Published in final edited form as:

Curr Opin Struct Biol. 2010 October ; 20(5): 575–583. doi:10.1016/j.sbi.2010.07.005.

Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan–protein complexes

Robert J Woods^{1,2} and Matthew B Tessier¹

¹Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Road, Athens, GA 30602, United States ²School of Chemistry, National University of Ireland, Galway, University Road, Galway, Ireland

Abstract

Modern computational methods offer the tools to provide insight into the structural and dynamic properties of carbohydrate–protein complexes, beyond that provided by experimental structural biology. Dynamic properties such as the fluctuation of inter-molecular hydrogen bonds, the residency times of bound water molecules, side chain motions and ligand flexibility may be readily determined computationally. When taken with respect to the unliganded states, these calculations can also provide insight into the entropic and enthalpic changes in free energy associated with glycan binding. In addition, virtual ligand screening may be employed to predict the three dimensional (3D) structures of carbohydrate–protein complexes, given 3D structures for the components. In principle, the 3D structure of the protein may itself be derived by modeling, leading to the exciting — albeit high risk — realm of virtual structure prediction. This latter approach is appealing, given the difficulties associated with generating experimental 3D structures for some classes of glycan binding proteins; however, it is also the least robust. An unexpected outcome of the development of algorithms for modeling carbohydrate–protein interactions has been the discovery of errors in reported experimental 3D structures and a heightened awareness of the need for carbohydrate-specific computational tools for assisting in the refinement and curation of carbohydrate-containing crystal structures. Here we present a summary of the basic strategies associated with employing classical force field based modeling approaches to problems in glycoscience, with a focus on identifying typical pitfalls and limitations. This is not an exhaustive review of the current literature, but hopefully will provide a guide for the glycoscientist interested in modeling carbohydrates and carbohydrate–protein complexes, as well as the computational chemist contemplating such tasks.

Introduction

When approaching the modeling of a carbohydrate–protein complex, it is essential to establish the goals of the project in advance. Many questions of fundamental importance in glycobiology can be rephrased or focused in terms of objectives that can be answered computationally. The questions thus defined enable an assessment to be made as to whether modeling is an appropriate tool, and if so, which of the computational approaches will be most likely to provide the answer.

The following questions are common and approachable computationally, with some rephrasing or clarification:

‘What does glycan X look like?’

‘How does glycan X bind to protein Y?’

‘Why does protein Y recognize glycan X but not glycan Z?’

‘What mutation can be made in the protein or the glycan to alter specificity or affinity?’

And of course,

‘How reliable are the theoretical predictions?’

These questions relate broadly either to issues associated with 3D structure (and dynamics) or to energies of the interactions. As NMR is frequently employed for characterizing glycans in solution, it is worth noting that the 3D structure of a flexible molecule derived by employing NMR data as restraints, will lead to an average 3D shape [1]. However, is the average shape actually relevant in recognition processes? Consider a snake wiggling on the ground. We can generally recognize it as a snake, in part by its static properties, but also by the way it moves. If the snake wiggles equally right and left, what is its average shape? If we saw this average shape, would we still recognize it as a snake?

To the extent that a glycan or polysaccharide may be similar to a snake, it may not have a single 3D shape that is universally representative of its spatial and temporal properties. Thus, the simple question ‘*What is the shape of my glycan?*’ needs to be rephrased, perhaps as, ‘*What shapes can the glycan adopt?*’. And of course some conformations will be energetically more favorable than others. So the question is better phrased as ‘*What conformational states does the glycan adopt and what are their populations?*’. Once the problem is defined precisely, several computational techniques may be brought to bear.

The snake analogy fails, however, to capture several properties of oligosaccharides that make them unique among biopolymers. Firstly, oligosaccharides (or glycans) are not confined to linear sequences, and may be highly branched, as in the case of the complex plant cell wall molecule rhamnogalacturonan II [2]. Secondly, the fundamental building blocks, monosaccharides, vary in chirality at the carbon centers leading to distinct variations in their chemical and structural properties. Oligosaccharides demonstrate low chemical diversity, but compensate with extreme conformational diversity, both spatially and temporally, as a result of these branching and configurational variations.

This article presents a summary of the basic strategies associated with modeling glycans and carbohydrate–protein complexes, with a focus on identifying typical approaches and limitations. Hopefully the review will provide a reference for standard practices to guide the glycoscientist interested in modeling these systems, as well as the computational chemist contemplating such a task.

Three general applications of computational methods to glycoscience will be considered:

1. 3D structure prediction for glycans in solution using molecular dynamics (MD) simulation.
2. Application of MD simulations to a carbohydrate–protein complex.
3. Application of automated ligand docking to predict the structure of a glycan–protein complex.

Computational techniques and methods

Over the past three decades, computational methods and computer performance have increased to the point where it is routinely possible to simulate proteins or protein–ligand complexes for up to the μ s timescale. Access to long simulation times enables computational convergence, which in turn facilitates non-anecdotal comparisons of computed and experimental properties. It is important to recognize that a converged theoretical answer is nevertheless limited by the accuracy of the theoretical method. For these reasons, it is essential to select the most appropriate simulation conditions and force field. It is also important to recognize that the most complete model is not necessarily superior to a simpler model; what may be more important is that, under the conditions of the simulation, the model will provide a converged result that can be tested.

