JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

COMMUNICATION

pubs.acs.org/JACS

Searching and Optimizing Structure Ensembles for Complex Flexible Sugars

Junchao Xia,*,† Claudio J. Margulis,‡ and David A. Case*,†

†BioMaPS Institute and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854, United States

‡Department of Chemistry, University of Iowa, Iowa City, Iowa 52242, United States

Supporting Information

ABSTRACT: NMR restrictions are suitable to specify the geometry of a molecule when a single well-defined global free energy minimum exists that is significantly lower than other local minima. Carbohydrates are quite flexible, and therefore, NMR observables do not always correlate with a single conformer but instead with an ensemble of low free energy conformers that can be accessed by thermal fluctuations. In this communication, we describe a novel procedure to identify and weight the contribution to the ensemble of local minima conformers based on comparison to residual dipolar couplings (RDCs) or other NMR observables, such as scalar couplings. A genetic algorithm is implemented to globally minimize the R factor comparing calculated RDCs to experiment. This is done by optimizing the weights of different conformers derived from the exhaustive local minima conformational search program, fast sugar structure prediction software (FSPS). We apply this framework to six human milk sugars, LND-1, LNF-1, LNF-2, LNF-3, LNnT, and LNT, and are able to determine corresponding population weights for the ensemble of conformers. Interestingly, our results indicate that in all cases the RDCs can be well represented by only a few most important conformers. This confirms that several, but not all of the glycosidic linkages in histo-blood group "epitopes" are quite rigid.

Carbohydrates play an important role in many molecular recognition phenomena. Their flexibility in solution is often important to their function[1,2] and has been investigated for several simple disaccharides, complex oligosaccharides. and polysaccharides.[3,5] Bush and co-workers have categorized their flexibility as that arising either from fluctuations within a single free energy minimum or due to transitions among different minima in glycosidic linkage space.[5]

For complex oligosaccharides, residual dipolar coupling (RDC)[6] measurements in anisotropic solution environments provide important global structural information. Calculating RDCs from structure requires knowing first the alignment tensor of a given model structure or rigid fragment. A single value decomposition (SVD) method[7] is often used to fit the alignment tensor to experimental data and the particular rigid domain.[8] However, this method is not suitable for flexible molecules for which a single global alignment tensor does not exist. Furthermore, deriving the alignment tensor using SVD requires at least 5 independent RDC values from each rigid structure fragment, and often significant uncertainty is present in the form of "structural noise". This is particularly problematic when only few dipolar couplings are available for fitting.[9]

Alternatively, for flexible systems, the alignment tensor has often been estimated from simulations.[10] For example, the PALES approach[10] estimates the alignment tensor by performing a Monte Carlo search of molecules in the vicinity of an infinite two-dimensional plate. Several other methods estimate the alignment tensor from 3D molecular conformation, using the radius of gyration tensor,[11] the moment of inertia,[12] or a direct integration in two or four dimensional space related to the Euler angles of molecular orientations.[13]

The idea of estimating the alignment tensor from molecular shape[11] has been applied in several research groups[14−21] to build ensembles of partially folded or unfolded proteins. We show here that broadly similar ideas, adapted to carbohydrates, provide remarkable insights into the conformational ensemble of oligosaccharides. Residual dipolar couplings in liquid crystal media have been utilized to determine the conformational structure of several carbohydrates.[2,4,5,22,23] However, significant challenges exist for the wide applicability of RDCs to study complex sugars.

Recently, the Margulis group has developed a fast structural prediction software (FSPS) to search for energy minima in glycosidic conformation space with the assistance of NMR data.[24] The general framework includes four major steps: (1) a coarse-grained systematic search in dihedral space for intramolecular clashes, (2) energy optimizations of sterically allowed conformers in the gas phase or in implicit solvent through an interface to external molecular modeling packages, (3) pooling large numbers of energy minimized structures into a smaller set of unique consensus structures that are conformationally and energetically similar, and (4) producing a ranking of these groups of conformers in comparison to calculated NMR observables such as NOEs, RDCs, or J couplings. The limitation with this approach is that so far the NMR observables have only been compared to those derived from individual conformers instead of against a properly weighted ensemble of conformers. This approach is destined to fail when significant flexibility is present.

