J Comput Aided Mol Des (2014) 28:1191–1204
DOI 10.1007/s10822-014-9794-3

Exploring the free-energy landscape of carbohydrate–protein complexes: development and validation of scoring functions considering the binding-site topology

Sameh Eid · Noureldin Saleh · Adam Zalewski · Angelo Vedani

Received: 7 July 2014 / Accepted: 4 September 2014 / Published online: 10 September 2014

Abstract Carbohydrates play a key role in a variety of physiological and pathological processes and, hence, represent a rich source for the development of novel therapeutic agents. Being able to predict binding mode and binding affinity is an essential, yet lacking, aspect of the structure-based design of carbohydrate-based ligands. We assembled a diverse data set comprising 273 carbohydrate–protein crystal structures with known binding affinity and evaluated the prediction accuracy of a large collection of well-established scoring and free-energy functions, as well as combinations thereof. Unfortunately, the tested functions were not capable of reproducing binding affinities in the studied complexes. To simplify the complex free-energy surface of carbohydrate–protein systems, we classified the studied proteins according to the topology and solvent exposure of the carbohydrate-binding site into five distinct categories. A free-energy model based on the proposed classification scheme reproduced binding affinities in the carbohydrate data set with an r^2 of 0.71 and root-mean-squared-error of 1.25 kcal/mol ($N = 236$). The improvement in model performance underlines the significance of the differences in the local micro-environments of carbohydrate-binding sites and demonstrates the usefulness of calibrating free-energy functions individually according to binding-site topology and solvent exposure.

Keywords Carbohydrates · Docking · Scoring function · Free energy

Introduction

Carbohydrates are involved in a broad spectrum of pathophysiological processes ranging from protein folding, bacterial adhesion, viral infection, cancer metastasis, inflammatory reactions, cell proliferation, and cell–cell communication [1, 2]. Carbohydrate research has gained considerable momentum in the past decade due to its potentially rewarding applications in therapeutics, drug delivery, diagnosis, and vaccine development [2–4]. Nevertheless, only a limited number of carbohydrate-based drugs have reached the market to date, and carbohydrates are still considered to be a relatively untapped source for new therapeutic agents [5]. The relatively slow development of carbohydrate-based therapeutics could be attributed to a number of factors; including the problematic synthesis of carbohydrate derivatives [6], inadequate pharmacokinetic profiles due to high water solubility [2], and the inherent low binding affinities (in the milli- to micro-molar range) of naturally occurring carbohydrates [5, 7]. Moreover, carbohydrates present a unique set of structural and energetic features that makes the accurate modeling of their properties a daunting task. Such features include: (1) complex stereochemistry and the high density of polar functional groups, which necessitates the accurate treatment of electrostatic interactions [8, 9], (2) the rich diversity of linear and branched structures formed by oligosaccharides as well as the multiple rotameric states of

Electronic supplementary material The online version of this article (doi:10.1007/s10822-014-9794-3) contains supplementary material, which is available to authorized users.

S. Eid · N. Saleh · A. Zalewski · A. Vedani
Department of Pharmaceutical Sciences, University of Basel,
Klingelbergstrasse 50, 4056 Basel, Switzerland

S. Eid (✉)
BioMed X Innovation Center, Im Neuenheimer Feld 583,
69120 Heidelberg, Germany
e-mail: eid@bio.mx

glycosidic bonds [8], (3) importance of the C–H···π interactions on the α-hydrophobic face of sugars [10–12], (4) the anomeric and exoanomeric effects [9, 13], and (5) the highly dynamic and relatively weak nature of carbohydrate–protein interactions [14, 15].

The increased interest in carbohydrate research over the past two decades has stimulated the development of computational tools specifically tuned for carbohydrate simulations. For instance, carbohydrate-specific force fields, e.g. GLYCAM06 [9], are increasing in number and quality and are being adopted more frequently in biomolecular simulations involving carbohydrate–macromolecule interactions [8]. However, the optimization of carbohydrate leads in drug discovery requires the correct identification of their native binding modes to macromolecular targets and the reliable estimation of binding affinities of putative complexes. Although a multitude of docking/scoring programs have achieved considerable success in reproducing crystal poses, the accurate prediction of binding affinity from these poses is still largely elusive [16, 17].

In addition to general utility scoring and free-energy functions [18–25] three attempts specifically dealing with the quantification of carbohydrate–protein binding are reported. In a first approach, Laederach and Reilly [26] employed a set of 30 carbohydrate–protein complexes to train an empirical model based on the AutoDock scoring function, plus a special term for hydrogen bond. The best performing model yielded a residual standard error of 1.4 kcal/mol in the training set. Later, Hill and Reilly [27] expanded this study to a training set of 115 complexes and introduced a novel entropic term that accounts for ligand's translational and rotational degrees-of-freedom. Starting from the AutoDock scoring function, they examined 288 different free-energy models and the best model (JA) achieved a root-mean-squared-error (RMSE) of 2.0 kcal/mol. The third approach was the sugar–lectin interactions and DoCKing (SLICK) scoring functions introduced by Kerzmann et al. [28], which employs a special term to account for C–H···π interactions [29]. The developed free-energy function predicted binding affinities in a training set of 20 lectin–sugar complexes within a maximum absolute error of 2.8 kJ/mol (0.7 kcal/mol). In an extended iteration of the study, the authors successfully redocked 17 out of 18 training complexes, with an average RMSD of 0.85 Å and an average absolute error of 3.6 kJ/mol (0.9 kcal/mol) in the binding free-energy estimate [30] Notably, the three attempts were derived by recalibrating an existing scoring function on training sets of carbohydrate–protein complexes.

Despite the relative abundance of methodologies for calculating different free-energy components, it would seem that we still lack a better understanding of why the traditional free-energy functions generally fail to yield good correlation with experimental results. In this study, we gathered and refined a large and diverse set of carbohydrate–protein complexes with experimentally determined binding affinities. We investigated a larger number of combinations of computational methods accounting for one or more of the free-energy components (e.g. force fields, scoring functions, solvent-accessible surface area, desolvation penalties, etc.). The employed methods vary in their theoretical derivation, degree of sophistication, and associated computational cost; from a simple integer representing the number of freely rotatable bonds in the ligand up to a sophisticated free-energy function employing an implicit solvent model such as MM/GBSA. The aim was to find the computational tools that could, either individually or in combination, serve as an objective free-energy function for carbohydrate–protein complexes. In addition, our study addressed two fundamental questions related to the quantification of carbohydrate–protein interactions: (1) the target-dependence of scoring functions [16, 19, 31, 32]; i.e. why is it that certain scoring functions could predict binding affinities accurately in some protein families and fail in others, and (2) the impact of the binding-site topology and solvent accessibility.

Results and discussion

Traditional approaches for estimating binding free energy

Our investigation started by assessing the performance of the Glide XP scoring function and the MM/GBSA method, as examples of well-established free-energy models, on our carbohydrate-specific data set. The evaluated free-energy functions showed poor correlations with the experimental binding affinities in our carbohydrate data set (Fig. 1; Fig. S1 in Online Resource 1 for AutoDock and MM/PBSA). Although this finding is disappointing, it is not by any means surprising. Despite the reported success of Glide and AutoDock in reproducing crystallographic conformations and database screening, they were shown to yield inaccurate binding affinity predictions in several protein families [16]. In general, the prediction accuracy of scoring functions employed in widely used docking programs is known to be system-dependent [16, 19, 31, 32]. On the other hand, performance of MM/GBSA and MM/PBSA in free-energy predictions was in most cases assessed on uniform data sets of ligands binding to the same protein [25, 33] or on relatively small data set of different proteins [23]. In the latter case, MM/GBSA and MM/PBSA were shown to exhibit target-dependent variation in prediction accuracy in a manner similar to the scoring functions employed in docking [23, 34, 35]. However, the apparent lack of