Force field selection

Force fields enable the prediction of molecular structures and potential energies using empirical parameters to describe electrostatic and van der Waals interactions, bond and angle vibrations, and internal torsion rotations (Eq. (1)). Some force fields are developed to be applicable to a broad range of molecules, while others describe molecular classes such as proteins, nucleic acids, lipids or carbohydrates. Several carbohydrate-specific force fields are available and have been recently reviewed in detail [3]. Most classical force fields share the functional form presented in Eq. (1); however, each differs in the philosophy and approach used in determining the values of the parameters (k , q , etc.) in the individual terms (Eqs. (2)–(6)).

$$V(r_1, \dots, r_N) = \sum_{i < j} V_{\text{Bonds}}(r_{ij}) + \sum_{i < j < k} V_{\text{Angles}}(\theta_{ijk}) + \sum_{i < j < k < h} V_{\text{Torsions}}(\varphi_{ijkl}) + \sum_{i < j} V_{\text{Coulomb}}(r_{ij}) + \sum_{i < j} V_{\text{vdW}}(r_{ij}) \quad (1)$$

where

$$V_{\text{Bonds}}(r_{ij}) = \frac{1}{2} k_{ij}^b (r_{ij} - r_{ij}^0)^2 \quad (2)$$

$$V_{\text{Angles}}(\theta_{ijk}) = \frac{1}{2} k_{ijk}^\theta (\theta_{ijk} - \theta_{ijk}^0)^2 \quad (3)$$

$$V_{\text{Torsions}}(\varphi_{ijkl}) = \frac{1}{2} k_\varphi (1 + \cos(n\varphi + \gamma)) \quad (4)$$

$$V_{\text{Coulomb}}(r_{ij}) = \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{\epsilon_r r_{ij}} \quad (5)$$

$$V_{\text{vdW}}(r_{ij}) = 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (6)$$

In selecting a force field for carbohydrate simulation, several criteria should be considered. Simulations with the selected force field should have been demonstrated to give rise to adequate agreement with experimental data for systems closely related to those under investigation. If there is any doubt, the onus is on the theoretician to demonstrate that the force field, employed under the proposed simulation conditions, performs well when compared to experimental data. Because glycans often populate multiple rotameric states in solution, particularly when 3-bond inter-glycosidic linkages are present, it is particularly important to select a force field that has been demonstrated to reproduce appropriate solution conformational properties, such as NMR observables.

Early carbohydrate force fields were not compatible with more established biomolecular force fields [4]. Currently, several of the major biomolecular force fields including AMBER [5], CHARMM [6,7], and GROMOS [8], contain parameters for carbohydrates that are compatible with the latest protein parameters, although fewer contain parameters for glycoproteins [5,9] or glycolipids [10–13]. In addition, the chemically reactive hydroxyl and amino groups in carbohydrates are frequently found derivatized in naturally occurring and synthetic glycans. Derivatives include sulfo-, and acetyl- moieties, particularly associated with glycosaminoglycans, such as heparin. Further rich sources for derivatized carbohydrates, which display methyl-, phosphoryl- or deoxy-modifications, are the polysaccharides present on many bacterial surfaces. Modeling the 3D structures of derivatized glycans requires a force field that is generalizable beyond the common monosaccharides [5]. To accommodate this diversity, the force field must have been developed with a rationale that enables derivatization while maintaining internal consistency, and, preferably without sacrificing applicability with other biomolecular classes, such as proteins. This issue is particularly important when applying computational methods in the design of carbohydrate-based mimetics or therapeutic agents [3].

Conformational sampling by MD simulation

‘What does glycan X look like?’

MD simulations describe time-dependent atomic motions by integrating Newton’s equations of motion, employing inter-atomic forces generated from the potential energy of the system. Although MD simulations can be performed with a quantum-mechanical Hamiltonian [14], for application to large systems (such as biomolecules) or to simulate motions over a long timescale (up to μ s), it is more common to compute the energies and forces from a classical mechanical force field.

A converged MD simulation provides an ensemble of 3D structures, whose average properties can be compared to experiment. It is also readily possible to compute the relative population of each conformational state observed in a MD simulation and use this information for comparison with experiment [15,16]. However, there are at least two caveats worth emphasizing in this regard. Firstly, a small error in the computed energy difference between the states can lead to a significant error in the predicted populations. Thus, the correct prediction of populations places the utmost demand on force field accuracy [17,18]. In addition, transitions between rotameric states may be infrequent, requiring MD simulation times of over 100 ns [19] or enhanced sampling [20] to achieve convergence. Nevertheless, MD simulations are effective at identifying the rotameric states, and this information alone can be sufficient to answer many questions. Secondly, the derivation of populations from experimental NMR observables (scalar 3-bond coupling constants, nuclear Overhauser enhancements, etc.) necessitates the application of several approximations and assumptions. Most notably, this deconvolution requires knowledge of the participating states and their 3D structures. Often idealized 3D structures are employed [21], and state identification is based on logical assumption. For these reasons, it is arguably better practice to compute the NMR

observables (rather than the state populations) from the MD data for direct comparison with the experimental data (see Table 1) [19,20]. Concomitantly, the 3D geometries for each state may be deduced from the MD data, and then employed to deconvolute the experimental data into state populations [19,20].