Because of their relevance to the immune system of infants,[25] and because several of these systems have already been the subject of detailed RDC as well as other NMR techniques studies[4,5,26,27] and computational studies,[24] we focus here on six

Received: June 13, 2011
Published: August 24, 2011

different human milk oligosaccharides, LNF-1, LNF-2, LNF-3, LND-1, LNnT, and LNT, shown in Scheme 1S (Supporting Information).

In this communication, we present a framework to search for the best conformational ensemble of oligosaccharides that, when properly weighted, match experimental RDC data in solution. We assume that each conformer within the ensemble has an alignment tensor and a corresponding set of RDC values and that the population averaged RDCs correspond to experimental values. Abandoning the philosophy of restrained MD simulations that match NMR constraints, we instead develop two independent programs using random walk Monte Carlo (RWMC) and a genetic algorithm (GA) to optimize the weights given to each conformer previously obtained from the exhaustive FSPS algorithm.[24] The total number of conformers derived from FSPS in each case (see ref 24) varies roughly from 1000 to 10 000, depending on the number of monosaccharide residues involved. These oligosaccharides thus provide a good test case: they are large enough to have significant flexibility yet small enough to permit a systematic exploration of the conformational space.

In a previous article,[24] we computed RDC values for each of these oligosaccharide conformers by deriving the alignment tensor from the gyration tensor of molecular shape[11] (see eqs 1S−4S in the Supporting Information). The R factor[28] between RDCs corresponding to individual conformers and those reported experimentally[5] were then obtained using eq 5S in the Supporting Information. In this way, a RDC ranking of R factors was constructed, with the smallest R value representing the best single conformational structure in comparison with experiments. In the current study on multiconformers, we have found that such preranking of individual structures is very useful to bias the initial condition of the GA or RWMC searches. This is crucial for fast convergence on such a large number of conformers (between 1000 and 10 000).

Results from our multiple-structure optimization of R factors, which assumes that each structure has an independent alignment tensor and a corresponding set of RDCs, are shown in Table 1 in comparison with experimental data from the group of Allen Bush[15] for human milk sugars LNF-1, LNF-2, LNF-3, LND-1, LNnT, and LNT. In each case, the averaged RDC value for the ith spin vector is calculated by weighting the result of individual conformers as described in eq 1

$$Q_i = \sum_{k=1}^{M} P_k Q_{ki} \qquad (1)$$

where Q_{ki} is the ith RDC value of the kth conformer included in the average and P_k is the probability weight of the kth structure.

Figure 1S compares the efficiency of the RWMC and GA. While an MC step is significantly faster than a generation of the GA, the GA converges to smaller values of the R factor. Because of this we only focus here on results derived from the GA. A set of checks (Table 4S and 5S, Figure 8S, 9S and 10S) in the Supporting Information give us confidence that our results are meaningful and unique. The tests show that populations can be recovered from calculated RDC's in a robust fashion, that the final results are independent of any starting guesses, that ensembles restricted to randomnly chosen subsets of the full space have poorer fits than the full calculation, and that the ability to converge on an ensemble degrades (as expected) as the number of experimental RDC's is reduced.

Table 1 displays R factors of calculated RDCs in comparison with experimental values[15] for human milk sugars LNF-1, LNF-2, LNF-3, LND-1, LNnT and LNT (see also Figure 2S and Table 1S in Supporting Information for detailed RDC values). The R factor of the single best conformer is contrasted against that obtained from the multistructure fitting algorithm (eqs 1 and 5S, Supporting Information). We see that the multistructure optimization improves the values of R factor, especially in the case of LND-1, LNnT, and LNT.

Table 1. R Factors of Calculated RDCs in Comparison with Experimental Data[5] for LNF-1, LNF-2, LNF-3, LND-1, LNnT, and LNT[a]

	LNF-1	LNF-2	LNF-3	LND-1	LNnT	LNT
BestS	0.188	0.137	0.365	0.207	0.094	0.101
BestM	0.176	0.120	0.337	0.130	0.051	0.055

[a] "bestS" denotes RDCs calculated from our best single conformer obtained from the FSPS algorithm; "bestM" corresponds to RDCs from the best multi-structure derived from our genetic algorithm.