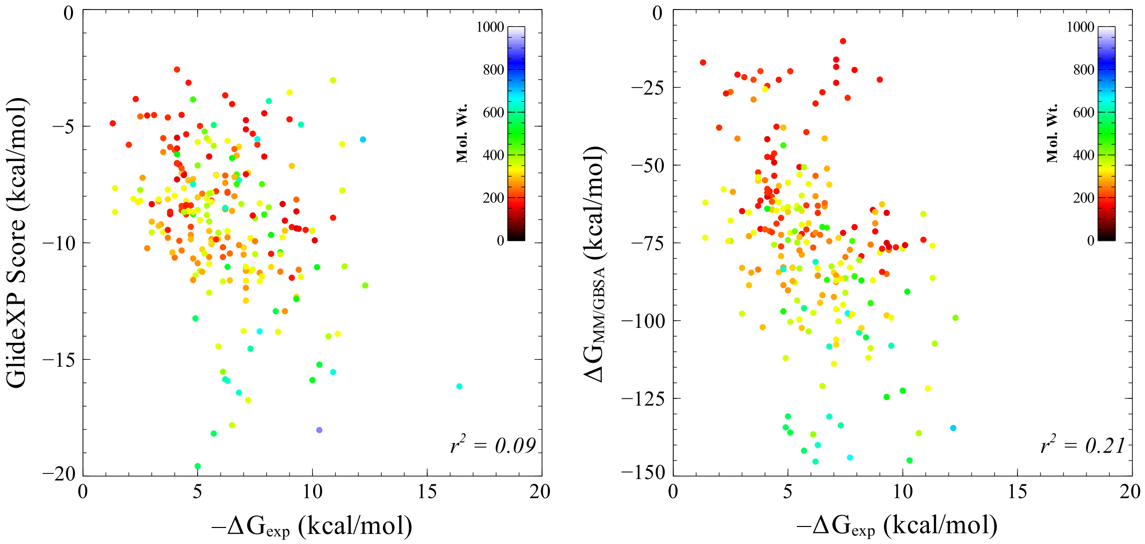

Fig. 1 Correlation plots of experimental free energies in the carbohydrate–protein data set versus Glide XP scoring function (*left*) and MM/GBSA free-energy function (*right*), points are *color-coded* according to the ligand's molecular weight ($N = 236$)

correlation in Fig. 1 is not dependent on the molecule size; i.e. the Glide XP and MM/GBSA energies incorrectly describe small rigid ligands and larger and more flexible ligands alike.

It is worth noting that both energy models were, at least to some extent, biased towards larger ligands awarding them higher scores (i.e. more negative values) in comparison to smaller ligands. Evidently, it has been reported that the binding free energy improves by ~ -1.5 kcal/mol for each non-hydrogen atom in the ligand up to a limit of 15, where it reaches a plateau [36, 37]. In addition, the solvent accessible surface that becomes buried when the ligand and the protein associate (i.e. contact area) is a major determinant of the strength of interaction [30, 38–40]. In our data set, however, no correlation was observed between binding affinities and ligand sizes or contact areas (Fig. S2, Online Resource 1). This could be attributed to the large diversity and the wide affinity range of the studied carbohydrate–protein complexes. The underlying physical model and mathematical formulation of the empirical scoring functions, e.g. Glide XP, differ significantly from those in the implicit solvent model of MM/GBSA free-energy function. Surprisingly, however, the energy scores of both methods correlate well with each other and suffer similarly from size-dependent bias in the calculated energies (Fig. S3, Online Resource 1).

It is important to note, however, that in the preliminary assessments above the four methods were used as black boxes and the calculated energies were used "as is" without parameter fitting to the carbohydrate data set. Previous studies on similar problems highlighted the difference in relative importance of certain components of binding free energy in carbohydrate–protein interactions. For example, Laederach and Reilly [26] reported that electrostatic interactions play a more important role in determining the affinity between a carbohydrate and a protein. Since the MM/GBSA model uses equal weights for the different energy components (electrostatic, vdW, etc.), it is crucial to introduce empirical weighting coefficients when applying it for carbohydrate–protein systems. Similarly, the coefficients employed in the evaluated scoring functions were optimized to reproduce the experimental affinities of specific training sets of 30 complexes in case of AutoDock [41] and 198 complexes in case of Glide XP [42]. Since the proteins employed to train these scoring functions are not necessarily carbohydrate binders, it would seem beneficial to recalibrate their coefficients for our carbohydrate-specific set.

Empirical free-energy functions

The use of linear regression models, or linear response models, is a recurring theme with several successful examples in the development of free-energy functions [43–46]; and the reported carbohydrate-specific scoring functions are, in fact, empirical models derived by recalibrating an existing scoring function on training sets of carbohydrate–protein complexes, with the occasional addition of terms to improve treatment of special interaction motifs, e.g. C–H···π interactions [26–28, 30]. The following Master Equation was employed as a testing device to assess different combinations of computational methods as potential free-energy models for carbohydrate–protein interactions.

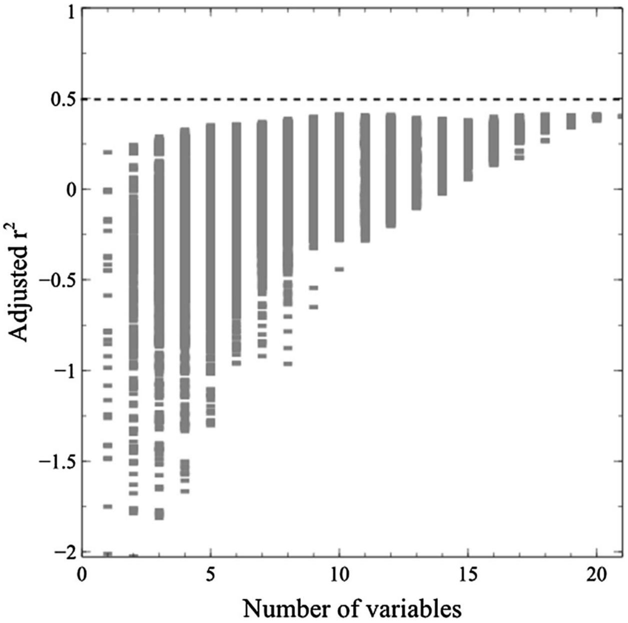

Fig. 2 Statistical assessment of the free-energy models resulting from the combinations of complex descriptors in the Master Equation (Fig. S4, Online Resource 1). The number of independent variables in the model is plotted on the *horizontal axis*, while the adjusted-r^2 as a measure of model predictive quality is plotted on the *vertical axis*. The *dotted line marks* the value of adjusted-$r^2 = 0.5$, which can be used as an arbitrary threshold delineating the potentially predictive models from the non-predictive models

$$\Delta G_{bind} = c_1\Delta G_{inter} + c_2\Delta G_{solv} + c_3\Delta G_{strain} + c_4\mathrm{T}\cdot\Delta S_{lig} + c_5\Delta G_{reward/penalty},$$

where ΔG_{inter} is the ligand–protein interaction energy, ΔG_{solv} is the desolvation penalty associated with binding, ΔG_{strain} is the conformational strain penalty, ΔS_{lig} is the entropy lost by the ligand upon binding, and $\Delta G_{reward/penalty}$ represent special rewards and penalties, e.g. the polar surface buried on binding. All permutations obtainable using different *complex descriptors* at each position in the Master Equation were evaluated (Fig. S4, Online Resource 1), aiming to investigate, as thoroughly as possible, the ability of the available repertoire of methodologies for modeling molecular interactions to formulate a reliable free-energy model for carbohydrate–protein systems. A total of 51,520 models were exhaustively enumerated and evaluated by linear fitting to the training set comprising 236 carbohydrate–protein complexes. The adjusted coefficient of determination (adjusted-r^2) was used to assess the quality of the resultant models.

The examined empirical models ranged in complexity from simple equations using a single predictor variable to complex equations using 21 variables. To our surprise, none of the assessed functions satisfactorily predicted binding affinities in our data set (Fig. 2). This was rather disappointing, since the employed pool of descriptors covered a very wide scope of structural and energetic features, including their ensemble averages from molecular dynamics (MD) simulations. It would seem, therefore, that contemporary molecular modeling methodologies with relatively low computational cost cannot be used reliably to predict binding affinity of carbohydrate–protein complexes.

Topological classification of carbohydrate-binding sites

Accounting for solvation effects is one of the most challenging issues in structure-based design. Methods combining force fields with implicit solvation model such as MM/PBSA and MM/GBSA are examples of rigorous methods with numerous successful applications in a variety of ligand–protein systems. Their performance, however, is known to be largely system-dependent [47, 48]. The physical model employed by both methods pictures the interacting molecules as zones of low dielectricity embedded in a continuum of high-dielectricity, i.e. the solvent. Among other factors, the limited accuracy of this model can be attributed to the difficulty in accurately defining the boundary between the two zones of differing dielectric properties [49–53]. Moreover, Hou et al. [23] demonstrated that MM/GBSA predictions are quite sensitive to the solute's dielectric constant. The authors recommended that the dielectric parameter 'should be carefully determined according to the characteristics of the protein/ligand binding interface'. Inaccuracy in the treatment of dielectric properties could result in errors in the final estimates of solvation contribution to the binding free energy. In principle, these errors would be relatively uniform in homogeneous sets and consequently have less negative impact on final free-energy estimates. In heterogeneous sets, however, binding sites exhibit larger variations in shape and solvent-accessibility. In such cases, the errors introduced by inaccurate dielectric boundary assignment will significantly vary with the topological features of the binding site, and hence have more detrimental effect on accuracy of the calculated free energies.