Other considerations

Aside from the initial 3D structure of the molecule under investigation, several other choices must be made when starting a MD-based modeling study. For example, solvent type (there are several models for water), system temperature and pressure, density, ion concentration, and protonation states of any ionizable groups must all be selected. The fundamental principle of MD — the Ergodic hypothesis — ensures that if the simulation has converged, the average properties of the system will also have converged [22]. Thus, at least in principle, the results do not depend on the initial conformation of the molecule. In practice, however, given the MD timescales currently accessible, statistical convergence can only be expected for motions that occur frequently on the sub- μ s timescale.

Conveniently, for carbohydrate modeling, the motions of the linkages between monosaccharides in a glycan (the glycosidic linkages) are often sampled well on the sub- μ s time scale. It is therefore feasible to achieve convergence of the conformational properties of a glycan based only on a reasonable initial conformation that is a conformation employing typical glycan linkage orientations. The practical impact of this feature is that it is possible to step directly from the primary structure of a glycan to a representative ensemble of 3D conformations. A convenient tool for generating 3D structures for glycans and glycoproteins, as well as the files necessary for MD simulation or ligand docking is GLYCAM-Web (<http://www.glycam.org>).

Currently, MD simulations with appropriately parameterized force fields, with suitable treatments of solvation, and performed for long enough to achieve convergence can generally identify the preponderant states populated by the glycan in solution. The accurate prediction of state populations remains a challenge for highly flexible 3-bond glycosidic linkages [23], but for the more common 2-bond case, such predictions are routinely accurate, principally because such linkages often populate only one rotamer to a significant extent. Carbohydrate flexibility therefore arises from two sources; long-lived transitions (ns time-scale) between stable rotameric states, and rapid librations (ps timescale) around stable rotameric states.

Computing binding free energies

‘What mutation can be made in the protein or the glycan to alter specificity or affinity?’

Binding energy calculations can be divided into two classes, namely those that compute an absolute affinity of a ligand for a receptor (direct binding), and those that predict the relative energy difference between ligands for a receptor (Figure 1).

Free energy difference calculations

Free energy difference (FED) methods often make use of the ability to convert one ligand into another over a series of small steps along a non-physical path [24] to compute a relative binding energy ($G_4 - G_3$ in Figure 1). These approaches are feasible computationally because each ligand is represented by a set of numerical force field parameters, and it is therefore feasible to interpolate mathematically between the two sets of descriptors. For example, in a classical force field description, chloride and bromide differ only in their van der Waals properties. It is therefore possible to define a system in which Cl^- in water can be

mutated to Br⁻ in water, *in a thermodynamically reversible manner*, by altering the van der Waals parameters for the ions from one state to the other in increments over the course of an MD simulation. Such approaches have the advantage that, under conditions of constant pressure and temperature (explicit water at 300 K), the integration of the step-wise energy differences equates to Gibbs free energy for the difference in binding of ligand one relative to ligand two (ΔG_{12} in Figure 1). The key limitation of FED methods is the need to maintain equilibrium along the transformation path, which necessitates very high levels of conformational sampling. FED calculations are computationally intensive, and their accuracy is roughly inversely proportional to the magnitude of the chemical change. FED methods are therefore often reserved for analysis of the effects of small structural changes in the ligand or the protein. Despite these limitations, the high level of accuracy achievable by FED approaches makes them appealing in the optimization of inhibitor design [25]. In the carbohydrate realm, these calculations have been employed to probe the effect of chemical modification of the glycan [26,27], mutations in the protein [28], and the influence of ions on binding affinity [29].

However, for comparing the binding properties of structurally diverse glycans, it is more common to employ more direct binding methods, which are often high throughput, but which conversely are less accurate than FED methods. Direct binding methods include virtual screening or automated docking, as well as approaches based on computing average molecular mechanical energies from post-processing of MD data.

Direct binding energy calculations

‘Why does protein Y recognize glycan X but not glycan Z?’

An MD trajectory can provide thousands of snapshots of the 3D structure as a function of simulation time, facilitating the derivation of average properties, such as binding energies. By decomposing the binding reaction into the traditional components associated with the physical process of ligand binding (Eqs. (7) and (8)), interaction energies can be computed by averaging the molecular mechanical energies for conformations selected from the MD data. The average molecular mechanical energies are often augmented with estimates of the desolvation energy from Poisson–Boltzmann or Generalized Born approximations (MM/PB/GBSA calculations), and may include estimates of the change in entropy.

$$\Delta G_{Total} = G_{Complex} - (G_{Ligand} + G_{Receptor}) \quad (7)$$

$$G = H_{MM} + G_{Desol} - T\Delta S_{Conf/Vib} \quad (8)$$

$$H_{MM} \approx V_{Coulomb} + V_{vdW} \quad (9)$$

In the MM/PB/GBSA approach, the enthalpic contribution to binding is approximated by the potential energy computed from the force field (Eq. (1)). Notably, the displacement of solvent waters from the binding site, as well as from the complementary surface of the ligand, which occurs during the physical process of binding, is extremely difficult to model accurately (Figure 1). The difficulty arises from the innumerable water configurations that each could influence the estimated affinity. In MM/PB/GBSA calculations, the most common practice is to include the waters in the simulation of the complex, but remove them from the trajectory before computing the energies. The effect of desolvation on binding energy is then approximated using an implicit water model (of the generalized Born or Poisson–Boltzmann type). Because the desolvation energy is generally unfavorable for

carbohydrate binding, in contrast to the direct electrostatic interactions formed with the ligand in the complex, the model employed to estimate the desolvation free energy can have a significant impact on the predicted net contribution from electrostatic interactions (Table 2). As none of the current PB or GB methods have been developed to specifically treat the amphipathic nature of carbohydrate interactions, this aspect of the MM/PB/GBSA method remains a source of considerable uncertainty. The entropic term can be estimated from an analysis of the change in the global rotational, translational, and vibrational modes of the components that occurs on binding [30]. In the case of flexible glycans, additional insight into the entropic contribution can come from an analysis of the changes in the low frequency internal rotations that occur in the ligand on binding [31–35].

There are several general conclusions that can be drawn from the reported MM/PB/GBSA calculations for glycan protein complexes (see Table 2).

1. The energies associated with direct electrostatic interactions (hydrogen bonds, etc.) between the protein and the glycan are strongly favorable. This is no surprise given the ubiquitous presence of hydrogen bonds in these complexes.
2. The energies required to desolvate the interacting surfaces are large and unfavorable. Again this is as expected given the amphipathic nature of the binding surfaces.
3. The energy to desolvate the protein and glycan contact surfaces considerably attenuates the favorable direct electrostatic interaction energies. This is perhaps less expected, and illustrates the dangers of rationalizing the strength of binding on the basis of simplistic metrics, such as counting hydrogen bonds.
4. Van der Waals attractions are often more significant than the combined electrostatic contributions, and can be the driving force for the stability of the complex.
5. Conformational entropies are unfavorable and increase roughly in proportion to the increase in binding energy from electrostatics and van der Waals. This is reflective of enthalpy–entropy compensation.

The MM/PB/GBSA approach illustrates several important concepts in modeling carbohydrate–protein interactions. Firstly, the complex is formed from a binding reaction between ligand and receptor, and failure to consider the energies associated with the unbound states will lead to inaccuracies. Secondly, modeling that ignores desolvation or entropic effects cannot reproduce absolute affinities, and at best only be able to rank relative binding energies for closely related ligands. In addition, since both desolvation and entropic effects are complex functions of molecular structure, their omission may also lead to incorrect relative ranking of ligand affinities.

Methods that compute the binding energy from an MD simulation necessarily require the pre-existence of a model of the complex. While the MM/PB/GBSA method may be used to compute the affinity of a ligand or the effect of a mutation [36] on the affinity, it is not useful in the generation of theoretical models for the complex. For this we must turn to automated ligand docking.

Virtual screening/automated docking

‘How does glycan X bind to protein Y?’

Each docking program operates slightly differently, but they share common features that enable them to:

1. Search for locations on the protein surface that lead to favorable interactions with the ligand.
2. Sample the conformational space of the ligand.
3. Compute the interaction energy (or score the binding) between the protein and ligand.

Computational docking requires two components, a 3D structure for the protein, which can come from a number of sources including X-ray diffraction, NMR spectroscopy, or homology modeling/protein threading, and a structure for the ligand. The interaction with the ligand relies on both the protein backbone fold in the region of the binding site, as well as on the orientation of the side chains in the binding site. One of the most significant limitations in docking is that it is generally performed while keeping the protein surface rigid, which prevents the consideration of the effects of induced fit within the binding site.

Interaction energy estimation from docking does not necessarily employ an energy function based purely on a classical force field. For example, the program AutoDock [37] applies adjustable weighting factors (W in Eq. (11)) and other empirical terms (for estimating desolvation or entropic penalties), which enable the theoretical affinities to be empirically calibrated against a reference set of binding energies for known ligand–protein complexes.

$$\Delta G = W_{vdW} \Delta G_{vdW} + W_{H-bond} \Delta G_{H-bond} + W_{Coulomb} \Delta G_{Coulomb} + W_{Entropic} \Delta G_{Entropic} + W_{Desol} \Delta G_{Desol} \quad (10)$$

$$\begin{aligned} \Delta G = & W_{vdW} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\ & + W_{H-bond} \sum_{i,j} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + E_{H-bond} \\ & + W_{Coulomb} \sum_{i,j} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} + W_{Entropic} N_{Tor} \\ & + W_{Desol} \sum_{i,j} S_i V_j \ell (-r_{ij}^2 / 2\sigma^2) \end{aligned} \quad (11)$$

Several carbohydrate-specific scoring functions have been proposed [38–40]; however, none have been tuned specifically for interactions with highly charged ligands, such as glycosaminoglycans. Moreover, to achieve generality, affinity scoring functions do not generally include accurate parameters for the internal energies of the ligand, and in the case of carbohydrates, this omission can readily lead to poses of the ligand in the binding site that are highly ranked, but which display conformations of the glycosidic linkages that violate expectations based on known structural preferences [41].