Figure 1 and Figures 3S through 7S in the Supporting Information show ϕ and ψ glycosidic dihedral angles for all conformers previously derived from the FSPS algorithm[4] and also the conformers with populations ($>1.0 \times 10^{-5}$) from the multistructure solutions derived from our GA optimization. In all cases, except perhaps for LNF-3, the RDC data can be very well represented by a subensemble of most important conformers. Even in the case of LNF-3 the result from the ensemble optimization is superior to that of a single structure. Perhaps one of the most important findings from this study is that the number of relevant conformers with significant weights are small for all studies sugars. Five in the case of LND-1, 3 for LNF-1, 4 for LNF-2, 5 for LNF-3, 5 for LNnT, and 5 for LNT, with all $\phi-\psi$ values listed in Table 2S (Supporting Information). The fact that there is only a small number of heavily weighted conformers is an indication that only certain regions in glycosidic space are likely to be important in solution. The reader should be aware that the FSPS is a coarse grained search algorithm in which two conformers are defined as different only if they meet certain angular and energetic difference criteria. It is possible that including conformers derived from local fluctuations around the FSPS generated energy basins may further reduce R factors.

From Table 2S (Supporting Information), we also note that conformers with weights ($>1.0 \times 10^{-5}$) coincide with individual conformations with low R factor. However, Figure 2 and Table 2S indicate that the best single conformer with the lowest R factor is not necessarily included in the subensemble of most relevant conformations derived from our GA. Only in the cases of LNF-1, LNF-2, and LNT does the single structure with the best individual R factor also have a significant contributions to the multistructure averaged RDCs values. In all cases, the weights of different important conformers that give rise to the best GA solution are quite different.

From the sequences depicted in Scheme 1S, we see that all six sugars have two common linkages at the reducing (right) end, β-D-GlcNAc-(1→3)-β-D-Gal, and β-D-Gal-(1→4)-β-D-Glc (link 3 and 4 for the first four sugars, link 2 and 3 for the other two). The remaining linkages constitute the histo-blood group epitopes: H type 3 for LNF-1, Lewis^a for LNF-2, Lewis^x for LNF-3, and Lewis^b for LND-1. The subensemble of most important conformers derived from our GA algorithm (green dots in Figure 1) as well as Figures 3S−7S (Supporting Information) indicate that the major conformational variability arises from two common

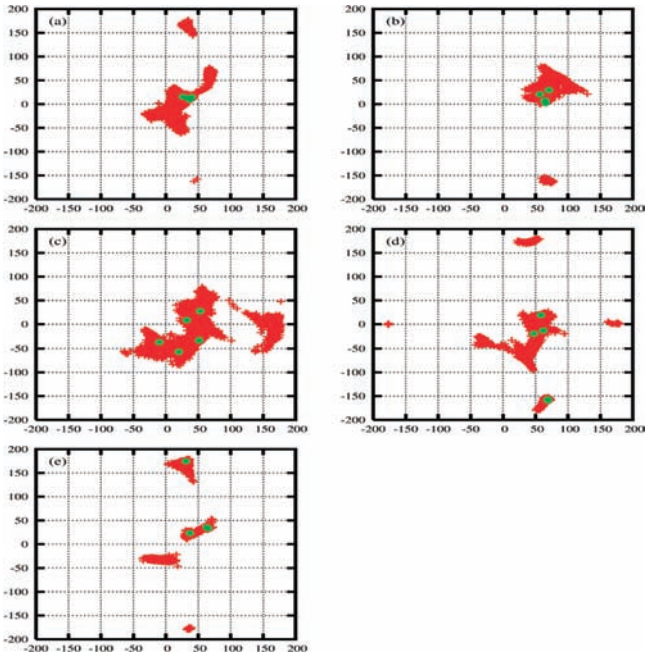

Figure 1. Distribution of conformations in $\phi-\psi$ glycosidic space in the case of LND-1. Red points represent the conformers generated by the FSPS algorithm. The green points are the conformations with highest weight derived from the GA multistructure fitting to RDC algorithm. (a) Link 1, α-L-Fuc-(1→2)-β-D-Gal, (b) Link 2, β-D-Gal-(1→3)-β-D-GlcNAc, (c) Link 3, β-D-GlcNAc-(1→3)-β-D-Gal, (d) Link 4, β-D-Gal-(1→4)-β-D-Glc, and (e) Link 5, α-L-Fuc-(1→4)-β-D-GlcNAc.

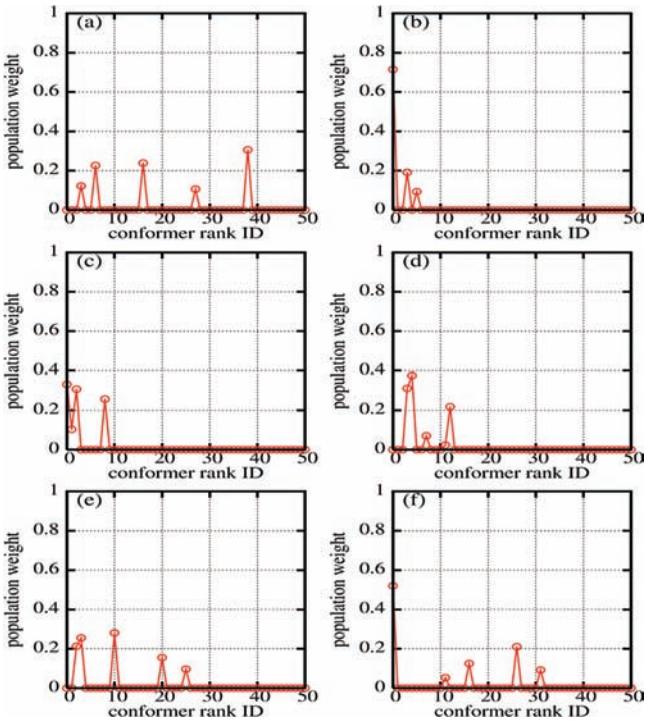

Figure 2. Population weights of structures in the ensemble that best matches experimental RDC data for human milk sugars: (a) LND-1, (b) LNF-1, (c) LNF-2, (d) LNF-3, (e) LNnT, and (f) LNT defined in Scheme 1S. Conformations with Rank IDs greater than 50 have population weights $<1.0 \times 10^{-5}$ and are not shown.

linkages (see Figure 1c and d). In contrast, the histo-blood group epitopes appear to have less conformational flexibility, namely, the distributions of $\phi-\psi$ values are restricted to small regions; see green dots in Figure 1a and b, as compared to those in Figure 1c and d. These results become more obvious as we perform visual check and rmsd calculation as follows. The 3D pictures of the most important conformers derived from the GA for all oligosaccharides studied are shown in Figure 3. In most of cases visual inspection of these most important conformers appear to indicate that epitopes have less conformational variability. This is most clear in the cases of LNT, LNnT and LNF-1. In Table 3S (Supporting Information), we show quantitative results that for all studied sugars the averaged RMSDs of epitopes are significantly smaller than that of common linkages. The picture of relatively small structural changes in histo-blood group epitopes and more flexible ones in the common linkages is consistent with the conclusions from previous RDC and NOE experiments[5,26] as well as with molecular dynamics simulations in explicit solvent using the CHARMm force field[26] and the OPLS-AA force field.[24] It is clear from Figure 3 that epitopes are not absolutely rigid. In particular, LND-1 and LNF-3 appear to have a more diverse set of relevant conformers.

On the basis of their RDC data,[5b] Martin-Pastor and Bush have proposed two best structures for each of the sugars in Scheme 1S. These authors justified their assignment on the reasonable approximation that the substructures defining the histo-blood groups are semirigid in solution. They then carried out a systematic search in the reduced dihedral space of the linkages that are not the histo-blood groups and that are common to the different oligosaccharides. While their results are very reasonable, in our study we did not have any constraints on the histo-blood

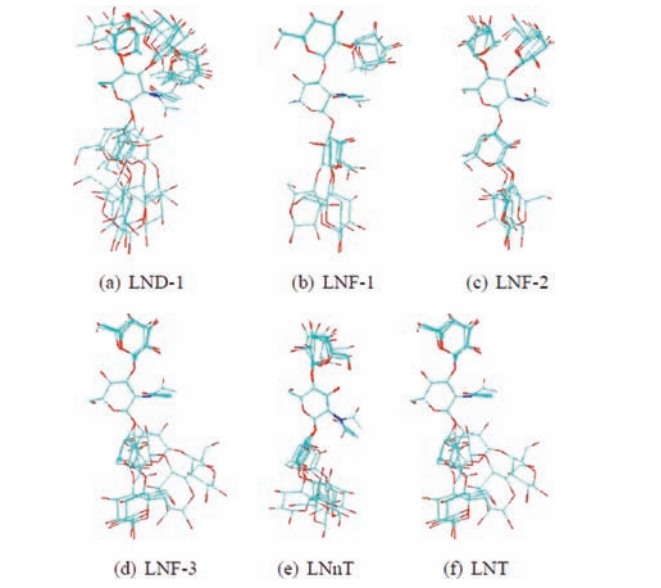

Figure 3. The best subensembles of conformers derived from our GA multistructure fitting. The glycosidic dihedral angles and populations are listed in Table 2S, Supporting Information. (All H atoms were deleted and all structures were aligned to β-D-GlcNAc.).

group epitopes. Instead, the subensemble of conformers that contribute the most to the averaged RDCs come from exploring the full dimensional space of all glycosidic linkages. That is why our weighted average over conformers exhibit significantly lower

R factors, as can be seen in Table 1. Furthermore, our approach provides not only important conformers but also population weights, which are crucial for predicting properties of flexible sugars.

In summary, we have created a program based on a genetic algorithm that is capable of generating the best set of statistical weights for conformers derived from the FSPS program or any other suitable conformational space sampling method. When the set of RDC values for computationally derived conformers are properly weighted, we obtain excellent agreement with experimental RDC values. We have used this algorithm to derive the subensemble of conformers that appears to be most important in the case of six different complex human milk sugars. The number of conformers chosen by the algorithm as having significant weights is small and provides an indication of which local minima are most important when these sugars are aligned in RDC studies. In our calculations, the alignment tensors of RDCs were estimated from the molecular shapes, which assumes that alignment is induced by steric factors. For other molecules and media, especially those with large charges, alignment by electrostatic forces might be dominant and their alignment tensors could be estimated by other methods.[10]

The GA program is totally independent from the FSPS conformation search program.[24] Accordingly, it is also applicable to structures obtained from other conformation search programs and even the trajectory conformations from standard molecular simulation packages. In addition, the general procedure is also applicable to other NMR observables, not just RDC measurements. We are planning to expand the code to perform predictions of conformer weights based on chemical shifts and J couplings. The resulting populations might also be used to calibrate force fields in molecular dynamics simulations.

■ ASSOCIATED CONTENT

Supporting Information. A description of the GA and RWMC algorithms, experimental and calculated RDC data, as well as the distributions of $\phi-\psi$ glycosidic angles. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
case@biomaps.rutgers.edu; junchao-xia@biomaps.rutgers.edu

■ ACKNOWLEDGMENT

This research was funded by NIH Grant GM45811 (D.A.C.) and by Grant 05-2182 from the Roy J. Carver Charitable Trust (C.J.M.).