The extent to which the carbohydrate-binding site is in continuity with the solvent bulk is governed by its shape and solvent accessibility, which in turn influences key parameters of the micro-environment where the intermolecular interaction takes place, e.g. dielectric properties. Nevertheless, analytical treatment of these parameters is practically unfeasible as it typically requires long converged conformational sampling in explicit solvent affinity, such as free-energy perturbation [54, 55] and thermodynamic integration [56]. However, the complexity of the free-energy landscape could, in principle, be simplified by defining families of binding site topologies within which the binding micro-environments are roughly identical.

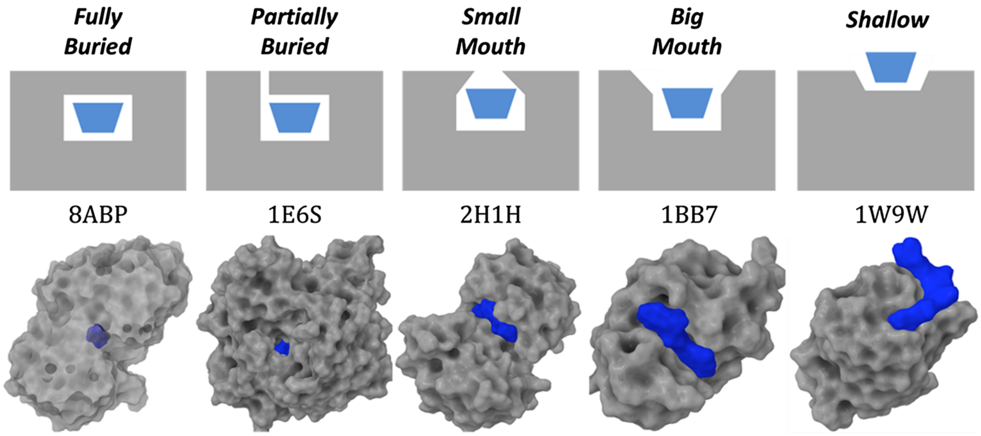

Fig. 3 Complexes were classified into five categories based on topology and solvent exposure of the carbohydrate-binding site. From *top* to *bottom*, the figure shows: category name; schematic representation of the category; PDB code for an example carbohydrate–protein complex; and the solvent-accessible surface representation of the example complex (*blue* ligand, *grey* protein). In the *left-most complex*, the protein surface is rendered transparent to show the completely buried ligand

Such topological classification could reduce the large and heterogeneous problem to a set of smaller more homogenous problems, for which simple free-energy formulations could be applied. Therefore, topologies of the carbohydrate–protein interfaces in the studied complexes were analyzed using DoGSite [57] combined with clustering and the complexes were allocated to one of five topological categories based on shape and degree of surface exposure of the binding site: fully buried, partially buried, small-mouth groove, big-mouth groove, and shallow (Fig. 3).

Figure 4 shows the distribution of key properties within the different binding site categories in our data set. As seen from the topmost plot, the proposed classification did not segregate complexes according to binding affinity, i.e. carbohydrate ligands could exhibit high or low affinity to their targets regardless of the binding-site topology. Complexes in the fully-buried category span similar range of binding affinities to those in the shallow category. There are, however, differences in molecular-weight distributions among the different categories. Fully-buried binding sites tend to accommodate smaller ligands while the three middle categories bind medium-sized ligands. On the other hand, fully exposed shallow binding sites can accommodate a wide range of ligand sizes including relatively large molecules. The area of the contact surface, however, follows a qualitatively different trend with the middle three binding categories exhibiting relatively larger interaction surfaces. The smaller average contact surfaces in fully buried binding sites could be justified by the small sizes of bound ligands in this category. Surprisingly, the shallow binding sites show on average contact surfaces of the same scale observed in case of the fully buried sites, although the former bind relatively larger ligands. This could indicate that ligands in shallow carbohydrate-recognition sites require relatively smaller contact areas to bind to their targets. This observation matches the picture of carbohydrate-binding proteins involved, for instance, in cell–cell communication, e.g. lectins, where the carbohydrate ligand is typically a large biopolymer interacting via a small di- or tri-saccharide motif at its tip. Finally, Glide XP seems to mirror the trends seen in molecular weights and contact surface areas. Glide XP tends to assign lower scores on average to ligands in the fully buried category (smaller ligands) and to those in the shallow category (small contact surface). This trend matches our earlier observation of the size-dependent bias in Glide XP scores.

The influence of categorization on the prediction accuracy of empirical scoring functions is presented in Fig. 5. It is obvious that independent training of the empirical free-energy functions for individual categories results in substantial improvement in prediction accuracy in contrast to training the models for the entire data set without categorization. A significant proportion of evaluated empirical scoring functions were capable of reproducing binding affinities of the training set with acceptable accuracy (adjusted-$r^2 > 0.6$). This result indicates that the problem at hand; i.e. predicting carbohydrate–protein binding affinities, is likely a collectively heterogeneous problem of smaller internally more homogeneous sub-problems. It is important to note, however, that the proposed classification scheme did not segregate the data set into distinct protein families (e.g. glycogen phosphorylases, neuraminidases, etc.), which could be inherently easier to model.

Free-energy models from the exhaustive search depicted in Fig. 5 (257,600 models resulting from 51,520 × 5 categories) were further analyzed to identify physically and statistically valid free-energy models. Firstly, scoring functions showing good prediction accuracy in *all*

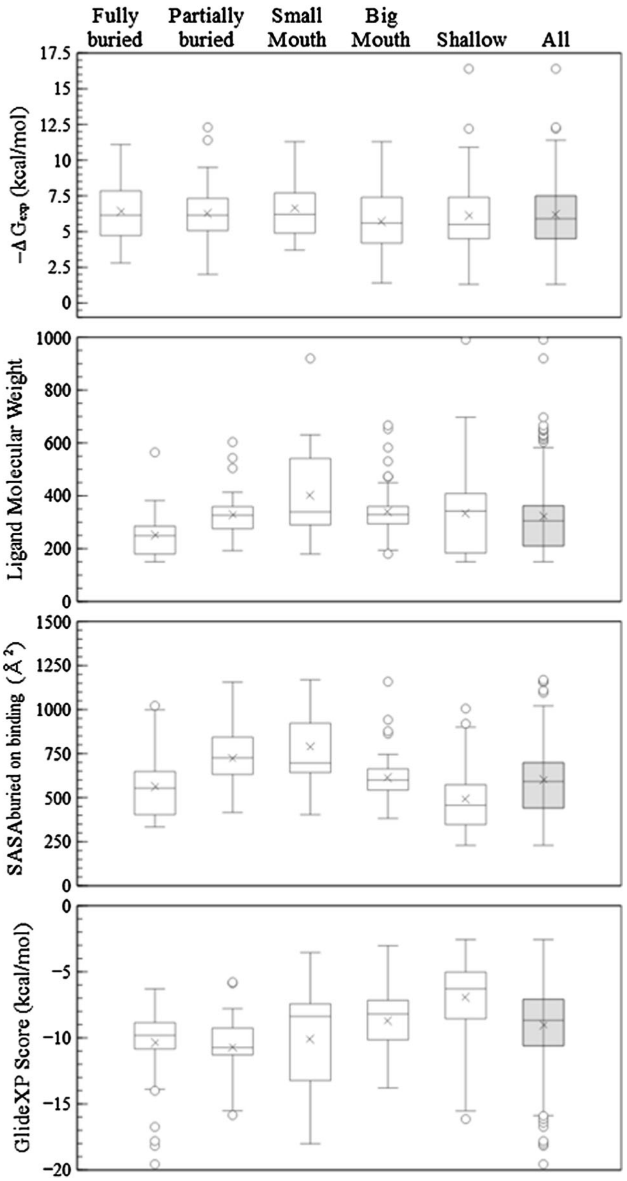

Fig. 4 Distribution of key properties within binding-site categories of the studied carbohydrate data set (*non-shaded box plots*) and the entire uncategorized data set (*shaded box plot*). Median indicated by *black bar*, average indicated by the *cross* marker. *Boxes* indicate the first (25 %) and third (75 %) quartiles. Whiskers plotted at ×1.5 interquartile range, roughly encompassing 99.7 % of the data (mean ± 3σ). *Circles* represent individual outliers larger than the upper/lower whiskers

