A Simple Method for Discovering Druggable, Specific Glycosaminoglycan-Protein Systems. Elucidation of Key Principles from Heparin/Heparan Sulfate-Binding Proteins

Aurijit Sarkar  
*Virginia Commonwealth University*, sarkara@mymail.vcu.edu

Umesh R. Desai  
*Virginia Commonwealth University*, urdesai@vcu.edu

Follow this and additional works at: http://scholarscompass.vcu.edu/medc_pubs

Part of the Pharmacy and Pharmaceutical Sciences Commons

Copyright: © 2015 Sarkar, Desai. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited
RESEARCH ARTICLE

A Simple Method for Discovering Druggable, Specific Glycosaminoglycan-Protein Systems. Elucidation of Key Principles from Heparin/Heparan Sulfate-Binding Proteins

Aurijit Sarkar, Umesh R. Desai*

Institute for Structural Biology, Drug Discovery & Development and Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond, Virginia, United States of America

* urdesai@vcu.edu

Abstract

Glycosaminoglycans (GAGs) affect human physiology and pathology by modulating more than 500 proteins. GAG-protein interactions are generally assumed to be ionic and nonspecific, but specific interactions do exist. Here, we present a simple method to identify the GAG-binding site (GBS) on proteins that in turn helps predict high specific GAG–protein systems. Contrary to contemporary thinking, we found that the electrostatic potential at basic arginine and lysine residues neither identifies the GBS consistently, nor its specificity. GBSs are better identified by considering the potential at neutral hydrogen bond donors such as asparagine or glutamine sidechains. Our studies also reveal that an unusual constellation of ionic and non-ionic residues in the binding site leads to specificity. Nature engineers the local environment of Asn45 of antithrombin, Gln255 of 3-O-sulfotransferase 3, Gln163 and Asn167 of 3-O-sulfotransferase 1 and Asn27 of basic fibroblast growth factor in the respective GBSs to induce specificity. Such residues are distinct from other uncharged residues on the same protein structure in possessing a significantly higher electrostatic potential, resultant from the local topology. In contrast, uncharged residues on nonspecific GBSs such as thrombin and serum albumin possess a diffuse spread of electrostatic potential. Our findings also contradict the paradigm that GAG-binding sites are simply a collection of contiguous Arg/Lys residues. Our work demonstrates the basis for discovering specifically interacting and druggable GAG-protein systems based on the structure of protein alone, without requiring access to any structure-function relationship data.

Introduction

Sulfated glycosaminoglycans (GAGs), such as heparan sulfate (HS), are nature’s most enigmatic biopolymers. Although made from a linear combination of simple saccharide rings, they display a staggering range of primary sequence diversity that surpasses the range possible for equivalent chains of other biopolymers. The occurrence of HS in planaria [1], the 2nd most primitive species of the animal kingdom, in a variably sulfated form closely matching that in
humans suggests these biopolymers play critical roles in multiple fundamental biological processes. It is now recognized that GAGs bind to hundreds of human proteins with implications in physiological as well as pathological processes such as hemostasis and thrombosis, wound repair and inflammation, neuronal growth and amyloidogenesis, angiogenesis and cancer, defense against microbes and infection [2,3]. Yet, precious little is understood about how these unique, linear polysaccharides recognize and modulate their targets.

Considerable effort has been expended in understanding GAG–protein interactions at an atomic level. The earliest attempt to rationally deduce GAG binding sites (GBSs) on proteins was that of Cardin and Weintraub, who identified ‘XBBXBX’ and ‘XBBBXXBX’ as GAG-recognition domains, where B and X represent basic and hydrophobic residues [4]. Later, these linear, $\alpha$-helical or $\beta$-strand-like segments were extended to include other secondary structural elements [5,6]. However, such linear elements imply divergent evolution of GBSs, whereas the large structural diversity of GAG-binding proteins (GBPs) suggests exactly the opposite. In fact, most GBPs do not follow the simplistic Cardin and Weintraub rules. Sophisticated computational tools are being developed for GBS identification [7–19]. While these approaches are successful within their set limits, a key question that remains unaddressed to date is the specificity of GAG–protein interactions [3,20,21], even though efforts have been made recently to do so [22,23]. For example, why are certain GAG–protein systems, e.g., heparin–antithrombin, highly specific [24,25], while others, e.g., heparin–thrombin [24,26,27], essentially nonspecific? More importantly, can ‘specific’ GAG–protein systems be more reliably predicted to help advance chemical biology and drug discovery?

Predicting the specificity of GAG–protein interactions is extremely challenging because of their reliance on long-range and non-directional Coulombic forces of attraction. Since multiple Arg/Lys and sulfates are involved in these interactions, many of which are redundant, most GAG–protein systems are traditionally forsaken as nonspecific. Yet, growing evidence suggests that GAGs display considerable non-ionic binding energy (10–40%) in recognizing their targets [24,28,29], which may arise from short-range and directional forces, such as hydrogen-bond(s), that induce ‘specific’ recognition. However, the exact origin of specificity among the multitude of interaction loci has been difficult to pinpoint.