■ REFERENCES

(1) (a) Cumming, D. A.; Carver, J. P. *Biochemistry* **1987**, *26*, 6664–6676. (b) Imberty, A.; Perez, S. *Chem. Rev.* **2000**, *100*, 4567–4588.
(2) Kiddle, G. R.; Homans, S. W. *Febs Lett.* **1998**, *436*, 128–130.
(3) (a) Almond, A.; DeAngelis, P. L.; Blundell, C. D. *J. Am. Chem. Soc.* **2005**, *127*, 1086–1087. (b) Eklund, R.; Lycknert, K.; Soderman, P.; Widmalm, G. *J. Phys. Chem. B* **2005**, *109*, 19936–19945. (c) Angulo, J.; Hricovini, M.; Gairi, M.; Guerrini, M.; de Paz, J. L.; Ojeda, R.; Martin-Lomas, M.; Nieto, P. M. *Glycobiology* **2005**, *15*, 1008–1015. (d) Henderson, T. J.; Venable, R. M.; Egan, W. *J. Am. Chem. Soc.* **2003**, *125*, 2930–2939. (e) Rundlof, T.; Venable, R. M.; Pastor, R. W.; Kowalewski, J.; Widmalm, G. *J. Am. Chem. Soc.* **1999**, *121*, 11847–11854.
(4) (a) Landersjo, C.; Jansson, J. L. M.; Maliniak, A.; Widmalm, G. *J. Phys. Chem. B* **2005**, *109*, 17320–17326. (b) Landersjo, C.; Hoog, C.; Maliniak, A.; Widmalm, G. *J. Phys. Chem. B* **2000**, *104*, 5618–5624.
(5) (a) Martin-Pastor, M.; Bush, C. A. *Biochemistry* **1999**, *38*, 8045–8055. (b) Martin-Pastor, M.; Bush, C. A. *Biochemistry* **2000**, *39*, 4674–4683. (c) Martin-Pastor, M.; Bush, C. A. *Carbohydr. Res.* **2000**, *323*, 147–155. (d) Ganguly, S.; Xia, J. C.; Margulis, C.; Stanwyck, L.; Bush, C. A. *Biopolymers* **2011**, *95*, 39–50.
(6) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111–1114.
(7) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334–342.
(8) (a) Fischer, M. W. F.; Losonczi, J. A.; Weaver, J. L.; Prestegard, J. H. *Biochemistry* **1999**, *38*, 9013–9022. (b) Bewley, C. A.; Clore, G. M. *J. Am. Chem. Soc.* **2000**, *122*, 6009–6016. (c) Mollova, E. T.; Hansen, M. R.; Pardi, A. *J. Am. Chem. Soc.* **2000**, *122*, 11561–11562.
(9) Zweckstetter, M.; Bax, A. *J. Biomol. NMR* **2002**, *23*, 127–137.
(10) (a) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679–690. Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791–3792.
(11) Almond, A.; Axelsen, J. B. *J. Am. Chem. Soc.* **2002**, *124*, 9986–9987.
(12) Azurmendi, H. F.; Bush, C. A. *J. Am. Chem. Soc.* **2002**, *124*, 2426–2427.
(13) (a) Fernandes, M. X.; Bernado, P.; Pons, M.; de la Torre, J. G. *J. Am. Chem. Soc.* **2001**, *123*, 12037–12047. (b) Berlin, K.; O'Leary, D. P.; Fushman, D. *J. Magn. Reson.* **2009**, *201*, 25–33.
(14) Esteban-Martin, S.; Fenwick, R. B.; Salvatella, X. *J. Am. Chem. Soc.* **2010**, *132*, 4626–4632.
(15) Bernado, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
(16) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099–13104.
(17) Marsh, J. A.; Baker, J. M. R.; Tollinger, M.; Forman-Kay, J. D. *J. Am. Chem. Soc.* **2008**, *130*, 7804–7805.
(18) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691–9702.
(19) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541–1542.
(20) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.
(21) Marsh, J. A.; Neale, C.; Jack, F. E.; Choy, W. Y.; Lee, A. Y.; Crowhurst, K. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2007**, *367*, 1494–1510.
(22) Tian, F.; Al-Hashimi, H. M.; Craighead, J. L.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 485–492.
(23) Venable, R. M.; Delaglio, F.; Norris, S. E.; Freedberg, D. I. *Carbohydr. Res.* **2005**, *340*, 863–874.
(24) (a) Xia, J. C.; Daly, R. P.; Chuang, F. C.; Parker, L.; Jensen, J. H.; Margulis, C. J. *J. Chem. Theory Comput.* **2007**, *3*, 1620–1628. (b) Xia, J. C.; Daly, R. P.; Chuang, F. C.; Parker, L.; Jensen, J. H.; Margulis, C. J. *J. Chem. Theory Comput.* **2007**, *3*, 1629–1643. (c) Xia, J. C.; Margulis, C. J. *J. Biomol. NMR* **2008**, *42*, 241–256. (d) Xia, J. C.; Margulis, C. J. *Biomacromolecules* **2009**, *10*, 3081–3088.
(25) (a) Morrow, A. L.; Ruiz-Palacios, G. M.; Jiang, X.; Newburg, D. S. *J. Nutr.* **2005**, *135*, 1304–1307. (b) Newburg, D. S.; Ruiz-Palacios, G. M.; Morrow, A. L. *Annu. Rev. Nutr.* **2005**, *25*, 37–58.
(26) Almond, A.; Petersen, B. O.; Duus, J. O. *Biochemistry* **2004**, *43*, 5853–5863.
(27) Martin-Pastor, M.; Canales, A.; Corzana, F.; Asensio, J. L.; Jimenez-Barbero, J. *J. Am. Chem. Soc.* **2005**, *127*, 3589–3595.
(28) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836–6837.