JCTC Journal of Chemical Theory and Computation

Sugar Folding: A Novel Structural Prediction Tool for Oligosaccharides and Polysaccharides 1

Junchao Xia,[†,§] Ryan P. Daly,[†,§] Feng-Chuan Chuang,[†] Laura Parker,[†] Jan H. Jensen,[‡] and Claudio J. Margulis*[,†]

Department of Chemistry, University of Iowa, Iowa City, Iowa 52242, and Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Received February 8, 2007

Abstract: This paper is the first in a series of two articles where we report the development of fast sugar structure prediction software (FSPS). To the best of our knowledge, this is the first automated tool for the systematic study of conformations of complex oligosaccharides in solution. In contrast to previously developed molecular builders such as POLYS (Engelsen, S. B.; Cros, S.; Mackie, W.; Perez, S. *Biopolymers* **1996**, *39*, 417−433) where only information about the minimum energy conformation of disaccharide pairs is considered in order to build larger oligosaccharides, this tool is based on a systematic search of dihedral conformational space, optimization of structures using implicit solvation models, explicit molecular dynamics simulations, NOE calculations, and a very powerful substructure recognition algorithm and database. Our FSPS can rapidly find minimum-energy conformers and rank them according to different criteria. Two such criteria are the energy of the conformers in implicit solvent and the root-mean-square deviation (RMSD) of computed NOEs with respect to experimental data. Even though experimental NOEs may result from an average over conformers instead of a single structure, we find that sorting according to NOE RMSD constitutes a better estimator for the global free-energy minimum structure in explicit solvent (i.e., the most likely structure in solution). In contrast, the lowest-energy structure in implicit solvent does not usually correspond to the free-energy minimum. A harmonic approximation to compute free energies of each conformer does not appear to reverse this conclusion, indicating that either explicit hydrogen bonding to the solvent or anharmonic effects in the free energy or both are fundamentally important. In the first article, we discuss our methodology and study, as a proof of concept, a simple substituted disaccharide. In the second article, we focus on two complex human milk oligosaccharides.

1. Introduction

Carbohydrates are powerful biological markers because they contain multiple asymmetric carbon centers and possess unique structures and chemical properties. Complex carbohydrates are involved in numerous molecular recognition phenomena because of their exquisite specificity in interacting with proteins and other recognition agents. Glycoconjugates (glycoproteins and glycolipids) are actively involved in biological functions like tumor immunology,[2] cell growth and differentiation,[3] signal transduction,[4] apoptosis,[5]spermatogenesis,[6] and T-cell activation.[7] Oligosaccharides are recognized by different enzymes and by a family of proteins called lectins. Usually, both enzymes and lectins only recognize a particular fold of the sugar. This is why being able to predict the conformation of oligosaccharides in solution is of utmost importance. Unfortunately, only a small set of sugar-binding proteins have been cocrystallized

* Corresponding author e-mail: claudio-margulis@uiowa.edu.
† University of Iowa.
‡ University of Copenhagen.
§ Equal contributions.

10.1021/ct700033y CCC: $37.00
Published on Web 06/14/2007

with their corresponding oligosaccharides. It is therefore desirable to have a computational tool in place that will predict the conformational structures that sugars can adopt in solution. This software also constitutes a powerful aid in the interpretation of nuclear Overhauser effect (NOE) spectra.

Characterization and a priori prediction of conformations of biologically relevant carbohydrates in solution is difficult. Typical tools such as UV and IR spectroscopy are not suitable to study these molecules, and NMR spectra only give information about sets of statistically averaged conformations on a millisecond time scale. Depending on the free-energy difference between conformers in solution, the NMR will be compatible with either a single structure or an ensemble of flexible structures. Therefore, a theoretical prediction of the oligosaccharide conformation is usually necessary to understand the NMR data. In the past, we[8,9] and several other groups (for example, see refs 10 and 11) have performed molecular dynamics (MD) simulations in order to predict and understand the conformation of carbohydrates in solution. However, this approach has significant drawbacks. The main problem with using molecular dynamics to study the configuration of complex carbohydrates is that only certain conformations of these molecules are visited during the duration of a typical MD run. The challenge is in some ways similar to that of protein folding. One does not expect to see a protein fold on a time scale accessible by computer simulations. This problem is in fact much more pronounced in the case of sugars because proteins are linear polymers while oligosaccharides are often branched and motion of the different branches is often strongly coupled. It is important to recognize that the problem is not simply related to the potential energy barrier between different sugar conformers. This energy is usually low and compatible with thermal fluctuations at room temperature; the problem for molecular dynamics simulations is related to entropy.[12,13] This is particularly evident in the case of branched saccharides or sacharides with linkage points that are adjacent.

Although much has been learned from research in the protein field where extensive libraries[14−28] of peptide rotamers are available, no such tools currently exist for oligosaccharide systems. In fact, the problem of designing a rotameric library is topologically much more complex in the case of sugars than in the case of peptides. Sugars have many different linkage points, and their allowed dihedral space not only depends on the linkages and identities of the two monosaccharide units but also on the possibility of branching. Furthermore, recognizing subtrees of connected rings within a larger tree is in itself a highly complex problem in graph theory. In this article, we describe how our newly developed tool overcomes several of these difficulties and delivers results that are very hard to obtain otherwise with current computational tools. This will become more evident in the second paper where we present our results for a pair of complex human milk sugar oligosaccharides. Several methods have been used for the structural predictions of oligosaccharides. Most commonly, molecular dynamics in explicit solvent is used in order to predict NMR or NOE data in solution. In our experience,[8,9] only disaccharides readily visit all possible free-energy minima during typical molecular dynamics runs at room temperature. Larger oligosaccharides are generally trapped in local basins for longer than tens of nanoseconds, the length of a typical MD run. This is particularly true in the case of branched sugars or sugars with adjacent linkage points.

