Published in final edited form as:

Methods Mol Biol. 2015 ; 1229: 289–314. doi:10.1007/978-1-4939-1714-3_24.

Designing ‘High-Affinity, High-Specificity’ Glycosaminoglycan Sequences Through Computerized Modeling

Nehru Viji Sankaranarayanan, Aurijit Sarkar, Umesh R. Desai, and Philip D. Mosier*

Department of Medicinal Chemistry, Virginia Commonwealth University, Richmond, VA, 23298, USA

Abstract

The prediction of high-affinity and/or high-specificity protein–glycosaminoglycan (GAG) interactions is an inherently difficult task, due to several factors including the shallow nature of the typical GAG binding site and the inherent size, flexibility, diversity and polydisperse nature of the GAG molecules. Here, we present a generally-applicable methodology termed Combinatorial Library Virtual Screening (CVLS) that can identify potential high-affinity, high-specificity protein–GAG interactions from very large GAG combinatorial libraries and a suitable GAG-binding protein. We describe the CVLS approach along with the rationale behind it and provide validation for the method using the well-known antithrombin–thrombin–heparin system.

Keywords

Glycosaminoglycan (GAG); Docking; Virtual Library; Virtual Screening (VS); Genetic Algorithm (GA); Molecular Interaction; Specificity

1. Docking of GAGs

Carbohydrates are being increasingly recognized as key players in a number of biological responses. Glycosaminoglycans (GAGs) are a special class of carbohydrates consisting of repeating hexosamine sugar and uronic acid/galactose units that are at the center of a number of these functions [1]. Most GAGs are sulfated, as exemplified by heparin (H), heparan sulfate (HS), chondroitin sulfate (CS) and dermatan sulfate (DS). Pattern of sulfation on GAGs are thought to be important for differential recognition of proteins, however, the extent of sequence-specific function has not been fully established [2].

The specificity of interaction between GAGs and proteins (or lack thereof) is the driving force for several pathways. For example, H/HS binds to antithrombin (AT) with high affinity and high specificity, while its interaction with thrombin (T) is known to be non-specific [3,4]. The AT–H/HS-mediated inhibition is a three-step procedure (Fig. 1). It begins with AT recognition by a specific pentasaccharide sequence on a long H/HS chain, known as the DEFGH sequence (or simply H5), leading to formation of a tight AT–H/HS complex, thus immobilizing AT and causing a conformational change that spatially orients AT. This is

*Address for Correspondence: Dr. Philip D. Mosier, Institute for Structural Biology and Drug Discovery, 800 E. Leigh Street, Suite 212, Richmond, VA 23219. Ph. (804) 828-7405; Fax (804) 827-3664; pdmosier@vcu.edu.

followed by a low-affinity (mostly ionic) binding of T at a distal site on the long H/HS chain, and interaction that is thought to be largely non-specific. This low-affinity and low-specificity interaction between H/HS and T allows the latter to “walk” along the polymeric chain and reach AT. In the final step, the AT–H/HS–T ternary complex results in inactivation of T *via* a covalent linkage between the AT reactive center loop (RCL) and S195 of the T catalytic triad. Thus, both the specificity of H/HS–AT interaction and the relative non-specificity of H/HS–T interaction are necessary for inhibition of T by AT.

The growing significance of GAGs in biology and the expectation that their specific interactions with proteins are the reason behind physiologic and pathologic responses, it is imperative to develop reliable methodologies for studying these phenomena. While approaches for studying carbohydrate–protein interactions have expanded considerably in the recent past [2], analytical approaches to study GAGs are made difficult by the very sulfate groups that bestow physiological functions to them. The high negative charge density makes it difficult to isolate pure, homogenous GAG sequences, which are mostly isolated as mixtures of chains with varied molecular weights and sulfation patterns. This poses challenges in elucidating their structural biology because techniques such as NMR and X-ray crystallography are highly dependent on the availability of pure homogenous samples. Chemical approaches for GAG synthesis have yielded only a limited number of homogeneous GAG sequences. This implies that the plethora of structurally diverse GAG sequences present in nature remain unstudied to date using conventional methodologies.

An atomistic dissection of GAG interactions with proteins is critical to fully elucidate the role of these important biopolymers, which may also lead to GAG sequences as therapeutic agents. Such information is typically provided by theoretical computational studies. Computational molecular modeling methodology typically addresses four major questions: (1) Which GAG sequences are likely to interact with which proteins in a specific manner? (2) What are their likely relative affinities? (3) What is the likely GAG binding site on proteins? and (4) What is the geometry of GAG binding in the binding site? We present the antithrombin–pentasaccharide system as a test case of the methodology developed in the past few years.

The earliest work in prediction of GAG-binding site on proteins was performed by Cardin and Weintraub [5], which was followed by other investigations [6-8]. Analysis of GAG-binding proteins revealed the presence of consensus primary amino acid sequences in some proteins. Other factors were also found to influence protein–GAG binding including the role of secondary and tertiary structure, the spatial separation and type (Arg/Lys) of basic amino acids, and surface exposure of amino acids containing donor nitrogens. In other words, the identification of GAG-binding sites was more related to the 3-D stereoelectronic composition of the protein surface and its complementarity to the GAG topology than its primary amino acid sequence. Automated docking and scoring has been used to predict how GAGs bind to proteins [9-14]. Many different docking routines have been explored including GOLD [15], AutoDock [16] and HADDOCK [17]. However, several challenges have to be addressed to ensure meaningful results. First, the conformational flexibility in GAG chains is much higher than in typical drug-like small molecules. In general, the effectiveness of docking protocols decreases exponentially with the number of rotatable

bonds, which implies that such methods do work well with longer GAG sequences. This situation is further complicated by amino acid side chain flexibility and exacerbated by higher-order protein structural changes that may occur upon GAG binding. Second, commercially available docking routines are primarily designed to predict how small molecules bind within well-defined deep pockets, whereas GAGs almost exclusively bind to shallow surfaces on proteins. Thus, docking of GAGs almost always begins by reducing the search space. For example, conformational space search is reduced by disallowing rotation around the glycosidic linkages and limiting sampling of ring conformations. Alternatively, an initial rigid body docking may be used as a screen to identify certain oligosaccharide sequences of interest before flexible docking is performed. A distinct advantage of the automated docking paradigm is that it can be made to run reasonably fast, thus allowing simultaneous virtual screening (VS) of many GAG sequences.

Prediction of binding affinities of GAG sequences for proteins may be accomplished either through the use of scoring functions or molecular dynamics (MD) simulations. An in-depth discussion of MD is beyond the scope of this work, but has been addressed elsewhere [18]. The presence of sulfates leads to the almost exclusive interaction of GAGs with positively charged areas of protein surfaces. This is a difficult hurdle in the path towards successful prediction of protein–GAG binding affinities because GAG-associated sulfates have, as of yet, not been especially well parameterized. Yet, force fields such as GLYCAM/AMBER [19,20], CHARMM [21–23] and GROMOS96 [24] do exist for carbohydrate MD simulations. Several MD simulations have been conducted on GAGs alone, and also protein–GAG systems [25,11,26–30]. Gandhi and Mancera [31] and Pichert, et al. [32] have also used MD simulations, combined with MM-PBSA (Molecular Mechanics Poisson Boltzmann Surface Area) and MM-GBSA (Molecular Mechanics General Born Surface Area) methods [33–35] to predict affinities of protein–GAG interactions.

A major advance would be enabled if a methodology that identifies likely ‘*high-affinity, high-specificity*’ GAG sequences is developed. Here, we present an algorithm for ‘high-affinity, high-specificity’ GAG sequences based on docking and scoring of virtual GAG libraries. We describe the steps involved in using this protocol with supporting rationale. Validation of the method is presented for antithrombin with heparin, the prototypical protein–GAG interaction in which a specific H/HS sequence is recognized.

2. Methodology

2.1 Introduction

The method presented here, known as **CVLS** (Combinatorial Virtual Library Screening), is based loosely on a two-stage virtual screening paradigm (Fig. 2). In the first stage or ‘filter’, the docking is performed relatively quickly, with 10,000 genetic algorithm (GA) iterations per docking run. In this way, many tens or hundreds of thousands of GAG sequences may be screened efficiently (on a large cluster), albeit at the possibility of few false negatives. In the second filter, a small fraction of the highest-scoring sequences are then re-docked under more rigorous conditions (100,000 GA iterations) and in triplicate. The top two highest-scoring solutions from each of the three docking runs are then compared. If these six solutions form a sufficiently tight cluster (backbone atom RMSD ≤ 2.5 Å; see Section 2.2.2

for backbone atom definition), the GAG sequence passes the second filter. Using the backbone atoms, which collectively serve as a maximum common subset of atoms common to all GAG sequences, allows sequences with different substituent patterns and conformations to be meaningfully compared.

A key operating principle in CVLS is that if the protein–GAG interaction is specific (meaning that the protein recognizes a specific GAG conformation and pattern of substituents), then the docking routine will return solutions that are consistently docked in a similar fashion (i.e. the RMSD among all solutions is sufficiently low). By this method, then, potentially ‘high affinity’ and ‘high specificity’ sequences can effectively be identified. It should be noted that if the size of the virtual library is sufficiently small, or the user's computing resources are sufficiently large, the first filter may be effectively removed, thus attenuating the false negative rate associated with the short-duration docking runs.

To provide an example that validates the method, a H/HS virtual library consisting of 46,656 hexasaccharide sequences is constructed and subjected to CVLS against the activated antithrombin structure from Protein Data Bank (PDB) ID = 1TB6 [36]. It is shown that the method is able to reproduce the heparin pentasaccharide also found in 1TB6. Although a specific example is provided here, it is also important to note that this methodology is potentially applicable to *any* class of GAG and *any* GAG-binding protein with an identified GAG-binding site whose coordinates are determined either experimentally or through homology modeling.

The CVLS algorithm as described here employs two well-established molecular modeling software packages, although other software packages and/or options could be used to achieve similar results. SYBYL-X (Tripos Associates, St. Louis, MO, hereafter referred to as SYBYL), its general-purpose molecular mechanics force field (Tripos Force Field; TFF), its associated scripting language known as SYBYL Programming Language (SPL) and its native file format (mol2) are used in all phases of the CVLS algorithm except for automated docking, which is handled by GOLDSuite (CCDC, Cambridge, UK, hereafter referred to as GOLD) [37] and its native scoring function (GOLDScore).

2.2. Building H/HS Virtual Libraries

2.2.1. Nomenclature of Monosaccharide Units and Disaccharide Building

Blocks—H/HS are composed of repeating disaccharide units consisting of β -D-glucuronic acid or α -L-iduronic acid and α -D-glucosamine residues (Fig. 3) [38]. Each monosaccharide unit is variably substituted with *N*-acetyl, *N*-sulfate and/or *O*-sulfate groups [39], giving rise to sequences that are phenomenally diverse. As H/HS sequences can be long, a simple, intuitive, extensible and compact naming convention for the sequences is required. Such a system was created based loosely on the symbolic system employed in GLYCAM [19]. The letter ‘Z’ is used for D-glucuronic acid, ‘u’ for L-iduronic acid and ‘Y’ for D-glucosamine. Similarly, ring conformations are encoded: ‘a’ for 1C_4 ; ‘b’ for 4C_1 and ‘c’ for 2S_0 (designations for ring conformation are derived from the Cremer–Pople ring puckering [40–42]). Glucosamine substituents are ‘H’ for an unsubstituted 2-position (NH_3^+), ‘C’ for *N*-acetyl, ‘2’ for *N*-sulfate, ‘3’ for 3-*O*-sulfate and ‘6’ for 6-*O*-sulfate. Uronic acid

substituents: 2 = 2-*O*-sulfate. Anomeric carbon configurations are encoded: 'A' for α and 'B' for β . These H/HS monosaccharide names used here are shown in Table 1.

2.2.2. Atom Numbering Convention—The relative orientation of a pair of adjacent monosaccharide residues is defined by two torsional angles at the glycosidic linkage, denoted by ϕ and ψ [43].

$$\phi = \text{O5} - \text{C1} - \text{O1} - \text{C4}' \quad (1)$$

$$\psi = \text{C1} - \text{O1} - \text{C4}' - \text{C5}' \quad (2)$$

The glycosidic linkage torsion angles are illustrated in Fig. 4. To facilitate the constrained docking of the H/HS oligosaccharides (*vide infra*), the atoms in the glycosidic linkage are systematically renumbered so that the glycosidic linkage atoms in each disaccharide have consistent numbering; i.e. O5 becomes atom number 5, C1 becomes atom number 6, O1 becomes atom number 7, C4' becomes atom number 8 and C5' becomes atom number 9 (see Fig. 5). In the complete H/HS oligosaccharide sequence, the atoms are analogously renumbered such that the atoms designating the glycosidic linkage span a continuous range from 1 to $n \times 5 + 3$, where n is the number of monosaccharides in the oligosaccharide sequence. The renumbering procedure is facilitated through the use of in-house SPL scripts.

SYBYL Line Notation (SLN), similar to Simplified Molecular-Input Line-Entry System (SMILES) [44], provides a convenient way to specify the unique topology of a molecule *via* a string of text and is the basis of powerful substructure search functionality within SYBYL. The numbering scheme and SLN used here to represent the H/HS hexasaccharide backbone atoms is shown in Fig. 5. The SLN representation of the backbone for any 1→4-linked GAG with n monomers may be created using a loop that runs from $i = 1$ to $(n-1)$ with an initial string 'HOC[3]COC', followed by $n-1$ copies of a second string '(OC[$i \times 5 + 3$]COC', an intermediate string '(OH)', $n-1$ copies of a fourth string 'CC@ $i \times 5 + 3$ ', and ending with a terminal string 'CC@3'. This convention allows facile calculation of the atom numbers in the SLN corresponding to the atoms defining the $n-1$ glycosidic linkage torsion angles in the GAG (colored red in Fig. 5). Additionally, the atom numbers of other backbone atoms may be generally specified, including the non-reducing end hydroxyl group oxygen (atom number 2) and hydrogen (atom number 1) and the reducing end hydroxyl group oxygen (atom number $n \times 5 + 2$) and hydrogen (atom number $n \times 5 + 3$).

2.2.3. Explicit Modeling of Ring Conformations—Ring conformations in the modeled GAGs are kept rigid to greatly simplify the docking of GAGs during the actual docking process. To this end, the pyranose rings of glucuronic acid (GlcA)- and glucosamine (GlcN)-based monosaccharides are assigned the ${}^4\text{C}_1$ conformation, while the rings of iduronic acid (IdoA)-based monosaccharides, which have two primary low-energy conformations, are modeled with either the ${}^1\text{C}_4$ or the ${}^2\text{S}_0$ conformation.

2.2.4. Atom and Bond Types and Representation of Functional Groups—The proper assignment of atom and bond types for the atoms and bonds of the GAG sequences is

very important for the automated docking program to correctly pair the atoms of the protein with the atoms of the GAG. GOLD, the docking program used here, reads files in the SYBYL mol2 format and also uses SYBYL atom and bond types internally; these will be described here. Unless noted otherwise, all bond types are '1' (single bonds).

The atom and bond types of the protein are taken from the input protein mol2 file as assigned automatically by SYBYL and should not normally be changed. For the GAG sequences, the backbone atom types are 'C.3' for the carbon atoms and 'O.3' for the oxygen atoms. For the free hydroxyl groups, the oxygen atoms are also of type 'O.3' and the hydrogen atoms have type 'H'. Both carboxylate and *O*-sulfate groups distribute their anionic charge over multiple atoms through delocalization. This in turn affects the electrostatic strength and nature of the hydrogen bonding ability of these groups. To ensure that the molecular modeling and docking programs correctly represent these interactions, SYBYL atom types are assigned such that the carboxylate carbon atoms have type 'C.2', *O*-sulfate sulfur atoms have type 'S.o2', the oxygen atom connecting the *O*-sulfate group to the GAG backbone has type 'O.3', and the terminal oxygen atoms of both carboxylate and *O*-sulfate groups have type 'O.co2'. In addition, the bonds connecting either the carboxylate carbon or the *O*-sulfate sulfur atom to each of the terminal oxygen atoms are of type 'ar' (aromatic) to denote the delocalized nature of the electrons in these groups. The sulfur–oxygen bond length should be 1.5 Å, as observed in high-resolution X-ray crystal structures. For primary amine groups, the nitrogen atom will usually be protonated at physiological pH and as such will have type 'N.4' with three hydrogen atoms of type 'H' attached. The *N*-acetyl group is an amide: The nitrogen is type 'N.am', the carbonyl carbon and oxygen atoms are 'C.2' and 'O.2' respectively and the CH₃ group has atoms of type 'C.3' and 'H'. The *N*-sulfate group nitrogen is 'N.3' and the remaining atom and bond types are as described for the *O*-sulfate group.

2.2.5. The 'Average Backbone' Approximation—The H/HS molecules possess both unusual rigidity and flexibility [39]. The unusual rigidity is due to the partial rotational freedom of the glycosidic linkage torsion angles; the majority of the flexibility in H/HS is due to the flexible ring conformation of IdoA, which can easily achieve both ¹C₄ and ²S_O conformations at equilibrium. A GAG backbone can be defined as the atoms of the pyranose rings and the glycosidic linkage atoms. On analyzing various available crystal structures [45–47] and also from the reports of various research groups [43], it was found that the glycosidic torsion angles fall within a relatively narrow range and suggest that they remain fairly constant irrespective of a change in substitution pattern around the inter-glycosidic bond. Thus, the 'average backbone' approximation, which utilized an average value of torsion for each inter-glycosidic linkage angle ϕ and ψ (Table 2), was employed. This greatly reduced the number of rotatable bonds in the GAG oligosaccharides, making the otherwise difficult or impossible task of docking of longer H/HS sequences (up to and including hexasaccharides) feasible. Also, in the process of building the H/HS sequences (both disaccharide and longer chains), energy-minimization was performed during which the average ϕ and ψ values were subjected to a restraining force constant of 0.01 kcal·mol⁻¹·deg⁻²

2.2.6. Building Monosaccharide Units—Construction of the monosaccharides (Table 1) may be accomplished in a number of ways, including sketching the structure manually and energy-minimizing or modifying a previously-determined set of atomic coordinates (*e.g.* from a crystal structure). A typical way of building the monosaccharide units “from scratch” using SYBYL will be briefly explained here. When SYBYL starts its main window is presented. The graphics window contains the menu bar, toolbar icons, and the display area. The console is docked below the display area. The general method of creating a new small molecule structure is to click ‘Edit > Molecule > Sketch’ to start the sketcher. Draw the structure (heavy atoms only) as a carbon skeleton in Draw mode, and then change the appropriate carbons to the correct heteroatom (nitrogen, oxygen or sulfur) using the Modify Atom Type mode. Exit the sketcher. Using the Modify Atom Type function from the main menu (‘Edit > Atom > Modify Atom Type’), change the SYBYL atom type of the atoms as described in Section 2.2.4 where necessary. Do the same for the SYBYL bond types (‘Edit > Bond > Modify Type’). Add hydrogen atoms and name the molecule (‘Edit > Hydrogens > Add All Hydrogens’). Assign a substructure name to the monosaccharide as shown in the first column of Table 1 (MODIFY SUBSTRUCTURE NAME command). Though not strictly necessary, naming the individual atoms (*e.g.* S3O for the 3-position *O*-sulfate sulfur atom) uniquely identifies each atom in a meaningful way (MODIFY ATOM NAME command). Energy-minimize the monosaccharide structures using the TFF with Gasteiger–Hückel charges, a fixed dielectric constant of 80 (to simulate an aqueous environment), and a non-bonded cutoff radius of 8 Å. Set the termination criteria for a maximum of 100,000 iterations subject to an energy gradient of 0.05 kcal/(mol·Å). Finally, export the structure as a mol2 file with the name of the monosaccharide contained in the file name.

Building β -D-glucuronic acid (GlcA) and its derivatives: Although it is possible to build the monosaccharide units ‘from scratch’, it is generally easier and less error-prone to start with an experimentally determined structure from the PDB. To build the β -anomer of D-glucuronic acid (GlcA; ZbB) in the 4C_1 conformation perform the following steps: Import the 1TB6 crystal structure into SYBYL. Select the substructure named GU1_445 corresponding to the ‘E’ residue of the H5 sequence and copy it to an empty molecule area (‘Edit > