbon of heparin. Those atoms would interact with hydroxyls if the ligand were a natural heparin. Hence, the energy evaluation of van der Waals force,

Table 4. Dependency of Cluster Analysis Accuracy on Distance Criterion *r*

	number of structures[a]			average RMSD[b] (Å)			best RMSD[c] (Å)			RMSD of selected model[d] (Å)		
protein	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å	2 Å	1 Å	0.5 Å
bFGF	73	29	13	0.98	0.85	1.14	0.66	0.66	0.96	0.94	0.66	1.12
antithrombin	37	24	9	2.60	2.62	2.59	2.35	2.37	2.45	2.60	2.61	2.59
annexin V−Ca(+)	14	10	5	1.65	1.33	15.23	0.72	0.73	15.03	1.76	1.51	15.23
annexin V−Ca(−)	19	16	9	0.77	0.62	0.61	0.4	0.4	0.4	0.56	0.57	0.57

[a] Number of structures categorized into the largest cluster by the first clustering. [b] Average RMSD of all structures in the cluster, relative to the crystal structure. [c] RMSD between the crystal structure and the closest one in the cluster. [d] RMSD between the crystal structure and the model selected by the cluster analysis.

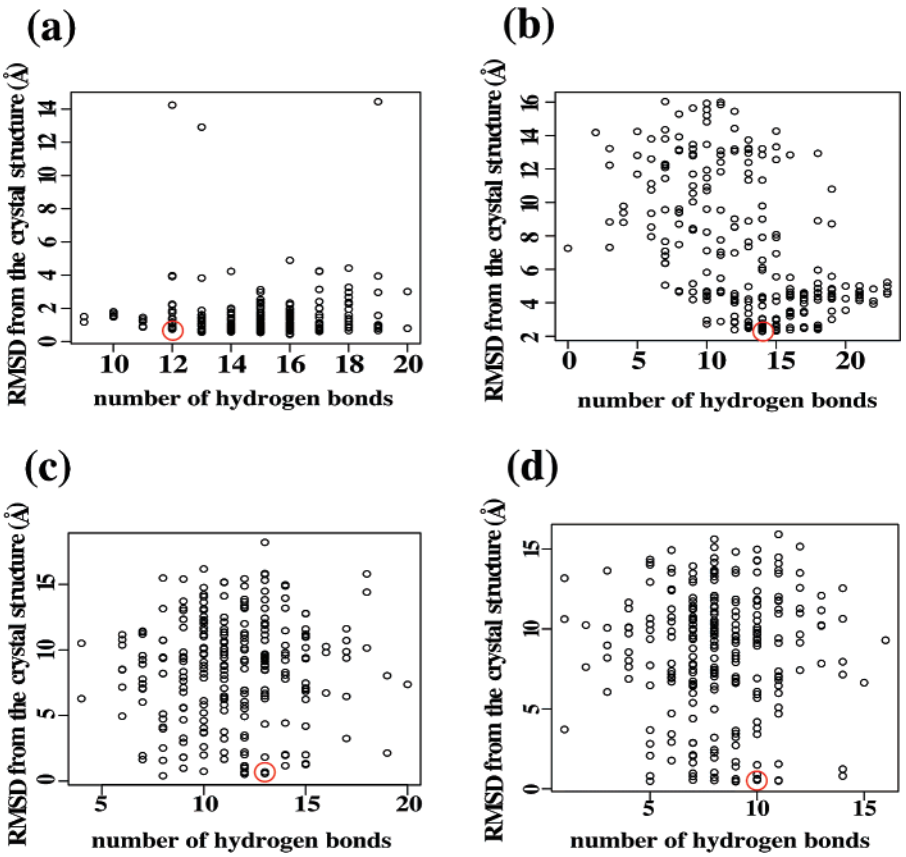

Figure 6. Comparison between the number of hydrogen bonds in protein−heparin complexes and the RMSD from the crystal structure. Each circle corresponds to one docking structure, and 256 circles appear in the respective graphs of (a)−(d). Data on the representative models are indicated by red circles. No significant correlation is observed between RMSD and number of hydrogen bonds. (a) bFGF, (b) antithrombin, (c) annexin V−Ca(+), (d) annexin V−Ca(−).

hydrogen bonding, and Coulomb force in the geometry search step is considerably different from that in other cases.

A study of the energetics of the interaction of bFGF with GAG by Thompson et al. showed that the electrostatic contribution of positively charged residues to the binding energy was only 30%.[47] They suggested that not only electrostatic interaction but also nonionic interaction, such as hydrogen bonding and van der Waals force, mainly contributed to the free energy for GAG−protein binding. We evaluated the contribution of hydrogen bonds in GAG−protein complexes (Figure 6). No significant correlation was found between RMSDs of docking structures, relative to the crystal, and the number of hydrogen bonds. This suggests that hydrogen bonding is not the only factor in GAG−protein binding and that van der Waals force interaction is also important. Since there are various factors to be considered for the binding free energy, it seems difficult to evaluate the affinity of GAGs for target proteins only from the energy in docking simulations. Consequently, a method for predicting the GAG−protein complex based on structures, namely the present approach, is needed.

In order to examine the causes for generation of structures not close to the crystal's, we focused on residues that are frequently located near the singular models but rarely interact with heparins in the crystal structure. The residues of the protein within 4.0 Å from the singular models were counted in each test case (Figure S2 in the Supporting Information). Acidic or hydrophobic residues were closely examined because their interaction with heparins cannot be straightforwardly explained. In the case of (A) bFGF, many singular models have interaction with K129 and G133. K129 is located in the binding site but rarely interacts with heparin in the crystal structure. Because a side chain of K129 extends outside, a heparin may be attracted to the residue. All 15

structures interacting with G133 also interact with Q134 and K135. These residues provide both a positive charge and hydrogen bond acceptors and donors; therefore, this area is a likely binding site for heparin (Figure S3(a) in the Supporting Information). In the case of (B) antithrombin, 14 residues before E42 are missing in the crystal structure for the simplex, 1E04. The N-terminus is therefore open and likely to interact with a sulfuric acid group of heparin. Attention should be given to the missing atoms when an intact crystal structure is employed for a receptor in docking simulations (Figure S3(b) in the Supporting Information). For case (C) for the Ca^{2+}-binding domain of annexin V, D66 is positioned near R61, and the side chains of D66 and R61 interact with each other. When a heparin is attracted to R61, the heparin will also be trapped near D66 (Figure S3(c) in the Supporting Information). Since Ca^{2+} ions counteract the negative charge of OD1 and OD2 of E70, the heparin interacting with S69 can be positioned near E70 (Figure S3(d) in the Supporting Information). For case (D) with no Ca^{2+}-associating domain of annexin V, OD1 and OD2 of D162 interact with the main chains of V201 and S202. That is, the positively charged side chain extends outside the protein. Therefore, a heparin is likely to approach D162 (Figure S3(e) in the Supporting Information). I245 and P246 are located at a loop near the target domain on the outside of the protein. Either pair of R205 and R206 or K284 and K288 interacts with heparin, and these residues are in the proximity of I245 and P246 (Figure S3(f) in the Supporting Information). Since those residues interact with singular models, we conclude that heparin is likely to be attracted to basic or nonpolar hydrophilic amino residues. In particular, a heparin has a very high binding affinity for Lys and Arg.

In the present study, GOLD 3.1 was not as accurate as AutoDock 3.0 in predicting the structure of GAG−protein complexes (Table 1 and Table S1 in the Supporting Information). Many heparin structures generated by GOLD 3.1 are apt to be bound to the inside of the proteins, despite the fact that heparins are incapable of being compactly packed inside the protein because of their strong negative charge. This can be explained from the GOLD scoring function for a geometry search.[48−50] In docking simulations by GOLD 3.1, the van der Waals term seems to have a particularly strong influence on the docking ligand structures. With an increase in the contact area, the van der Waals contribution becomes large. Hence, the stabilization energy for ligand binding is estimated to be smaller when a ligand adheres to the surface of a protein than when a ligand is inside a protein. In addition, it might be disadvantageous for GOLD to estimate the large electrostatic interaction between a ligand and positively charged residues. Therefore, GOLD 3.1 might be inadequate for docking simulations of GAGs such as heparins because the binding site is on the surface of a protein and the binding is highly influenced by charges.

Coulomb force is an explicit factor of the scoring function in a geometry search of AutoDock 3.0. Therefore, even negatively charged GAGs are evaluable as the ligand in docking simulations. In the present study, the binding sites of docking simulations were determined on the basis of crystal structures. If the binding site is unidentified, calculation of the surface electrostatic potential of a protein will be helpful in searching for probable binding sites for GAGs.

Conclusion

By performing (1) a docking simulation with AutoDock 3.0, (2) energy minimization with AMBER8, and (3) cluster analysis, it is possible to model the complex of a heparin and a GAG-binding protein. An adequate structure for the complex is predictable by this approach if the unliganded protein structure is available. The van der Waals force, hydrogen bonding, and Coulomb force are of considerable importance in the GAG−protein binding; therefore, incorporation of all these terms in docking simulations is highly desirable. The scoring function in the geometry search of AutoDock 3.0 contains all three of these terms; hence, AutoDock 3.0 is appropriate for GAG−protein docking simulations, although careful consideration should be given to the point that the strong influence of Lys and Arg of a target protein sometimes leads to generation of inadequate binding structures.

Acknowledgment. This work was supported by the Japan Science and Technology Agency.

Supporting Information Available: Tables S1−S3 and Figures S1−S3. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Paul, N.; Rognan, D. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 521.

(2) Ewing, T. J. A.; Kuntz, I. D. *J. Comput. Chem.* **1997**, *18*, 1175.

(3) Hoffmann, D.; Kramer, B.; Washio, T.; Steinmetzer, T.; Rarey, M.; Lengauer, T. *J. Med. Chem.* **1999**, *42*, 4422.

(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727.

(5) Lindahl, U.; Lidholt, K.; Spillman, D.; Kjellén, L. *Thromb. Res.* **1994**, *75*, 1.

(6) Yayon, A.; Klagsbrun, M.; Esko, J. D.; Leder, P.; Ornitz, D. M. *Cell* **1991**, *64*, 841.

(7) Prestrelski, S.; Fox, G. M.; Arakawa, T. *Arch. Biochem. Biophys.* **1992**, *293*, 314.

(8) Bjork, I.; Lindahl, V. *Mol. Cell. Biochem.* **1982**, *48*, 161.

(9) Goodsell, D. S.; Morris, G. M.; Halliday, R. S.; Huey, R. *J. Comput. Chem.* **1998**, *19*, 1639.

(10) Goodsell, D. S.; Olson, A. J. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 95.

(11) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293.

(12) Holland J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(13) Goodford, P. J. *J. Med. Chem.* **1985**, *28*, 849.

(14) Boobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. *J. Med. Chem.* **1989**, *32*, 1083.

(15) Wade, R. C.; Clark, K. J.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 140.

(16) Wade, R. C.; Goodford, P. J. *J. Med. Chem.* **1993**, *36*, 148.

(17) Goodsell, D. S.; Morris, G. M.; Olson, A. J. *J. Mol. Recognit.* **1996**, *9*, 1.

(18) Bitomsky, W.; Wade, R. *J. Am. Chem. Soc.* **1999**, *121*, 3004.

(19) Case, D. A.; Darden, T. A.; et al. . *AMBER, version 8*; Department of Pharmaceutical Chemistry, University of California: San Francisco, CA, 2004.

(20) Faham, S.; Hileman, R. E.; Fromm, J. R.; Linhardt, R. J.; Rees, D. C. *Science* **1996**, *271*, 1116.

(21) Ago, H.; Kitagawa, Y.; Fujishima, A.; Matsuura, Y.; Katsube, Y. *J. Biochem.* **1991**, *110*, 360.

(22) Jin, L.; Abrahams, J. P.; Skinner, R.; Petitou, M.; Pike, R. N.; Carrell, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14683.

(23) Skinner, R.; Abrahams, J. P.; Whisstock, J. C.; Lesk, A. M.; Carrell, R. W.; Wardell, M. R. *J. Mol. Biol.* **1997**, *266*, 601.

(24) Capila, I.; Hernaiz, M. J.; Mo, Y. D.; Mealy, T. R.; Campos, B.; Dedman, J. R.; Linhardt, R. J.; Seaton, B. A. *Structure* **2001**, *9*, 57.

(25) Swairjo, M. A.; Concha, N. O.; Kaetzel, M. A.; Dedman, J. R.; Seaton, B. A. *Nat. Struct. Biol.* **1995**, *2*, 968.

(26) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

(27) Woods, R. J.; Dwek, R. A.; Edge, C. J.; Fraser-Reid, B. *J. Phys. Chem.* **1995**, *99*, 3832.

(28) Kirschner, K. N.; Woods, R. J. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10541.

(29) Frisch, M. J.; Trucks, G. W.; et al. *Gaussian 03, Revision C.02*; Gaussian Inc.: Wallingford, CT, 2004.

(30) Cornell, W. D.; Cieplak, P.; Bayly, C.; Kollman, P. A. *J. Am. Chem. Soc.* **1993**, *115*, 9620.

(31) Ode, H.; Neya, S.; Hata, M.; Sugiura, W.; Hoshino, T. *J. Am. Chem. Soc.* **2006**, *128*, 7887.

(32) Sato, Y.; Hata, M.; Neya, S.; Hoshino, T. *J. Phys. Chem.* **2006**, *110*, 22804.

(33) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(34) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824.

(35) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics Modell.* **1996**, *14*, 33.

(36) Sato, Y.; Hata, M.; Neya, S.; Hoshino, T. *J. Phys. Chem.* **2006**, *110*, 22804.

(37) Ode, H.; Matsuyama, S.; Hata, M.; Hoshino, T.; Kakizawa, J.; Sugiura, W. *J. Med. Chem.* **2007**, *50*, 1768.

(38) Ode, H.; Matsuyama, S.; Hata, M.; Neya, S.; Kakizawa, J.; Sugiura, W.; Hoshino, T. *J. Mol. Biol.* **2007**, *370*, 598.

(39) Walters, P.; Dolata, M.; Babel, S. *A Molecular Structure Information Interchange Hub*; Department of Chemistry, University of Arizona: Tucson, AZ (accessed Aug 22, 2005).

(40) Shortle, D.; Simons, K. T.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11158.

(41) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153.

(42) Zagrovic, B.; Pande, V. S. *Biophys. J.* **2004**, *87*, 2240.

(43) AutoDock3.0.5_USGuide.pdf. Molecular Graphics Lab. http://www.scripps.edu/mb/olson/doc/autodock (accessed Nov 23, 2005).

(44) Azra, P.; Cindy, G.; Kenneth, A. J.; Xue-Jun, H.; Robert, J. L. *Glycobiology* **1995**, *5*, 83.

(45) Jin, L.; Abrahams, J. P.; Skinner, R.; Petitou, M.; Pike, R. N.; Carrell, R. W. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14683.

(46) Basten, J.; Jaurand, G.; Olde-Hanter, B.; Duchaussoy, P.; Petitoub, M.; van Boeckel, C. A. A. *Bioorg. Med. Chem. Lett.* **1992**, *2*, 905.

(47) Thompson, L. D.; Pantoliano, M. W.; Springer, B. A. *Biochemistry* **1994**, *33*, 3831.

(48) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43.

(49) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 457.

(50) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 609.

CT700029Q