Careful assessment of the likelihood of the predicted poses in terms of glycan conformation is essential, and it is common that the most highly ranked pose is not the most consistent with experimental data. One approach to resolving this dilemma is to analyze the probability of the docking poses in terms of their statistical occurrence. For example, docking programs such as AutoDock provide more than one docked conformation, enabling similar poses to be clustered in terms of their 3D position on the protein surface. In many cases, the most highly clustered pose is more consistent with experiment than the highest ranked pose. This suggests that, while docking can be a powerful tool for predicting the 3D structures of carbohydrate–protein complexes, rescoring, such as by MM/PB/GBSA methods may be required to more accurately rank affinities [42].

Automated docking is typically performed in either of two scenarios; for ligand discovery (virtual screening of libraries of diverse ligands) or for binding mode prediction (docking of known ligands). With regard to ligand discovery, it should be noted that automated docking does not permit the ligand not to bind. That is, docking does not mimic the reality of the binding process, but instead forces the ligand to adhere to the protein. It might be considered to be equivalent to performing the binding experiment at infinite concentration of ligand. Obviously this aspect can lead to false positive results and places a heavy burden on the scoring function. Again, an alternative assessment may be provided through an examination of the pose clustering. For example, a ligand that is predicted to adhere to the protein in numerous surface locations, is likely displaying non-specific 'binding', whereas a ligand that is more tightly clustered in a single region is likely to be a true binder. In these assessments, it is clear that if the binding site is restricted only to a localized region of the protein surface, the clustering will be skewed, and it can be difficult to separate real from false positive binders.

To determine the binding mode of a known ligand, the docking is usually focused in the region of the known binding site. Given that docking will always provide a structure for the complex, in addition to the results from scoring and clustering analyses, it is important to confirm that the docked complex is consistent with experimental data for the same or closely related systems. Useful corroborating data include contact sites identified by saturation transfer difference (STD) or transferred nuclear Overhauser enhancement (tr-NOE) NMR experiments [43,44], protein mutagenesis studies [45], and epitope mapping studies employing closely related congeners of the ligand [46]. Confidence in the results can also be built by including negative controls in the docking set; unless the scores for known non-binding ligands are significantly different from those predicted for known binders, then the scoring function alone cannot be employed with confidence.

Validation

'How reliable are the theoretical predictions?'

In the preceding sections, several aspects of validation have been mentioned, including comparison with solution NMR data for glycans in solutions and comparison with crystallographic or NMR data for complexes. Much effort has been spent over the past decade and a half on retrospective calculations of glycans and their complexes [47–49]; that is, on establishing the extent to which simulations can reproduce existing experimental data. These calculations have not been performed purely for proof of principle, but have been driven by the fact that NMR data are often insufficient to uniquely characterize the 3D structures of glycans in solution or of their complexes. Computational glycoscience has been driven by a need for objective methods capable of determining experimentally consistent 3D structures [50,51]. The accuracy of the conformational properties of glycans generated by MD simulation is now well established, and appropriate experimental data sets for use in validating the predictions are abundant. Binding energy calculations applied to carbohydrate–protein complexes have also become commonplace; however, their accuracy is highly dependent on the computational method employed and on the reliability of the 3D structure of the complex employed in the calculation. The potential now exists to employ these simulations and energy calculations predicatively to aid in many aspects of glycoscience, such as in the design of ligands, the development of protein mutation strategies, and the interpretation of sparse experimental data.

The adage that everyone trusts an experimental result, except the person who performed the experiment; while no one trusts a theoretical result, except the person who performed the calculation, clearly fails to recognize the fact that there are errors associated with both experimental measurements and theoretical predictions. In addition to errors in precision,

significant sources of error include the approximations employed in converting experimental data to either conformations or populations, and in generating experimental observables from theoretical data. To minimize these errors, it may be preferable to convert the theoretical data to the experimental observables (NOEs, J-couplings, etc.) rather than to derive 3D structures from sparse experimental data [20]. Further, as noted by Lütke [52], there is unusually high level of structural and annotation errors in the X-ray structures of carbohydrate–protein complexes and glycoproteins. These errors arise in part from limitations in carbohydrate residue and linkage nomenclature [53] employed in the Protein Data Bank (<http://www.rcsb.org/pdb/>), as well as from a lack of adoption of carbohydrate-specific tools for structure curation or refinement validation [54]. Computational methods can therefore play an important role in identifying potential experimental errors, and theoreticians must be cognizant of the existence of such errors when comparing to experimental data sets.

Confidence in all computational results can be built not only through comparison with 3D structural data, but also through the ability of the model to reproduce or rationalize binding data. However, no matter how complex or sophisticated the model, it is always an approximation to reality, and as noted by Box [55], it is important to remember that:

“all models are wrong; the practical question is how wrong do they have to be to not be useful.”

Conclusions

A well-executed simulation (suitable force field and water model with equilibrated sampling) can provide accurate and reproducible results that give unmatched insight into the molecular mechanisms that underlie biological processes. The use and application of these methods in glycoscience requires an awareness of the challenges inherent in each technique, which it is hoped that this brief review has provided.

Theoretical methods for identifying glycan conformations, predicting glycan-protein binding thermodynamics, assessing the per-residue or per-atom contributions to binding, predicting binding modes, and screening potential ligands, are all being applied to develop deeper understanding of glycan structure and recognition. Much work is ongoing in the field of computational glycoscience to tune existing theoretical methods for application to glycans, in terms of force field development [3], to establish the accuracy of energy analyses from MD data, and to establish the suitability of automated docking in the prediction of carbohydrate binding modes.

Given the difficulties associated with carbohydrate synthesis and their 3D structural characterization using traditional experimental methods, a tremendous opportunity exists for the predictive application of computational methods. While MD simulations of glycans in solution are remarkably reliable, the same cannot be said for automated docking of glycans. Nevertheless, the fact that when docking is performed under optimal theoretical conditions, (specifically, re-docking the carbohydrate in its bound conformation to the crystal structure from which it came), it can give excellent agreement with experiment [56], provides encouragement for further studies. Continued development in the area of solvent approximations, scoring function optimization, rescoring protocols, and pose selection for automated glycan docking will be required before these methods can make their full impact, but such methods offer the potential to be transformative.

Acknowledgments

Funding is gratefully acknowledged from the National Institutes of Health (RR005351 and GM055230) and from the Science Foundation of Ireland (07/RPI/B1321).

References

1. Cumming DA, Carver JP. Virtual and solution conformations of oligosaccharides. *Biochemistry*. 1987; 26:6664–6676. [PubMed: 3427035]
2. O'Neill MA, Ishii T, Albersheim P, Darvill AG. Rhamnogalacturonan II: structure and function of a borate cross-linked cell wall pectic polysaccharide. *Annu Rev Plant Biol*. 2004; 55:109–139. [PubMed: 15377216]
3. Fadda E, Woods RJ. Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects. *Drug Discovery Today*.
4. Stuike-Prill R, Meyer B. A new force-field program for the calculation of glycopeptides and its application to a heptacosapeptide–decasaccharide of immunoglobulin G₁. *Eur J Biochem*. 1990; 194:903–919. [PubMed: 2269309]
5. Kirschner KN, et al. Glycam06: a generalizable biomolecular force field, carbohydrates. *J Comput Chem*. 2008; 29:622–655. [PubMed: 17849372]
6. Guvench O, et al. Charmm additive all-atom force field for glycosidic linkages between hexopyranoses. *J Chem Theory Comput*. 2009; 5:2353–2370. [PubMed: 20161005]
7. Hatcher E, Guvench O, MacKerell AD. Charmm additive all-atom force field for aldopentofuranoses, methylaldopentofuranosides, and fructofuranose. *J Phys Chem B*. 2009; 113:12466–12476. [PubMed: 19694450]
8. Lins RD, Hunenberger PH. A new gromos force field for hexopyranose-based carbohydrates. *J Comput Chem*. 2005; 26:1400–1412. [PubMed: 16035088]
9. Bosques CJ, Tschampel S, Woods RJ, Imperiali B. Effects of glycosylation on peptide conformation: a synergistic experimental and computational study. *J Am Chem Soc*. 2004; 126:8421–8425. [PubMed: 15237998]
10. DeMarco ML, Woods RJ. Atomic-resolution conformational analysis of a ganglioside in a lipid bilayer and its implications for ganglioside–protein interactions at the plasma membrane surface. *Glycobiology*. 2009; 19:344–355. [PubMed: 19056784]
11. Tessier MB, DeMarco ML, Yongye AB, Woods RJ. Extension of the Glycam06 biomolecular force field to lipids, lipid bilayers and glycolipids. *Mol Simulat*. 2008; 34:349–364.
12. Shroll RM, Straatsma TP. Molecular structure of the outer bacterial membrane of *Pseudomonas aeruginosa* via classical simulation. *Biopolymers*. 2002; 65:395–407. [PubMed: 12434428]
13. Lins RD, Straatsma TP. Computer simulation of the rough lipopolysaccharide membrane of *Pseudomonas aeruginosa*. *Biophys J*. 2001; 81:1037–1046. [PubMed: 11463645]
14. Biarnes XA, Planas A, Rovira C, Laio A, Parrinello M. The conformational free energy landscape of D-glucopyranose. Implications for substrate preactivation in α -glucoside. *J Am Chem Soc*. 2007; 129:10686–10693. [PubMed: 17696342]
15. Vlieghehart, JFG.; Woods, RJ., editors. Recent Advances. Washington: American Chemical Society; 2006. Nmr Spectroscopy and Computer Modeling of Carbohydrates; p. 22ACS Symposium Series
16. Gonzalez-Outeriño J, Kadirvelraj R, Woods RJ. Structural elucidation of type III Group B streptococcus capsular polysaccharide using molecular dynamics simulations: the role of sialic acid. *Carbohydr Res*. 2005; 340:1007–1018. [PubMed: 15780265]
17. Spiwok V, Tvaroska I. Metadynamics modelling of the solvent effect on primary hydroxyl rotamer equilibria in hexopyranosides. *Carbohydr Res*. 2009; 344:1575–1581. [PubMed: 19539272]
18. Kirschner KN, Woods RJ. Solvent interactions determine carbohydrate conformation. *Proc Natl Acad Sci USA*. 2001; 98:10541–10545. [PubMed: 11526221]
19. Gonzalez-Outeriño J, Kirschner KN, Thobhani S, Woods RJ. Reconciling solvent effects on rotamer populations in carbohydrates: a joint MD and NMR analysis. *Can J Chem*. 2006; 84:569–579.