In the past, our group[8,9] has used the following reasonable scheme to study conformations of complex oligosaccharides in solution: First, a search through dihedral space of each independent isolated disaccharide pair is constructed. Long molecular dynamics runs for each component disaccharide are performed to obtain free-energy minima. Second, in order to build all possible oligosaccharides, a combinatorial approach is used in which all possible free-energy minima of the component disaccharides are combined to form all possible oligosaccharide structures. In most cases, many of the combinations are disallowed because of steric clashes or bad hydrogen-bond energetics, and only several combinations are obtained. In principle, this seems like a very large combinatorial problem, but in fact, for biologically relevant oligosaccharides, only tens or hundreds of structures need to be scrutinized. Unlike the case of proteins or polypeptides, sugar monomers are much bulkier, and therefore many conformations are disallowed, particularly when they are branched. The third and final step is to study the dynamics of the different oligosaccharide conformers in order to find which of these are stable in solution.

Even though the approach described above is reasonable, there are two main problems with it. First, it is very time-consuming. It requires long molecular dynamics for all component disaccharide pairs and further molecular dynamics of the resulting oligosaccharides. Second, this sampling method assumes that no other structure except those that correspond to free-energy minima of each disaccharide pair will be minima in the case of the oligosaccharide. This, although reasonable, may preclude the existence of other free-energy minima that appear due to stabilization through interaction between monomer units that are nonadjacent (i.e., stabilization due to secondary structure). This type of interaction could potentially be very common in the case of branched oligosaccharides particularly near crowded linkage points. In section 2, we describe a fast alternative method that overcomes these difficulties.

2. Simulation Methods

We have developed a completely automatic tool to study sugar molecules. A considerable part of a systematic search program involves the elucidation of the topology of polysaccharides by using ring perception techniques. Much work has gone into the development of algorithms for the determination of the smallest set of smallest rings and other ensembles of rings representative of a chemical structure.[29] However, the problem of ring recognition in carbohydrates is simpler in that compound rings are an exception to the norm in carbohydrate chemical structure. As a result, a more efficient ring perception algorithm can be used. The primary motivation in ring perception is to enumerate the dihedral degrees of freedom from glycosidic linkages in the molecule to allow a search of the conformational space. [Our ring perception algorithm is implemented through the use of graph

theoretical methods, and we have developed a C++ graph class to deal specifically with all aspects of saccharide ring topology recognition. Atomic coordinates are initially loaded from a generic XYZ file which contains no residue information. A graph object is initialized with vertices and edges which correspond to the atoms and bonds in an oligosaccharide. Subsequently, we derive the ring topology of a complex oligosaccharide by performing a series of depth-first and breadth-first searches of the graph structure. The linkages between different rings (monosaccharides) and the connectivity of side chains are derived using similar methods. Within the graph object, atoms (vertices) are stored with specific information which helps to expedite these processes.] The proverbial systematic search algorithm attempts to visit every possible conformation. Such an approach quickly becomes unfeasible even for the most efficient algorithms on the fastest machines. One way to dampen the effect of combinatorial explosion is to minimize the search space for each dihedral degree of freedom during iteration. Sugar residues, in particular those oligosaccharides relevant to biology, lend themselves very well to such a procedure because the allowed conformations of the ϕ and ψ dihedral angles of a glycosidic linkage are generally constrained to within approximately 30% or less of the total space. In a pentasaccharide with four glycosidic linkages, for example, the overall required search space is reduced to $0.30^4 = 0.0081 = 0.81\%$ of its unfiltered size. The reduction is substantially more important in the case of complex branched oligosaccharides for which our methodology is intended. In fact, the allowed number of conformers could be smaller for larger sugars than for smaller ones. This is the case for the oligosaccharides discussed in the second paper.

Our systematic approach can be described as a set of sequential steps:

1. The first step involves complex ring perception.[29] The input is an arbitrary "xyz" file. No atom typing or residue database is required.

2. A molecule is decomposed into its component oligosaccharides fragments. These fragments are checked against a database (which is currently being populated) using a sophisticated subtree matching algorithm as described in section 2.1. If a fragment of the molecule has already been studied, then no systematic search is carried out on it. This will save significant amounts of computational time in the future when the database has many entries. This obviously amounts to a sophisticated version of a rotameric library in which monomers are not simply linked sequentially, but the effect of branching and adjacency is considered through bonds as well as through space. Rotameric libraries for proteins usually only have information about pairs of residues; our approach will store information about larger sugar subfragments. This is feasible since the number of sugar monomers in a typical biologically relevant oligomer is much smaller than the number of amino acids in a protein and since monosaccharides are in general considerably bulkier than amino acids. This procedure is most useful in the case when branching is present since sterics will significantly restrict conformation space and, consequently, the number of configurations in space that we need to store. The case of linear sugars with nonadjacent linkages is the least interesting to us since the number of configurations to store becomes exponentially large as the size of the oligosaccharide grows.