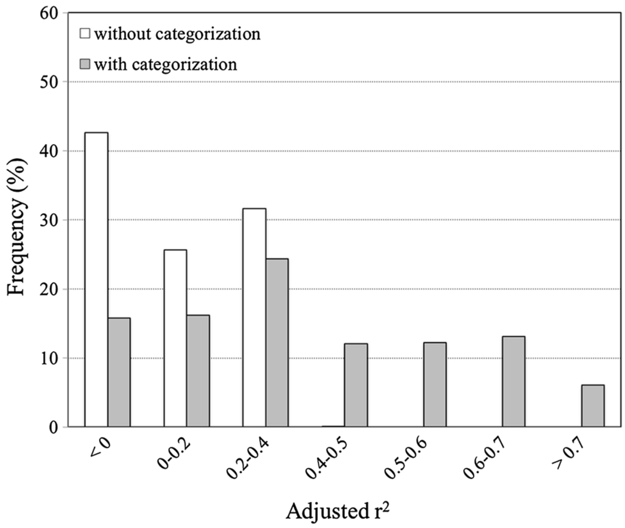

Fig. 5 Comparison of the performance of free-energy models derived from the Master Equation on the uncategorized data set and after categorization according to binding-site topology. The *vertical axis* shows the fraction of all assessed models with adjusted-r^2 in the range defined in the *horizontal axis*

Model GA1

$$-\Delta G_{bind} = c_1 E^{Glide}_{Coul} + c_2 E^{Glide}_{vdw} + c_3 SASA^{non-polar}_{buried} + c_4 SASA^{polar}_{buried} + c_5 N_{rot} + c_6 Q_{lig}$$

Model GA2

$$-\Delta G_{bind} = c_1 E^{MMFF}_{Coul} + c_2 E^{MMFF}_{vdw} + c_3 E^{MMFF}_{solvation} + c_4 SASA^{non-polar}_{buried} + c_5 SASA^{polar}_{buried} + c_6 N_{rot} + c_7 N_{inonized\ groups} + c_8 \Delta G^{MM/GBSA}_{Self-Contacts}$$

Model GA3

$$-\Delta G_{bind} = c_1 E^{OPLS}_{Coul} + c_2 E^{OPLS}_{vdw} + c_3 E^{OPLS}_{solvation} + c_4 SASA^{ligand}_{buried} + c_5 SASA^{receptor}_{buried} + c_6 N_{rot} + c_7 N_{inonized\ groups}$$

Fig. 6 Free-energy models showing the best performance after statistics and physics-based filtering

categories and exhibiting no co-linearity within the employed descriptors were kept. Secondly, models exhibiting regression coefficients that made no physical sense, e.g. entropic penalty or ligand strain energy contributing favorably to affinity, were excluded. Finally, the remaining models were subjected to stringent statistical tests including cross-validation and y-scrambling. Results of the statistical quality-based and physics-based filtering are summarized in Fig. S5 in Online Resource 1. The best performing free-energy models are listed in Fig. 6, and results of their statistical validation are shown in Table 1 (Details for models GA2, and GA3 are given in Table S1 in Online Resource 1). Models GA2d and GA3d were developed by replacing terms in the corresponding *static* models, GA2 and GA3 with the corresponding MD-derived averages (Fig. S8, Online Resource 1). Despite the evident fluctuations in the calculated interaction energies along MD simulations (Fig. S9, Online Resource 1), the use of

Table 1 Results of statistical validation for the best performing free-energy models GA1, GA2, and GA3 and the corresponding models GA2d, and GA3d using ensemble averages from MD simulations

Model	Category	N	r^2	RMSE	MUE	q^2_{LOO}	q^2_{LKO}	y-scrambling
GA1	Fully buried	58	0.67	1.25	1.04	0.53	0.52	−0.11 (−0.39, 0.16)
	Partially buried	32	0.68	1.26	0.98	0.57	0.54	−0.05 (−0.59, 0.48)
	Small mouth	29	0.82	0.89	0.76	0.70	0.67	−0.22 (−0.83, 0.32)
	Big mouth	47	0.70	1.32	1.03	0.57	0.56	−0.11 (−0.46, 0.23)
	Shallow	70	0.71	1.32	1.02	0.63	0.63	−0.30 (−0.68, 0.15)
	Pooled	236	0.71	1.25	0.99	0.60	0.59	n/a
	Uncategorized	236	0.25	2.02	1.57	0.18	0.18	n/a
GA1^{rc}	Fully buried	58	0.31	1.91	1.52	0.04	0.01	n/a
	Partially buried	32	0.34	1.79	1.44	−0.36	−0.44	n/a
	Small mouth	29	0.38	1.73	1.39	−0.27	−0.34	n/a
	Big mouth	47	0.32	1.85	1.46	−0.01	−0.05	n/a
	Shallow	70	0.29	1.93	1.51	0.07	0.05	n/a
	Pooled	236	0.35	1.88	1.48	−0.11	−0.15	n/a
GA2	Pooled	236	0.76	1.14	0.91	0.61	0.58	n/a
GA2d	Pooled	236	0.48	1.73	1.28	0.00	−0.13	n/a
GA3	Pooled	236	0.73	1.20	0.93	0.59	0.56	n/a
GA3d	Pooled	236	0.62	1.39	1.09	0.32	0.25	n/a

GA1^{rc} show the results for model GA1 when complexes are randomly allocated to binding site topological categories (average of 100 runs)

N: number of carbohydrate–protein complexes in the category; r^2: coefficient of determination; RMSE: root-mean-squared error (kcal/mol); MUE: mean unsigned error (kcal/mol); q^2: cross-validation r^2; LOO: leave-one-out cross-validation; LKO: leave-*k*-out cross-validation (*k* chosen so that the data set is divided into seven equal subsets); y-scrambling: r^2 values resulting from randomly assigning experimental free energy values amongst the training set complexes, average(minimum, maximum) r^2 values from 100 scrambling cycles

dynamic averages of interaction energies had a negative impact on the prediction quality of the free-energy models (Table 1), which could indicate that longer and more extensive simulations are required [23, 47, 58].

The GA1 model exhibited the best balance between complexity and comprised Columbic and van der Waals interaction energies from the Glide XP scoring function, two solvent-accessible surface area terms accounting for the non-polar and polar solvent-accessible surface area (SASA) that becomes buried on binding, and two reward/penalty terms for the number of rotatable bonds (N_{rot}) and formal charge of the ligand (Q_{lig}). Statistical performance of the model is summarized in Table 1. The GA1 model reproduced binding free energies within topological categories with r^2 values ranging from 0.67 to 0.82, RMSE from 0.89 to 1.32 kcal/mol and mean unsigned errors of 0.76–1.04 kcal/mol in the predicted free energies. Results of leave-one-out and leave-*k*-out cross-validation confirm robustness and internal consistency of the model. In the leave-*k*-out cross-validation, the *k* is chosen such that in each cycle one-*seventh* of the training set is removed then predicted using the model trained for the remaining complexes. The perturbation introduced by removing one-seventh of the complexes is more significant compared to removing a single complex in leave-one-out cross-validation. The leave-*k*-out cross-validation, therefore, is a more stringent test for model robustness. Finally, randomization of experimental affinities across carbohydrate–protein complexes in each category resulted in a substantial drop in quality prediction.

To assess the overall performance of the GA1 free-energy model, prediction errors were *pooled* from the five binding site topological categories. The GA1 model reproduces binding free energies in the entire data set within RMSE of 1.25 kcal/mol, which corresponds to a factor of 10-off from experimental values. Prediction accuracy of the GA1 model is substantially reduced when applied to the entire uncategorized data set. Notably, the GA1 model did not exhibit the size-dependent bias observed in the traditional scoring functions (Fig. S6, Online Resource 1). Furthermore, Fig. 7 presents the influence of the proposed categorization scheme on the performance of the GA free-energy model. The GA1 Model does not seem to exhibit systematic over- or under-estimations in the predicted ΔG values. However, it shows a slight bias in the plot of residuals against experimental ΔG values (Fig. S7, Online Resource 1), i.e. some high affinity ligands are underestimated while some low affinity ligands are overestimated. On the other hand, in the range $3.0 \leq \Delta G_{bind} \leq 12.0$ kcal/mol, the residuals are more evenly distributed with no clear bias.