The scientific community has historically focused on site directed mutagenesis of residues such as arginine, lysine and histidine, often present within the GBS (Table 1) and which will likely possess a positive charge under physiological conditions. However, GBSs also possess

<table>
<thead>
<tr>
<th>Protein [PDB code]</th>
<th>Arg/Lys in GBS</th>
<th>Other polar residues in GBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antithrombin [1tb6, [30]]</td>
<td>Arg 46, 47, 129, 132, 136; Lys 114, 125, 133, 275</td>
<td>Asn 45</td>
</tr>
<tr>
<td>Thrombin [1tb6, [30]]</td>
<td>Arg 93, 101, 126, 165, 233; Lys 236, 240</td>
<td>Asn 184; Gin 256; His 87</td>
</tr>
<tr>
<td>Basic fibroblast growth factor (FGF2) [1fq9]</td>
<td>Arg 120Lys 26, 119, 125, 129, 135</td>
<td>Asn 27</td>
</tr>
<tr>
<td>3-O-sulfotransferase 1 (HS3ST1) [3uan]</td>
<td>Arg 67, 72, 197, 268, 276Lys 68, 123, 171, 173, 274</td>
<td>Asn 89, 167; Gin 163; His 92</td>
</tr>
<tr>
<td>2-O-sulfotransferase 1 (HS2ST1) [4ndz]</td>
<td>Arg 80, 184, 189, 190, 288Lys 111, 284, 289</td>
<td>Asn91, 108, 112; His 106, 140, 142</td>
</tr>
</tbody>
</table>

Table 1. Polar residues present in various GBSs. These residues form direct interactions with GAGs, as evidenced by analysis of crystal structures.
polar residues such as asparagine and glutamine [29]. Histidine is a special case because it may or may not be charged under physiological conditions (pKₐ ~ 6.9) and its protonation state during GAG-binding isn’t always clear, but Asn and Gln are certainly never positively charged. Therefore, the reason for presence of Asn and Gln in GAG-binding sites remains unclear, even though it is firmly established that they interact with GAGs in many systems (Table 1). We asked a fundamental question: can these uncharged, polar residues have something to do with specificity?

We have uncovered that uncharged residues, such as Asn or Gln, help identify the GBS and also segregate specific GAG-protein systems from nonspecific ones. Our computational results in conjunction with structural and biochemical results show that an unusual constellation of ionic and non-ionic residues constituting the GAG-binding site is necessary for high specificity interaction. These principles can help parse proteins, including those that follow the Cardin–Weintraub rule [4] as well as those that do not, based on their specificity of interaction with GAGs. Based on our findings, we propose that Asn/Gln reduce desolvation penalties during the mostly electrostatic GAG-protein interactions, in addition to hydrogen bonding to sulfates and carboxylates on the GAG. Our results challenge the traditional paradigm that GAG-binding sites on proteins are located in contiguous segments, such as helices and/or turns and that specificity arises from Arg and Lys residues. Our work is expected to fundamentally change the landscape of discovery of ‘highly specific’ GAG-binding sites on proteins, which should greatly assist with identification of druggable GAG-protein systems for designing novel GAGs and GAG mimetics as drugs.

Results

Electrostatic potential alone at electropositive residues does not always identify a GAG binding site

Traditionally, a cluster of electropositive residues has been assumed to form a GBS, suggesting that the electrostatic potential (G_ES) at Arg/Lys residues should unequivocally identify GBSs. To quantitatively assess whether this expectation is correct, we calculated G_ES on Arg/Lys residues of representative heparin-binding proteins including antithrombin, thrombin, FGF2, HS2ST1, HS3ST1 and HS3ST3A1. The G_ES are represented on 2DSE plots, which we devised as new tools for easy quantitative visualization of energy at any locus, e.g., nitrogen donor atom of an Arg or Lys, on the protein surface in two-dimensions (see Methods and Supplementary Methods). Briefly, the position of a locus (such as an atom) on a protein surface is projected from 3D space onto a 2D plane to obtain a scatterplot. The area of each data point in this scatterplot is scaled in size to reflect the G_ES manifested at the atom by the rest of the structure. Therefore, the larger spots represent locations most conducive to binding negatively charged entities such as GAGs. Our expectation, as explained above, was that Arg/Lys of the GBS would demonstrate the highest (most negative) G_ES.