20. Yongye AB, Foley BL, Woods RJ. On achieving experimental accuracy from molecular dynamics simulations of flexible molecules: aqueous glycerol. *J Phys Chem A*. 2008; 112:2634–2639. [PubMed: 18311953]
21. Rockwell GD, Grindley BT. Effect of solvation on the rotation of hydroxymethyl groups in carbohydrates. *J Am Chem Soc*. 1998; 120:10953–10963.
22. Allen, MP.; Tildesley, DJ. *Computer Simulation of Liquids*. Oxford: Oxford Science Publications; 1987.
23. Yongye AB, et al. The conformational properties of methyl A-(2,8)-di/trisialosides and their N-acyl analogs: implications for anti-*Neisseria meningitidis* B vaccine design. *Biochemistry*. 2008; 47:12493–12514. [PubMed: 18954144]
24. Jorgensen WL, Thomas LL. Perspective on free-energy perturbation calculations for chemical equilibria. *J Chem Theory Comput*. 2008; 4:869–876. [PubMed: 19936324]
25. Pearlman DA, Charifson PS. Are free energy calculations useful in practice? A comparison with rapid scoring functions for the P38 map kinase protein system. *J Med Chem*. 2001; 44:3417–3423. [PubMed: 11585447]
26. Kadirvelraj R, Foley BL, Dyekjaer JD, Woods RJ. Involvement of water in carbohydrate–protein binding: concanavalin a revisited. *J Am Chem Soc*. 2008; 130:16933–16942. [PubMed: 19053475]
27. Pathiaseril A, Woods RJ. Relative energies of binding for antibody–carbohydrate–antigen complexes computed from free-energy simulations. *J Am Chem Soc*. 2000; 122:331–338. [PubMed: 17211491]
28. Nam K, Verdine GL, Karplus M. Analysis of an anomalous mutant of Mutm DNA glycosylase leads to new insights into the catalytic mechanism. *J Am Chem Soc*. 2009; 131:18208–18209. [PubMed: 19961158]
29. Lawrenz M, et al. Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins Struct Funct Bioinf*. 2010; 00:000–10.
30. Case DA, et al. The amber biomolecular simulation programs. *J Comput Chem*. 2005; 26:1668–1688. [PubMed: 16200636]
31. Xu D, et al. Distinct glycan topology for avian and human sialopentasaccharide receptor analogues upon binding different hemagglutinins: a molecular dynamics perspective. *J Mol Biol*. 2009; 387:465–491. [PubMed: 19356594]
32. Kadirvelraj R, et al. Understanding the bacterial polysaccharide antigenicity of *Streptococcus agalactiae* versus *Streptococcus pneumoniae*. *PNAS*. 2006; 103:8149–8154. [PubMed: 16705032]
33. Bryce RA, Hillier IH, Naismith JH. Carbohydrate–protein recognition: molecular dynamics simulations and free energy analysis of oligosaccharide binding to concanavalin A. *Biophys J*. 2001; 81:1373–1388. [PubMed: 11509352]
34. Ford MG, Weimar T, Kohli T, Woods RJ. Molecular dynamics simulations of galectin-1-oligosaccharide complexes reveal the molecular basis for ligand diversity. *Proteins: Struct Funct Genet*. 2003; 53:229–240. [PubMed: 14517974]
35. Liu Z, Zhang Y. Molecular dynamics simulations and Mm–Pbsa calculations of the lectin from snowdrop. *J Mol Model*. 2009; 15:1501–1507. [PubMed: 19449157]
36. Wimmerová M, Mishra NK, Pokorná M, Ko a J. Importance of oligomerisation on *Pseudomonas aeruginosa* lectin-Ii binding affinity, in silico and in vitro mutagenesis. *J Mol Mod*. 2009; 15:673–679.
37. Morris, GM., et al. *Autodock*. La Jolla: Scripps Research Institute; 1999.
38. Kerzmann A, Fuhrmann J, Kohlbacher O, Neumann D. Balldock/slick: a new method for protein–carbohydrate docking. *J Chem Inform Model*. 2008; 48:1616–1625.
39. Kerzmann A, Neumann D, Kohlbacher O. Slick—scoring and energy functions for protein–carbohydrate interactions. *J Chem Inf Model*. 2006; 46:1635–1642. [PubMed: 16859295]
40. Laederach A, Reilly PJ. Specific empirical free energy function for automated docking of carbohydrates to proteins. *J Comput Chem*. 2003; 24:1748–1757. [PubMed: 12964193]
41. Rao, VSR.; Qasba, PK.; Balaji, PV.; Chandrasekaran, R. *Conformations of Carbohydrates*. Amsterdam: Harwood Academic; 1998.

