Glycobiology vol. 24 no. 12 pp. 1323–1333, 2014
doi:10.1093/glycob/cwu077
Advance Access publication on July 21, 2014

Toward a robust computational screening strategy for identifying glycosaminoglycan sequences that display high specificity for target proteins

Nehru Viji Sankaranarayanan[2] and Umesh R Desai[1,2]

[2]Department of Medicinal Chemistry and Institute for Structural Biology and Drug Discovery, Virginia Commonwealth University, 800 E. Leigh Street, Suite 212, Richmond, VA 23219, USA

Received on January 30, 2014; revised on July 3, 2014; accepted on July 15, 2014

Glycosaminoglycans (GAGs) interact with many proteins to regulate processes such as hemostasis, cell adhesion, growth and differentiation and viral infection. Yet, majority of these interactions remain poorly understood at a molecular level. A major reason for this state is the phenomenal structural diversity of GAGs, which has precluded analysis of specificity of their interactions. We had earlier presented a computational protocol for predicting "high-specificity" GAG sequences based on combinatorial virtual library screening (CVLS) technology. In this work, we expand the robustness of this technology through rigorous studies of parameters affecting GAG recognition of proteins, especially antithrombin and thrombin. The CVLS approach involves automated construction of a virtual library of all possible oligosaccharide sequences (di- to octasaccharide) followed by a two-step selection strategy consisting of "affinity" (GOLD score) and "specificity" (consistency of binding) filters. We find that "specificity" features are optimally evaluated using 100 genetic algorithm experiments, 100,000 evolutions and variable docking radius from 10 Å (disaccharide) to 14 Å (hexasaccharide). The results highlight critical interactions in H/HS oligosaccharides that govern specificity. Application of CVLS technology to the antithrombin–heparin system indicates that the minimal "specificity" element is the GlcA*p* (1 → 4)GlcN*p*2S3S disaccharide of heparin. The CVLS technology affords a simple, intuitive framework for the design of longer GAG sequences that can exhibit high "specificity" without resorting to exhaustive screening of millions of theoretical sequences.

Keywords: glycosaminoglycans / heparin/heparan sulfate / molecular docking / specificity / virtual screening

Introduction

Glycosaminoglycans (GAGs) interact with numerous proteins to regulate various physiological and pathological processes such as hemostasis, cell adhesion, growth factor signaling, coagulation, viral pathogenesis and protease regulation (Capila and Linhardt 2002; Gandhi and Mancera 2008). Heparin (H) and heparan sulfate (HS), members of the GAG superfamily, are composed of alternating 1 → 4-linked glucosamine (GlcN*p*) and uronic acid (UA*p*) residues [either glucuronic acid (GlcA*p*) or iduronic acid (IdoA*p*)] that are incompletely modified through sulfation, acetylation and epimerization reactions (Shriver et al. 2012). These modifications can produce 48 distinct disaccharides, of which 23 have been found in nature to date (Esko and Selleck 2002). Further, the IdoA*p* residue can exist in multiple conformations, especially $^{1}C_{4}$ and $^{2}S_{O}$ (Mulloy and Forster 2000), that can interconvert easily in solution to enhance structural possibilities. Thus, combinatorial arrangements of the several configurational and conformational variations possible at the monosaccharide level generate millions of distinct H/HS sequences. Although this massive library of natural H/HS sequences offers a major advantage for enhancing the probability of protein recognition, it also presents difficulties in deciphering detailed structure–function relationships for these interesting biopolymers.

A majority of GAG sequences bind to proteins through nonspecific interactions because practically any collection of positive charges on a protein surface tends to recognize a sulfated GAG chain. Interactions that rely only on electrostatics, e.g. Coulombic, operate over longer distance than those that rely on hydrogen bond or van der Waals forces [Coulomb forces have a r^{-1} relationship (r = distance between two ions), whereas van der Waals forces have a r^{-3} to r^{-6} dependence]. This implies that unless a GAG sequence forms multiple hydrogen-bonding or equivalent interactions, the protein–GAG system will exhibit poor specificity. In fact, not many GAG sequences exhibit characteristics of high specificity. The prototypic high affinity, high-specificity GAG sequence is the heparin pentasaccharide sequence DEFGH (Figure 1A), which contains several key sulfate groups including a 3-*O*-sulfate on the central GlcN*p* residue, that recognizes antithrombin (AT) (Jin et al. 1997; Desai et al. 1998). Other sequences likely to exhibit high specificity include a heparin octasaccharide binding to glycoprotein gD of herpes simplex virus-1 (Copeland et al. 2008) and a dermatan sulfate hexasaccharide binding to heparin cofactor II

[1]To whom correspondence should be addressed: Tel: +1-804-828-7328; Fax: +1-804-827-3664; e-mail: urdesai@vcu.edu

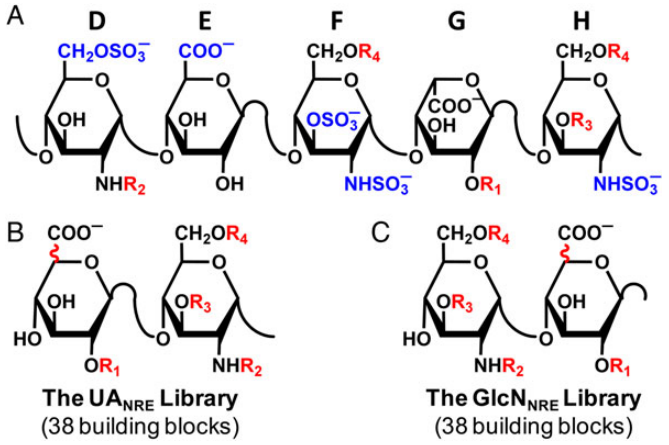

Fig. 1. (**A**) Structure of natural pentasaccharide DEFGH, where D, E, F, G and H labels refer to historical assignment of residue labels (Desai et al. 1998). Residue D forms the non-reducing end, while H is at the reducing end of the polysaccharide chain. Groups highlighted in blue are critical for high-affinity interaction with antithrombin. R_1, R_2, R_3 and R_4 groups are variable groups. (**B**) The UA$_{NRE}$ library of oligosaccharides (di- to octasaccharide) has a GlcA*p* or IdoA*p* residue at the non-reducing end and ends with a GlcN*p* residue at the reducing end. (**C**) The GlcN$_{NRE}$ library of oligosaccharides has a GlcN*p* residue at the non-reducing end and either a GlcA*p* or IdoA*p* residue at the reducing end. R_1, R_2, R_3 and R_4 variations, UA*p* epimerization variation and conformational variations ($^1C_4$ or ^2S_O) for IdoA*p* residue generate 38 disaccharide building blocks.

(Maimone and Tollefsen 1990; Raghuraman et al. 2010). Several other GAG sequences may exhibit specificity of protein recognition (Capila and Linhardt 2002; Gandhi and Mancera 2008) and are awaiting rigorous characterization.

Biochemical and/or biological studies based on gene knockouts of biosynthetic enzymes support the contention that should be many specific, or selective, GAG sequences within the millions present naturally to induce the biological changes in a spatiotemporal manner (Esko and Selleck 2002). Yet, obtaining a library of thousands of homogenous, synthetic GAG sequences is difficult, which precludes rigorous identification of sequences that exhibit high specificity. Likewise, a large library (e.g. >10,000) of homogenous sequences using biosynthetic enzymes has not been developed as yet and microarray-based identification of specific sequences from the mixture of GAG isolates from nature requires much further development.

Under these conditions, computational library screening approaches (Raghuraman et al. 2006, 2010; Agostino et al. 2014) offer considerable promise for identifying sequences that exhibit high specificity. The rapid increase in computational power has enabled simultaneous screening of GAG sequences so as to afford detailed understanding of structure–function relationships. Yet, modeling GAGs is also fraught with challenges. The computational power is still not sufficient to address the entire theoretical conformational search space of GAGs. For example, a simple, unsulfated H/HS disaccharide (GlcA*p*(1 → 4)GlcN*p*) possesses 11 rotatable bonds, which may require a study of up to 10^{17} bond rotations (Assuming that conformational energetics are calculated every 10°, 36 rotations will have to be analyzed for every rotatable bond. For the unsulfated disaccharide containing 11 rotatable bonds, $36^{11} = 1.32 \times 10^{17}$ rotations must be analyzed to understand the entire conformational space.) to understand conformational energetics. For sulfated disaccharides (and higher oligomers), the conformational search space is much larger. Another fundamental challenge is the difficulty of computationally parsing the rare specific GAG interactions from the horde of non-specific interactions. The primary contributory factor to this state is the surface exposure of GAG-binding sites on proteins that induces a primarily electrostatic recognition (Desai 2013).

Despite these difficulties, several computational approaches have been presented in the literature. For example, standard molecular dynamics and docking were used to deduce the pentasaccharide binding site on antithrombin (Grootenhuis and van Boeckel 1991; Bitomsky and Wade 1999). More recently, Agostini et al. (2014) have presented a study of different docking protocols in mapping the binding geometry of GAGs on several proteins. Likewise, we have presented an algorithm that attempted to identify "high affinity, high-specificity" GAG sequences based on two tandem, logical filters including (1) the GOLD score (the "affinity" filter) and (2) consistency of binding (the "specificity" filter) (Raghuraman et al. 2006, 2010). The dual-filter strategy rapidly sorted a small library of H/HS hexasaccharide sequences binding to AT into "specific" and "non-specific" sequences. Yet, this first-generation approach was limited in scope. Herein, we expand the robustness of our GOLD-based combinatorial virtual library screening (CVLS) algorithm by establishing a set of parameters necessary for identification of "high-specificity" sequences, if any, from a haystack of more than 100,000 GAG sequences. Our modified CVLS approach can be readily applied to all possible H/HS sequences ranging from disaccharide to hexasaccharide, irrespective of the number of rotatable bonds. Interestingly, our new CVLS protocol identifies that the minimal sequence capable of recognizing AT with high-specificity is the GlcA*p*(1 → 4)GlcN*p*2S3S disaccharide (or alternatively the "EF" disaccharide), which is much smaller than that identified through synthetically prepared tri- and pentasaccharide variants (Desai et al. 1998). Finally, the new CVLS algorithm affords for the first time a rational framework for designing longer H/HS sequences that bind AT with high specificity without resorting to exhaustive screening of all possible sequences.

Methods

Software

SYBYLX 1.3 (Tripos Associates, St. Louis, MO) was used for molecular visualization, minimization and for preparation of protein structures from the Protein Data Bank. GOLD, v5.1 (Jones et al. 1997) was used for molecular docking experiments. GAG sequences were built combinatorially in an automated manner using in-house SPL (SYBYL Programming Language) scripts.

GAG library generation

The first step in the CVLS approach is the generation of two libraries (Figure 2). Starting from either a UA*p* or a GlcN*p* at the non-reducing end (NRE), so as to give either a UA$_{NRE}$ or GlcN$_{NRE}$ library, respectively, appropriate number of GlcA*p* (in

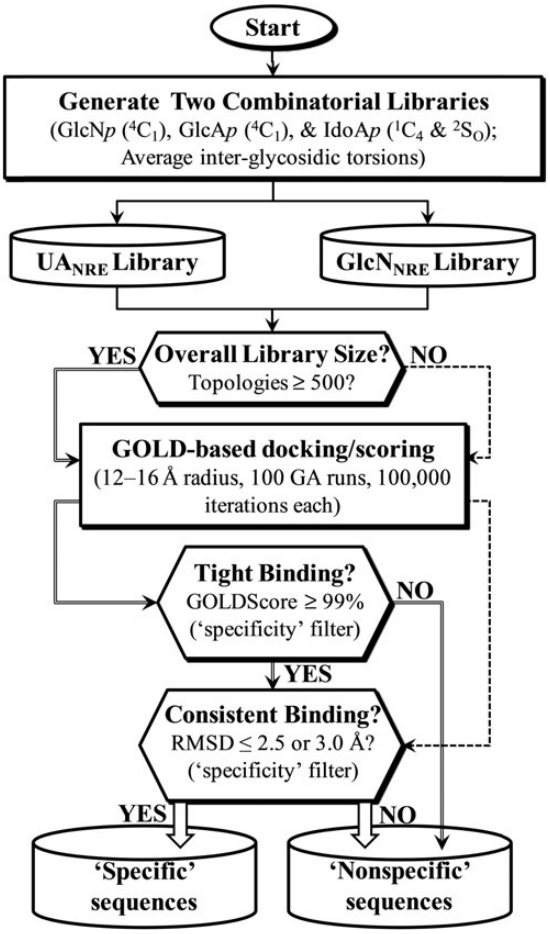

Fig. 2. Combinatorial virtual library screening (CVLS) protocol used to study the AT– H/HS interaction. The CVLS protocol assessed the interaction of two H/HS libraries (UA_{NRE} and $GlcN_{NRE}$) using a dual-filter strategy that relied on the geometric convergence filter (RMSD) to assess specificity of binding.

$^4C_1$ ring pucker), IdoA*p* (in either $^1C_4$ or ^2S_O) and GlcN*p* (in $^4C_1$) residues were added in a combinatorial manner to generate each library of desired chain length (di-, tetra-, hexa- or octasaccharide). The co-ordinates for the two libraries were generated in an automated fashion with a series of SPL scripts and a set of 38 naturally occurring, disaccharide building blocks belonging to each of the UA*p*-GlcN*p* (Figure 1B) and GlcN*p*-UA*p* (Figure 1C) series (Esko and Selleck 2002). Herein, the different monosaccharide units are substituted with *N*-acetyl, *N*-sulfate or *O*-sulfate groups (Mulloy and Forster 2000), which gives rise to unique sequences. To name each unique H/HS sequence, the symbolic representation employed in the GLYCAM (Kirschner et al. 2008) designation was used. Briefly, the letter "Z" was used for GlcA*p*, "u" for IdoA*p* and "Y" for GlcN*p*. Similarly, ring conformations were encoded as "a" for $^1C_4$, "b" for $^4C_1$ and "c" for ^2S_O conformations (Cremer and Pople 1975; Forster and Mulloy 1993; Rao et al. 1998). Substituents on rings were represented as "H" (for unsubstituted 2-position), "C" (for *N*-acetyl), "2" (for *N*- or *O*-sulfate), "3" (for 3-*O*-sulfate) and "6" (for 6-*O*-sulfate). Anomeric carbon configuration was encoded as "A" for α and "B" for β. This monosaccharide nomenclature is also shown in Table I. Analysis of the available crystal structures showed that the interglycosidic torsions ϕ_H (O5–C1–O1–C4′) and ψ_H (C1–O1–C4′–C5′) fall within a relatively narrow range and are essentially invariant irrespective of the substitution pattern (Jin et al. 1997; McCoy et al. 2003; Johnson et al. 2006; Pol-Fachin and Verli 2008). Thus, average bond torsions, shown in Table II, were used for interglycosidic linkages. The disaccharide building blocks were then used to build the desired library using an SPL script following which each sequence was minimized, in an automated manner. Thus, the two UA_{NRE} and $GlcN_{NRE}$ libraries contained a total of $2 \times (38 \times 38 \times 38) = 109{,}744$ unique H/HS hexasaccharide sequences. Likewise, the di- and tetrasaccharide libraries consisted of 76 and 2888 unique sequences, respectively.

Table I. Naming convention for the H/HS monosaccharides and the 38 disaccharide building blocks derived from them

Name[a]	Name	Conf.[b]	Anomer	Disaccharide building blocks		
uaA	IdoA*p*	$^1C_4$	α-	ZbB-YbCA	ua2A-YbC6A	ua2A-YbH3A
ua2A	IdoA*p*2S	$^1C_4$	α-	ZbB-YbC6A	uc2A-YbC6A	uc2A-YbH3A
ucA	IdoA*p*	^2S_O	α-	ZbB-Yb2A	uaA-Yb2A	ua2A-YbH36A
uc2A	IdoA*p*2S	^2S_O	α-	Zb2B-Yb2A	ua2A-Yb2A	uc2A-YbH36A
Yb2A	GlcN*p*2S	$^4C_1$	α-	ZbB-Yb26A	uc2A-Yb2A	ucA-Yb2A
Yb23A	GlcN*p*2S3S	$^4C_1$	α-	Zb2B-Yb26A	uaA-Yb26A	uc2A-Yb23A
Yb26A	GlcN*p*2S6S	$^4C_1$	α-	ZbB-Yb23A	ucA-Yb26A	
Yb236A	GlcN*p*2S3S6S	$^4C_1$	α-	Zb2B-Yb23A	ua2A-Yb26A	
Yb26A	GlcN*p*2S6S	$^4C_1$	α-	ZbB-Yb236A	uc2A-Yb26A	
YbCA	GlcN*p*2Ac	$^4C_1$	α-	ZbB-YbHA	uaA-Yb23A	
YbC6A	GlcN*p*2Ac6S	$^4C_1$	α-	uaA-YbCA	ucA-Yb23A	
YbHA	GlcN*p*	$^4C_1$	α-	ucA-YbCA	ua2A-Yb23A	
YbH3A	GlcN*p*3S	$^4C_1$	α-	ua2A-YbCA	uaA-Yb236A	
YbH36A	GlcN*p*3S6S	$^4C_1$	α-	uc2A-YbCA	ucA-Yb236A	
ZbB	GlcA*p*	$^4C_1$	β-	uaA-YbC6A	ua2A-Yb236A[c]	
Zb2B	GlcA*p*2S	$^4C_1$	β-	ucA-YbC6A	uc2A-Yb236A[c]	

[a]Symbols: Z = *D*-GlcA*p*, u = *L*-IdoA*p*, Y = *D*-GlcN*p*. Ring conformations: a = $^1C_4$; b = $^4C_1$; c = ^2S_O. Substituents: H = No substitution at position 2; Ac = *N*-acetyl, S = sulfate; anomer configuration: A = α, B = β.
[b]Conformation.
[c]Disaccharides modeled in addition to those presented by Esko and Selleck (2002).

Table II. Average torsion across the 1 → 4 interglycosidic bonds used in this CVLS study

Disaccharide building block	Φ (O5–C1–O1–C4′)	Ψ (C1–O1–C4′–C5′)
GlcA*p*(1 → 4)GlcN*p*	−81.8	−114.0
IdoA*p*(1 → 4)GlcN*p*	−87.7	−128.3
GlcN*p*(1 → 4)GlcA*p*	91.1	−151.6
GlcN*p*(1 → 4)IdoA*p*	87.4	−132.3

Preparation of AT and GAG sequences for docking

The coordinates for the activated form of AT were extracted from the crystal structure of the ternary AT–pentasaccharide–thrombin complex (Brookhaven Protein Data Bank entry 1TB6) (Li et al. 2004). Likewise, thrombin coordinates were extracted from both the 1TB6 structure and 1XMN structure (chains A and B) (Carter et al. 2005). Hydrogen atoms were added in SYBYL X1.3 and the structure minimized with fixed heavy-atom co-ordinates using the Tripos forcefield for a maximum of 5000 iterations subject to a termination gradient of 0.05 kcal/(mol Å). Energy minimization of the modeled GAG, AT and thrombin structures was performed using the Tripos forcefield with Gasteiger-Hückel charges, a fixed dielectric constant of 80 and a non-bonded cutoff radius of 8 Å.

Docking of the GAG sequences

Molecular docking of the library of sequences onto the activated form of AT was performed using GOLD v.5.1 (Jones et al. 1997). GOLD is a "soft docking" method that implicitly handles local protein flexibility by allowing a small degree of interpenetration, or van der Waals overlap, of ligand and protein atoms. GOLD also optimizes the positions of hydrogen-bond donating atoms on Ser, Thr, Tyr, Lys and Arg residues as part of the docking process. The binding site in AT was defined to cover key heparin-binding residues including Lys11, Arg13, Arg46, Arg47, Trp49, Lys114, Phe121, Lys125, Arg129 and Arg132 (Jin et al. 1997; Desai 2005; Pike et al. 2005). Similarly, the binding site in thrombin was defined to encompass basic residues of exosite 2 including Arg93, Arg101, Arg126, Arg165, Arg233, Lys236 and Lys240. The grid center was defined as the center of the enclosure containing these residues.

For the GAG sequences, the interglycosidic bonds were constrained. In addition to the two libraries for each chain length, docking was also performed for the heparin pentasaccharide sequence (Figure 1A) using the set of optimized parameters. GOLD starts with a population of 100 arbitrarily docked ligand orientations, evaluates them using a scoring function (the GA "fitness" function) and improves their average "fitness" by an iterative optimization procedure that is biased towards high scores. As the initial population is selected at random, several such GA runs are required to more reliably predict correct bound conformations. The optimized parameters included 100 GA runs for each sequence docked onto a binding site of 5–16 Å radius, depending upon chain length and number of rotatable bonds, using a maximum of 100,000 iterations that are continuously evaluated by the GOLD score and/or root-mean-square difference (RMSD) between top-ranked solutions. Collectively, these 100 GA runs form one docking experiment from which the top two solutions were considered for further analysis. Experiments were minimally performed in triplicate, which would yield at least six solutions. To enhance efficiency, the GA was set to preterminate if the top two ranked solutions were within 2.5 Å RMSD. A one or two-step docking protocol was utilized depending on the library size, as described in Figure 2. If the library contained <500 sequences, the analysis relied only on the RMSD between the top six ranked solutions obtained in three independent experiments for every sequence. If the number of sequences was much >500, a two-step protocol involving selection of the most promising sequences based on their GOLD scores (the "affinity" filter) followed by redocking of the selected sequences and evaluation of RMSD between the top-ranked solutions (the "specificity" filter). Docking was driven by $\text{GOLDScore} = \text{HB}_{\text{EXT}} + 1.375 \times \text{VDW}_{\text{EXT}}$ equation (HB_{EXT} and VDW_{EXT} are the non-bonded intermolecular hydrogen bond and van der Waals terms, respectively) to prioritize different poses, as reported earlier (Raghuraman et al. 2006).

Results and discussion

Development of a more robust CVLS approach for identifying "high-specificity" sequences

Our previous CVLS protocol could be applied to a small library of H/HS sequences (~7000 sequences) and screened a limited domain of 3D space to assess "specificity" of binding (Raghuraman et al. 2006). However, the repertoire of nature's sequences is huge, of which the majority are never studied rigorously. Thus, a more robust CVLS protocol that can be optimally applied to a larger H/HS library is desirable. Considering the dramatic increase in conformational and configurational space with the size of the library, it was important to define the optimal parameter set that ensures reproducible and accurate outcome. Hence, we selected a representative group of 65 H/HS hexasaccharide sequences using the following rationale. Fifteen hexasaccharide sequences contained the DEF structure, which is known to be the key high-specificity element of the DEFGH pentasaccharide. This group was identified as G_{DEF}. Likewise, 15 hexasaccharide sequences were selected to contain either the EFG or the FGH structure. These were identified as G_{EFG} and G_{FGH} groups. The G_{FGH} group was expected to be the most non-specific group based on literature reports (Desai et al. 1998). Two additional groups of 10 library members each were constructed to assess specificity elements. The G_{USU} group contained fewer sulfate groups than the G_{DEF} group, whereas the G_{SSS} contained more sulfate groups (Supplementary data, Table S1).

Because the size of the library in this limited study was small (<500 members), we utilized the one-step CVLS protocol (Figure 2) to assess how well a particular sequence binds to AT in multiple docking runs. A low RMSD (<2.50 Å) between multiple interaction poses in a GA-based (The GA-based approach incorporates domain swaps and mutations so as to screen essentially the entire 3D space for each sequence. In other words, a GA-based run attempts to dock an H/HS sequence onto the binding site with all possible conformations/configurations so as to identify the most optimal fit. This enhances the probability of identifying interactions arising from key structural features on

the H/HS sequence) search suggests high consistency of binding. When only a few sequences present in a library display such high consistency of binding, it suggests specific binding with the protein. In contrast, sequences that bind with poor consistency (RMSD > 2.50 Å) possess no structural features that uniquely recognize the target and hence are likely to be non-specific. Thus, based on the literature (Desai et al. 1998; Capila and Linhardt 2002; Raghuraman et al. 2006), the G_{DEF} group should exhibit the highest proportion of sequences that display RMSD < 2.5 Å (high specificity), while the G_{FGH} and G_{USU} group should exhibit relatively poor specificity. These predictions allowed a rigorous assessment of CVLS parameters that impact the robustness of the protocol including the number of GA runs (10 or 100) and the size of docking radius.

Figure 3 shows CVLS results for the five groups of H/HS hexasaccharide sequences. For each group, the proportion of sequences displaying RMSD < 2.5 Å was found to be highest at 14 Å and 100 GA runs. More importantly, G_{DEF} displayed higher proportion "specific" sequences, as expected. In contrast, the "non-specific" group G_{FGH} displayed much lower proportion mirroring biochemical results (Desai et al. 1998). The G_{USU} group displayed the least specificity, as one would expect based on the absence of key sulfate groups.

Interestingly, a 16 Å radius failed to reliably identify "specific" sequences. This was an unusual result considering that search with 14 Å radius identifies majority of sequences. Most probably, the failure of 16 Å radius search arises from incomplete conformational search imposed by the limitation in the maximum number of GA iterations (100,000) used in this study. It is well established that an increase in docking radius dramatically increases the conformational search space, which in turn requires orders of magnitude increase in the number of GA iterations. At the other extreme, the reason why searches with 8 and 10 Å radii failed is most likely due to the size of the binding site becoming too small for hexasaccharide sequences at these radii. With regard to the number of GA iterations, the results show that 10,000 runs appear to not reliably perform a thorough search of the 3D space. We did not screen an order of magnitude increase in GA iterations, i.e. 1,000,000 runs, because it would be time-wise not feasible. In combination, the results for hexasaccharide sequences show that an arbitrary choice of docking parameters may not yield optimal results within the allocated time frame or make take inexorbitant time to complete. Thus, GAG docking experiments should be approached with measured steps.

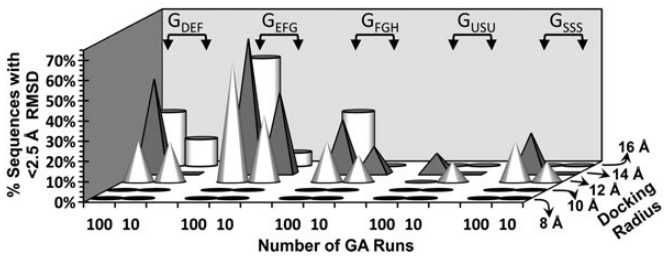

Fig. 3. Optimization of GOLD-based CVLS protocol for H/HS hexasaccharide sequences. Parameters including docking radius (8–16 Å), number of GA runs (either 10 or 100) and number of iterations (10,000 or 100,000; not shown here) were evaluated in a rigorous manner for groups of H/HS hexasaccharide sequences including G_{DEF}, G_{EFG}, G_{FGH}, G_{USU} and G_{SSS}, which refer to 15 sequences containing the DEF structure, 15 sequences containing the EFG structure, 15 sequences containing the FGH structure, 10 sequences containing suboptimal level of sulfation and 10 sequence containing higher level of sulfation, respectively. The structures of the 65 hexasaccharide sequences studied are listed in Supplementary data, Table S1.

Inference of structure on the specificity of recognition

The above study also led to two key advances with regard to H/HS–AT interaction. (1) It is generally assumed that high sulfation level of a GAG chain ensures protein binding. This arises from the non-directional nature of electrostatic forces, which favor recognition of practically any collection of basic residues. In direct contrast to this assumption, the G_{SSS} group, which comprises of sequences that are more sulfated on average than the G_{DEF} group (see Supplementary data, Table S1), displays very few "specific" sequences (Figure 3). In fact, this proportion is even lower than that for G_{FGH}, which is known to be non-specific in solution, and just slightly better than that for G_{USU}. Clearly, for high-specificity GAG–protein systems hyper sulfation destroys specificity of interaction. (2) The G_{EFG} group has not been studied well in the literature, except for tetrasaccharide EFGH (Petitou et al. 1997; Desai et al. 1998). The CVLS results show that G_{EFG} group displays the highest proportion of sequences with RMSD < 2.5 Å. This proportion is even higher than that for G_{DEF} suggesting that the EFG trisaccharide of the DEFGH sequence may be a better cause of specific recognition of AT. This aspect is further addressed below.

Application of the CVLS protocol to a large H/HS library

Can the above more robust CVLS protocol reliably identify specific sequences from a large library? To address this question, a combinatorial library of H/HS hexasaccharide sequences was generated from all possible monosaccharide residues found in nature (Figure 1B and C, Table I), except for the rare free GlcN*p*. As in our earlier study, the library considered the conformational flexibility of IdoA*p* residues in an explicit manner through the inclusion of both the $^{1}C_{4}$ and $^{2}S_{O}$ conformations and utilized the "average backbone" geometry of the interglycosidic bonds (Raghuraman et al. 2006). Yet, a significant advance over the earlier study was the explicit consideration of sequences with two different NREs. This led to two libraries, named as UA_{NRE} and $GlcN_{NRE}$ libraries, which were treated independently for computational purposes (Figure 2). Considering that AT binding is sequence specific, the UA_{NRE} and $GlcN_{NRE}$ libraries may yield different results. Each library contained 54,872 distinct hexasaccharide sequences made combinatorially from 38 disaccharide building blocks, which were generated in a fully automated manner.

The CVLS protocol for the larger library consisted of two steps. The first step involved docking each of the >50,000 hexasaccharides onto AT using the parameters developed above and analyzing the poses using GOLDScore (see Methods). The highest ranking 54 sequences (or top 0.1%) from each library were re-docked in three independent runs and the top two poses of each run were utilized for assessing consistency of binding. The RMSD between the six poses of every sequence was calculated and sequences displaying values <2.5 Å were identified as most promising from the perspective of specificity. This

analytical approach eliminates the need for a reference co-crystal structure, which may not be possible for many GAG–protein systems, and thus, greatly expands the applicability of CVLS protocol.

Our CVLS protocol affords deduction of binding specificity from both biological and chemical considerations. Biological specificity refers to a unique mode of interaction in the binding site among many possible modes, whereas chemical specificity refers to a unique ligand sequence among the many sequences available. Because the CVLS protocol operates on a large library and attempts to identify "needles in a haystack", the final identified GAG sequences capture features of chemical specificity. Likewise, the GA-based identification on one binding mode binding (i.e. low RMSD value) from among many likely captures features of biological specificity. Yet, the CVLS protocol cannot be expected to capture all the chemical and biological specificity features of a GAG–protein system in few sequences because the GAG conformational search space is enormous. However, the dual-filter protocol greatly enhances the probability of rapidly and accurately identifying "specific" GAG sequences.

Of the 54 sequences identified from each of the UA_{NRE} and $GlcN_{NRE}$ hexasaccharide libraries following the 1st filter, 10 and 24 sequences, respectively, satisfied the 2nd filter (Table III). These poses were then compared with the pentasaccharide DEFGH geometry of the co-crystal structure (Li et al. 2004) and found to be essentially identical (not shown). Further, the hydrogen-bonding analysis using LIGPLOT (Wallace et al. 1995) showed that the sequences identified by CVLS bound to AT in a manner similar to DEFGH. The 34 sequences represent a significant increase in identification of specific sequences from the 10 such sequences identified in our earlier work from a library of 6859 (Raghuraman et al. 2006). This highlights the enhanced robustness of the new algorithm. More importantly, ua2A-YbCA-ua2A-YbCA-uaA-Yb26A, or alternatively IdoA*p*2S-GlcN*p*2Ac-IdoA*p*2S-GlcN*p*2Ac-IdoA*p*-GlcN*p*2S6S, a false-negative hexasaccharide identified in our first attempt, was effectively eliminated by in this more careful protocol.

A comparison of the binding geometry of the 10 and 24 sequences from the UA_{NRE} and $GlcN_{NRE}$ libraries reveals an interesting insight. Although each sequence is distinct, the binding poses attempt to satisfy a core group of interactions arising from the D, E and F residues. This induces a shift in frame of one residue between UA_{NRE} and $GlcN_{NRE}$ sequences (Figure 4). This also explains why more high-specificity sequences were identified from the $GlcN_{NRE}$ library in comparison to the UA_{NRE} library. The difference in frame between the two libraries affords interactions with the extended heparin-binding site residues for the $GlcN_{NRE}$ library, which are not realized by sequences of the UA_{NRE} library (Figure 4B, Table III). These additional interactions contribute to the specificity of binding.