3. Items not previously studied or stored are separated into monosaccharide residues and side chains. We then perform a systematic grid search for the allowed $\phi-\psi$ pairs for each residue linkage and side-chain linkages. The angular increment can be arbitrarily chosen; we have used $10-20°$ for residue linkages and $60-120°$ for side-chain linkages. After a clash check using a hard sphere criterion, we obtain corresponding steric Ramanchandran maps for all residue and side-chain linkages. Clash checks are only performed between atoms in different residues, not within the same ring.

4. The oligosaccharide is reconstructed by reassembling the linkages one by one at corresponding allowed conformations. At this point, depending on the size of the structure pool, coarse graining can be applied to constrain the number of candidate structures. For example, four neighboring points in Ramachandran space will become a new point which is calculated as the geometric center of the allowed points. As opposed to other clustering schemes, this coarse-graining method is unlikely to miss small isolated regions in configuration space.

5. After obtaining the sterically allowed conformations, we perform energy minimizations using an implicit solvent model. In this paper and in the second paper, we have used different software programs[30−33] and force fields[34,35] to achieve this.

6. We pool the minimized structures into unique conformational families. We consider that two conformations have a unique structure if the energy difference $\Delta E < 5.0$ kcal/mol and the difference in each of the dihedral angles is $<10°$. We keep the structure with lowest energy in each family, and we define this as a "unique" conformer.

7. Unique conformers can be sorted on the basis of different criteria. We have used an energy rank as well as a rank based on the root-mean-square deviation (RMSD) between experimental NOE values for proton pairs on different monosaccharide units and our computed values for the unique structures. Our approach for computing NOEs is the same as that used by Cumming and Carver,[36,37] which is based on the model-free approach.[38,39]

8. Finally, we run short 5 ns molecular dynamics simulations in explicit solvent in order to gauge the stability of each of the different unique conformers and in order to compute time-averaged NOEs. [A structure is deemed stable if after 5 ns of simulation glycosidic angles have not changed to a different local minimum. Clearly, these short simulations only indicate whether a structure is in a deep local minimum as compared to *KT* and not whether the structure is at a global free-energy minimum. Much more expensive procedures such as parallel tempering can be applied if accurate relative free energies between different conformers of complex oligosaccharides generated by our fast sugar structure prediction software (FSPS) are sought.]

2.1. Sub-Tree Recognition and Database. Because of the possibility of structural branching in sugars, the act of querying a database in search of a set of structures most

similar to the molecule of interest is highly complex. A practical carbohydrate database query protocol was presented by Aoki et al.[40−43] While their method makes use of a scoring algorithm to sort matches, the method we use accomplishes the same task through the use of a slightly more generalized algorithm which effectively solves the maximum common subgraph problem for trees with labeled nodes and edges.

The first methods to solve the graph isomorphism problem were mostly set theoretic in nature.[44,45] More recently, some researchers have focused on using the eigenvalues and eigenvectors associated with the adjacency matrix of a molecule as a means to distinguish it from others. However, this method has two limitations. First of all, it does not explicitly take vertex labels (i.e., atom types) into account. Second of all, it is only capable of verifying exact matches.

The methodology in our FSPS makes use of association graphs to solve this problem[46] by using an approach developed by Hopfield which is described in ref 47. In order to make a molecular-structure-based database query practical, in this method, a routine which allows the comparison of two structures to find any substructure in common is used. The structural information stored in our database is a residue graph, which is simply the graph generated by viewing each residue in an oligosaccharide as a vertex and each glycosidic linkage as an edge between vertices. Residue graphs are the objects which are compared when a query is made to the database. This type of comparison is made by solving the graph isomorphism problem, which seeks to find the maximum subgraph (i.e., the subgraph containing the largest number of vertices) present in both graphs. Because the vast majority of biological oligosaccharides contain no compound ring structure, the resulting residue graphs are tree graphs since they are presumed to contain no rings (or cycles). The algorithm used in this work finds the maximum subtree common to both oligosaccharides and is derived from a method by Jain and Wysotski.[47] This method is dependent upon the generation of an association graph, which is basically a map from one residue graph to another (see Jain and Wysotski[47]). The generation of the association graph is a process in and of itself and can be optimized independently of the actual search. The main criterion in its optimization is to make it as small as possible and with as few edges as possible. Once the association graph is generated, a neural network algorithm is used to find the maximum clique[47] in the association graph. Maximum cliques correspond to subsets of vertices in the association graph which map residues from a new oligosaccharide to those of the structures already stored in our database. The resulting map points out common substructures between the molecule from the database and the molecule of interest. If a match is returned which is 100% of the size of a molecule in the database, then the sterically allowed conformational space stored with this entry in the database is used as an admissible search space for the mapped portion of the molecule of interest.

The association graph method has the advantage of being highly customizable. In addition to the ability to take into account atom types, other information such as atom chirality and bond type can be used to further eliminate possible matches. This procedure greatly shrinks the size of the association graph. New conformations are constructed on the basis of stored vectors containing coupled $\phi-\psi$ information for each linkage of the oligosaccharide in the database. The rest of the oligosaccharide is assembled by avoiding clashes with the database fragment. In the second paper,[48] we describe the use of this method for the analysis of conformations of complex milk sugars.

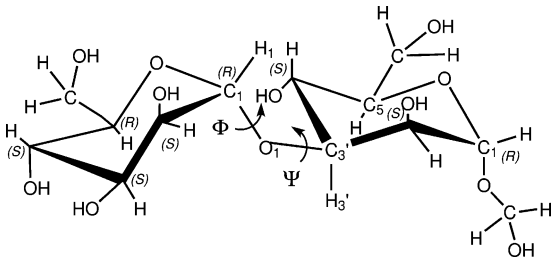

Figure 1. Schematic representation of the α-D-Man-(1→3)-α-D-Man-O-Me disaccharide molecule. The two dihedrals angles are defined as $\phi = H_1-C_1-O_1-C'_3$ and $\psi = C_1-O_1-C'_3-H'_3$.

3. Results and Discussion: A Simple but very Important Test Problem

The simplest possible example that can be used in order to exemplify our procedure and to test the accuracy and validity of the different approximations involved is a substituted disaccharide. We have chosen α-D-Man-(1→3)-α-D-Man-O-Me, with the schematic representation shown in Figure 1, because molecular dynamics simulations in explicit solvent can be converged to probe its full free-energy landscape on a several nanosecond time scale. Furthermore, this sugar has been well-studied by means of the nuclear Overhauser effect,[49] relaxed potential energy surfaces through an extensive molecular mechanics (MM) scheme,[50,51] and also as a fragment of an oligosaccharide via molecular dynamics.[52] The full free-energy landscape is not easily accessible for complex oligosaccharides like those studied in the second paper. In the case of the current paper, by having the full free-energy landscape of the molecule as a function of glycosidic dihedral angles, we are able to probe which sorting criteria is best (sequential step 7 in section 2) for our FSPS. Furthermore, because of the small system size, high-level ab initio calculations using an implicit solvent model can be carried out to thoroughly test the accuracy of molecular mechanics energetic predictions.

3.1. Implicit and Explicit Solvent, Force Fields, and ab Initio Calculations. What Matters and What Does Not for the Correct Prediction of Sugar Structures in Solution. *3.1.1. Using MM3 with TINKER.* Figure 2 shows the distribution of unique structures from our systematic search in $\phi-\psi$ glycosidic space using MM3[34] with TINKER.[30,31] The two dihedral angles are defined as $\phi = H_1-C_1-O_1-C'_3$ and $\psi = C_1-O_1-C'_3-H'_3$ as shown in Figure 1. $\phi-\psi$ torsion angles have been adjusted in steps of 10° over the whole angular space. At each sterically allowed point, an energy minimization was performed using the generalized Born surface area (GBSA) implicit solvent model.[53,54] Rotations were also performed for the hydroxymethyl group. Figure 2a displays the distribution of unique conformations

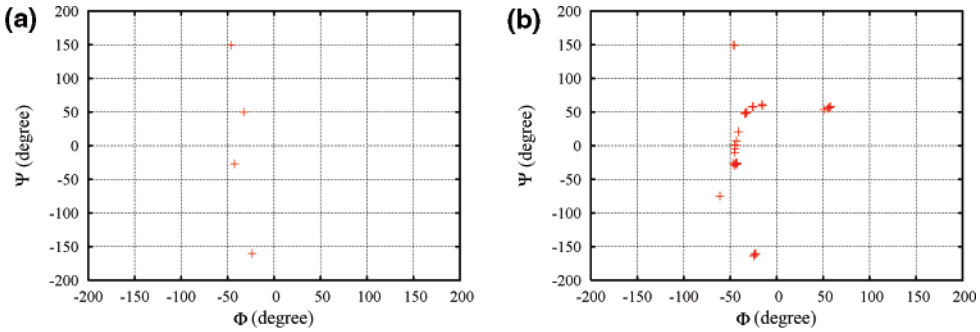

Figure 2. Distribution of unique conformations in $\phi-\psi$ glycosidic space in the case of (a) no side-chain rotations and (b) with 120° rotations of the first dihedral angle on any side chain with at least two rotable dihedral angles. Side-chain rotation reveals more local minima. Energy minimizations were performed using TINKER[30,31] with the MM3 force field[34] and GBSA implicit solvent model.[53,54]

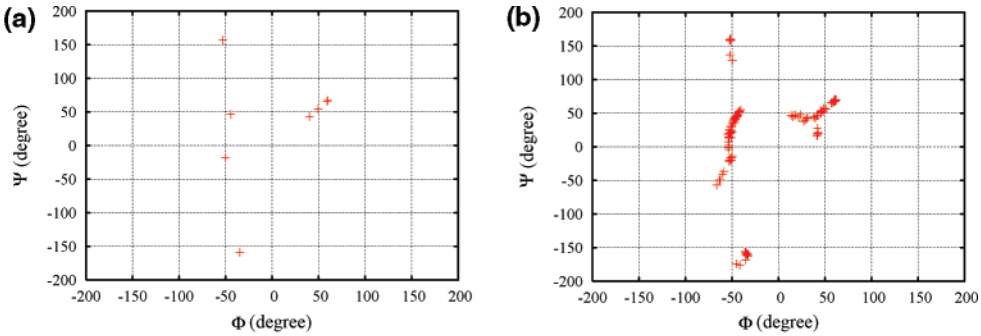

Figure 3. Same as Figure 2 except that the energy minimization is done using GROMACS with the OPLS-AA force field in the gas phase.