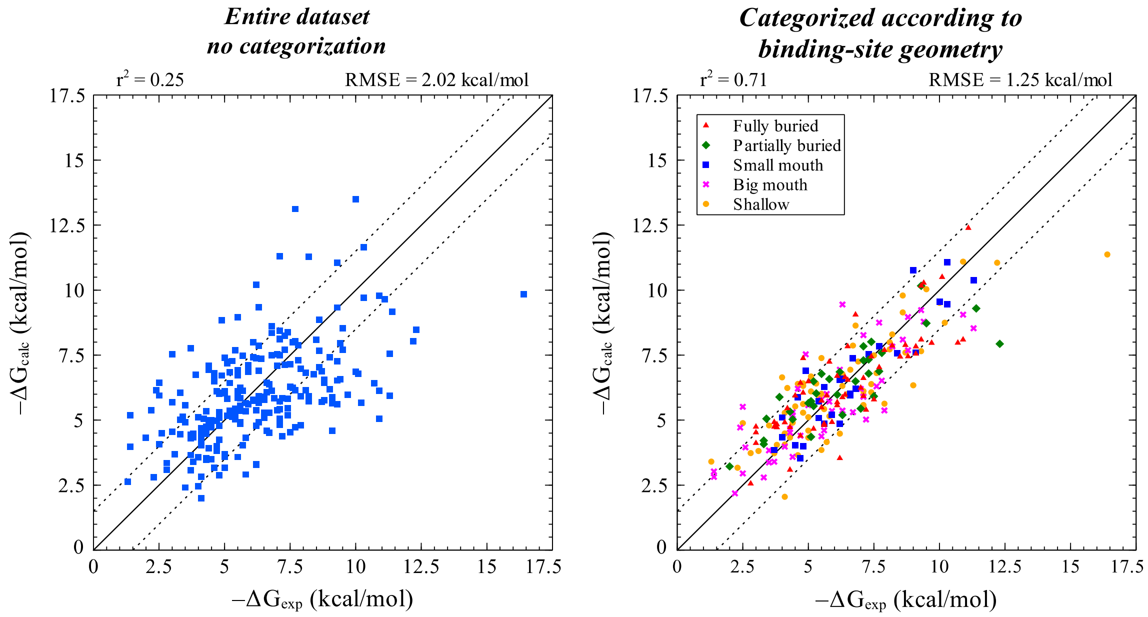

Fig. 7 Distributing the carbohydrate–protein data set into binding site topological categories according to the proposed classification scheme leads to a substantial improvement in the performance of the GA1 empirical free-energy model ($N = 236$). *Dashed lines* mark tenfold deviations from experimental binding affinity

Table 2 Average contributions of individual free-energy components in the GA1 free-energy model to the total binding free energy in different binding site topological categories

Category	E_{vdw}^{Glide}	E_{Coul}^{Glide}	$SASA_{buried}^{non-polar}$	$SASA_{buried}^{polar}$	N_{rot}	Q_{lig}
Fully buried	2.87 ± 1.06	7.85 ± 2.64	−0.78 ± 0.34	−0.96 ± 0.28	−1.90 ± 0.69	−0.68 ± 1.73
Partially buried	1.97 ± 0.53	8.48 ± 4.05	2.64 ± 0.86	−3.13 ± 0.79	−2.66 ± 1.01	−0.93 ± 2.64
Small mouth	−4.25 ± 2.07	2.80 ± 1.82	9.48 ± 4.09	4.57 ± 1.58	−5.24 ± 2.55	−0.81 ± 0.80
Big mouth	5.64 ± 1.75	3.34 ± 1.22	1.02 ± 0.33	1.84 ± 0.45	−6.15 ± 2.32	0.01 ± 0.10
Shallow	2.05 ± 1.26	3.92 ± 1.34	1.32 ± 0.70	1.04 ± 0.34	−2.16 ± 1.20	−0.01 ± 0.13

Values are given as mean ± SD in kcal/mol

The improvement in the performance of the GA1 model could be a mere consequence of reducing the dimensionality of the problem from the total of 236 complexes in the complete data set to smaller subsets of 29–70 complexes per category. To examine this possibility, carbohydrate–protein complexes were randomly allocated to five dummy categories having the same sizes of the binding-site topological categories disregarding the actual binding-site topology. The GA1 model was then applied to the resultant categories and its performance was evaluated. Average performance results from 100 category-randomization runs are presented in Table 1. The apparent deterioration of the GA1 model performance confirms that mixing complexes with differing binding site topologies in small categories is not alone sufficient to yield useful free-energy correlations. This further confirms the relevance of actual binding site topology in defining the free-energy response surface within categories and also verifies the validity of the proposed classification scheme.

Since the GA1 free-energy model was fitted five times, once for each binding site topological category, five sets of empirical weighting coefficients were obtained. The empirical coefficients are listed in Table 2 after multiplying each of them by the mean and the standard deviation of the corresponding energy components for each category. The resulting values are the mean (±SD) of the free energy contributed by each component in the GA1 model to the total binding free energy within individual categories. As seen from Table 2, the values of the average energy contributions (and the underlying empirical weighting coefficients) show evident category-dependent variations. Interpretation of these coefficients, however, could be complicated by their unavoidable dependence on the training set and the inherent complexity of the free-energy landscape. Nevertheless, a couple of interesting trends can be noted. Firstly, the contribution of electrostatic interactions to the total free energy is relatively larger in the fully buried and partially buried categories. This could be attributed to the differences in rewards for releasing the more trapped water molecules in these two categories compared to the relatively more easily exchangeable waters in the remaining categories. Secondly, the existence of charged groups (reflected by the formal charge of the ligand, Q_{lig}) is associated with moderate penalty in the fully buried, partially buried and small mouth categories. In the big mouth and shallow categories, however, the contribution of Q_{lig} to binding free energy is nearly negligible. This could be justified by the expected higher cost for moving charges from the bulk solvent to the protein interior in the former three categories, while in the latter two categories the formal charge could interact with the solvent to some extent. It is also noteworthy that the contribution of electrostatic interactions to the binding free energy is roughly similar to those of vdW interactions, which is in agreement with the JA model reported by Hill and Reilly on the expanded carbohydrate data set [27].

Conclusion

The increasing interest in carbohydrate-based therapeutics in the past few decades has intensified the need for reliable and efficient molecular modeling tools specifically dealing with quantification of carbohydrate–protein interactions. We thoroughly investigated the performance of well-established computational methodologies on a specially curated set of 236 diverse carbohydrate–protein crystal structures with known binding affinity. Although the descriptor pool (with approximately 170 entries) extends across a significant portion of the potential solution space, none of the assessed models satisfactorily predicted the binding affinities in our data set. Binding site topologies were clustered and the complexes in our data set were allocated into five topological categories based on the shape and degree of surface exposure of the carbohydrate-binding site: fully buried, partially buried, small-mouth groove, big-mouth groove, and shallow. Free-energy models independently fitted for individual categories exhibited a substantial improvement in prediction accuracy. The best performing free-energy model (GA1 model) exhibited an overall r^2 of 0.71 and a RMSE of 1.25 kcal/mol in the predicted binding affinity (corresponding to a factor of 10 in the affinity). The results would seem to indicate that topological classification could be used to reduce the large and heterogeneous problem to a set of smaller more homogenous problems, for which simple free-energy formulations could be applied.

Despite the known difficulties in calculating binding affinities for carbohydrate–protein complexes, this study have achieved three important goals. First, a high-quality binding affinity data set for a large and diverse collection carbohydrate–protein complexes has been compiled and thoroughly revised. Second, we proposed a rigorous function for predicting binding affinity from the atomic configuration of carbohydrate–protein complexes. Finally, we propose classification of carbohydrate-binding proteins according to the topology and surface exposure of the binding site. Differences between the free-energy models individually calibrated for each topological class reflect the differences in the nature of the local binding micro-environments. Although it might be difficult to fully explain how such differences might affect the shape of the free-energy response surface, the results of this study show how these differences complicate the free-energy prediction problem and demonstrate the usefulness of calibrating free-energy functions individually according to binding-site topology and surface exposure.

Computational methods

Preparing carbohydrate–protein complexes

Compiling the data set

A pool of ligand–protein complexes was gathered by mining three databases: the Protein Data Bank for structural information, and Binding MOAD [59] and Binding-DB [60] for binding affinities. Complexes used previously in similar studies were also included [26–28, 30]. The crude collection was refined to a data set of 273 entries of reviewed experimental affinities for carbohydrate–protein complexes (a detailed listing is given in Table S3 in Online Resource 1). Some complexes were excluded during the structure preparation step due to uncertainties in geometry or the inability of common force fields to handle some ligand atoms (cf. Online Resource 1). The final data set employed in the study of free-energy models contained 236 complexes. The employed set comprised 90 unique proteins (corresponding to 65 unique SCOP and 43 unique CATH domain classes) and 175 unique carbohydrate ligands (cf. Fig. S10 in Online Resource 1 for more details). All binding affinity values were converted to binding free energies (ΔG, kcal/mol) using the thermodynamic master equation $\Delta G = -\mathrm{RT}ln\mathrm{K}$.

Preprocessing complexes

All ligand–protein complexes were retrieved from the Protein Data Bank (www.pdb.org) and processed using

Maestro's Protein Preparation Wizard (Maestro, version 9.2, 2011, Schrödinger, LLC, New York). All hydrogen atoms in the input structures were deleted, bond orders were automatically assigned, and hydrogens were added accordingly. Water molecules within 5.0 Å from non-standard residues (e.g. ligands, cofactors, metals) were kept and all other water molecules were deleted. Missing side chains were completed and optimized using Prime (Prime, version 3.0, 2011, Schrödinger, LLC, New York).

Multiple ligand copies

When a complex exhibited multiple chains with several copies of the ligand molecule in the asymmetric unit, the individual chains were superimposed and heavy-atom RMSDs were computed for the ligand and the surrounding residues. In most complexes all the copies had RMSD values within 1.0 Å; in which case the first chain having a resolved ligand was used and its chain identifier was noted. Complexes where ligand copies differed significantly in conformation and/or orientation in the binding site, i.e. RMSD $>$ 1.0 Å were discarded (examples: 1A0T and 1JZ7). In some complexes, the ligand had two overlapping representations, mostly resulting from the α- and β-anomers being simultaneously resolved in the binding pocket. Unless the affinity measurement explicitly refers to the β-anomer, the α-anomer was used in subsequent computations and the β-anomer copy was deleted. In some complexes there was a ligand copy in an allosteric binding site, as indicated in the original publication of the PDB structure. In such cases, we confirmed that the measured affinity was competitive by revisiting the respective publication, and subsequently deleted the allosteric copy of the ligand (examples: 2QN8 and 2QNB). Before proceeding, we made sure that each complex had one, and only one, ligand copy. Relevant processing notes—e.g. retained chains in case of multiple-chain PDB's, deleted ligand copies, etc.—are given in Table S3 in Online Resource 1.

Covalent structure and protonation

Each ligand's chemical structure was cross-checked against the corresponding primary citation and inconsistencies resulting from incorrect bond order assignments were corrected manually. Protonation and tautomeric states for all HET groups were automatically assigned using Epik [61]. We used the protonation state of the ligand whenever it was explicitly mentioned in the original publication; otherwise the top-ranked suggestion from Epik was used. At this stage, fully-atomistic models of all 236 ligand–protein complexes, each having a unique ligand molecule with revised chemical structure and protonation state, were ready for the subsequent analyses.

Geometry optimization

The geometry and orientation of all added hydrogen atoms were exhaustively sampled for optimal H-bond formation, including any necessary flipping of glutamine, asparagine, and histidine side chains. Finally, each complex was refined by full minimization using OPLS_2005 force field as implemented in Schrödinger's MacroModel (MacroModel, version 9.9, 2011, Schrödinger, LLC, New York). Minimization was set to converge within heavy-atom RMSD of 0.3 Å from the input geometry to avoid significant deviations from the experimental geometry.

Complex descriptors

A complex descriptor is a quantity measuring some geometric or energy-based feature of a given ligand–protein complex. In the context of this study, they serve as the building blocks of the investigated empirical scoring functions (cf. Table S2 in Online Resource 1 and Online Resource 2).

Non-bonded interaction energies from force fields

The first force field employed in this study was OPLS_2005, the MacroModel implementation of the OPLS-All-Atom force field [62]. Optimized potentials for liquid simulations (OPLS) was originally optimized for protein simulations [63], and later upgraded to the all-atom variant OPLS-AA [64], then extended to *carbohydrates* by refitting some of the parameters to ab initio results for complete hexopyranoses [65] and by applying additional scaling factors for the 1.5 and 1.6 electrostatic interactions [66]. Moreover, OPLS-AA-driven MD simulations have been successfully employed for studying carbohydrate–protein interactions [67, 68]. The second force field employed in this study was MMFFs, MacroModel implementation of the MMFF94s force field [69–71]. The Merck molecular force field (MMFF) was parameterized using a wide variety of chemical systems, and targets simulations of small molecules as well as proteins and biological systems. The MMFF94s variant enforces planarity around sp^2 hybridized nitrogens. The chemical classes included in MMFF94 core parameterization do not include carbohydrates, though. We included the MMFFs as a general-utility biomolecular force field to compare its performance against OPLS-AA, which has been optimized for carbohydrates. The non-bonded interaction energy components (electrostatic, van der Waals, and solvation) were calculated for each complex by performing a single-point energy calculation using the respective force field on the ligand–protein complex, the protein alone, and the ligand alone according to the formula:

$$E_{non-bonded} = E_{complex} - (E_{ligand} + E_{protein})$$

MM/GBSA and MM/PBSA free-energy functions

The combined Molecular Mechanics/implicit solvent models such as the Generalized Born Surface Area (MM/GBSA) and the Poisson–Boltzmann Surface Area (MM/PBSA) approaches offer a good compromise between computational efficiency and accurate treatment of solvation effects [72, 73]. In the current study, MM/GBSA computation were performed in Schrödinger's Prime, using the VSGB 2.0 energy model [74] to calculate the GBSA contribution and the OPLS-AA force field to calculate the molecular-mechanics energy [64–66]. The VSGB 2.0 model includes physics-based correction terms for improved handling of π–π stacking, hydrogen-bonding interactions, hydrophobic interactions, and self-contacts of the side chains of certain residues. Moreover, the VSGB 2.0 model employs a Surface Generalized Born (SGB) model [75, 76] in conjunction with a variable dielectric (VD) treatment to account for polarization effects from protein side chains by varying the internal dielectric constants from 1.0 to 4.0 [77].

For MM/PBSA calculations, carbohydrate–protein complexes were prepared with the Leap module of the AMBER 12 suite [78] using the AMBER 99SB force-field [79]. Prior to processing, structures were minimized with the Sander module (25 cycles). The MMPBSA.py script was used for all energy calculations [80]. Ions and water molecules were removed and the ionic strength was set to 0.15 M. The PB equation was solved numerically by the *pbsa* program. The MM/GBSA and MM/PBSA-derived ΔG_{bind} and their components employed as complex descriptors are listed in Table S2 in Online Resource 1.

Glide XP and AutoDock scoring functions

We included two well-established scoring functions as sources for complex descriptors in our study; namely Glide XP and AutoDock. Glide (Grid-based Ligand Docking with Energetics) is a widely used docking software [81], which has been successfully employed to predict and rank binding configurations of carbohydrate ligands to protein targets [82–84]. The scoring function employed in Glide is based on the empirical ChemScore function [85] and has two variants; Glide SP (Standard Precision) and Glide XP (eXtra Precision). Glide XP has numerous specific reward and penalty terms and covers a wider range of ligand–protein interaction motifs, which makes it more suitable for our study [42]. Glide (Glide, version 5.7, 2011, Schrödinger, LLC, New York, NY) was used to calculate the docking scores for the studied complexes. Scores were computed using two modes: (1) the *in place* mode, where the input ligand coordinates are used directly for scoring, and (2) the *refine input* mode, where the input ligand coordinates are optimized in the field of the receptor prior to scoring.

The second scoring function considered in this study was the AutoDock empirical scoring function [41, 86]. AutoDock has been used in several studies for modeling and quantification of ligand–protein interactions [19, 82, 84] and has provided the basis for two empirical carbohydrate-specific free-energy models [26, 27]. The AutoDock scoring function employs the change in solvent-accessible surface area of non-polar ligand atoms to account for the solvation contribution [41]. AutoDock scores for the studied complexes were computed using the scoring function implemented in AutoDock 4.2 [87].