Table III. H/HS hexasaccharide sequences from two libraries of 54,872 sequences each that satisfied the dual-filter CVLS strategy for antithrombin

No.	Hexasaccharide sequence[a]	GoldScore[b]	No. of H-bonds[c]
UA_{NRE} library			
1	ZbB-Yb26A-ZbB-Yb236A-ucA-Yb2A	137.81	13
2	uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	136.76	13
3	ucA-YbC6A-Zb2B-Yb23A-ucA-Yb2A	133.15	14
4	ua2A-YbH36A-ZbB-Yb236A-ucA-Yb2A	132.83	12
5	ucA-Yb26A-ZbB-Yb23A-ucA-YbCA	132.66	14
6	ucA-YbC6A-ZbB-Yb236A-ucA-Yb2A	132.42	14
7	ZbB-Yb26A-ZbB-Yb236A-ucA-YbCA	132.39	12
8	uaA-Yb23A-ZbB-Yb236A-ucA-Yb2A	131.63	12
9	ucA-Yb26A-ZbB-Yb236A-ucA-Yb2A	131.25	14
10	ua2A-Yb23A-ZbB-Yb236A-ucA-Yb236A	129.54	14
$GlcN_{NRE}$ library			
1	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB	147.76	12
2	Yb26A-ua2A-Yb26A-ZbB-Yb236A-ZbB	142.34	11
3	Yb2A-ua2A-YbC6A-ZbB-Yb236A-ZbB	141.45	11
4	Yb23A-uaA-YbC6A-ZbB-Yb23A-ZbB	139.57	14
5	Yb2A-ua2A-Yb26A-Zb2B-Yb236A-ZbB	138.59	14
6	Yb26A-uaA-YbC6A-ZbB-Yb236A-ZbB	138.57	10
7	Yb26A-ua2A-Yb236A-ZbB-Yb236A-ZbB	136.99	10
8	Yb236A-ua2A-YbC6A-ZbB-Yb236A-ZbB	135.44	11
9	Yb2A-ua2A-Yb26A-Zb2B-Yb23A-ZbB	134.74	10
10	Yb26A-uaA-YbC6A-ZbB-Yb236A-uc2A	134.55	13
11	Yb2A-uaA-Yb26A-ZbB-Yb236A-ZbB	134.3	12
12	Yb26A-ua2A-Yb26A-Zb2B-Yb23A-ZbB	134.23	12
13	Yb2A-uaA-Yb236A-ZbB-Yb236A-uc2A	133.72	11
14	Yb26A-ua2A-Yb26A-ZbB-Yb23A-ZbB	133.08	12
15	Yb23A-ua2A-YbC6A-ZbB-Yb236A-uc2A	132.82	11
16	Yb236A-ua2A-YbC6A-ZbB-Yb23A-ZbB	132.60	10
17	Yb2A-ua2A-Yb26A-ZbB-Yb236A-ZbB	132.54	12
18	Yb26A-uaA-YbC6A-ZbB-Yb23A-uc2A	132.43	11
19	Yb236A-ua2A-Yb236A-ZbB-Yb236A-ZbB	131.50	13
20	Yb23A-uc2A-Yb26A-Zb2B-Yb236A-ZbB	131.21	13
21	YbC6A-ZbB-Yb236A-ucA-Yb2A-ua2A	130.78	16
22	Yb23A-uaA-YbC6A-ZbB-Yb23A-uc2A	130.29	9
23	Yb23A-ua2A-Yb236A-ZbB-Yb236A-ZbB	129.75	11
24	Yb236A-ua2A-Yb26A-Zb2B-Yb236A-uc2A	129.06	13

[a]See definitions of residue labels and substitution in Table I.
[b]Refers to modified GoldScore, as defined in Methods.
[c]Number of hydrogen bonds calculated using LIGPLOT.

Importance of appropriate restriction on conformational search space

The results with hexasaccharide sequences indicated that CVLS does not work well with too large a conformational search space, as demonstrated by an optimal docking radius of 14 Å. We reasoned that presumably this would hold true for oligosaccharides with other lengths too. Thus, we docked tetrasaccharide sequences containing the DEF structure onto AT using docking radius ranging from 8 to 16 Å under otherwise identical conditions. Theoretically, each tetrasaccharide sequence should bind with a low RMSD because of the presence of the DEF scaffold. Yet, the results revealed that consistent docking was best obtained with a radius of 12 Å (Figure 5). For disaccharides, we studied all 78 sequences and found that a docking radius of 10 Å was most optimal. Thus, the results highlight the importance of selecting appropriate docking radius, which is surrogate for restricting conformational search space, for identifying "needles in a haystack". For H/HS sequences, it appears that di-, tetra- and hexasaccharide sequences are best studied by using docking radius of 10, 12 and 14 Å, respectively.

CVLS predicts the "minimal" H/HS sequence that exhibits high specificity for AT

Although pentasaccharide DEFGH (Figure 1A) is recognized as the minimal AT-specific sequence, detailed biochemical study with pentasaccharide variants led to the conclusion that

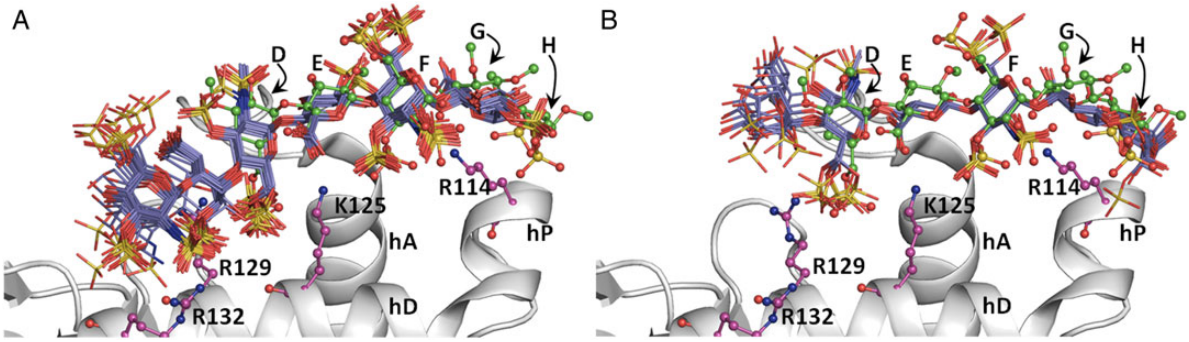

Fig. 4. CVLS predicted hexasaccharide sequences from the $GlcN_{NRE}$ (**A**) and UA_{NRE} libraries (**B**) containing 54,872 sequences each. Shown are overlays of the docked poses of hexasaccharide sequences that bind AT with "high specificity" by satisfying the dual-filter strategy. (**A**) Twenty-four sequences (blue sticks) from the $GlcN_{NRE}$ library and (**B**) 24 sequences (blue sticks) from UA_{NRE} hexasaccharide library. Helices A (hA), D (hD) and P (hP) of antithrombin are shown in ribbon form and residues Arg132, Arg129, Lys125 and Arg114 are shown in ball and stick display. The crystal structure of DEFGH in green ball and sticks display is shown to highlight correspondence with the CVLS predicted poses.

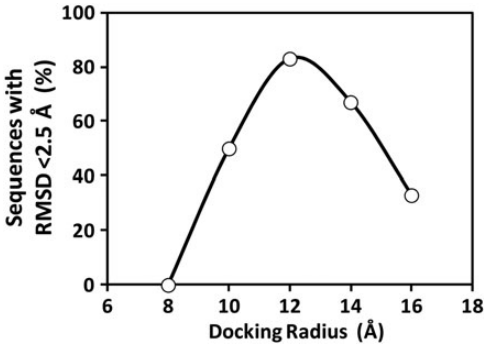

Fig. 5. Optimization of docking radius to be used in CVLS of H/HS tetrasaccharide sequences. A library of 17 DEF containing tetrasaccharides was assessed for "specificity" of binding using varying docking radius. See Supplementary data, Figure S1 for structural poses of the specific tetrasaccharide sequences.

trisaccharide DEF was minimally needed (Desai et al. 1998). The development of the current more robust CVLS protocol presented an opportunity to assess this computationally using the libraries of tetrasaccharide (2888 sequence), disaccharide (76 sequences) and monosaccharides (15 sequences). The results showed that 14 UA_{NRE} and 16 $GlcN_{NRE}$ tetrasaccharide sequences (Supplementary data, Figure S1) and 3 UA_{NRE} (and none $GlcN_{NRE}$) disaccharide sequences satisfied the RMSD filter (Figure 6). Each of the three disaccharide sequences contained the GlcA*p*(1 → 4)GlcN*p*2S3S structure (the EF disaccharide, see Supplementary data, Table S3), which was also present in the high-specificity 30 tetrasaccharides (see Supplementary data, Table S2) identified from the library of 2888 sequences. Reducing the size further to monosaccharides eliminated interaction specificity completely. Likewise, neither the GlcN*p*6S(1 → 4)GlcA*p* sequence nor the GlcN*p*2S3S (1 → 4)GlcA*p* sequence, i.e. neither the DE nor the FE sequence, satisfied the consistency of binding filter (Figure 6B). The results suggested that the exquisite specificity features displayed by the H/HS–AT system arises from the EF disaccharide motif.

The above deduction, in fact, has experimental support. Our detailed studies using stopped flow fluorimetry (Desai et al. 1998) on a group of variants of pentasaccharide DEFGH have shown that removal of residue D (as in tetrasaccharide EFGH) or residues G & H (as in trisaccharide DEF) does not affect specificity of AT recognition. But elimination of residues D and E (as in trisaccharide FGH) resulted in complete loss in specificity. Thus, key elements of specific recognition of antithrombin are actually resident in the disaccharide sequence EF. Thus, these CVLS studies significantly advance understanding on the structural basis of specificity with regard to the antithrombin–heparin system.

Although disaccharide sequence EF may be all that is necessary for specific recognition of AT, this does not imply residues D, G and H of pentasaccharide DEFGH are not important. This work shows that if the EF sequence is not present, then the sequence(s) display(s) multiple modes of binding (i.e. lack biological specificity). The function of D, G and H residues is to contribute binding energy. This is the reason why pentasaccharide DEFGH sequence is a better pharmaceutical agent than tetrasaccharide EFGH or trisaccharide FGH.

A simple CVLS-based algorithm for designing longer H/HS sequences

Considering that many proteins bind longer GAG sequences (Capila and Linhardt 2002; Gandhi and Mancera 2008), we sought to assess the application of our CVLS approach to the design of H/HS sequence(s) longer than hexasaccharide. Yet, designing longer sequences de novo is considerably more challenging. For example, a de novo library of octasaccharide sequences built from the 38 natural disaccharide building blocks would consist of 38^4, or 2,085,136, unique sequences for each of the two libraries, UA_{NRE} and $GlcN_{NRE}$. Exhaustive screening at this scale is not possible in a reasonable timeframe with current computational power. Hence, we developed a reductionist approach. We reasoned that the 3D space on either side of the most specific hexasaccharide sequence(s) could be explored using our CVLS protocol to derive the most optimal octasaccharide sequence(s).

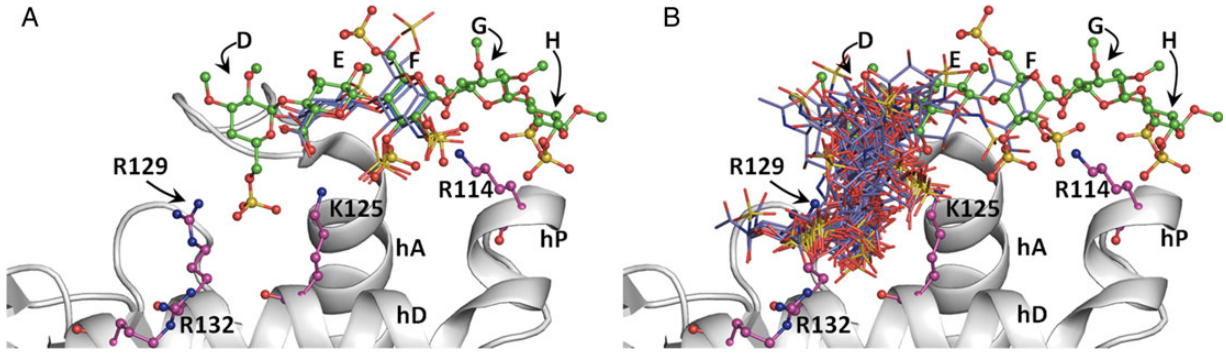

Fig. 6. CVLS predicted disaccharide sequences from the UA_{NRE} (**A**) and $GlcN_{NRE}$ libraries (**B**). Only three disaccharide sequences (blue sticks), each related to EF structure of pentasaccharide DEFGH, were found to satisfy the dual-filter strategy from the library of 78 UA_{NRE} sequences (**A**). None of the 78 possible disaccharides (blue sticks) belonging to the $GlcN_{NRE}$ library bound antithrombin with "high specificity". Helices A (hA), D (hD) and P (hP) of antithrombin are shown in ribbon form and residues Arg132, Arg129, Lys125 and Arg114 are shown in ball and stick display. The crystal structure of DEFGH in green ball and sticks display is shown to highlight correspondence with the CVLS predicted poses.

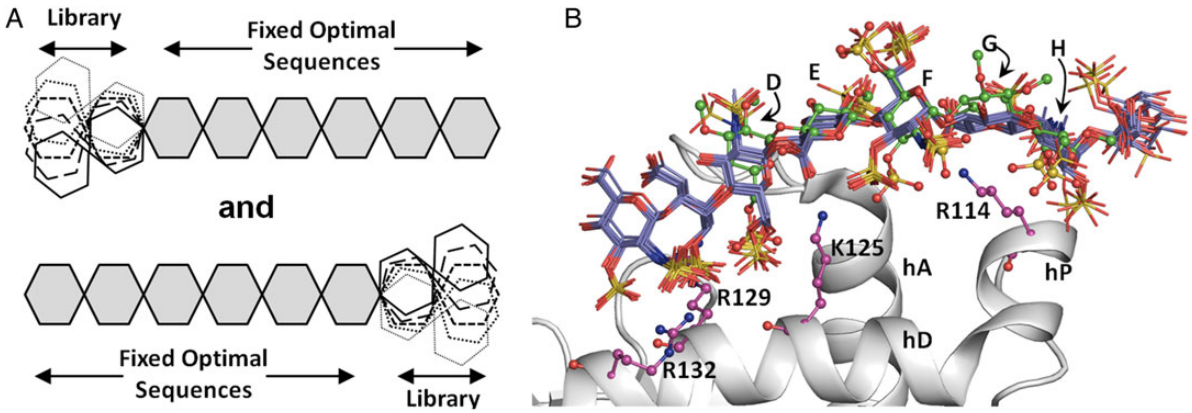

Fig. 7. (**A**) Incremental neighborhood optimization strategy used in the design of octasaccharide sequences. Five hexasaccharide sequences that were found to be most optimal from each of the two libraries (UA_{NRE} and $GlcN_{NRE}$) were selected and then 38 disaccharide sequences of UA_{NRE} and $GlcN_{NRE}$ libraries were appended at either the reducing or the non-reducing ends. The total of $5 \times 38 \times 4 = 760$ octasaccharide sequences were then studied using the dual-filter CVLS protocol. (**B**) Ten octasaccharide sequences from the $GlcN_{NRE}$ library and six sequences from the UA_{NRE} library satisfied the CVLS dual-filter criteria. Helices A (hA), D (hD) and P (hP) of antithrombin are shown in ribbon form and residues Arg132, Arg129, Lys125 and Arg114 are shown in ball and stick display. The crystal structure of DEFGH in green ball and sticks display is shown to highlight correspondence with the CVLS predicted poses.

To test this design algorithm, the docked poses of the five best ranked hexasaccharide sequences from the UA_{NRE} library, which satisfied the dual filters, were selected and 38 disaccharide blocks were combinatorially attached to either the NRE or the RE to derive two octasaccharide libraries of 190 sequences each (Figure 7A). Likewise, the same procedure was applied to the five best $GlcN_{NRE}$ hexasaccharide sequences to prepare two libraries of 190 sequences each containing disaccharide extensions at either the NRE or the RE. The octasaccharide sequences were built in an automated manner, their ring conformations and interglycosidic torsions assessed for consistency, and then sequences docked onto AT using the dual-filter CVLS protocol described in Figure 2. Following the application of the first filter, the 10 best ranked sequences of the 190 were selected for the consistency of binding analysis.

Although each sequence in these four libraries contained the DEFGH sequence, which is theoretically expected to bind to AT with 100% efficiency, the CVLS results suggest a striking preference for the type of library. No octasaccharide sequence satisfied the "specificity" filter for UA_{NRE} and $GlcN_{NRE}$ libraries to which disaccharides were added at the RE and NRE, respectively. In contrast, 6 and 10 octasaccharide sequences passed the consistency of binding filter from the UA_{NRE} and $GlcN_{NRE}$ libraries possessing disaccharides at the NRE and RE, respectively (Figure 7B, Table IV). The docked poses indicate that longer H/HS sequences possess additional interactions with residues of the extended heparin-binding site of AT, especially Arg132. More importantly, the results reveal for the first time the exact order of residues flanking DEFGH that contribute to AT binding and specificity (see Table IV).

The computational results also provide structural basis for biochemical studies observed earlier by Belzar et al. (2000). In their work, extension of the DEFGH sequence at the RE end was found to induce a frameshift in binding. We find that extension at

Table IV. H/HS octasaccharide sequences that satisfied the CVLS strategy for antithrombin

No.	Octasaccharide sequence[a]	GoldScore[b]	No. of H-bonds[c]
UA_{NRE} library			
1	Zb2B-Yb2A-ZbB-Yb26A-ZbB-Yb236A-ucA-Yb2A	145.53	16
2	ucA-Yb26A-uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	137.10	13
3	uc2A-Yb26A-uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	136.19	14
4	uaA-Yb236A-uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	134.42	14
5	uaA-Yb2A-uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	131.93	13
6	uc2A-YbH3A-uc2A-Yb26A-ZbB-Yb236A-ucA-Yb2A	130.27	12
$GlcN_{NRE}$ library			
1	Yb2A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-Yb26A-Zb2B	149.99	13
2	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-YbH3A-ua2A	149.62	13
3	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-YbCA-uaA	148.30	10
4	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-Yb26A-ua2A	147.56	14
5	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-Yb23A-uc2A	146.40	13
6	Yb2A-ua2A-YbC6A-ZbB-Yb236A-ZbB-Yb236A-ZbB	146.09	15
7	Yb2A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-Yb236A-ua2A	141.69	11
8	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-Yb23A-ZbB	141.67	12
9	Yb23A-ua2A-Yb26A-Zb2B-Yb236A-ZbB-YbCA-ZbB	138.82	11
10	Yb2A-ua2A-YbC6A-ZbB-Yb236A-ZbB-Yb26A-ua2A	136.58	9

[a]See definitions of residue labels and substitution in Table I.
[b]Refers to modified GoldScore, as defined in Methods.
[c]Number of hydrogen bonds calculated using LIGPLOT.

both the NR and RE ends induces an equivalent shift in the frame of the sequences when compared with pentasaccharide DEFGH (Supplementary data, Figure S2). This frameshift arises to induce better interaction of the extended sequence with AT, which measurably enhances the binding affinity. Also in a recent study, an octasaccharide sequence with two 3-*O*-sulfated GlcN*p* residues was reported to induce higher affinity to AT in comparison to pentasaccharide DEFGH (Guerrini et al. 2013). A review of our CVLS results (see Table IV) shows that nine such sequences were identified as the most optimal and specific octasaccharide sequences.

Thus overall, the success of the CVLS algorithm implies that this approach could be used to design longer H/HS sequences one disaccharide block at a time. This is the first simple and intuitive approach to computationally design longer GAG sequences. The approach obviates extensive library screening and may be termed as incremental neighborhood optimization strategy.

Application of the CVLS protocol to identify key heparin sequences binding to thrombin —

To assess the applicability of the CVLS approach to proteins other than AT, we studied the thrombin–heparin system. Thrombin is a key protease of the blood coagulation cascade and is widely recognized as interacting with heparin in a non-specific manner (Olson et al. 1991; Mosier et al. 2012). Heparin binds to thrombin in anion-binding exosite 2 (Carter et al. 2005). A crystal structure of heparin–thrombin complex has been reported ("1XMN"; Carter et al. 2005), which presents two different modes of binding within exosite 2 for the common heparin hexasaccharide, further alluding to lack of specificity presented in the literature.

Each biochemical study performed to date with the thrombin–heparin system has relied either on the most common heparin sequence, i.e. $(IdoAp2S\text{-}GlcNp2S6S)_n$, or on a mixture of unfractionated heparin sequences. We reasoned that because our CVLS approach affords a rigorous study of every sequence possible in nature, it would be better suited to assess elements of specificity in heparin binding to thrombin. Thus, the optimized CVLS approach deduced above was applied to thrombin using the library of >50,000 hexasaccharide sequences. Application of the "affinity" filter led to identification of the best 54 sequences (top 0.1%), which were re-docked to assess the consistency of binding. Only one sequence, i.e. IdoA*p*2S-GlcN*p*2S6S-IdoA*p*2S-GlcN*p*2S-IdoA*p*-GlcN*p*2S3S6S, satisfied the second filter (Table V). The pose of this sequence was compared with that of the sequence reported in the heparin–thrombin co-crystal structure and found to be essentially identical (Figure 8A). Thus, the CVLS technology was able to correctly predict the preferred sequence and its binding geometry. It is instructive to note that the GOLDScore for this sequence is only ~120 (Table V), whereas in comparison, the GOLDScore for DEFGH containing sequences binding to AT is much higher (Table III). This implies that the predicted affinity of the thrombin binding sequence is relatively low.

Table V. H/HS hexasaccharide sequences that satisfied the CVLS strategy for thrombin

No.	Hexasaccharide sequence[a]	GoldScore[b]	No. of H-bonds[c]
1XMN			
1	ua2A-Yb26A-uc2A-Yb2A-uaA-Yb236A	121.48	14
1TB6			
1	ua2A-Yb23A-uc2A-YbC6A-ua2A-Yb26A	136.13	13
2	ua2A-Yb26A-ucA-Yb23A-ua2A-Yb26A	130.38	13
3	ua2A-YbC6A-uc2A-YbC6A-ua2A-YbH36A	129.36	12
4	ua2A-YbCA-uc2A-Yb26A-ua2A-YbH36A	128.91	14

[a]See definitions of residue labels and substitution in Table I.
[b]Refers to modified GoldScore, as defined in Methods.
[c]Number of hydrogen bonds calculated using LIGPLOT.

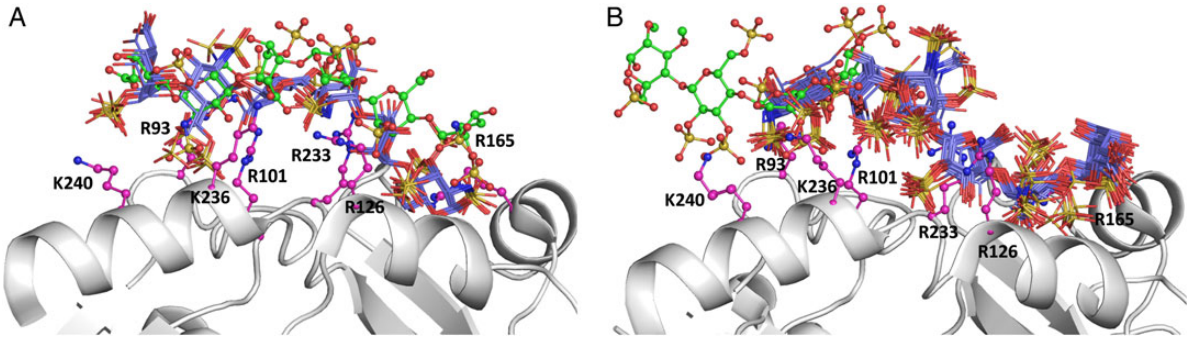

Fig. 8. CVLS predicted hexasaccharide sequences that bind with a consistent mode onto exosite 2 of thrombin. (**A**) Multiple poses of the only sequence (blue sticks) that satisfied the dual-filter CVLS strategy for 1XMN crystal structure of thrombin and (**B**) poses of the four sequences (blue sticks) identified for the 1TB6 structure. Thrombin surface and key exosite 2 residues are shown in ribbon and ball and stick representations, respectively. Heparin present in the co-crystal structure is shown as green ball and sticks.

To assess these results further, we applied the CVLS technology to the 1TB6 thrombin structure (Li et al. 2004). This structure presents thrombin in a ternary complex with AT and heparin. A key aspect of this structure is that thrombin is bound to a polymeric heparin sequence and therefore presents slightly altered exosite 2 electrostatic surface features (not shown). Once again, the best 54 "affinity" filtered sequences from the >50,000 studied in the first stage were processed for the consistency of binding in multiple docking experiments. Only four sequences were identified in this study (Figure 8B, Table V). All the four sequences bind with relatively high consistency but weaker GOLDScore. The binding pose of these four sequences matches well with that of highly sulfated heparin monomers observed in the 1TB6 structure. The hydrogen-bonding analysis using LIGPLOT (Wallace et al. 1995) suggested favorable interactions of the sequences with key residues of exosite 2 including Arg93, Arg101, Arg126, Arg165, Arg233, Lys236 and Lys240, as expected. Interestingly, all four hexasaccharide sequences identified by CVLS displayed a repeating disaccharide unit of $(IdoA\ –GlcN)_3$ with variations in the position of sulfate groups. This is essentially identical to the common repeating sequence present in unfractionated heparin. Thus, the heparin–thrombin system turns out to be chemically non-specific, although biologically, a distinct mode of sulfated hexasaccharide binding in exosite 2 is evident.

Significance

Our interest in designing and exploring the huge database of H/HS sequences arose from our previous work (Raghuraman et al. 2006), where we used a limited hexasaccharide library (6859 sequences). Most docking approaches to date focus primarily on the affinity of interaction and minimally on the specificity of interaction (Kitchen et al. 2004). Our genetic algorithm-based approach places major emphasis on the specificity of interaction, which is more challenging to determine for H/HS. Our approach allows screening of considerably large conformational space and attempts to reduce the number of false positives.

This work represents the first, large-scale combinatorial GAG library screening to date. Our work demonstrates that such large library screening is feasible for GAG sequences, especially if a high-resolution crystal structure of the target protein is available. Considering the success achieved for AT and thrombin, two prototypic "highly specific" and "non-specific" GAG-binding proteins, respectively, we expect that CVLS approach may be more generally applicable to other proteins, especially for other serpins such as heparin cofactor II and protein C inhibitor, for which specificity features remain poorly defined. Theoretically, the CVLS technology should be applicable to any GAG–protein system assuming that appropriate validating solution experiments can be performed to assess predictions.

Our CVLS strategy utilized fairly stringent criteria for selection. While the affinity filter selected the upper 0.1% of sequences, the specificity filter was set to select only those sequences that satisfy self-consistency 100% of the time. It was possible to use high filtering stringency because the AT–H/HS system is a biochemically well-studied system. For other less understood systems, such stringent criteria may eliminate potentially useful information, which implies that appropriate relaxation in criteria may be necessary to introduce. One important deduction from our work is that GAG docking onto proteins should be approached with caution and care. We cannot assume that "one-size-fits-all" approach typically used in analyzing protein–ligand interactions will work well for GAG–protein studies as demonstrated by the observation that a higher docking radius does not necessarily result in higher probability of an outcome. This caution becomes even more important for relatively non-specific GAG–protein interactions for which it becomes difficult to estimate the validity of a result.

This work also presents the first algorithm, the incremental neighborhood optimization strategy, to design longer GAG sequences. This approach significantly reduces the computational cost and enhances efficiency over de novo design of a longer GAG sequence. Yet, the approach is expected to work for GAG–protein systems that exhibit high specificity. For non-specific system, it remains to be seen whether the incremental strategy provides meaningful results.

Finally, we expect our CVLS technology to be especially useful in the design of pharmaceutically useful agents. For example, the deduction of minimal "EF" disaccharide sequence as the origin of specificity implies that small GAG mimetics

containing the EF domain should function as specific AT activators. Such small GAG mimetics should be possible to design computationally.

Supplementary data

Supplementary data for this article are available online at http://glycob.oxfordjournals.org/.

Funding

The computational facilities are provided by National Center for Research Resources (award Number S10RR027411). The work is supported by a research grant P01 HL107152 provided to URD from National Heart, Lung and Blood Institute.

Acknowledgements

We thank Drs Aurijit Sarkar and Philip Mosier of VCU for helpful discussions and suggestions.

Conflict of interest statement

None declared.

Abbreviations

AT, antithrombin; GAG, glycosaminoglycan; GlcA*p*, glucuronic acid; GlcN*p*, glucosamine; H, heparin; HS, heparan sulfate; IdoA*p*, iduronic acid;NRE, non-reducing end; RMSD, root-mean-square difference; SPL, SYBYL Programming Language; UA*p*, uronic acid.

References

Agostino M, Gandhi NS, Mancera RL. 2014. Development and application of site mapping methods for the design of glycosaminoglycans. *Glycobiology* (in press).

Belzar KJ, Dafforn TR, Petitou M, Carrell RW, Huntington JA. 2000. The effect of a reducing-end extension on pentasaccharide binding by antithrombin. *J Biol Chem*. 275:8733–8741.

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.

Capila I, Linhardt RJ. 2002. Heparin-protein interactions. *Angew Chem Int Ed*. 41:391–412.

Carter WJ, Cama E, Huntington JA. 2005. Crystal structure of thrombin bound to heparin. *J Biol Chem*. 280:2745–2749.

Copeland R, Balasubramaniam A, Tiwari V, Zhang F, Bridges A, Linhardt RJ, Shukla D, Liu J. 2008. Using a 3-O-sulfated heparin octasaccharide to inhibit the entry of herpes simplex virus type 1. *Biochemistry*. 47: 5774–5783.

Cremer D, Pople JA. 1975. A general definition of ring puckering. *J Am Chem Soc*. 97:1354–1358.

Desai UR. 2005. Antithrombin activation and designing novel heparin mimics. In: Garg HG, Linhardt RJ, Hales CA, editors. *Chemistry and Biology of Heparin and Heparan Sulfate*. Amsterdam: Elsevier Science. p. 483–512.

Desai UR. 2013. The promise of sulfated synthetic small molecules as modulators of glycosaminoglycan function. *Future Med Chem*. 5:1363–1366.

Desai UR, Petitou M, Bjork I, Olson ST. 1998. Mechanism of heparin activation of antithrombin. Role of individual residues of the pentasaccharide activating sequence in the recognition of native and activated states of antithrombin. *J Biol Chem*. 273:7478–7487.

Esko JD, Selleck SB. 2002. Order out of chaos: Assembly of ligand binding sites in heparan sulfate. *Annu Rev Biochem*. 71:435–471.

Forster MJ, Mulloy B. 1993. Molecular dynamics study of iduronate ring conformations. *Biopolymers*. 33:575–588.

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

Grootenhuis PDJ, van Boeckel CAA. 1991. Constructing a molecular model of the interaction between antithrombin III and a potent heparin analogue. *J Am Chem Soc*. 113:2743–2747.

Guerrini M, Elli S, Mourier P, Rudd TR, Gaudesi D, Casu B, Boudier C, Torri G, Viskov C. 2013. An unusual antithrombin-binding heparin octasaccharide with an additional 3-*O*-sulfated glucosamine in the active pentasaccharide sequence. *Biochem J*. 449:343–351.

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

Johnson DJ, Li W, Adams TE, Huntington JA. 2006. Antithrombin-S195A factor Xa-heparin structure reveals the allosteric mechanism of antithrombin activation. *EMBO J*. 25:2029–2037.

Jones G, Willett P, Glen RC, Leach AR, Taylor R. 1997. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 267:727–748.

Kirschner KN, Yongye AB, Tschampel SM, Gonzalez-Outeirino J, Daniels CR, Foley BL, Woods RJ. 2008. GLYCAM06: A generalizable biomolecular force field. Carbohydrates. *J Comput Chem*. 29:622–655.

Kitchen DB, Decornez H, Furr JR, Bajorath J. 2004. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Disc*. 3:935–949.

Li W, Johnson DJ, Esmon CT, Huntington JA. 2004. Structure of the antithrombin-thrombin-heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nat Struct Mol Biol*. 11:857–862.

Maimone MM, Tollefsen DM. 1990. Structure of a dermatan sulfate hexasaccharide that binds to heparin cofactor II with high affinity. *J Biol Chem*. 265:18263–18271.

McCoy AJ, Pei XY, Skinner R, Abrahams JP, Carrell RW. 2003. Structure of beta-antithrombin and the effect of glycosylation on antithrombin's heparin affinity and activity. *J Mol Biol*. 326:823–833.

Mosier PD, Krishnasamy C, Kellogg GE, Desai UR. 2012. On the specificity of heparin/heparan sulfate binding to proteins. Anion-binding sites on antithrombin and thrombin are fundamentally different. *PLoS ONE*. 7:e48632.

Mulloy B, Forster MJ. 2000. Conformation and dynamics of heparin and heparan sulfate. *Glycobiology*. 10:1147–1156.

Olson ST, Halvorson HR, Björk I. 1991. Quantitative characterization of the thrombin-heparin interaction. Discrimination between specific and nonspecific binding models. *J Biol Chem*. 266:6342–6352.

Petitou M, Barzu T, Herault JP, Herbert JM. 1997. A unique trisaccharide sequence in heparin mediates the early step of antithrombin III activation. *Glycobiology*. 7:323–327.

Pike RN, Buckle AM, le Bonniec BF, Church FC. 2005. Control of the coagulation system by serpins. Getting by with a little help from glycosaminoglycans. *FEBS J*. 272:4842–4851.

Pol-Fachin L, Verli H. 2008. Depiction of the forces participating in the 2-O-sulfo-alpha-L-iduronic acid conformational preference in heparin sequences in aqueous solutions. *Carbohydr Res*. 343:1435–1445.

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.

Raghuraman A, Mosier PD, Desai UR. 2010. Understanding dermatan sulfate-heparin cofactor II interaction through virtual library screening. *ACS Med Chem Lett*. 1:281–285.

Shriver Z, Capila I, Venkataraman G, Sasisekharan R. 2012. Heparin and heparan sulfate: Analyzing structure and microheterogeneity. *Handb Exp Pharmacol*. 207:159–176.

Wallace AC, Laskowski RA, Thornton JM. 1995. LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*. 8:127–134.