Glycobiology Advance Access published May 23, 2014



# Development and application of site mapping methods for the design of glycosaminoglycans

*Mark Agostino, Neha S. Gandhi and Ricardo L. Mancera**

School of Biomedical Sciences, CHIRI Biosciences, Curtin University, GPO Box U1987, Perth WA 6845, Australia

Keywords: glycosaminoglycans/interaction analysis/molecular docking/site mapping/virtual screening

## Abstract

Glycosaminoglycans (GAGs) are complex polysaccharides involved in a wide range of biological signalling events, as well as being important as biological structural materials. Despite the ubiquity and importance of GAG-protein interactions in biological systems and potentially as therapeutic targets, detailed structures of such interactions are sparse in availability. Computational methods can provide detailed structural knowledge of these interactions; however, they should be evaluated against suitable test systems prior to their widespread use. In this study, we have investigated the application of automated molecular docking and interaction mapping techniques to characterizing GAG-protein interactions. A series of high-resolution X-ray crystal structures of GAGs in complex with proteins was used to evaluate the approaches. Accurately scoring the pose fitting best with the crystal structure was a challenge for all docking programs evaluated. The site mapping technique offered excellent prediction of the key residues involved in ligand recognition, comparable to the best pose and improved over the top ranked pose. A design protocol incorporating site- and ligand-based mapping techniques was developed and applied to identify GAGs capable of binding to acidic fibroblast growth factor (aFGF). The protocol was able identify ligands known to bind to aFGF and accurately able to predict the binding modes of those ligands when using a known ligand-binding conformation of the protein. This study demonstrates the value of mapping-based techniques in identifying specific GAG epitopes recognized by proteins and for GAG-based drug design.

## Introduction

Glycosaminoglycans (GAGs) are complex, linear polysaccharides that are generally sulfated (with the exception of hyaluronan) and involved in a wide range of biological processes (Gandhi, N.S. and Mancera, R.L. 2008). GAGs such as hyaluronan, chondroitin sulfate, dermatan sulfate and keratan sulfate have well-established roles as biological structural materials, forming key components of cartilage, synovial fluid, and corneal tissue. There is, however, emerging evidence of their role in the manifestation of cancer and brain and spinal cord recovery after injury. Although typically associated with blood coagulation, heparin and heparan sulfate have been shown to play roles in development and differentiation, cancer metastasis and inflammation. In the case of GAG binding to antithrombin, only the recognition of a very specific sulfation pattern results in inhibition of the coagulation cascade (Jin, L., et al. 1997). Thus, there exists great potential to develop molecules that mimic the biological activity of GAGs for the development of biological nanomaterials, novel anticancer agents and novel anti-inflammatory agents (Gandhi, N.S. and Mancera, R.L. 2010b).

As is the case for other types of carbohydrate-protein complexes (Agostino, M., et al. 2012, Agostino, M., et al. 2009b, Agostino, M., et al. 2011b), limited high quality structural information is available for GAG-protein complexes. A key factor contributing to the lack of high quality structures is the high flexibility of carbohydrates. GAGs are particularly flexible compared to other types of carbohydrates (Jin, L., et al. 2005, Pogány, P. and Kovács, A. 2009, Silipo, A., et al. 2008), such as Lewis antigens (Yuriev, E., et al. 2005) and pig xenoantigens (Yuriev, E., et al. 2009). Furthermore, unlike most other carbohydrate residues of biological significance, iduronic acid (IdoA) and its 2-*O*-sulfated form (IdoA(2S)), both frequently found in GAGs, can readily interconvert between chair ($^1C_4$) and skew-boat ($^2S_0$) conformations (Gandhi, N.S. and Mancera, R.L. 2010a). This is due to both low energy barriers and similar energies between these states (Sattelle, B.M., et al. 2010). The most likely conformation in solution appears to vary depending on the sulfation pattern of the residue and the position of the residue within a GAG chain (Ferro, D.R., et al. 1990). Due to the flexibility of the IdoA ring, different proteins

may recognize specific conformations of IdoA, which in turn, affects the understanding of GAG-protein recognition, both in experimental and computational structural studies (Coombe, D.R., et al. 2008).

In addition to the difficulties posed to X-ray diffraction methods by the flexibility of GAGs, GAG-protein complexes pose some additional challenges. GAGs are notoriously difficult to obtain at suitable levels of purity for crystallographic studies. Many biochemical studies use enzymatically cleaved products purified to a particular length of GAG, but rarely a defined composition (Ostrovksy, O., et al. 2002, Pye, D.A., et al. 2000, Taylor, K.R., et al. 2005). Heparinase cleavage results in the formation of an unsaturated $\Delta^4$-uronic acid (ΔUA/ΔUA(2S)) at the non-reducing end of the cleavage product, thus preventing the ability to distinguish between iduronic acid and glucuronic acid in the original GAG sequence. Chemical synthesis of GAGs has been achieved, but like other strategies to carbohydrate chemical synthesis, it involves a complex series of reactions utilizing specific protecting group strategies to achieve the desired product. Nonetheless, the chemical synthesis of fondaparinux, a heparin mimetic used to treat thrombosis, is routinely performed for industrial production.

Given the difficulties associated with experimentally studying GAG-protein interactions, numerous groups have used computational techniques, particularly molecular docking and molecular dynamics simulations, in conjunction with experimental techniques to elucidate the structural basis of GAG-protein recognition (Canales, A., et al. 2006, Nieto, L., et al. 2011, Pichert, A., et al. 2012, Sapay, N., et al. 2011). Several studies validating the use of particular docking approaches have been published over the last decade or so (Bitomsky, W. and Wade, R.C. 1999, Bytheway, I. and Cochran, S. 2004, Samsonov, S.A., et al. 2011, Takaoka, T., et al. 2007); however, limited attempts have been made to perform a comparison of a wide range of tools against a wide range of cases. Docking studies performed to date have generally placed little emphasis on structural validation of the docking method (Costa, M.G.S., et al. 2010, Mulloy, B. and Forster, M.J. 2008, Torrent, M., et al. 2011). AutoDock is a popular choice for performing molecular docking of GAGs and earlier versions of it have been shown to perform well in validation studies (Gandhi, N.S., et al. 2008, Gandhi, N.S., et al. 2012, Gandhi, N.S. and Mancera, R.L. 2011, Gandhi, N.S. and Mancera, R.L. 2012). However, in general, docking scoring

functions perform poorly at predicting carbohydrate-protein complexes, and further analysis is required to identify the most likely binding mode from the set of possible solutions (Agostino, M., et al. 2010).

The site mapping technique has previously been demonstrated in a wide range of protein-ligand recognition scenarios, particularly carbohydrate-protein recognition, to provide comparable or superior performance to the top ranked pose obtained from molecular docking in predicting the key protein residues involved in ligand recognition (Agostino, M., et al. 2012, Agostino, M., et al. 2011a, Agostino, M., et al. 2009b, Agostino, M., et al. 2011b). This technique utilizes the interactions taking place in a set of up to 100 ranked docking solutions, which are tallied according to both the protein residue with which they occurred and the type of interaction (hydrogen bonding or van der Waals interaction). For each type of interaction, the tallies are then converted to percentages (relative to the total number of interactions of the type) and sorted from most frequently to least frequently contacted. The cumulative sum of the percentages going down the lists is then computed and all residues occurring above given a cumulative sum cutoff are deemed important for recognition. Using an appropriate validation set of crystallographic complexes, the cumulative sum cutoffs can be optimized for a given type of protein-ligand system by computation of site maps across an exhaustive set of cutoffs to determine the combination that gives rise to best prediction of the contacts in the crystallographic complexes. The site mapping procedure and its corresponding optimization protocol have recently been released as the AutoMap package (Agostino, M., et al. 2013). The technique appears to give the best quality predictions in cases that are difficult to study by molecular docking alone, such as carbohydrate-lectin recognition (Agostino, M., et al. 2011b) and ganglioside-antibody recognition (Agostino, M., et al. 2012). In these respective cases, difficulties in applying molecular docking are due to the generally shallow nature of the protein binding sites (carbohydrate-lectin recognition) and the negative charge and high flexibility of the ligand (ganglioside-antibody recognition). As GAG-protein interactions generally feature flexible, anionic ligands binding to shallow protein sites, they embody both of these problems, and thus may be better studied using the site mapping technique rather than molecular docking alone. Furthermore,

ligand-based mapping techniques have also been developed and may offer some further insight into recognition in a ligand design context (Agostino, M., et al. 2010).

We report the first validation study using structural data of a computational approach using a combination of molecular docking simulations and mapping techniques. We first evaluate a range of molecular docking programs for their ability to reproduce the GAG binding mode in a series of high resolution X-ray crystal structures of GAG-protein complexes (Table I). We then demonstrate the application of the site mapping approach to identifying key protein residues involved in GAG recognition. Finally, we demonstrate for the first time the development and application of a ligand design strategy incorporating the mapping techniques (Figure 1), specifically demonstrated here for investigating the recognition of GAG ligands by acidic fibroblast growth factor (aFGF). This protein target was chosen as it is the only case to our knowledge for which a comprehensive library of GAG fragments of defined composition and sufficiently small size for docking has been evaluated experimentally, with both high affinity ligands and their complexes with the protein determined (Hu, Y.-P., et al. 2012). Although the design strategy has been demonstrated solely against aFGF, it is likely to be extendible to other types of carbohydrates and potentially non-carbohydrate ligands.

## Results

*Evaluation of cognate molecular docking*

In general, most docking programs had trouble in producing accurate GAG poses in the range of test cases (Figure 2, Table S1). For at least one test case, each program failed to identify any docking solutions, with the exception of FRED, which was able to identify docking solutions for every test case. Where reasonably accurate poses could be obtained, they were rarely highly ranked by the docking scoring functions of each program. Furthermore, no specific pattern to the successful cases could be drawn, unlike in previous carbohydrate-protein docking investigations, where trends relating docking success to the length of the carbohydrate, the types of linkages in the carbohydrate, and the binding site topography could clearly be established (Agostino, M., et al. 2009a, Agostino, M., et al. 2012, Agostino,

M., et al. 2011b). While it is likely that the size of the ligands affected the docking performance, there is no evidence within our test set to suggest that docking a small ligand (e.g., disaccharide) is any easier than docking a larger ligand (e.g., tetra/pentasaccharide). Instead, poor docking performance is more likely to be associated with binding site topography and the chemical functionality of GAGs and their target proteins. The target proteins of the test set generally feature large and flat binding sites, thus making it difficult to draw specific conclusions about the effect of binding site topography. However, previous studies on carbohydrate-protein docking indicate that the method usually proceeds better when docking to cavity-like binding sites, such as those in antibodies (Agostino, M., et al. 2009a) and poorly when docking to large extended binding sites, such as in lectins (Agostino, M., et al. 2011b). Previous studies also indicate that longer, more flexible ligands (such as those feature 1→6/2→8 linkages) are also more challenging to dock (Agostino, M., et al. 2012).

The interactions between sulfated GAGs and proteins are generally dominated by electrostatic interactions (including charge-assisted hydrogen bonds) between lysine, arginine or histidine residues and the sulfate groups of GAGs. However, due to the high degree of sulfation of GAGs, as well as the presence of numerous positively charged residues in the binding site of the protein, it is generally difficult to score the correct binding pose accurately. This is somewhat related to the issues previously observed in accurately ranking the binding modes of acidic carbohydrates when bound to antibodies (Agostino, M., et al. 2012), which also feature numerous positively charged residues in their binding sites. Essentially, the docking program can "misdirect" the placement of the entire ligand structure by determining that a specific pairing of negative and positively charged groups – other than the experimentally observed pairing – is a more favorable arrangement. The combined effect of a reasonably flat binding site containing many positively charged residues is the observation of docking results that are generally much worse than that observed for either carbohydrate-lectin docking or antibody recognition of acidic carbohydrates.

Despite the difficulties associated with molecular docking of GAGs to proteins, AutoDock, Glide and FITTED were reasonably consistent in producing at least one moderately accurate pose within their top

100 ranked poses for each of the test cases. Given the overall high level of success of GOLD in predicting carbohydrate-protein complexes in previous investigations (Agostino, M., et al. 2009a, Agostino, M., et al. 2012, Agostino, M., et al. 2011b), it was anticipated that it would be the best performer, but it is instead one of the worst performers, second only to DOCK. AutoDock was the best performer overall, a result somewhat at odds with our previous investigations into carbohydrate-protein recognition, but consistent with that previously reported for GAG-protein interactions (Samsonov, S.A., et al. 2011, Takaoka, T., et al. 2007). As previously noted, the difference in performance between GOLD and AutoDock is most likely due to the differences in the way electrostatics are treated by these programs; the AutoDock scoring function provides an explicit treatment of electrostatic interactions (Huey, R., et al. 2007), whereas the GOLD scoring function does not (Verdonk, M.L., et al. 2003).

These findings highlight the importance of carrying out evaluation studies to select the most appropriate docking program for a given molecular recognition investigation, and the variation in performance that can occur when seemingly subtle aspects of the investigated system are changed. A major caveat to the use of AutoDock is its inability to consider any more than 32 rotatable bonds during docking calculations, which excludes its application to carbohydrates larger than tetrasaccharides. Thus, for these cases, where AutoDock cannot be applied, alternatives must be sought. While FITTED does not appear to have limitations imposed on ligand size or the number of rotatable bonds, it is not capable of consistently providing docking poses for every test case (Table S1). Both FRED and Glide can dock the cases unable to be docked by AutoDock and perform reasonably comparably across the test cases. However, FRED is the most appropriate second choice for cases where AutoDock cannot be applied successfully, as it is the only program that provides results for every member of the test set. The major caveat to the use of FRED is that it does not perform conformational searching of the ligand during docking; the search must be performed by an external tool. This is problematic for two reasons. Firstly, ligands larger than tetrasaccharides will require very long conformational searches in order to sample a large enough conformational space, thus may not be represented adequately. Secondly, it is possible that different approaches for conformational searching could give rise to different conformational ensembles

(Musafia, B. and Senderowitz, H. 2010). Therefore, the consistency of the predicted binding poses with FRED is likely to be highly dependent on the approach to conformer generation used and the performance of FRED observed must be interpreted with this in mind.

It is important to note once again that the top ranked pose predicted by *any* program rarely afforded the best fit to the crystal structure. This highlights the need for alternative scoring strategies when investigating GAG-protein interactions using docking.

*Site mapping of glycosaminoglycan-protein interactions*

Since AutoDock was found to be the best performing program, it was used to provide poses as input for site mapping. In the cases that could not be evaluated using AutoDock, the docking poses predicted by FRED were used as input for site mapping. Initially, the hydrogen bonding and van der Waals cutoffs for site mapping were optimized using the evaluation systems. It was determined that the optimal cutoffs were 90% for hydrogen bonding interactions and 50% for van der Waals interactions (Figure 3, Table S2). This is similar to the cutoffs observed for the recognition of acidic carbohydrates by antibodies (Agostino, M., et al. 2012). This similarity is most likely a result of GAGs also featuring acidic functionality.

The use of site mapping to predict the key protein residues involved in recognition afforded a statistically significant improvement ($p < 0.05$ in two-tailed t-test) over considering the top ranked pose alone (Figure 4, Table S3). Furthermore, a slight improvement in the mean $F_1$ score of the site maps is observed when compared to that of the best poses. These findings suggest that considering information from multiple poses is required to describe GAG recognition by proteins accurately. Like the molecular docking predictions, it was difficult to determine any patterns that led to either the increased likelihood of success or failure of the technique. It appears that the length of the GAG fragment can influence the quality of prediction, but this does not appear to influence prediction quality in a consistent manner. The effect is most dramatic when comparing the two annexin A2 structures and the two basic fibroblast growth factor (bFGF) structures (Figures S1 and S2). For bFGF, the interactions made by the tetrasaccharide (PDB code 1BFB) could be specifically reproduced by the site map far more effectively

than the equivalently generated site map using the hexasaccharide (PDB code 1BFC). In the case of annexin A2, the effect of length is the reverse to that of bFGF: site mapping with the pentasaccharide (PDB code 2HYV) gave a superior result than site mapping with the tetrasaccharide (PDB code 2HYU).

*Evaluation of design and binding mode prediction strategy against aFGF*

The design and binding mode prediction strategy was found to perform quite well for predicting the disaccharides capable of binding to aFGF and the corresponding binding modes when applied to the aFGF complex structures (PDB codes 3UD8 and 3UD9). The initial energy-based selection identified six molecules from the possible twenty-four (Table II), with three of these molecules (GlcNS(3S)α(1→4)IdoA(2S), GlcNS(6S)α(1→4)IdoA(2S) and GlcNS(3S,6S)α(1→4)IdoA(2S)) known to bind to aFGF (Hu, Y.-P., et al. 2012). The six molecules yielded approximately 150 binding modes in docking to each of the complex structures. Almost half of these in each case had an $SF_1$ Score greater than 0.9. For the poses generated by docking to the 3UD8 structure, cluster analysis indicated four clusters with at least two members and an average $EF_1$ Score greater than 1.0 (Table S4). For the poses generated by docking to the 3UD9 structure, cluster analysis indicated seven clusters with at least two members and an average $EF_1$ Score greater than 1.0 (Table S5). Binding modes from all six ligands are still represented in the clusters derived from docking to the 3UD8 structure, while the clustering analysis has eliminated poses from two non-binding ligands when applied to the poses obtained from docking to the 3UD9 structure. In both cases, at least one cluster is scored significantly better than the other clusters by ClusScore and the top scoring cluster is the closest to the crystallographic binding mode (Figure 5). In the cluster selected from the 3UD8 data, one binding mode from each of the three known ligands identified by the initial energy-based selection is represented, while in the cluster selected from the 3UD9 data, only two of the three initially identified known ligands are represented. Nonetheless, this demonstrates that, given an appropriately induced protein structure, the design strategy can identify correct ligands for the correct reason.

In order to investigate whether the design strategy could be applied to a protein conformation not induced by a ligand, we then attempted to apply it to two conformations of the native structure of aFGF

(PDB code 1BAR, chains A and B). Prior to applying the design strategy to the native structures, the four structures were superposed to identify any changes in the overall protein conformation that may hinder the identification of accurate binding modes (Figure S3). Like most other GAG-binding proteins (Gandhi, N.S. and Mancera, R.L. 2008), aFGF features numerous arginine and lysine residues in the binding site, which are important for GAG recognition but also particularly flexible and difficult to accurately model. Superposing the four structures revealed that the conformation of Lys113 could significantly affect accurate ligand docking in chain A of PDB 1BAR, as the terminal amine is placed such that it overlaps with the bound GAG in the 3UD8 and 3UD9 structures. Other nearby residues that vary in their placement between the set of structures include Lys118, Arg122 and Lys128. As these residues point directly into the binding pocket, they may also affect ligand placement in docking.

The initial application of the design strategy to each of the native aFGF conformations yielded similarly scoring clusters (Table S6) and selected binding modes moderately distant from the complex structure. In applying the strategy to chain A of 1BAR, binding modes were selected that were dramatically different in overall ligand placement. However, the *N*-sulfate groups of the GlcNS residues of the selected binding modes were placed between the sulfate groups of the co-crystallized ligands (Figure 6a), indicating that the importance of that region of the receptor for sulfate binding can be correctly identified. In applying the strategy to chain B of 1BAR, binding modes are identified where the 6-*O*-sulfate of the GlcNS residues is placed to overlap with the 2-*O*-sulfate of the IdoA(2S) residues of the co-crystallized ligands (Figure 6b). This again highlights that the importance of that region of the receptor for sulfate binding can be correctly identified. However, the failure to identify the crystallographic binding mode is most likely due to differences between the protein conformations in the native and complexed states. Four binding modes in two clusters were obtained when the design strategy was applied to chain A of 1BAR, while ten binding modes in four clusters were obtained when the design strategy was applied to chain B of 1BAR; the side chain prediction procedure was applied to all of these complexes.

In applying the design strategy to the new conformers of each chain of 1BAR, clearly preferred clusters could be obtained, as indicated by the observation of top scoring clusters with significantly better ClusScores than the lower scoring clusters (Table S7). The binding modes obtained in each case are not identical to the crystallographic binding mode, but bear some striking resemblances (Figure 7). The best ranked cluster obtained featured binding modes of the ligands GlcNS(3S)α(1→4)IdoA(2S) and GlcNS(3S,6S)α(1→4)IdoA(2S) (Figure 7a), both which are known to bind to aFGF with high affinity (Hu, Y.-P., et al. 2012). The binding modes of these ligands feature the 3-*O*-sulfate groups positioned almost identically to that in the crystallographic binding mode. However, the *N*-sulfate groups have been positioned to interact with the side chain of Lys112 and the resampled Arg122, rather than becoming more buried in the protein to interact with the side chains of Lys118 and Asn18 as in the crystal structure complex. Nonetheless, the *N*-sulfate-binding region is still identified as important for sulfate binding and the 2-*O*-sulfate of the IdoA(2S) residues in the binding modes is placed in an identical position to the *N*-sulfate group in the crystallographic binding mode, permitting it to interact with the side chain of Lys118.

After applying the design strategy to the new conformers of chain B chain of 1BAR, the best ranked cluster featured binding modes only for GlcNS(3S,6S)α(1→4)IdoA(2S). The binding modes obtained were reversed compared to the crystallographic binding modes, but also "shifted" with respect to these (Figure 7b,c). The GlcNS residues of the determined binding modes fitted well with those of the crystallographic binding modes, however, the location of the *N*-sulfates and the 3-*O*-sulfates were reversed between the determined binding modes and the crystallographic binding modes. This again highlights the ability of the approach to identify the location of sulfate-binding regions on the aFGF site correctly.

## Discussion

Although there are some differences in the methodology employed in previous studies validating GAG docking compared to our study, our findings agree with previous observations that AutoDock performs well for docking GAGs to proteins, at least in terms of producing the correct binding poses. It has been previously noted that this improvement is most likely due to the explicit treatment of electrostatics in the scoring function of AutoDock, as well as improved ligand sampling by the Lamarckian genetic algorithm (Takaoka, T., et al. 2007). A general issue with the use of AutoDock to perform molecular docking and site mapping is that one is limited to studying GAG fragments as long as tetrasaccharides, due to the rotatable bond limit of the program. However, many GAG-recognizing proteins utilize large surfaces to recognize much larger fragments (Gandhi, N.S. and Mancera, R.L. 2008). The recently developed Computational Carbohydrate Grafting technique (CCG) (Tessier, M.B., et al. 2012) excellently complements our approach presented in this paper: our approach allows the accurate placement of at least a key interacting portion of a carbohydrate, while CCG can be used to grow the remainder of the carbohydrate. This could be a particularly useful strategy in identifying some of the more unusual potential binding modes of large GAG chains (Gandhi, N.S. and Mancera, R.L. 2011, Johnson, D.J., et al. 2006, Krieger, E., et al. 2004).

In optimizing the site mapping technique to analyze GAG-protein recognition, the hydrogen bonding and van der Waals cutoffs obtained were found to be similar to those obtained for ganglioside-antibody binding (Agostino, M., et al. 2012). This is most likely due to the functional group similarities between gangliosides and GAGs, the predominant one being the presence of negatively charged groups: carboxylates in the case of gangliosides and both carboxylates and sulfates in the case of GAGs. The density of sulfates and indeed negative charge on GAGs is unrivalled by any other natural molecule (Gandhi, N.S. and Mancera, R.L. 2008). While GAGs do not feature residues that contain prominent hydrophobic faces, unlike blood group-related carbohydrates (e.g., galactose, fucose), molecular mechanics-based calculations of the free energies of binding suggest that both van der Waals and

hydrogen bonding/electrostatic interactions make energetic contributions of a similar order of magnitude in GAG-protein interactions (Gandhi, N.S. and Mancera, R.L. 2009). Detailed examination of the epitope mapping data may reveal the precise mechanism of van der Waals interactions in GAG-protein interactions.

Although the GAG binding modes determined against native aFGF incorporating flexibility into the model do not match directly with the crystallographic complexes, the degree of similarity among these binding modes is comparable to the similarity between carbohydrates modelled in dual conformations in several crystallographic complexes. In the complex of rat mannose protein A with Manα(1→3)Man (PDB 1KX0) (Ng, K.K.-S., et al. 2002), the same atoms of the terminal α-mannose contact the calcium in the binding site in both conformations, however, the orientation of these around the calcium is reversed from one conformation to the other. The rings of the terminal α-mannose in both cases overlay well with one another, however, the difference in its placement results in a very large shift in the placement of the second mannose. Similar binding modes of α-mannose have been observed in complexes with both langerin (PDB 3P5F) (Feinberg, H., et al. 2011) and DC-SIGN (PDBs 2IT5 and 2IT6) (Feinberg, H., et al. 2007). The difference between the two observed binding modes in these cases is similar to that observed here between the crystallographic binding modes of the GAG disaccharides in complex with aFGF and the binding modes predicted by the application of the mapping-based strategy, specifically, that shown in Figure 7b (and to an extent, that shown in Figure 7c). In the complexes of *Talaromyces emersonii* Cel7A with cellobiose (PDB 3PFX) and cellotetraose (PDB 3PFZ), a cellobiose molecule can be observed in both cases to be modelled in two different conformations, each one shifted from the other within the enzyme by the length of approximately half a residue. The differences between these binding modes are similar to those between the predicted and crystallographic binding modes shown in Figure 7a, which are shifted to a similar degree with respect to one another. Thus, the types of differences between the predicted and experimental structures observed here for aFGF have clear experimental analogues, and therefore could represent potential alternative binding modes for these carbohydrates with aFGF. Certainly, GAG binding to aFGF can occur in multiple ways, with

alternative conformations being observed in separate X-ray crystallography and NMR experiments (Canales, A., et al. 2006, DiGabriele, A.D., et al. 1998), although despite the variation in overall binding modes, sulfate groups are consistently placed, as we have generally observed here.

A key limitation of the approaches presented here is that no attempt to assign the ring conformation of iduronic acid is made. The choice of ring conformation can have a dramatic impact on docking accuracy (Samsonov, S.A. and Pisabarro, M.T. 2013). Iduronic acid typically exists as either a $^{1}C_{4}$ chair or a $^{2}S_{0}$ skew-boat, generally with a significant bias towards the chair state (Gandhi, N.S. and Mancera, R.L. 2010a). In our docking and site mapping validations, we retained the ring conformations of the original ligand, as they were modeled in the crystal structure complexes. In the validation of the design strategy, iduronic acid was assumed to be in the chair conformation. While this assumption was valid for aFGF, it is not a strictly valid assumption to make for all GAG-binding proteins. It is currently unclear as to the best approach to identifying ring conformations in iduronic acid; an adaptation of the dynamic site mapping approach to assessing multiple ring conformations may be an appropriate strategy to consider (Agostino, M., et al. 2012). However, from this study it is clear that if the binding ring conformations are known, as well as the correct protein conformation, the developed design and binding mode prediction strategy gives very accurate results, even in the absence of correct ranking by the docking scoring function.

A second limitation of the approaches presented here is that knowledge of the approximate location of the GAG binding site is required; a global scan of the protein structure to identify prospective sites, such as that performed by FTMap (Brenke, R., et al. 2009) or Schrödinger's SiteMap (Halgren, T.A. 2009), is not performed in our current protocol. Where possible, experimental knowledge should be used to guide the identification of binding sites and subsequent application of the approaches presented here. However, caution should be exercised in interpreting the results of experimental studies. For instance, in the cases of RANTES (Shaw, J.P., et al. 2004) and SDF-1α (Murphy, J.W., et al. 2007), X-ray crystallography, NMR and biochemical studies provide conflicting evidence for the placement of GAG binding sites on these proteins. Since the conditions for obtaining the X-ray structures of RANTES and

SDF-1α deviate significantly from physiological conditions, there is the potential that GAG-binding sites identified on these proteins are not biologically relevant, and thus utilizing these sites for structure-based design may result in erroneous conclusions. Incorporating an appropriately evaluated method for performing a global binding site scan into the protocol may help to address issues in locating the GAG binding site caused by conflicting experimental evidence.

## Materials and methods

*Selection of evaluation systems*

GAG-protein complexes for use as evaluation systems were selected by searching the Protein Data Bank for structures with ligands containing sulfated glucosamine and iduronic acid residues. The structures selected contained heparin fragments, between two and six residues in length, in complex with non-enzymatic proteins at moderate to high resolution (≤ 2.5 Å) (Table I). Throughout the entire manuscript, all carbohydrates are linked via α(1→4) linkages, unless otherwise specified.

*Evaluation of cognate molecular docking*

Glide 5.7 (Friesner, R.A., et al. 2004, Schrödinger, LLC 2011), AutoDock 4.2 (Huey, R., et al. 2007), DOCK 6.5 (Lang, P.T., et al. 2009), FRED 2.2.5 (McGann, M. 2011), FITTED 3.5 (Corbeil, C.R., et al. 2007) and GOLD 5.1 (Verdonk, M.L., et al. 2003) were each evaluated for their ability to reproduce the crystallographic binding mode of the evaluation systems. Ligands were treated with full flexibility, except for the carbohydrate rings, that were maintained in the conformation observed in the crystal structure in each case. No protein flexibility was considered. The detailed settings used for Glide, AutoDock, DOCK and GOLD were identical to those employed in previous investigations (Agostino, M., et al. 2012, Agostino, M., et al. 2011b), with some minor adjustments to the settings used for AutoDock and GOLD. Protein and ligand structure preparation was also identical to previous investigations. The Lamarckian genetic algorithm was employed for binding mode generation in AutoDock (Morris, G.M., et al. 1998). The option to retain diverse solutions in GOLD was disabled, due to a significant increase in computation time in applying the algorithm in the current version of GOLD. Instead of generating diverse solutions during the docking, the resulting poses were clustered

using the cluster script in the Silico toolkit (Chalmers, D.K. and Roberts, B.P. 2011). For FITTED 3.5, the structures prepared for use with DOCK were used as input. The default settings were used to carry out the setup and docking, with the exception of the number of genetic algorithm runs, which was set to 200 to be equivalent to that used for AutoDock and GOLD. As the FRED setup was reasonably more complex, it is described in the following section.

For all programs, the root-mean-square deviation (RMSD) to the crystallographic binding mode was used to evaluate the success of molecular docking. Poses with an RMSD of 2.5 Å or less were deemed highly accurate, while those between 2.5-5.0 Å were deemed moderately accurate. RMSDs are reported for the top-ranked pose according to the scoring function of the respective programs (referred to as the "top pose"), as well as for the pose within the set of results with the lowest RMSD with respect to the crystallographic pose (referred to as the "best pose").

*Setup of FRED docking runs*

Although FRED utilizes a rigid docking approach, conformational flexibility of the ligand can be considered during docking by providing multiple ligand conformations. Conformational searches in Macromodel 9.9 were performed for each ligand. The Monte Carlo Multiple Minimum (MCMM) method was employed in all searches. Automatic setup of the ligands was initially performed with manual adjustment of the following search: torsion rotations in rings were not considered, and all ring closure and torsion check parameters were removed. These settings were selected to maintain the original ring conformations observed in the respective crystal structure. The searches used extended torsional sampling. Mirror images were not retained to preserve the stereochemistry of the GAG fragments. The searches were set to conclude after completing 2000 steps per rotatable bond, or a maximum of 100000 steps, whichever was reached first. An energy window of 50 kJ/mol was used for saving structures. The Polak-Ribiere conjugate gradient algorithm (PRCG) was used to perform energy minimization on the conformers. Minimization terminated after 1000 cycles or upon reaching a gradient of 0.1 kJ/(mol Å). Redundant conformers were eliminated using the Redundant Conformer Elimination tool in Macromodel, using a 2.0 Å cutoff applied to heavy atoms.

The FRED Receptor GUI was used to set up the protein for docking. The co-crystallized ligand was used to guide the box size and placement. A high quality site shape potential was generated within the box. The outer contour was adjusted such that the entire ligand could be accommodated within it; the inner contour was adjusted according to the ratio between the volumes of the initially determined outer and inner contours. The orientation of the site shape was not optimized to preserve the adjustments made to the contours and to ensure that the protein coordinates remained constant for subsequent RMSD calculations performed on the ligand. Constraints were not used. A trial docking was performed with the bound ligand to ensure that the site was set up correctly and could accommodate the ligand.

For the production run, the Chemgauss3 scoring function was used (McGann, M.R., et al. 2003), with further optimization by Chemgauss3. Two hundred alternate binding poses were retained in each docking run.

*Site mapping of GAG-protein interactions*

The AutoMap procedure was used to perform site mapping for each GAG-protein system, the full details of which are described elsewhere (Agostino, M., et al. 2013). As in previous studies (Agostino, M., et al. 2012, Agostino, M., et al. 2011b), the program which gave rise to the most accurate poses overall, regardless of their ranking, was used to provide input for site mapping. MLP was used to determine the interactions between each ligand pose and the target protein. The collected interactions, as well as the list of known interacting residues obtained from the crystal structures, were passed to OPTCUTOFF to determine optimal hydrogen bonding and van der Waals cutoffs for site mapping of GAG-protein interactions. SITEMAP was then used to determine the site maps and identify the key interacting residues using the optimized cutoffs identified.

The ability of the site maps and given binding modes to accurately predict residues involved in recognition is assessed using the $F_1$ score, which is defined by the following expression:

$$F_1 = 2 \times \frac{r \times p}{r + p}$$

where $r$ is the recall and $p$ is the precision. Recall and precision are computed according to the following expressions:

$$r = \frac{TP}{TP + FN}$$

$$p = \frac{TP}{TP + FP}$$

where $TP$ is the number of residues involved in interactions in the crystallographic complex, $FN$ is the number of true positive residues failed to be identified as contacts and $FP$ is the number of residues falsely identified as contacts. An $F_1$ Score of 1 indicates precise recall of the residues involved in interactions in the crystallographic complex.

*Development of a computational design strategy for GAG epitope and binding mode prediction*

A computational design strategy for the prediction of GAG binding epitopes and their corresponding binding modes was devised, incorporating the site mapping strategies (Figure 2). The strategy incorporates energy-based selection of binding epitopes, followed by mapping-based analysis of the binding modes of the selected epitopes, clustering of the binding modes and an energy-based selection from the binding mode clusters.

The docked binding modes of each ligand are analyzed to determine the median value of the mean binding energy of highly populated binding mode clusters for that ligand; in this study, a cluster is considered to be highly populated if it contains five or more members. The ligands are then ranked according to this value and the top 25% of ranked ligands are selected for further analysis. Site mapping analysis is carried out on the selected ligands. A series of site maps is generated – one for each ligand, generated by the poses of that ligand – using the optimized hydrogen bonding and van der Waals cutoffs. The degree of fit of each ligand pose to the corresponding site map is then computed; this quantity is termed the *$SF_1$ Score* and is computed in an identical fashion to the $F_1$ score. However, in

this case, the true positives are the residues identified from site mapping, rather than a crystallographic complex, and the false positives/negatives defined based on the definition of the true positives. Poses with an $SF_1$ Score of greater than 0.9 are selected for cluster analysis, thus selecting poses that correlate highly with the site map.

Prior to performing cluster analysis on the selected poses, epitope mapping analysis is carried out on the selected ligands, using a revised procedure to that previously published (Agostino, M., et al. 2010). In this case, the contribution of a binding mode to the epitope hydrogen bonding and van der Waals counts is weighted according to the $SF_1$ score for that binding mode. For instance, if a binding mode has an $SF_1$ score of 0.67, then each contact made by an epitope atom will count as 0.67 to the interaction tally for that epitope atom. Thus, poses that better fit the site maps carry more weight when tallying the contributions of interacting epitope atoms. A 50% cutoff is used to identify the important epitope points involved in both hydrogen bonding and van der Waals interactions. Epitope scores are produced, specifically, an *$hbEF_1$ Score* and an *$nbEF_1$ Score*, calculated in an identical fashion to the $SF_1$ score, using the hydrogen bonding epitope maps and the van der Waals epitope maps respectively to provide the definition of true and false positives. These scores are summed to give the final *$EF_1$ Score*, which can be a maximum of 2 (as it is the sum of two $F_1$ scores).

Cluster analysis is performed on the poses selected by the $SF_1$ Score filter. The clusters are determined by comparing the coordinates of the carbohydrate rings of each of the selected poses. Poses with carbohydrate rings within an RMSD of 1.5 Å of one another are deemed to belong to the same cluster. Clusters containing a single pose are discarded. The average $EF_1$ score for the cluster is computed and clusters with an average $EF_1$ below 1.0 are also discarded. The mean energy of the cluster is determined, as well as the standard deviation of the cluster energy. The final score reported and used to rank the clusters (*ClusScore*) is the negative of the mean energy of the cluster divided by the standard deviation of the cluster energy:

$$ClusScore = \frac{-\mu_E}{\sigma_E}$$

A lower limit on the standard deviation of 0.125 is applied to prevent two-membered clusters with members very close in energy from being scored overly favorably.