Table 1. The Potential Energy Difference (ΔE) and Free Energy Difference (ΔG) (kcal/mol) of Each Unique Conformations in Figure 2a Using Different MM and QM Procedures and Basis Sets[a]

		conf. 1	conf. 2	conf. 3	conf. 4
TINKER(MM3) with GBSA	(ϕ,ψ)	(−32.5, 49.6)	(−42.7, −26.8)	(−23.6, −160.4)	(−46.4, 149.4)
	ΔE	0.0	1.44	1.85	4.79
GROMACS(OPLS-AA) gas phase	(ϕ,ψ)	(−44.3, 46.4)	(−50.1, −18.3)	(−34.6, −159.1)	(−52.8, 157.0)
	ΔE	0.0	2.81	3.47	4.51
GAMESS gas phase	(ϕ,ψ)	(−36.5, 48.8)	(−50.8, −27.9)	(−28.9, −150.1)	(−42.7, 158.0)
B3LYP/6 − 31*G*(*d*,*p*)	ΔE	0.0	1.99	0.82	3.62
(nvib = 2)	ΔG	0.0	2.71	0.85	4.63
GAMESS implicit solvent	(ϕ,ψ)	(−36.1, 45.9)	(−50.7, −29.9)	(−28.2, −150.7)	(−43.5, 159.2)
B3LYP/6 − 31*G*(*d*,*p*)PCM	ΔE	0.0	0.63	0.96	4.04
MD GROMACS(OPLS-AA)	(ϕ,ψ)	(−38.7, 49.2)	(−52.0, −14.0)		
with explicit solvent	ΔG	0	−3		

[a] MM calculations with MM3 and OPLS-AA show that conformation 1 is the global energy minimum. QM calculations also find that conformation 1 is the global energy minimum. Note that QM calculations in the gas phase change the energy ordering of conformations 2 and 3. On the contrary, MD simulations in explicit solvent and experiments reveal that conformation 2 is in fact the global free-energy minimum.

obtained without any side-chain rotation. Only four minimized conformations are found. However, several other minima are shown in Figure 2b as we rotate the first dihedral angle of the hydroxymethyl group. The latter distribution is consistent with the adiabatically relaxed potential energy surface of Imberty et al. using the MM2 force field.[50] Including a full dihedral search for all hydroxyl groups instead of only the hydroxymethyl group produces more mimina (results not shown); however, this is expensive and does not appear, at least in this particular case, to significantly modify the energy ordering of the conformers.[50]

3.1.2. Using GROMACS with the OPLS-AA Force Field and ab Initio Calculations with GAMESS. At the time of our simulations, GROMACS[32,33] did not offer an implicit solvent option. The OPLS-AA potential[35] appears to show more local minima than MM3 as can be appreciated in Figure 3; however, these extra minima are at much higher energies. The four main unique conformations found from our MM3 TINKER calculations are similar to those predicted by OPLS-AA and appear to keep the same relative energy ordering as shown in Table 1.

We have also analyzed the relative potential energies of these four unique conformations by quantum mechanical (QM) calculations using GAMESS[55,56] (Table 1). All QM calculations in the gas phase and in implicit solvent[57,58] appear to indicate that conformation 1 is the lowest-energy

Table 2. Comparison Between Observed and Calculated NOE Values from the α-D-Man-(1→3)-α-D-Man-O-Me Disaccharide[a]

		NOE observed		NOE calculated									
				conf. 1		conf. 2		conf. 3		conf. 4		MD	
proton 1	proton 2	absolute	relative	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.	abs.	rel.
H_1	H_2	0.11	1.0	0.12	1.0	0.12	1.0	0.12	1.0	0.13	1.0	0.09	1.0
	H'_3	0.11	1.0	0.18	1.5	0.13	1.08	0.0	0.0	0.0	0.0	0.13	1.4
	H_2	0.0	0.0	0.0	0.0	−0.01	−0.08	0.01	0.08	0.14	1.08	0.0	0.0
	H'_4	0.01	0.1	0.04	0.33	0.0	0.0	0.17	1.42	0.0	0.0	0.0	0.0
H'_2	H'_1	0.065	1.0	0.11	1.0	0.11	1.0	0.10	1.0	0.11	1.0	0.10	1.0
	H_5	0.04	0.60	0.02	0.18	0.08	0.73	0.0	0.0	0.0	0.0	0.08	0.8
RMSD				1.3		0.65		6.56		1.78		0.63	

[a] The four conformations are those in Figure 2a, and the experimental data are from Reference 49. Clearly, of all unique structures, conformation 2 has the closest NOE values to experimental data as demonstrated by its RMSD. The time-averaged NOE values from all MD trajectories in Figure 4 are close to that of conformation 2.

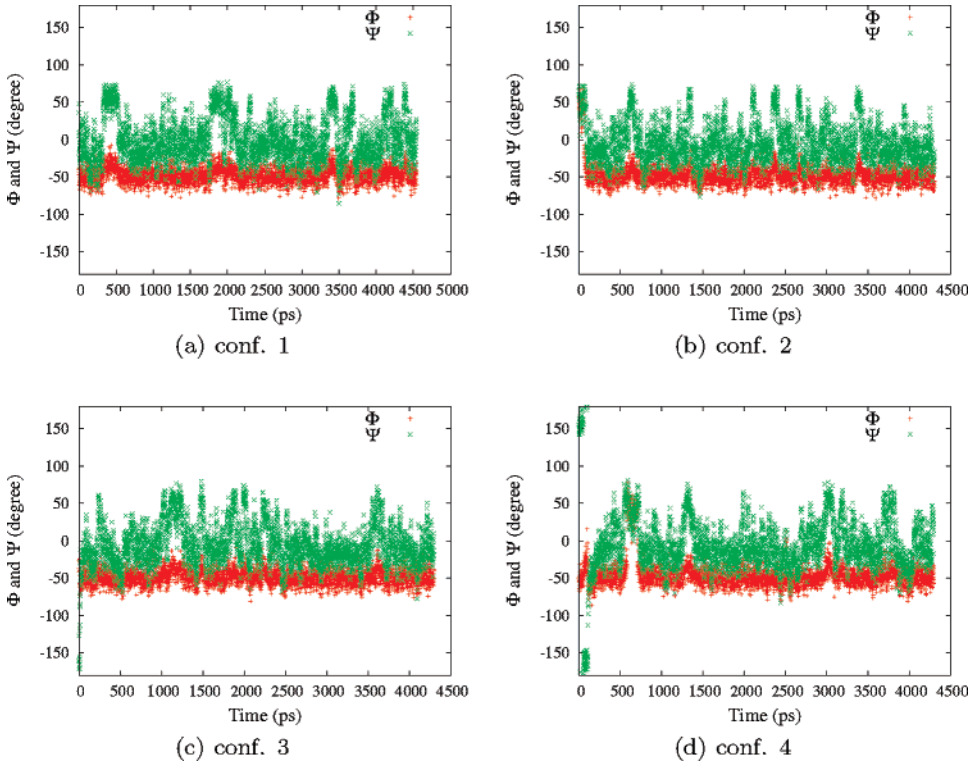

Figure 4. Time evolutions of ϕ and ψ for different unique conformations as shown in Figure 2a using the OPLS-AA force field and explicit SPC water. Conformations 3 and 4 are not preferred, and conformation 2 is visited more frequently than conformation 1.

minimum, similarly to what we found in the case of the MM calculations. We also note that QM calculations in the gas phase change the energy ordering of conformations 2 and 3. This does not occur when we use an implicit solvent model. We have also computed ab initio free energies (Table 1) using a harmonic approximation. These calculations also predict conformation 1 to be the one with the lowest free energy.

3.1.3. The NOE Sorting Criteria. The goal of the FSPS is to predict the most likely structure or structures of oligosaccharides in solution. In order to evaluate and compare the four unique conformations (Figure 2) found by our algorithm, we computed their corresponding NOEs using the procedure described by Cumming and Carver[36,37] from the model-free approach.[38,39] We find that, even though, under the approximations used here, conformation 1 is the global-energy and free-energy minimum in implicit solvent, conformation 2 has in fact a NOE closest to the experimental data as shown in Table 2!

The fact that conformation 2 is indeed the most likely structure in solution is confirmed by our molecular dynamics simulations in explicit solvent. In fact, we predict a free-energy difference of about 3 kcal/mol between conformation 2 and conformation 1 (see subsection 3.1.4). This indicates that gas-phase energies, energies in implicit solvent, or free energies computed using a harmonic approximation may not be an adequate estimator for the most likely structure in solution. This may be due to anharmonic effects or to hydrogen bonding with the solvent. It is well-known that sugars easily form structures stabilized by water-mediated hydrogen bonds.[8]

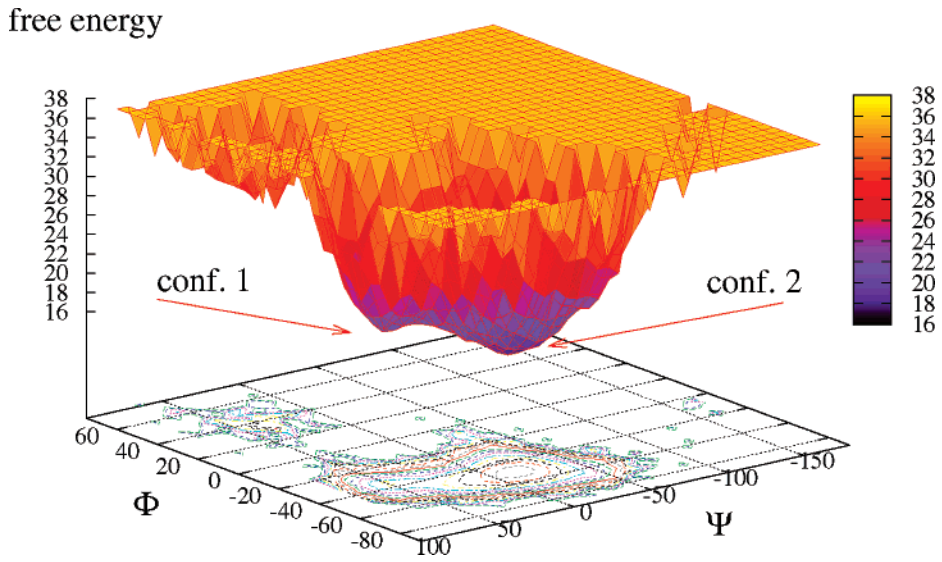

Figure 5. Free energy calculated from the probability distribution of $\phi-\psi$ obtained from the time evolutions shown in Figure 4. Conformation 2 is the global free-energy minimum in solution, while conformation 1 is the global energy minimum in implicit solvent and the gas phase. The free-energy difference between these two minima is about 3 kcal/mol at 298 K.

Since NOE values correspond to an average over an ensemble of structures, in general, there is not a one-to-one correspondence between a set of NOE values and a particular sugar conformation. However, it appears that, at least in the case of this sugar, free-energy differences in solution between conformers are large enough that the NOE values are dominated by only one conformational structure. Our FSPS can easily enumerate all reasonable minimum-energy structures. It takes a matter of seconds to sort these structures according to their root-mean-square deviation with respect to experimental NOEs. On the basis of our results in this article and those in the second paper, it appears that this sorting criterion is a reliable estimator for the most likely structures in solution.

3.1.4. The Free-Energy Landscape in Explicit Solvent. We used GROMACS[32,33] with the OPLS-AA force field[35] in explicit simple point charge (SPC) water[59] to model the dynamics of our system. We started runs from each of the four different unique conformations in Table 1. Figure 4 shows the time evolution of ϕ and ψ for each run. Regardless of the initial conformation, it is obvious that the molecule readily transfers between conformations 1 and 2 when the system reaches equilibrium. Conformations 3 and 4 are not preferred in water. It is also clear from the plot that molecules spend more time in conformation 2 than in conformation 1. The time-averaged NOE values from these four MD trajectories shown in Figure 4 are close to that of unique conformation 2 as shown in Table 2 and closely coincide with experiments.

From the time evolution of the dihedral angles, we computed the free energy $f = -KT \ln P(\phi,\psi)$, where $P(\phi,\psi)$ is the probability distribution of $\phi - \psi$. In Figure 5, we see that unique conformation 2 is indeed the global free-energy minimum with a free-energy difference at 298 K of about 3 kcal/mol with respect to unique conformation 1, which is a metastable state. Hence, free-energy calculations from explicit MD coincide with the very inexpensive a priori prediction of our FSPS on the basis of the deviation of single unique structure NOE values with respect to experiments.

4. Conclusions

Much can be learned from the α-D-Man-(1→3)-α-D-Man-O-Me system since it has been fully experimentally characterized and since MD time scales are suitable to correctly capture the relative probability of all minima and therefore the corresponding free-energy landscape. It is clear that, in order to predict which conformer is the most likely in solution only on the basis of energetics, the correct relative probability (i.e., the free-energy landscape) of the conformers must be obtained. This probability landscape was accessible in this case because the molecule in question is relatively small and the dynamics is ergodic on the time scale of our simulations. For larger sugars, particularly branched sugars or sugars with adjacent linkage points, this brute-force approach is simply not viable.

Our automatic structure prediction algorithm was able to capture all corresponding energy minima in a tiny fraction of the time required to carry out molecular dynamics simulations long enough to sample them. A simple sorting criterion based on energies or free energies in implicit solvent was not adequate to establish a ranking for these conformers in solution. On the other hand, given the experimental NOEs, a ranking can be devised on the basis of the RMSD between these and those computed from our unique structures. The systematic search algorithm combined with the RMSD sorting criteria provides an accurate definition for the lowest free-energy structure without the need to run any expensive MD simulations. Identifying structure 2 as the most likely configuration in solution (even though its predicted energy in an implicit solvent was higher than that of structure 1) took a minute fraction of the time required to carry out the MD simulations which later confirmed the result.

Our approach provides a viable way to analyze the structure of oligosaccharides since, in our experience, for

sugars with six or seven arbitrarily connected rings, the most relevant energy minima can be obtained within a time scale of hours. By comparing the NOEs of each of these structures against experiments, it is fairly easy to establish a ranking of structures in solution. In the second paper, we show that our algorithm is able to capture many more stable local minima than those previously found by carrying out explicit solvent MD simulations. We will also show that our sorting criteria indeed capture the most likely structures in solution. These results are very promising, and we hope that the study of complex oligosaccharides will become easier as our database of fragments becomes larger.

Acknowledgment. This research was funded by Grant #05-2182 from the Roy J. Carver Charitable Trust awarded to C.J.M. and by the Skou Fellowship from the Danish Natural Sciences Research Council awarded to J.H.J.

References

(1) Reference deleted in press.

(2) Wölfl, M.; Batten, W. Y.; Posovszky, C.; Bernhard, H.; Berthold, F. *Clin. Exp. Immunol*. **2002**, *130*, 441−448.

(3) Schaade, L.; Thomssen, R.; Ritter, K. *Z. Naturforsch., C: J. Biosci.* **2000**, *55*, 1004−1010.

(4) Simons, K.; Toomre, D. *Nat. Rev. Mol. Cell Biol.* **2000**, *1*, 31−39.

(5) Simon, B. M.; Malisan, F.; Testi, R.; Nicotera, P.; Leist, M. *Cell Death Differ*. **2002**, *9*, 758−767.

(6) Takamiya, K.; Yamamoto, A.; Furukawa, K.; Zhao, J.; Fukumoto, S.; Yamashiro, S.; Okada, M.; Haraguchi, M.; Shin, M.; Kishikawa, M.; Shiku, H.; Aizawa, S.; Furukawa, K. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 12147−12152.

(7) Nohara, K.; Ozawa, H.; Taic, T.; Sajib, H.; Fujimakia, H. *Biochim. Biophys. Acta* **1997**, *1345*, 207−214.

(8) Veluraja, K.; Margulis, C. J. *J. Biomol. Struct. Dyn.* **2005**, *23*, 101−111.

(9) Margulis, C. J. *J. Phys. Chem. B* **2005**, *109*, 3639−3647.

(10) Almond, A.; Petersen, B. O.; Duus, J. *Biochemisty* **2004**, *43*, 5853−5863.

(11) Woods, R. *Glycoconjugate J.* **1998**, *15*, 209−216.

(12) Boone, M. A.; Striegel, A. M. *Macromolecules* **2006**, *39*, 4128−4131.

(13) Striegel, A. M. *Macromolecules* **2003**, *125*, 4146−4148.

(14) Gordon, D. B.; Hom, G. K.; Mayo, S. L.; Pierce, N. A. *J. Comput. Chem*. **2003**, *24*, 232−243.

(15) Desmet, J.; Spriet, J.; Lasters, I. *Proteins: Struct. Funct. Genet.* **2002**, *48*, 31−43.

(16) Kono, H.; Saven, J. G. *J. Mol. Biol.* **2001**, *306*, 607−628.

(17) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct. Funct. Genet.* **2000**, *40*, 389−408.

(18) Petrella, R. J.; Lazaridis, T.; Karplus, M. *Folding Des.* **1998**, *3*, 353−377.

(19) Huang, E. S.; Koeh, P.; Levitt, M.; Pappu, R. V.; Ponder, J. W. *Proteins. Struct. Funct. Genet.* **1998**, *33*, 204−217.

(20) Roland, L.; Dunbrack, J.; Cohen, F. E. *Protein Sci.* **1997**, *6*, 1661−1681.

(21) Bower, M. J.; Cohen, F. E.; Dunbrack, R. L. *J. Mol. Biol.* **1997**, *267*, 1268−1282.

(22) DeMaeyer, M.; Desmet, J.; Lasters, I. *Folding Des.* **1997**, *2*, 53−66.

(23) Marcusa, E.; Kellera, D. A.; Shibataa, M.; Ornsteinb, R. L.; Rein, R. *Chem. Phys.* **1996**, *204*, 157−171.

(24) Desjarlais, J. R.; Handel, T. M. *Protein Sci.* **1995**, *4*, 2006−2018.

(25) Roland, L.; Dunbrack, J.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543−574.

(26) Koehl, P.; Delarue, M. *Nat. Struct. Biol*. **1995**, *2*, 163−170.

(27) Leach, A. R. *J. Mol. Biol.* **1994**, *235*, 345−356.

(28) Kolinski, A.; Godzik, A.; Skolnick, J. *J. Chem. Phys.* **1993**, *98*, 7920−7433.

(29) Balducci, R.; Pearlman, R. S. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 822−831.

(30) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(31) Jay, W.; Ponder, F. M. R. *J. Comput. Chem.* **1987**, *8*, 1016−1024.

(32) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem*. **2005**, *26*, 1701−1718.

(33) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Biol.* **2001**, *7*, 306−317.

(34) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *J. Am. Chem. Soc.* **1989**, *111*, 8551−8566.

(35) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *117*, 11225−11236.

(36) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6664−6676.

(37) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6676−6683.

(38) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.

(39) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559−4570.

(40) Ueda, N.; Aoki-Kinoshita, K. F.; Yamaguchi, A.; Akutsu, T.; Mamitsuka, H. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1051−1064.

(41) Yamaguchi, A.; Aoki, K. F.; Mamitsuka, H. *Inf. Process. Lett*. **2004**, *92*, 57−63.

(42) Aoki, K. F.; Ueda, N.; Yamaguchi, A.; Akutsu, T.; Kanehisa, M.; Mamitsuka, H. *Sigmod Rec.* **2004**, *33*, 33−38.

(43) Aoki, K. F.; Yamaguchi, A.; Ueda, N.; Akutsu, T.; Mamitsuka, H.; Goto, S.; Kanehisa, M. *Nucleic Acids Res.* **2004**, *32*, W267−W272.

(44) McKay, B. D. *Congr. Numerantium* **1981**, *30*, 45−87.

(45) Ullmann, J. R. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.

(46) Pelillo, M.; Siddiqi, K.; Zucker, S. W. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1105−1120.

(47) Jain, B. J.; Wysotzki, F. *Neurocomputing* **2005**, *63*, 45−67.

(48) Xia, J.; Daly, R. P.; Chuang, F.-C.; Parker, L.; Jensen, J. H.; Margulis, C. J. **2007**, *4*, 1629−1643.

(49) Brisson, J. R.; Carver, J. P. *Biochemistry* **1983**, *22*, 3680−3686.

(50) Imberty, A.; Tran, V.; Pérez, S. *J. Comput. Chem*. **1989**, *11*, 205−216.

(51) Imberty, A.; Gerber, S.; Tran, V.; Pérez, S. *Glycoconjugate J*. **1990**, *7*, 27−54.

(52) Woods, R. J.; Pathiaseril, A.; Wormald, M. R.; Edge, C. J.; Dwek, R. A. *Eur. J. Biochem*. **1998**, *258*, 372−386.

(53) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(54) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett*. **1995**, *246*, 122−129.

(55) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Koseki, S.;Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347−1363.

(56) Gordon, M. S.; Schmidt, M. W. Advances in Electronic Structure Theory: GAMESS a Decade Later. In *Theory and Applications of Computational Chemistry, the First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005.

(57) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032−3041.

(58) Li, H.; Jensen, J. H. *J. Comput. Chem.* **2004**, *25*, 1449−1462.

(59) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981.

CT700033Y