42. Graves AP, et al. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J Mol Biol.* 2008; 377:914–934. [PubMed: 18280498]
43. Haselhorst T, et al. Std Nmr spectroscopy and molecular modeling investigation of the binding of N-acetylneuraminic acid derivatives to rhesus rotavirus Vp8*core. *Glycobiology.* 2007; 17:68–81. [PubMed: 16973731]
44. Bhunia A, et al. Consistent bioactive conformation of the Neu5aca(2-3)Gal epitope upon lectin binding. *ChemBioChem.* 2008; 9:2941–2945. [PubMed: 18850605]
45. Iobst ST, et al. Binding of sugar ligands to Ca^{2+} -dependent animal lectins. *J Biol Chem.* 1994; 269:15505–15511. [PubMed: 8195194]
46. Bundle DR, et al. Molecular recognition of a Salmonella trisaccharide epitope by monoclonal antibody Se155-4. *Biochemistry.* 1994; 33:5172–5182. [PubMed: 7513555]
47. Yui T, Ogawa K. Computer modeling of carbohydrate molecules. *Trends Glycosci Glycotechnol.* 2005; 17:159–176.
48. Imberty A, Perez S. Structure: Conformation, and dynamics of bioactive oligosaccharides: theoretical approaches and experimental validations. *Chem Rev.* 2000; 100:4567–4588. [PubMed: 11749358]
49. Woods RJ. Three-dimensional structures of oligosaccharides. *Curr Opin Struct Biol.* 1995; 5:591–598. [PubMed: 8574693]
50. Sayers EW, Prestegard JH. Solution conformations of a trimannoside from nuclear magnetic resonance and molecular dynamics simulations. *Biophys J.* 2000; 79:3313–3329. [PubMed: 11106634]
51. Bock K. The preferred conformation of oligosaccharides in solution inferred from high resolution Nmr data and hard sphere exo-anomeric calculations. *Pure Appl Chem.* 1983; 55:605–622.
52. Lütke T. Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr D.* 2009; 65:156–168. [PubMed: 19171971]
53. DeMarco ML, Woods RJ. Structural glycobiology: a game of snakes and ladders. *Glycobiology.* 2008; 18:426–440. [PubMed: 18390826]
54. Lütke T, Von der Lieth C-W. Pdb-Care (Pdb Carbohydrate Residue Check): a program to support annotation of complex carbohydrate structures in Pdb files. *BMC Bioinformat.* 2004; 5:1–6.
55. Box, GEP.; Draper, NR. *Empirical Model-Building and Response Surfaces.* Wiley; 1987.
56. Agostino M, et al. Molecular docking of carbohydrate ligands to antibodies: structural validation against crystal structures. *J Chem Inf Model.* 2009; 49:2749–2760. [PubMed: 19994843]

Figure 1.

Schematic representation of the binding of two related ligands (red and yellow) to the same protein (light blue) releasing waters (blue) from the interacting surfaces. In direct binding energy calculations (docking, MM/PB/GBSA) G_1 and G_2 are computed. In FED calculations, the difference in the binding energies of the two ligands G_{12} is computed by performing the computational alchemical mutations G_3 and G_4 . Note $G_{12} = G_1 - G_2 = G_4 - G_3$.

Table 1

Computationally derived and experimentally measured NMR coupling constants versus computed and NMR-derived rotamer populations for aqueous glycerol [20].

From 1 μ s MD		NMR		
NMR Observables (Hz)
$^3J_{AB}$	6.3 ± 1.0	6.3 ± 0.2
$^3J_{AC}$	4.0 ± 0.8	4.4 ± 0.1		
Rotameric States	Populations from MD (%)	Derived Populations (%)		
$\alpha\gamma$	35	28–30		
$\alpha\beta$	28	20–21		
$\alpha\alpha$	17	18–21		
$\beta\gamma$	15	15–17
$\gamma\gamma$	3	10–12		
$\beta\beta$	2	5		

Table 2

Major interaction energy components (kcal/mol) for several representative carbohydrate–protein complexes.

	Galectin/disaccharide [34]	Con A/trisaccharide [33]	Snowdrop lectin/trisaccharide [35]	Antibody/bacterial oligosaccharide [32]	Hemagglutinin/pentasaccharide [31]
$V_{Electrostatic}$	-67.5	-90.7	-119.2	-167.5	-148.0
G_{Desolv}	11.7	69.2	26.5	228.8	163.3
Total electrostatics	-52.2	-21.5	-92.7	61.3	15.3
V_{vdw}	-17.5	-27.6	-42.5	-126.9	-43.3
$-T S_{Vib}$	18.3	16.4	52.7	77.6	30.2