Converse to our expectations, Fig 1 reveals that not all GBS Arg/Lys (Table 1) carry a high G_ES. This is in direct contrast to common empirical assumptions. For example, G_ES for Arg46, Lys114 and Arg129 of antithrombin were significantly higher than other basic residues present in the heparin-binding site such as Arg47, Lys125, Arg132, Lys133, Arg136 and Lys275 (Fig 1A). In fact, Lys114 and Arg129 are known to be crucial for heparin binding [25,31,32], but Arg46 is not. Likewise, Arg93 stands out as the only residue with high G_ES for thrombin, surpassing two residues (Arg97 and Arg101) known to play key role in heparin binding, (Fig 1B) [26,33]. Arg120 of FGF2 is only one of several, similarly contributing residues of the GBS [34,35] but demonstrates significantly higher G_ES (Fig 1C). Similar results were obtained in calculations for HS3ST3A1, HS3ST1 and HS2ST1, three enzymes of the heparan sulfate
Fig 1. $G_{ES}$ at arginines and/or lysines does not identify the GBS on a protein. $G_{ES}$ for Arg/Lys residues are mapped using 2DSE plots for multiple GAG-binding proteins including (a) antithrombin, (b) thrombin, (c) FGF2, (d) HS3ST3A1, (e) HS3ST1 and (f) HS2ST1. The maps reveal that GAG-binding site Arg/Lys residues may not always possess high $G_{ES}$ and not all Arg/Lys with high $G_{ES}$ on a protein are part of the GAG-binding site.

doi:10.1371/journal.pone.0141127.g001
biosynthesis pathway. Although Arg166, Arg274 and Arg325 of HS3ST3A1 display high $G_{ES}$ (Fig 1D), the latter two are not known to be part of the GBS [36]. Likewise, Arg223 of HS3ST1 shows strong $G_{ES}$ (Fig 1E) but is not known to be part of the GBS [37]. Finally, Arg184 of HS2ST1 demonstrates high $G_{ES}$ as well (Fig 1F) but is not known to be important for binding GAGs [35,38]. This implies that not all Arg/Lys residues on a GAG-binding protein’s surface carry high electrostatic potential. More importantly, not all Arg/Lys possessing high electrostatic potential contribute to GAG binding. In fact, for some proteins, multiple loci of high $G_{ES}$ are observed in disparate locations, e.g., HS3ST3A1 (Fig 1D) and HS3ST1 (Fig 1E), and not all are part of the GBS [36,37]. Thus, $G_{ES}$ at electropositive Arg and Lys residues alone does not always define GAG binding and specificity.

Desolvation is an important factor in GAG–protein binding

Considering that $G_{ES}$ alone cannot define binding specificity, we set to identify other parameters that may contribute to specificity of GAG interactions. We hypothesized that to a first approximation the overall binding energy ($\Delta G$) of a GAG–protein system is a composite of electrostatic ($\Delta G_{ES}$) and desolvation ($\Delta G_{DS}$) free energy components. Additional terms, such as vibrational entropy and van der Waals’ energy [39,40], are likely to contribute but were not considered in this first approximation because it was our goal to identify GAG-binding sites from the structure of protein alone. Changes in vibrational entropy and van der Waal’s energy during binding will vary with nature of the cognate GAG binding partner, which precludes inclusion of these terms at this time. To test our hypothesis that $\Delta G_{ES}$ and $\Delta G_{DS}$ are the major factors determining GAG-protein interactions, we studied the antithrombin–heparin pentasaccharide and thrombin–heparin systems, for which a large body of solution experimental data is available [41–45]. For both systems, $\Delta G_{ES}$ varies linearly ($R^2$~0.6–0.7) with change in free energy of GAG binding ($\Delta \Delta G_{OBS}$) suggesting an important role for electrostatic forces to the interaction (Fig 2A and 2B; see Tables 2 and 3). Yet, while statistically significant ($p<0.05$), the correlation is moderate at best, which quantitatively confirms that electrostatics alone does not sufficiently address binding. To assess whether the release of solvent molecules is an important contributor in GAG-protein interactions, desolvation energy ($\Delta G_{DS}$) was calculated using the Poisson-Boltzmann Surface Area (PBSA) method [46] (see Methods). For both proteins, $\Delta G_{DS}$ demonstrated an inverse correlation (Fig 2C and 2D) with $\Delta \Delta G_{OBS}$ ($R^2$~0.6–0.7). Interestingly, electrostatic forces ($\Delta \Delta G_{ES}$) and desolvation forces ($\Delta \Delta G_{DS}$) were found to be directly opposed to each other ($R^2 = 0.99, p<0.05$) (Fig 2E and 2F). More importantly, the magnitude of desolvation effects was substantial, which implies that desolvation energy cannot be discarded and is an important driver of GAG-protein association.

Since desolvation opposes electrostatic interactions, we hypothesized that residues capable of (a) producing non-ionic interactions such as hydrogen bonds (which also possess a partial covalent character to compensate for less favorable ionic interactions) with GAGs, and (b) lowering desolvation costs during such interactions, would play an important role in such interactions. Considering that GAGs possess several hydrogen bond acceptors, e.g., hydroxyls, carboxylates and sulfates, it is likely that asparagine and glutamine residues may serve as hydrogen bond donors.

GAG-protein sites are consistently identified by $G_{ES}$ at uncharged polar residues

It has been long known that GBSs do possess uncharged polar residues [29]. Table 1 provides some typical examples. Their importance has mostly been construed from their presence in co-crystal structures and yet they have not been evaluated routinely by site-directed mutagenesis.
However, whenever they have been, significant reduction in binding and activity are observed. For example, the Gln255Ala mutant of HS3ST3A1, a key enzyme that introduces a rare 3-sulfate group of the herpes simplex virus 1 recognition motif, demonstrates a >99% loss in activity [36]. Similarly, the Gln163Ala and Asn167Ala mutants of HS3ST1, the enzyme responsible for introduction of the 3-sulfate in unit F of DEFGH, show a ~60% drop in activity [37,47]. Likewise, Asn27, Gln123 and Gln134 of FGF2 have also been implicated in heparin binding, of which Asn27 makes a significant enthalpic contribution [34]. However, the reason for such
observations has typically not been clear. Given the predominant roles played by Arg/Lys in GAG-protein binding, it seems unlikely that Asn/Gln alone would form motifs capable of binding GAGs. We predicted that a combined effect manifested by both, electropositive and uncharged residues, may provide us with clues about the roles of such residues in GAG-protein interactions.

Electrostatic interactions are effective at long range and therefore Arg/Lys may also affect the environment of other residues in a protein structure. For example, GES at Asn and Gln residues may be significantly altered. Fig 3 demonstrates GES at Asn and Gln residues (and other neutral hydrogen bonding donors) in the form of 2DSE plots. While GES calculated at basic Arg and Lys residues failed to reliably identify the GBS (Fig 1), the same calculated for neutral hydrogen bond donors identified the GBS consistently (Fig 3). The highest GES for any given protein structure is always present at the GBS, as observed for antithrombin, thrombin, FGF2, HS3ST3A1, HS3ST1, HS2ST1 and serum albumin (Fig 3 and S1 Fig). The identification of the GBS was independent of specific or nonspecific nature of the protein. Quite clearly, the critical positioning of basic residues near uncharged hydrogen bond donors creates a unique

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\Delta G_{ES}$ (kcal/mol)</th>
<th>$\Delta \Delta G_{ES}$ (kcal/mol)</th>
<th>$\Delta G_{DS}$ (kcal/mol)</th>
<th>$\Delta \Delta G_{DS}$ (kcal/mol)</th>
<th>$K_d$ (nM)</th>
<th>$\Delta \Delta G_{OBS}$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wt</td>
<td>-199.8</td>
<td>0</td>
<td>184.6</td>
<td>0</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>K125Q</td>
<td>-125.1</td>
<td>74.7</td>
<td>116.6</td>
<td>-68.0</td>
<td>12</td>
<td>0.4</td>
</tr>
<tr>
<td>K136T</td>
<td>-200.0</td>
<td>-0.25</td>
<td>185.0</td>
<td>0.4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>N135A</td>
<td>-198.9</td>
<td>0.8</td>
<td>183.7</td>
<td>-0.9</td>
<td>2</td>
<td>-0.6</td>
</tr>
<tr>
<td>N135A/R129Q$^a$</td>
<td>-150.7</td>
<td>48.2</td>
<td>139.2</td>
<td>-45.4</td>
<td>1800</td>
<td>3.4</td>
</tr>
<tr>
<td>N135A/R129H$^a$</td>
<td>-151.7</td>
<td>47.2</td>
<td>139.8</td>
<td>-43.9</td>
<td>820</td>
<td>2.9</td>
</tr>
<tr>
<td>N135A/K114A$^a$</td>
<td>-85.4</td>
<td>113.5</td>
<td>80.2</td>
<td>-103.5</td>
<td>1800000</td>
<td>7.6</td>
</tr>
<tr>
<td>R132M</td>
<td>-156.7</td>
<td>43.0</td>
<td>145.2</td>
<td>-39.4</td>
<td>89</td>
<td>1.6</td>
</tr>
<tr>
<td>K133M</td>
<td>-159.4</td>
<td>40.4</td>
<td>147.0</td>
<td>-37.6</td>
<td>171</td>
<td>2.0</td>
</tr>
</tbody>
</table>

$^a$ Calculated from a comparison with the N135A mutant.

$^b$ $K_d$ values were obtained from references [41–44]. $\Delta \Delta G_{OBS}$ values were calculated using the formula $\Delta G = RT \log (K_d^1/K_d^2)$ where $K_d^1$ is for mutant and $K_d^2$ is for wt.

$^c$ Calculated from data available in reference [43].

doi:10.1371/journal.pone.0141127.t002

Table 3. Effect of electrostatic interactions on binding of various thrombin mutants with heparin.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>$\Delta G_{ES}$ (kcal/mol)</th>
<th>$\Delta \Delta G_{ES}$ (kcal/mol)</th>
<th>$\Delta G_{DS}$ (kcal/mol)</th>
<th>$\Delta \Delta G_{DS}$ (kcal/mol)</th>
<th>$K_d$ (nM)</th>
<th>$\Delta \Delta G_{OBS}$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wt</td>
<td>-503.1</td>
<td>0</td>
<td>461.9</td>
<td>0</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>K169E</td>
<td>-275.9</td>
<td>73.8</td>
<td>392.4</td>
<td>-69.4</td>
<td>150</td>
<td>0.31</td>
</tr>
<tr>
<td>R175E</td>
<td>-401.6</td>
<td>101.5</td>
<td>366.8</td>
<td>-95.1</td>
<td>570</td>
<td>1.11</td>
</tr>
<tr>
<td>R233E</td>
<td>-337.3</td>
<td>165.8</td>
<td>311.5</td>
<td>-150.4</td>
<td>720</td>
<td>1.25</td>
</tr>
<tr>
<td>K236E</td>
<td>-275.9</td>
<td>227.2</td>
<td>255.5</td>
<td>-206.4</td>
<td>4800</td>
<td>2.38</td>
</tr>
<tr>
<td>K240E</td>
<td>-282.6</td>
<td>220.5</td>
<td>262.5</td>
<td>-199.4</td>
<td>1000</td>
<td>1.44</td>
</tr>
<tr>
<td>R233E-K240E$^a$</td>
<td>-117.1</td>
<td>220.2</td>
<td>111.8</td>
<td>-199.7</td>
<td>16000</td>
<td>3.11</td>
</tr>
</tbody>
</table>

$^a$ Calculated from comparison with the R233E mutant.

$^b$ These values were reported in reference [45] as equilibrium constant for thrombin dissociating from the ternary antithrombin-thrombin-heparin complex. Since the mutations are at exosite II, where heparin binds (and not antithrombin), these quantify thrombin-heparin interactions. $\Delta \Delta G$ values were calculated using the formula $\Delta G = RT \log (K_d^1/K_d^2)$ where $K_d^1$ is for mutant and $K_d^2$ is for wt.

doi:10.1371/journal.pone.0141127.t003
Identifying High Specificity GAG-Protein Binding Sites

(a) Antithrombin

(b) Thrombin

(c) FGF2

(d) HS3ST3A1

(e) HS3ST1

(f) HS2ST1

$G_{ES}$ (kcal/mol)

$(x^2 + y^2)^{1/2}$
environment which is preferentially bound by GAGs. Thus, G_ES at neutral hydrogen bond donors can be called G\textsubscript{GAG}-binding. As will be demonstrated below, G\textsubscript{GAG}-binding can segregate specific GBSs from nonspecific ones.

Uncharged polar residues of specific GBSs possess significantly higher G\textsubscript{GAG}-binding than the rest of the protein

Antithrombin, FGF2 and sulfotransferases possess specific GBSs while thrombin and serum albumin are known to be nonspecific in nature [25,27,30–32,36,37,35,38,47–49]. All these proteins possess basic and uncharged polar residues in their binding site. We can correctly identify their GBSs by calculating G\textsubscript{GAG}-binding at uncharged hydrogen bond donors (vide supra), but can we segregate specific proteins from nonspecific ones? By first principles, a specific GBS will be an area on the protein surface that possesses significantly higher G\textsubscript{GAG}-binding compared to the rest of the protein. GAGs will therefore preferentially bind to that location rather than anywhere else. For antithrombin, the location of Asn45 possesses a much larger G\textsubscript{GAG}-binding (Fig 3A) than any other area on its surface, indicating that GAGs will preferentially bind to it. Replacement of Asn45 with Ala, Arg or Lys completely alters the unique electrostatic environment observed in the wt structure (S2 Fig) suggesting Asn45 is a “hot spot” for binding the heparin pentasaccharide. In contrast, thrombin did not demonstrate any such unique neutral hydrogen-bond donor (Fig 3B) despite the presence of many basic residues in its GBS. Likewise, bovine serum albumin demonstrated no outstanding neutral donor loci (S1 Fig). It is also known to interact with heparin nonspecifically [49].

An analysis of other GAG-interactors also confirmed our ability to identify such loci; the 2DSE plots for neutral hydrogen-bond donors of FGF2, HS3ST3A1 and HS3ST1 show focused G\textsubscript{GAG}-binding at highly localized regions (Fig 3C, 3D and 3E) similar to AT (Fig 3A). The same observations are made at various dielectric constants and by employing the CHARMM charge model [50,51], (S3 and S4 Figs). Note that using the more accurate CHARMM charge model gave us a better representation of HS2ST1’s specificity (S3 Fig). Thus, GAGs prefer to engage loci containing a strategically positioned neutral hydrogen bond donors near positively charged residues. If the G_ES at a few neutral hydrogen bond donors (G\textsubscript{GAG}-binding) is significantly high in comparison with other regions of the protein surface, then it is a specific GBS. 2DSE plots are a convenient way of identifying the GBS and visualizing its specificity.

Further, a simple statistical distribution of G\textsubscript{GAG}-binding across any given protein can also be used to segregate specific GBSs from nonspecific ones, as we have done in Fig 4. Here, G\textsubscript{GAG}-binding was visualized using violin plots. The specific proteins, antithrombin, FGF2, HS3ST1 and HS3ST3A1 possess at least one location where G\textsubscript{GAG}-binding is significantly higher (more negative) than the average over the entire surface, as delineated by the long tails in the plots. Conversely, the nonspecific proteins thrombin and serum albumin do not possess similar tails, which suggests G\textsubscript{GAG}-binding is rather uniform across most of their surfaces and GAGs will not prefer binding at one location more than any other. This explains why they are nonspecific.

Fig 3. 2DSE plots for G\textsubscript{ES} at neutral hydrogen bond donors. GAGs bind neutral hydrogen bond donors on the protein that possess significantly high G\textsubscript{ES}. (a) Asn45 of antithrombin GAG-binding site possesses the highest G\textsubscript{ES} within the structure. (b) In contrast, the nonspecific thrombin GAG-binding site demonstrates a diffused G\textsubscript{ES}. Similarly, significantly high G\textsubscript{ES} are observed at (c) Asn27 of the FGF2 GBS; Asn27Ala mutation affects GAG-binding (ΔΔG~1.2 kcal/mol) almost as much as K125A (ΔΔG~1.7 kcal/mol), which had the largest effect. (d) Asn255 of the HS3ST3A1 GAG-binding site; the N255A mutant is inactive, and (e) Gln163 of HS3ST1; Gln163Ala mutant loses ~65% activity. (f) Diffused G\textsubscript{ES} of HS2ST1 may represent its ability to bind low-sulfated GAGs. However, Asn91 and 112 of the HS2ST1 GAG-binding site possess a potential higher than His106, mutation of which is already known to affect GAG-binding.

doi:10.1371/journal.pone.0141127.g003
Pharmacophoric basis for targeting specific GBSs

What causes a biased GGAG-binding to exist in specific GAG-binding proteins? Close visualization of the antithrombin/FGF2-heparin interactions explains the origins of the unique environment surrounding Asn45 of antithrombin and Asn27 of FGF2 (Fig 5A and 5B). Both proteins possess positively charged basic residues very close to their respective neutral hydrogen bond donors, causing a large GGAG-binding at these locations. This is the unique feature responsible for GAG-binding specificity elicited in them. Similar structural features are also observed in the sulfotransferase enzymes (Fig 5C–5E). Such features are a hallmark of specific GBSs.

GAG sequences that target specific proteins can be identified by considering which mono- or disaccharides will bind these pharmacophoric features. The identification of GAG sequences that may bind with specificity to a given protein is an additional challenge in GAG-protein interactions. The structural diversity of GAGs requires docking of large GAG libraries onto protein surfaces, but can be used to reasonably represent GAG-protein interactions [8,10,13,15,16,18,40,52–54]. A single residue can be crucial for interactions in GAG-protein systems, e.g. antithrombin binds only heparin containing GlcNS3S [55] and FGF2 binds only IdoA2S-containing heparin [56]. Molecular modeling studies are also most successful at identifying binding partners for such systems where evidence of specific binding exists [18,52,54]. The Combinatorial Virtual Library Screening (CVLS) approach employs an “affinity filter” and a “specificity filter” to successfully identify GAG sequences that bind proteins [15,16,52]. Similar approaches can be refined to a great extent by focused searches targeting the pharmacophoric features identified by our method. First, our method can be used to identify hot spots, followed by docking only mono- or disaccharide libraries to identify preferred motifs that bind
them. The CVLS algorithm can then be used to dock focused GAG oligosaccharide libraries enriched in those preferred mono- or disaccharides. Such a comprehensive approach would pave the path for design of chemical probes and drugs.

Discussion

Our computational results in conjunction with structural and biochemical studies indicate that (i) not all Arg/Lys carry equal and high electrostatic potential; (ii) not all Arg/Lys possessing high electrostatic potential are located in the GAG-binding site; and (iii) not all GAG-binding sites contain a neutral hydrogen-bond donor with significant electrostatic potential. On the contrary, high electrostatic potential is induced at neutral residues by nearby basic residues to generate GAG-binding sites with high specificity of recognition. In fact, we have found that considering the electrostatic potential that exists at neutral hydrogen bond donors not only depicts the correct GBS, but also reveals its specificity. Nature appears to engineer a few strategic neutral hydrogen-bond donors within the highly positively charged domain to induce specificity. At a fundamental level, this represents a major advance over the current paradigm of GAG–protein interactions, which are thought to originate only from Arg/Lys residues.

At a fundamental level, specificity of GAG–protein interaction can be of two types—‘biological’ specificity, which refers to how unique is the geometry of GAG-binding on the protein, and ‘chemical’ specificity, which refers to how unique is the sequence of a GAG that is recognized by the protein. Although most researchers assume that both biological and chemical specificities are identical for GAG–protein interaction, it is not so. For example, thrombin

---

**Fig 5. The structural basis for existence of hot spots in GBSs.** Nature has designed specific GBSs by placing neutral hydrogen bond donors such as the ND2 and NE2 atoms of Asn and Gln respectively in close proximity to charged Arg or Lys residues, as seen in (a) antithrombin, (b) FGF2, (c) HS3ST3A1, (d) HS3ST1 and (e) HS2ST1. This close proximity maximizes the GES at these residues, thereby generating a specific GBS. Not all atoms are displayed, for the sake of visual clarity.

doi:10.1371/journal.pone.0141127.g005
recognizes multiple GAG sequences, which can bind in different orientations on the protein [26] indicating that thrombin–GAG system is neither biologically specific nor chemically specific. In contrast, antithrombin recognizes a unique GAG sequence, which binds in a unique binding geometry indicating that antithrombin–GAG system is both biologically and chemically specific [24,25,28]. By the same token, proteins exhibiting intermediate forms of specificity, e.g., chemically nonspecific but biologically specific, and vice versa, may also exist. While concrete evidence of such systems is lacking, we predict that we will find such systems with increase in our ability to assess specificity of GBs. The method we present here will certainly add to such abilities.

Operationally, this knowledge implies that high specificity GAG-binding sites on apo-protein structures will contain neutral hydrogen-bond donors with high (i.e., more negative) electrostatic potential compared to the rest of the protein. The advanced understanding can be used to develop a simple two-step algorithm for identification of proteins that can bind GAGs with reasonably high level of specificity (Fig 6).

The first step would involve calculation of electrostatic potential at every neutral hydrogen bond donor locus followed by statistical analysis of GES gradients to identify ‘hot spots’ i.e., neutral loci among multiple Arg/Lys residues that presents high electrostatic potential. Hot spots can be identified easily using 2DSE plots. Such hot spots likely engineer high GAG-binding energy through ease of desolvation penalties and hydrogen bond formation, inducing unique GAG-binding orientation and specificity of interaction. Further biophysical investigation is required to confirm the importance of desolvation energy in eliciting specificity of binding, but our ability to elucidate specific and nonspecific GBs is already a critical advancement because it allows identification of specific GAG-binding proteins. The presence of a specific GB on a protein likely signifies a role for GAGs in regulating that protein. Mapping specific GBs onto biological pathways will identify novel biological roles for GAGs. Furthermore, we now have the ability to rationally target biological pathways to design GAG-based chemical probes and drugs.

Our work is expected to fundamentally change the study of GAG-protein interactions because it minimizes reliance on approximations such as linear secondary structure sequences [4] or molecular modeling/docking on regions of high positive charge density [7–19]. Because the approach can work on most high resolution structures (1.5–3.0 Å) being reported today, it is likely to ease the discovery of ‘druggable’ GAG-binding proteins, which has been a major stumbling block. In fact, our method will be of major use in rationally guiding high-throughput docking of GAG sequences [15,16,52] for identifying novel therapeutically relevant GAGs and mimetics that can specifically bind to target protein(s) and also in computational analysis of the proteome to identify the GAG interactome.

Methods

2D-Surface Energy (2DSE) Plots

To the best of our knowledge, 2DSE plots have not been described in the literature. 2DSE plots enable quantitative visualization of 3D energy distribution on a protein surface in two dimensions. Here, the position of a locus on a protein surface is projected from 3D space onto a 2D plane to obtain a scatterplot, in which the area of the dot is scaled to reflect the GES. The foundation of this novel graphical display is as follows. If $i$ is the $i^{th}$ atom at point $(x_i, y_i, z_i)$ in 3D space and $N$ the total number of atoms in the binding site, then total energy $G$ is the sum of energies at each $i^{th}$ atom located at $(x_i, y_i, z_i)$ coordinate. PDB files are a typical example where the atomic $(x, y, z)$ coordinates are provided for proteins. Each atom can therefore be mapped onto the abscissa–ordinate plane by plotting $(x^2 + y^2)^{1/2}$ against $z$. In this projection, the
expression \((x^2 + y^2)^{1/2}\) is the distance of an atom/residue from the origin on the \(x-y\) plane. Further, the size of each point in the plot can be scaled relative to each other to compare their contributions towards binding. Thus, the scatter plot helps visualize the energetics (e.g., \(G_{ES}\)) in 2D arising from \(loci\) in the 3D. Finally, each scatter point can be scaled using an exponential
function to represent the change in energetics at individual loci in comparison to the standard, as defined by the well-established Eq 2.

$$\Delta G_2 - \Delta G_1 = RT \ln (K_{D,2}) - RT \ln (K_{D,1})$$  (1)

$$\frac{K_{D,2}}{K_{D,1}} = e^{\frac{\Delta \Delta G}{RT}}$$  (2)

Here, $K_D$ represents affinity ($\Delta G = RT \ln K_D$) and can be substituted by a factor that can be correlated with affinity. $G_E$ is one such factor known to influence GAG-protein interactions. Therefore, a direct comparison of affinities afforded by $G_E$ at two different points becomes possible using Eq 2. Typical code for generating 2DSE plots is provided in Supplementary Methods.

GAG-Protein Interaction Energy and Desolvation Energy Calculation using Poisson-Boltzmann Surface Area (PBSA) Method

Electrostatic interaction energy ($\Delta G_{E_S}$) and desolvation energy ($\Delta G_{D_S}$) involved in antithrombin-heparin or thrombin-heparin interactions were calculated using the Poisson-Boltzmann Surface Area (PBSA) method implemented in Openeye’s ZAP toolkit [46,57]. Bondi van der Waal’s radii [58], MMFF94 [59] charges and atom-centered Gaussians [46] were used for these calculations for which we assumed an inner dielectric constant of 2 [60] and an outer dielectric constant of 80. The pdb code 1tb6 [30] provided structural information for the calculations. $\Delta G_{E_S}$ was calculated using the formula $\Delta G_{E_S} = \Delta G_{E_S, \text{complex}} - (\Delta G_{E_S, \text{protein}} + \Delta G_{E_S, \text{GAG}})$. Similarly, $\Delta G_{D_S}$ was calculated using $\Delta G_{D_S} = \Delta G_{D_S, \text{complex}} - (\Delta G_{D_S, \text{protein}} + \Delta G_{D_S, \text{GAG}})$.

Statistical Analysis

Statistical analysis was performed using SigmaPlot (sigmaplot.com), or R statistical environment (r-project.org), and the ggplot2 module (ggplot2.org). Statistical significance was calculated using a two-tailed ANOVA on data found to be normal using a Shapiro-Wilk test and demonstrating constant variance. In case of unequal variances, the non-parametric Spearman rank order correlation test was used.

Calculation of $G_{E_S}$ at Individual Residues

Multi-body $G_{E_S}$ calculations were performed using the formula in Fig 6 implemented in the form of a Python script (Supplementary Methods). These calculations were performed at multiple dielectric constants, and also using the CHARMM charge model [50,51] to ensure that the results were similar.

Supporting Information

S1 Fig. 2DSE plots for serum albumin. (a) the $G_{E_S}$ at Arg/Lys residues reflects the relatively hydrophobic nature of albumin (it is known to bind several hydrophobic ligands to reduce their bioavailability) and (b) the $G_{E_S}$ at neutral hydrogen bond donors on serum albumin resembling the same map for thrombin (Fig 1B).

(TIF)

S2 Fig. Mutation of Asn45 to Ala, Arg or Lys alters the electrostatic environment of the protein. When compared to Fig 1A, these 2DSE plots clearly demonstrate that the “hot spot”
at Asn45 cannot exist even on mutation to Arg/Lys.

(TIF)

S3 Fig. The effect of various dielectric constants on potentials. $G_{ES}$ on neutral H-bond donors were recalculated using dielectric constants (a) 2, (b) 3, (C) 4 and (d) 10. Clearly, the trend remains exactly the same.

(TIF)

S4 Fig. Effect of using CHARMM charges on potentials. CHARMM charges were used to calculate $G_{ES}$ on neutral hydrogen bond donors for all the cases. Clearly, while the $G_{ES}$ value may change, there is still a clear distinction between specific and non-specific proteins. However, HS2ST1 now seems far more specific than with MMFF94 charges, in line with expectations that the enzyme will possess a specific GBS.

(TIF)

S1 File. An R script that creates 2DSE plots

(R)

S2 File. The GAG-Binding Site Predictor written in python. NOTE: This will require file 3 installed properly to function.

(PY)

S3 File. A short molecule input/output module written in python.

(PY)

Acknowledgments

The authors thank Openeye Scientific Software, NM for providing software licenses. We also thank the availability of research resources from National Center for Research Resources (S10 RR027411) to Virginia Commonwealth University.

Author Contributions

Conceived and designed the experiments: AS URD. Performed the experiments: AS. Analyzed the data: AS URD. Contributed reagents/materials/analysis tools: AS. Wrote the paper: AS URD. Designed and performed the computational studies: AS. Conceived and supervised the project: URD. Interpreted the observations: AS URD.

References


52. Sankaranarayanan NV, Desai UR. Toward a robust computational screening strategy for identifying glycosaminoglycan sequences that display high specificity for target proteins. Glycobiology. 2014;0: 1–11.