*Evaluation of design and binding mode prediction strategy with aFGF*

The design and binding mode prediction strategy was evaluated by adapting a recent experimental study whereby a comprehensive library of GAG disaccharides was prepared and tested against aFGF (Hu, Y.-P., et al. 2012). A GAG library exclusively featuring biosynthetically accessible disaccharides was used as the ligand library for screening (Raghuraman, A., et al. 2006). As in the experimental studies on aFGF, only disaccharides with glucosamine derivatives at the non-reducing end were screened. Iduronic acid residues were exclusively examined in the $^{1}C_{4}$ conformation. Four structures of aFGF were used to evaluate the approach. These were its complexes with GlcNS(3S)α(1→4)IdoA(2S)α(1-OMe) (PDB code 3UD8) and GlcNS(3S,6S)α(1→4)IdoA(2S)α(1-OMe) (PDB code 3UD9) (Hu, Y.-P., et al. 2012), as well as the native structure, which features two molecules of aFGF in slightly differing conformations (PDB code 1BAR) (Zhu, X., et al. 1991). The aFGF structures were prepared using the Protein Preparation Wizard in Maestro, with missing side-chains modeled using Prime Side Chain Prediction. Docking was carried out using AutoDock 4.2, as this was the program identified as performing the best for docking GAGs (see Results). In order to speed up the docking calculations, multilevel parallel AutoDock 4.2 (mpAD4) was employed for screening the library (Norgan, A.P., et al. 2011), using the same settings for each docking run as the cognate docking evaluation.

*Incorporation of protein flexibility into design and binding mode prediction strategy*

An adaptation of the Prime Refinement step of Schrödinger's Induced Fit Protocol was applied to introduce protein flexibility (Sherman, W., et al. 2006). All of the complexes selected by applying the design strategy to the native aFGF structures were subject to Prime Side Chain Prediction in the presence of the ligand. Protein residues within 5.0 Å of the ligand were selected for the refinement procedure. The disaccharide library was then docked to each of the newly generated protein conformers. The steps for selecting the ligands and generating $SF_1$ and $EF_1$ scores were carried out individually for

each protein conformer, while the final cluster analysis to select the poses performed on the collected set of results.

## Acknowledgements

This work was supported by an NHMRC Early Career Fellowship (GNT1054245) to MA and a Curtin Early Career Research Fellowship to NSG.

## References

Agostino M, Jene C, Boyle T, Ramsland PA, Yuriev E. 2009a. Molecular docking of carbohydrate ligands to antibodies: structural validation against crystal structures. *J. Chem. Inf. Model.*, 49:2749-2760.

Agostino M, Mancera RL, Ramsland PA, Yuriev E. 2013. AutoMap: a tool for analyzing protein-ligand interactions using multiple ligand binding modes. *J. Mol. Graph. Model.*, 40:80-90.

Agostino M, Ramsland PA, Yuriev E. 2012. Antibody recognition of cancer-related gangliosides and their mimics investigated using *in silico* site mapping. *PLoS One*, 7:e35457.

Agostino M, Sandrin MS, Thompson PE, Ramsland PA, Yuriev E. 2011a. Peptide inhibitors of xenoreactive antibodies mimic the interaction profile of the native carbohydrate antigens. *Biopolymers*, 96:193-206.

Agostino M, Sandrin MS, Thompson PE, Yuriev E, Ramsland PA. 2009b. In silico analysis of antibody-carbohydrate interactions and its application to xenoreactive antibodies. *Mol. Immunol.*, 47:233-246.

Agostino M, Sandrin MS, Thompson PE, Yuriev E, Ramsland PA. 2010. Identification of preferred carbohydrate binding modes in xenoreactive antibodies by combining conformational filters and binding site maps. *Glycobiology*, 20:724-735.

Agostino M, Yuriev E, Ramsland PA. 2011b. A computational approach for exploring carbohydrate recognition by lectins in innate immunity. *Front. Immunol.*, 2:23.

Bitomsky W, Wade RC. 1999. Docking of glycosaminoglycans to heparin-binding proteins: validation for aFGF, bFGF, and antithrombin and application to IL-8. *J. Am. Chem. Soc.*, 121:3004-3013.

Brenke R, Kozakov D, Chuang G-Y, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. 2009. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics*, 25:621-627.

Bytheway I, Cochran S. 2004. Validation of molecular docking calculations involving FGF-1 and FGF-2. *J. Med. Chem.*, 47:1683-1693.

Canales A, Lozano R, Lopez-Mendez B, Angulo J, Ojeda R, Nieto PM, Martin-Lomas M, Gimenez-Gallego G, Jimenez-Barbero J. 2006. Solution NMR structure of a human FGF-1 monomer, activated by a hexasaccharide heparin-analogue. *FEBS J.*, 273:4716-4727.

Chalmers DK, Roberts BP. 2011. Silico - a Perl molecular modelling toolkit.

Coombe DR, Stevenson SM, Kinnear BF, Gandhi NS, Mancera RL, Osmond RI, Kett WC. 2008. Platelet endothelial cell adhesion molecule (PECAM-1) and its interactions with glycosaminoglycans. 2. Biochemical analyses. *Biochemistry*, 47:4863-4875.

Corbeil CR, Englebienne P, Moitessier N. 2007. Docking ligands into flexible and solvated macromolecules. 1. Development and validation of FITTED 1.0. *J. Chem. Inf. Model.*, 2007:435.

Costa MGS, Batista PR, Shida CS, Robert CH, Bisch PM, Pascutti PG. 2010. How does heparin prevent the pH inactivation of cathepsin B? Allosteric mechanism elucidated by docking and molecular dynamics. *BMC Genomics*, 11.

DiGabriele AD, Lax I, Chen DI, Svahn CM, Jaye M, Schlessinger J, Hendrickson WA. 1998. Structure of a heparin-linked biologically active dimer of fibroblast growth factor. *Nature*, 393:812-817.

Feinberg H, Castelli R, Drickamer K, Seeberger PH, Weis WI. 2007. Multiple modes of binding enhance the affinity of DC-SIGN for high mannose *N*-linked glycans found on viral glycoproteins. *J. Biol. Chem.*, 282:4202-4209.

Feinberg H, Taylor ME, Razi H, McBride R, Knirel YA, Graham SA, Drickamer K, Weis WI. 2011. Structural basis for langerin recognition of diverse pathogen and mammalian glycans through a single binding site. *J. Mol. Biol.*, 405:1027-1039.

Ferro DR, Provasoli A, Ragazzi M, Casu B, Torri G, Bossennec V, Perly B, Sinaÿ P, Petitou M, Choay J. 1990. Conformer populations of L-iduronic acid residues in glycosaminoglycan sequences. *Carbohydr. Res.*, 195:157-167.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shaw DE, Shelley M*, et al.* 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, 47:1739-1749.

Gandhi NS, Coombe DR, Mancera RL. 2008. Platelet endothelial cell adhesion molecule 1 (PECAM-1) and its interactions with glycosarninoglycans: 1. Molecular modeling studies. *Biochemistry*, 47:4851-4862.

Gandhi NS, Freeman C, Parish CR, Mancera RL. 2012. Computational analyses of the catalytic and heparin-binding sites and their interactions with glycosaminoglycans in glycoside hydrolase family 79 endo-beta-d-glucuronidase (heparanase). *Glycobiology*, 22:35-55.

Gandhi NS, Mancera RL. 2008. The structure of glycosaminoglycans and their interactions with proteins. *Chem. Biol. Drug. Des.*, 72:455-482.

Gandhi NS, Mancera RL. 2009. Free energy calculations of glycosaminoglycan-protein interactions. *Glycobiology*, 19:1103-1115.

Gandhi NS, Mancera RL. 2010a. Can current force fields reproduce ring puckering in 2-*O*-sulfo-α-L-iduronic acid? A molecular dynamics simulation study. *Carbohydr. Res.*, 345:689-695.

Gandhi NS, Mancera RL. 2010b. Heparin/heparan sulphate-based drugs. *Drug. Discov. Today*, 15:1058-1069.

Gandhi NS, Mancera RL. 2011. Molecular dynamics simulations of CXCL-8 and its interactions with a receptor peptide, heparin fragments, and sulfated linked cyclitols. *J. Chem. Inf. Model.*, 51:335-358.

Gandhi NS, Mancera RL. 2012. Prediction of heparin binding sites in bone morphogenetic proteins (BMPs). *BBA - Proteins Proteomics*, 51:1374-1381.

Halgren TA. 2009. Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, 49:377-389.

Hu Y-P, Zhong Y-Q, Chen Z-G, Chen C-Y, Shi Z, Zulueta MML, Ku C-C, Lee P-Y, Wang C-C, Hung S-C. 2012. Divergent synthesis of 48 heparan sulfate-based disaccharides and probing the specific sugar-fibroblast growth factor-1 interaction. *J. Am. Chem. Soc.*, 134:20722-20727.

Huey R, Morris GM, OIson AJ, Goodsell DS. 2007. A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.*, 28:1145-1152.

Jin L, Abrahams JP, Skinner R, Petitou M, Pike RN, Carrell RW. 1997. The anticoagulant activation of antithrombin by heparin. *Proc. Natl. Acad. Sci. U.S.A.*, 94:14683-14688.

Jin L, Barran PE, Deakin JA, Lyon M, Uhrín D. 2005. Conformation of glycosaminoglycans by ion mobility mass spectrometry and molecular modelling. *Phys. Chem. Chem. Phys.*, 7:3464-3471.

Johnson DJ, Langdown J, Li W, Luis SA, Baglin TP, Huntington JA. 2006. Crystal structure of monomeric native antithrombin reveals a novel reactive center loop conformation. *J. Biol. Chem.*, 281:35478-35486.

Krieger E, Geretti E, Brandner B, Goger B, Wells TN, Kungl AJ. 2004. A structural and dynamic model for the interaction of interleukin-8 and glycosaminoglycans: Support from isothermal fluorescence titrations. *Proteins*, 54:768-775.

Lang PT, Brozell SR, Mukherjee S, Pettersen ET, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID. 2009. DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA*, 15:1219-1230.

McGann M. 2011. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, 51:578-596.

McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK. 2003. Gaussian docking functions. *Biopolymers*, 68:76-90.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. 1998. Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J. Comput. Chem.*, 19:1639-1662.

Mulloy B, Forster MJ. 2008. Application of drug discovery software to the identification of heparin-binding sites on protein surfaces: a computational survey of the 4-helix cytokines. *Mol. Simul.*, 34:481-489.

Murphy JW, Cho Y, Sachpatzidis A, Fan C, Hodsdon ME, Lolis E. 2007. Structural and functional basis of CXCL12 (stromal cell-derived factor-1 alpha) binding to heparin. *J. Biol. Chem.*, 282:10018-10027.

Musafia B, Senderowitz H. 2010. Biasing conformational ensembles towards bioactive-like conformers for ligand-based drug design. *Expert Opin. Drug Discov.*, 5:943-959.

Ng KK-S, Kolatkar AR, Park-Snyder S, Feinberg H, Clark DA, Drickamer K, Weis WI. 2002. Orientation of bound ligands in mannose-binding proteins: implications for multivalent ligand recognition. *J. Biol. Chem.*, 277:16088-16095.

Nieto L, Canales A, Gimenez-Gallego G, Nieto PM, Jimenez-Barbero J. 2011. Conformational Selection of the AGA*IA(M) Heparin Pentasaccharide when Bound to the Fibroblast Growth Factor Receptor. *Chem.-Eur. J.*, 17:11204-11209.

Norgan AP, Coffman PK, Kocher J-PA, Katzmann DJ, Sosa CP. 2011. Multilevel parallelization of AutoDock 4.2. *J. Cheminform.*, 3:12.

Ostrovksy O, Berman B, Gallagher J, Mulloy B, Fernig DG, Delehedde M, Ron D. 2002. Differential effects of heparin saccharides on the formation of specific fibroblast growth factor (FGF) and FGF receptor complexes. *J. Biol. Chem.*, 277:2444-2453.

Pichert A, Samsonov SA, Theisgen S, Thomas L, Baumann L, Schiller J, Beck-Sickinger AG, Huster D, Pisabarro MT. 2012. Characterization of the interaction of interleukin-8 with hyaluronan, chondroitin sulfate, dermatan sulfate and their sulfated derivatives by spectroscopy and molecular modeling. *Glycobiology*, 22:134-145.

Pogány P, Kovács A. 2009. Conformational properties of the disaccharide building units of hyaluronan. *Carbohydr. Res.*, 344:1745-1752.

Pye DA, Vivès RR, Hyde P, Gallagher JT. 2000. Regulation of FGF-1 mitogenic activity by heparan sulfate oligosaccharides is dependent on specific structural features: differential requirements for the modulation of FGF-1 and FGF-2. *Glycobiology*, 10:1183-1192.

Raghuraman A, Mosier PD, Desai UR. 2006. Finding a needle in a haystack: development of a combinatorial virtual screening approach for identifying high specificity heparin/heparan sulfate sequence(s). *J Med Chem*, 49:3553-3562.

Samsonov SA, Pisabarro MT. 2013. Importance of IdoA and IdoA(2S) ring conformations in computational studies of glycosaminoglycan-protein interactions. *Carbohyd. Res.*, 381:133-137.

Samsonov SA, Teyra J, Pisabarro MT. 2011. Docking glycosaminoglycans to proteins: analysis of solvent inclusion. *J. Comput.-Aided Mol. Des.*, 25:477-489.

Sapay N, Cabannes E, Petitou M, Imberty A. 2011. Molecular modeling of the interaction between heparan sulfate and cellular growth factors: bringing pieces together. *Glycobiology*, 21:1181-1193.

Sattelle BM, Hansen SU, Gardiner J, Almond A. 2010. Free energy landscapes of iduronic acid and related monosaccharides. *J. Am. Chem. Soc.*, 132:13132-13134.

Schrödinger, LLC. 2011. Glide, version 5.7. New York, NY.

Shaw JP, Johnson Z, Borlat F, Zwahlen C, Kungl A, Roulin K, Harrenga A, Wells TN, Proudfoot AE. 2004. The X-ray structure of RANTES: heparin-derived disaccharides allows the rational design of chemokine inhibitors. *Structure*, 12:2081-2093.

Sherman W, Day T, Jacobson MP, Freisner RA, Farid R. 2006. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.*, 49:534-553.

Silipo A, Zhang Z, Cañada FJ, Molinaro A, Linhardt RJ, Jimenez-Barbero J. 2008. Conformational analysis of a dermatan sulfate-derived tetrasaccharide by NMR, molecular modeling, and residual dipolar couplings. *ChemBioChem*, 9:240-252.

Takaoka T, Mori K, Okimoto N, Neya S, Hoshino T. 2007. Prediction of the structure of complexes comprised of proteins and glycosaminoglycans using docking simulation and cluster analysis. *J. Chem. Theory Comput.*, 3:2347-2356.

Taylor KR, Rudisill JA, Gallo RL. 2005. Structural and sequence motifs in dermatan sulfate promoting fibroblast growth factor-2 (FGF-2) and FGF-7 activity. *J. Biol. Chem.*, 280:5300-5306.

Tessier MB, Grant OC, Heimburg-Molinaro J, Smith D, Jadey S, Gulick AM, Glushka J, Deutscher SL, Rittenhouse-Olson K, Woods RJ. 2012. Computational screening of the human TF-glycome provides a structural definition for the specificity of anti-tumor antibody JAA-F11. *PLoS One*, 8:e54874.

Torrent M, Victoria Nogues M, Boix E. 2011. Eosinophil cationic protein (ECP) can bind heparin and other glycosaminoglycans through its RNase active site. *J. Mol. Recognit.*, 24:90-100.

Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. 2003. Improved protein-ligand docking using GOLD. *Proteins*, 52:609-623.

Yuriev E, Agostino M, Farrugia W, Christiansen D, Sandrin MS, Ramsland PA. 2009. Structural biology of carbohydrate xenoantigens. *Expert Opin. Biol. Ther.*, 9:1017-1029.

Yuriev E, Farrugia W, Scott AM, Ramsland PA. 2005. Three-dimensional structures of carbohydrate determinants of Lewis system antigens: implications for effective antibody targeting of cancer. *Immunol. Cell. Biol.*, 83:709-717.

Zhu X, Komiya H, Chirino A, Faham S, Fox GM, Arakawa T, Hsu BT, Rees DC. 1991. Three-dimensional structures of acidic and basic fibroblast growth factors. *Science*, 251:90-93.

## Figure Legends

**Figure 1.** Flowchart of design and binding mode prediction strategy.

**Figure 2.** Comparison of the performance of docking programs at predicting GAG binding modes. Assessment of docking performance is made for each test case docked by each program by identifying the pose with the lowest root-mean-squared deviation (RMSD) of heavy atom coordinates relative to the crystallographic binding mode. The complete data is presented in Table S1.

**Figure 3.** Optimization of hydrogen bonding and van der Waals cutoffs for site mapping of GAG-protein interactions. Assessment of accuracy of the prediction of binding residues is made using the $F_1$ score, as described in the Methods.

**Figure 4.** Comparison of the performance of methods for identifying the interacting protein residues involved in ligand recognition. Outliers are excluded from the plots. Assessment of accuracy of the prediction of binding residues is made using the $F_1$ Score, as described in the Methods.

**Figure 5.** Poses from best scoring binding mode clusters obtained after applying design strategy to aFGF structures, overlaid with co-crystallized ligand. a) PDB 3UD8. b) PDB 3UD9. Carbons are colored to the following legend: black – crystallographic ligand, green – GlcNS(6S)-IdoA(2S), blue – GlcNS(3S,6S)-IdoA(2S), grey – GlcNS(3S)-IdoA(2S).

**Figure 6.** Overlay of structures obtained by applying the design strategy to the unliganded aFGF and the crystallized ligands. a) Chain A of PDB 1BAR. b) Chain B of PDB 1BAR. Figure legend:

black – crystallized ligands (from PDBs 3UD8 and 3UD9), green – binding modes from best ranked clusters according to ClusScore. Full cluster details are provided in the Supplementary Information.

**Figure 7.** Best structures according to ClusScore following incorporation of protein flexibility in design strategy. a) Complex of GlcNS(3S)-IdoA(2S) with aFGF (PDB 1BAR chain A) conformation induced by the 36th ranked pose (according to AutoDock) of GlcNS(3S,6S)-IdoA(2S). b) Complex of GlcNS(3S,6S)-IdoA(2S) with aFGF (PDB 1BAR chain B) conformation induced by the 2nd ranked pose (according to AutoDock) of GlcNS(3S,6S)-IdoA. c) Complex of GlcNS(3S,6S)-IdoA(2S) with aFGF (PDB 1BAR chain B) conformation induced by the 10th ranked pose of (according to AutoDock) of GlcNS(3S,6S)-IdoA(2S). All figures overlaid with the crystallographic complex of GlcNS(3S)-IdoA(2S) with aFGF (PDB 3UD8). Figure legend: black – crystallographic complex; white – docked complexes; green atoms – sulfur atoms in crystallographic ligand; yellow atoms – sulfur atoms in docked ligands; green dashes – electrostatic interactions in crystallographic complex; yellow dashes – electrostatic interactions in docked complexes. Carbohydrate residue labels are coloured according to sulfur colours.

## Tables

**Table I.** GAG-protein complexes used for method validation

| PDB ID | Protein | Protein type | Ligand[a] | Resolution (Å) |
|---|---|---|---|---|
| 1BFB | Basic fibroblast growth factor | Growth factor | ΔUA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 1.90 |
| 1BFC | Basic fibroblast growth factor | Growth factor | ΔUA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 2.20 |
| 3UD8 | Acidic fibroblast growth factor | Growth factor | GlcNS(3S)-IdoA(2S)[b] | 2.37 |
| 3UD9 | Acidic fibroblast growth factor | Growth factor | GlcNS(6S)-IdoA(2S)[b] | 2.34 |
| 1GMN | NK1 | Growth factor | IdoA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S)-IdoA(2S) | 2.30 |
| 1G5N | Annexin V | Annexin | ΔUA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 1.90 |
| 2HYU | Annexin A2 | Annexin | ΔUA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 1.86 |
| 2HYV | Annexin A2 | Annexin | ΔUA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S)-IdoA(2S) | 1.42 |
| 1U4L | RANTES | Chemokine | ΔUA(2S)-GlcNS(6S) | 2.00 |
| 1U4M | RANTES | Chemokine | ΔUA(2S)-GlcNS | 2.00 |
| 2NWG | CXCL12 | Chemokine | ΔUA(2S)-GlcNS(6S)[c] | 2.07 |
| 1QQP | Foot-and-mouth disease virus | Virus particle | IdoA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S)-IdoA(2S) | 1.90 |
| 1ZBA | Foot-and-mouth disease virus serotype A10 61 | Virus particle | GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 2.00 |
| 2WNU | C1q | Complement component | IdoA(2S)-GlcNS(6S) | 2.30 |

| | | | | |
|---|---|---|---|---|
| 3B9F | Thrombin | Serine protease[d] | IdoA(2S)-GlcNS | 1.60 |
| 3QMK | Amyloid precursor-like protein 1 | Amyloid precursor protein | IdoA(2S)-GlcNS(6S)-IdoA(2S)-GlcNS(6S) | 2.21 |

[a]All linkages in ligands are α(1→4) linkages. [b]Ligand capped with α(1-OMe) group. [c]Ligand bound at two sites in structure. [d]This protein is non-enzymatic towards heparin.

**Table II.** Molecules initially selected by screening of GAG disaccharides against aFGF structures

| | Structure | | | |
|---|---|---|---|---|
| | 3UD8 | | 3UD9 | |
| Rank | Ligand[a] | Score[b] | Ligand[a] | Score[b] |
| 1 | GlcNS(3S,6S)-IdoA(2S)* | -9.83 | GlcNS(3S,6S)-IdoA(2S)* | -9.31 |
| 2 | GlcNS(6S)-IdoA(2S)* | -9.33 | GlcNS(6S)-IdoA(2S)* | -8.55 |
| 3 | GlcNS(3S,6S)-GlcA | -8.79 | GlcNS(3S,6S)-IdoA | -8.44 |
| 4 | GlcNS(3S)-IdoA(2S)* | -8.76 | GlcNS(3S,6S)-GlcA | -8.30 |
| 5 | GlcNS(3S,6S)-IdoA | -8.36 | GlcNS(3S)-GlcA(2S) | -8.08 |
| 6 | GlcNS(3S)-GlcA(2S) | -8.22 | GlcNS(3S)-IdoA(2S)* | -8.06 |

[a]Ligands marked with an asterisk indicate those known to bind to aFGF. [b]Score computed as the median value of the mean binding energy of the highly populated clusters of that ligand obtained from docking. Results are ranked according to score.