Entropic penalty

Change in entropy upon ligand–protein association is probably the most elusive component of the binding free energy. Commonly, a constant penalty is assigned for each freely rotatable bond in the ligand, ranging in value from 0.4 to 1.0 kcal/mol [20]. We also included the entropic term proposed by Hill and Reilly, which employs an empirical coupling coefficient, ξ, to account for loss of translational and rotational degrees-of-freedom upon binding [27]. Moreover, we included the entropic penalty term employed in Glide scoring function, which accounts for the residual ligand mobility by applying the penalty only to bonds expected to be frozen in the bound conformation [85]. Finally, we used the rigid-rotor harmonic oscillator approximation to estimate the changes in vibrational, rotational, and translational components of ligand's entropy upon binding (MacroModel, 2011, Schrödinger, LLC, New York).

Characterization of binding sites

Changes in the polar and non-polar molecular surfaces play a key role in ligand–protein interactions [20, 38–40]. To account for these changes, several SASA components were calculated in Maestro using a water-sized spherical probe (radius = 1.4 Å) scanning the surface of the analyzed molecule(s) at 0.1 Å spaced grid points (cf. Table S2 and Fig. S11 in Online Resource 1). To characterize the topology of carbohydrate-binding sites, the studied complexes were analyzed using DoGSite [57]. DoGSite employs a 3D Difference-of-Gaussian filter to identify and characterize binding pockets and splits identified pockets into subpockets, thereby allowing a refined structural description of the topology of active sites. DoGSite captures the key topological features binding sites including volume, surface area (total, protein-contact, and solvent exposed), pocket depth, ligand coverage, and pocket coverage. Carbohydrate–protein complexes were allocated into five non-overlapping categories by applying the Density Based Spatial Clustering of Applications with Noise (DBSCAN) unsupervised clustering algorithm [88] to the pool of SASA and DoGSite descriptors (cf. Online Resource 2).

Ligand-based descriptors

A number of ligand-derived descriptors were included to represent potentially relevant structural and energetic features, in our descriptor pool. The molecular weight and number of heavy atoms of the ligand were included to compensate for the potential size bias observed in scoring function [19], e.g. by penalizing large ligands and/or rewarding relatively smaller ligands [42]. We also included descriptors to account for ligand internal strain; defined as the energetic cost paid for forcing the relaxed unbound conformation of the ligand to assume the bioactive conformation. The relaxed conformation could be taken to be the *nearest local minimum* found in by typical energy minimization or to the *global minimum* [89]. The global minima for the studied carbohydrate ligands were obtained through an exhaustive conformational search using MacroModel, setting the maximum number of generated conformers to 5,000 and employing a wide energy window (40.0 kcal/mol) for conformer rejection. In addition, the SM8 quantum mechanical aqueous continuum solvation model [90] was employed to estimate ligands' desolvation penalties. The computation was carried out on the crystallographic ligand conformation using B3LYP density functional and the 6-31G** basis set in Jaguar (version 7.8, Schrödinger, LLC, New York, NY). We also employed SM8 solvation free energy weighted according to the ligand's buried surface area to account for partial ligand desolvation, particularly for ligands bound close to the surface.

Statistical validation

Empirical free-energy models investigated in this study were linear combinations of terms each representing a component of the free-energy change associated with binding.

$$\Delta G_{bind} = c_1\Delta G_1 + c_2\Delta G_2 + \cdots$$

The experimental binding affinity, ΔG_{bind}, is the dependent (or response) variable (y) while the complex descriptors, ΔG_i's, constitute the independent (or predictor) variables (x's). Standard multiple linear regression was used to derive the weighting coefficients, c_i's, by fitting the linear equation(s) to experimental binding affinities. All generated models were subjected to rigorous validation using traditional statistical methods; including coefficient of determination r^2, cross-validation r^2 (q^2), scrambling of response variable (binding affinity), as well as random allocation of the complexes to topological sub-categories (cf. Online Resource 1 for details). In all cases, models lacking physicochemical sense were not considered.

Acknowledgments This research was financially supported by the Swiss National Foundation (Grant No. 315230_122478) and the Department of Pharmaceutical Sciences, University of Basel. We thank Prof. Beat Ernst and Dr. Martin Smieško for the most valuable discussions.

References

1. Cummings RD (2009) The repertoire of glycan determinants in the human glycome. Mol BioSyst 5(10):1087–1104
2. Magnani JL, Ernst B (2009) Glycomimetic drugs—a new source of therapeutic opportunities. Discov Med 8(43):247–252
3. Shukla RK, Tiwari A (2011) Carbohydrate molecules: an expanding horizon in drug delivery and biomedicine. Crit Rev Ther Drug Carrier Syst 28(3):255–292
4. van Ree R (2002) Carbohydrate epitopes and their relevance for the diagnosis and treatment of allergic diseases. Int Arch Allergy Immunol 129(3):189–197
5. Ernst B, Magnani JL (2009) From carbohydrate leads to glycomimetic drugs. Nat Rev Drug Discov 8(8):661–677
6. Galan MC, Benito-Alifonso D, Watt GM (2011) Carbohydrate chemistry in drug discovery. Org Biomol Chem 9(10):3598–3610
7. Cipolla L, Araújo AC, Bini D, Gabrielli L, Russo L, Shaikh N (2010) Discovery and design of carbohydrate-based therapeutics. Expert Opin Drug Discov 5(8):721–737
8. Foley BL, Tessier MB, Woods RJ (2012) Carbohydrate force fields. Wiley Interdiscip Rev Comput Mol Sci 2(4):652–697
9. Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL et al (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. J Comput Chem 29(4):622–655
10. Laughrey ZR, Kiehna SE, Riemen AJ, Waters ML (2008) Carbohydrate–pi interactions: what are they worth? J Am Chem Soc 130(44):14625–14633
11. Brandl M, Weiss MS, Jabs A, Sühnel J, Hilgenfeld R (2001) C-H…pi-interactions in proteins. J Mol Biol 307(1):357–377
12. Muraki M (2002) The importance of CH/pi interactions to the function of carbohydrate binding proteins. Protein Pept Lett 9(3):195–209
13. Tvaroska I, Carver JP (1998) The anomeric and exo-anomeric effects of a hydroxyl group and the stereochemistry of the hemiacetal linkage. Carbohydr Res 309(1):1–9
14. Fadda E, Woods RJ (2010) Molecular simulations of carbohydrates and protein–carbohydrate interactions: motivation, issues and prospects. Drug Discov Today 15(15–16):596–609
15. Tempel W, Tschampel S, Woods RJ (2002) The xenograft antigen bound to *Griffonia simplicifolia* lectin 1-β(4). X-ray crystal structure of the complex and molecular dynamics characterization of the binding site. J Biol Chem 277(8):6615–6621
16. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH et al (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931

17. Huang SY, Grinter SZ, Zou X (2010) Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. Phys Chem Chem Phys 12(40):12899–12908
18. Cheng T, Li X, Li Y, Liu Z, Wang R (2009) Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model 49(4):1079–1093
19. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL (2004) Assessing scoring functions for protein–ligand interactions. J Med Chem 47(12):3032–3047
20. Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl 41(15):2644–2676
21. Grosdidier S, Fernández-Recio J (2009) Docking and scoring: applications to drug discovery in the interactomics era. Expert Opin Drug Discov 4(6):673–686
22. Guimaraes CRW (2011) Direct comparison of the MM-GB/SA scoring procedure and free-energy perturbation calculations using carbonic anhydrase as a test case: strengths and pitfalls. J Chem Theory Comput 7(7):2296–2306
23. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. J Chem Inf Model 51(1):69–82
24. Núñez S, Venhorst J, Kruse CG (2010) Assessment of a novel scoring method based on solvent accessible surface area descriptors. J Chem Inf Model 50(4):480–486
25. Rastelli G, Del Rio A, Degliesposti G, Sgobba M (2010) Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. J Comput Chem 31(4):797–810
26. Laederach A, Reilly PJ (2003) Specific empirical free energy function for automated docking of carbohydrates to proteins. J Comput Chem 24(14):1748–1757
27. Hill AD, Reilly PJ (2008) A Gibbs free energy correlation for automated docking of carbohydrates. J Comput Chem 29(7):1131–1141
28. Kerzmann A, Neumann D, Kohlbacher O (2006) SLICK–scoring and energy functions for protein–carbohydrate interactions. J Chem Inf Model 46(4):1635–1642
29. Fernández-Alonso MC, Cañada FJ, Jiménez-Barbero J, Cuevas G (2005) Molecular recognition of saccharides by proteins. Insights on the origin of the carbohydrate–aromatic interactions. J Am Chem Soc 127(20):7379–7386
30. Kerzmann A, Fuhrmann J, Kohlbacher O, Neumann D (2008) BALLDock/SLICK: a new method for protein–carbohydrate docking. J Chem Inf Model 48(8):1616–1625
31. Mooij WT, Verdonk ML (2005) General and targeted statistical potentials for protein–ligand interactions. Proteins 61(2):272–287
32. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56(2):235–249
33. Srivastava HK, Sastry GN (2012) Molecular dynamics investigation on a series of HIV protease inhibitors: assessing the performance of MM-PBSA and MM-GBSA approaches. J Chem Inf Model 52(11):3088–3098
34. Guimarães CRW, Mathiowetz AM (2010) Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations. J Chem Inf Model 50(4):547–559
35. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. J Comput Chem 32(5):866–877
36. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. Proc Natl Acad Sci USA 96(18):9997–10002
37. Neumann D, Lehr CM, Lenhof HP, Kohlbacher O (2004) Computational modeling of the sugar–lectin interaction. Adv Drug Deliv Rev 56(4):437–457
38. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. Nature 319(6050):199–203
39. Pace CN (1992) Contribution of the hydrophobic effect to globular protein stability. J Mol Biol 226(1):29–35
40. Wang J, Wang W, Huo S, Lee M, Kollman PA (2001) Solvation model based on weighted solvent accessible surface area. J Phys Chem B 105(21):5055–5067
41. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK et al (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19(14):1639–1662
42. Friesner R, Murphy R, Repasky M, Frye L, Greenwood J, Halgren T et al (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J Med Chem 49(21):6177–6196
43. Aqvist J, Marelius J (2001) The linear interaction energy method for predicting ligand binding free energies. Comb Chem High Throughput Screen 4(8):613–626
44. Hansson T, Marelius J, Aqvist J (1998) Ligand binding affinity prediction by linear interaction energy methods. J Comput Aided Mol Des 12(1):27–35
45. Marelius J, Ljungberg KB, Aqvist J (2001) Sensitivity of an empirical affinity scoring function to changes in receptor–ligand complex conformations. Eur J Pharm Sci 14(1):87–95
46. Wesolowski SS, Jorgensen WL (2002) Estimation of binding affinities for celecoxib analogues with COX-2 via Monte Carlo-extended linear response. Bioorg Med Chem Lett 12(3):267–270
47. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M (2005) Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 48(12):4040–4048
48. Pearlman DA (2005) Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. J Med Chem 48(24):7796–7807
49. Bordner AJ, Huber GA (2003) Boundary element solution of the linear Poisson–Boltzmann equation and a multipole method for the rapid calculation of forces on macromolecules in solution. J Comput Chem 24(3):353–367
50. Boschitsch AH, Fenley MO (2004) Hybrid boundary element and finite difference method for solving the nonlinear Poisson-Boltzmann equation. J Comput Chem 25(7):935–955
51. Davis ME, McCammon JA (1991) Dielectric boundary smoothing in finite difference solutions of the poisson equation: an approach to improve accuracy and convergence. J Comput Chem 12(7):909–912
52. Fogolari F, Brigo A, Molinari H (2002) The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology. J Mol Recognit 15(6):377–392
53. Neves-Petersen MT, Petersen SB (2003) Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules—applications in biotechnology. Biotechnol Annu Rev 9:315–395
54. Jorgensen WL, Thomas LL (2008) Perspective on free-energy perturbation calculations for chemical equilibria. J Chem Theory Comput 4(6):869–876
55. Zwanzig RW (1954) High-temperature equation of state by a perturbation method. I. Nonpolar gases. J Chem Phys 22(8):1420–1426
56. Rodinger T, Howell PL, Pomes R (2005) Absolute free energy calculations by thermodynamic integration in four spatial dimensions. J Chem Phys 123(3):34104–34111
57. Volkamer A, Griewel A, Grombacher T, Rarey M (2010) Analyzing the Topology of active sites: on the prediction of pockets and subpockets. J Chem Inf Model 50(11):2041–2052

58. Genheden S, Ryde U (2010) How to obtain statistically converged MM/GBSA results. J Comput Chem 31(4):837–846
59. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P et al (2008) Binding MOAD, a high-quality protein–ligand database. Nucleic Acids Res 36(database issue):D674–D678
60. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res 35(database issue):D198–D201
61. Shelley J, Cholleti A, Frye L, Greenwood J, Timlin M, Uchimaya M (2007) Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. J Comput Aided Mol Des 21(12):681–691
62. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105(28):6474–6487
63. Jorgensen WL, Tirado-Rives J (1988) The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110(6):1657–1666
64. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118(45):11225–11236
65. Damm W, Frontera A, Tirado-Rives J, Jorgensen WL (1997) OPLS all-atom force field for carbohydrates. J Comput Chem 18(16):1955–1970
66. Kony D, Damm W, Stoll S, Van Gunsteren WF (2002) An improved OPLS-AA force field for carbohydrates. J Comput Chem 23(15):1416–1429
67. Sharma A, Vijayan M (2011) Influence of glycosidic linkage on the nature of carbohydrate binding in β-prism I fold lectins: an X-ray and molecular dynamics investigation on banana lectin–carbohydrate complexes. Glycobiology 21(1):23–33
68. Margulis CJ (2005) Computational study of the dynamics of mannose disaccharides free in solution and bound to the potent anti-HIV virucidal protein cyanovirin. J Phys Chem B 109(8):3639–3647
69. Halgren TA (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comput Chem 17(5–6):490–519
70. Halgren TA (1999) MMFF VI. MMFF94s option for energy minimization studies. J Comput Chem 20(7):720–729
71. Halgren TA (1999) MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. J Comput Chem 20(7):730–748
72. Wang JM, Hou TJ, Xu XJ (2006) Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. Curr Comput-Aided Drug Des 2(3):287–306
73. Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L et al (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. Acc Chem Res 33(12):889–897
74. Li J, Abel R, Zhu K, Cao Y, Zhao S, Friesner RA (2011) The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. Proteins 79(10):2794–2812
75. Ghosh A, Rapp CS, Friesner RA (1998) Generalized born model based on a surface integral formulation. J Phys Chem B 102(52):10983–10990
76. Yu Z, Jacobson MP, Friesner RA (2006) What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. J Comput Chem 27(1):72–89
77. Zhu K, Shirts MR, Friesner RA (2007) Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects. J Chem Theory Comput 3(6):2108–2119
78. Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE et al (2012) AMBER 12. University of California, San Francisco
79. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 65(3):712–725
80. Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012) MMPBSA.py: an efficient program for end-state free energy calculations. J Chem Theory Comput 8(9): 3314–3321
81. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749
82. Agostino M, Jene C, Boyle T, Ramsland P, Yuriev E (2009) Molecular docking of carbohydrate ligands to antibodies: structural validation against crystal structures. J Chem Inf Model 49(12):2749–2760
83. Alexacou K-M, Hayes JM, Tiraidis C, Zographos SE, Leonidas DD, Chrysina ED et al (2008) Crystallographic and computational studies on 4-phenyl-N-(beta-D-glucopyranosyl)-1H-1,2,3-triazole-1-acetamide, an inhibitor of glycogen phosphorylase: comparison with alpha-D-glucose, N-acetyl-beta-D-glucopyranosylamine and N-benzoyl-N′-beta-D-glucopyran. Proteins 71(3):1307–1323
84. Nurisso A, Kozmon S, Imberty A (2008) Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III. Mol Simul 34(4):469–479
85. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 11(5):425–445
86. Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semi-empirical free energy force field with charge-based desolvation. J Comput Chem 28(6):1145–1152
87. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS et al (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. J Comput Chem 30(16):2785–2791
88. Ester M, Kriegel H, Sander J, Xu X (eds) (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Second international conference on knowledge discovery and data mining, AAAI Press
89. Perola E, Charifson PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. J Med Chem 47(10):2499–2510
90. Marenich AV, Olson RM, Kelly CP, Cramer CJ, Truhlar DG (2007) Self-consistent reaction field model for aqueous and nonaqueous solutions based on accurate polarized partial charges. J Chem Theory Comput 3(6):2011–2033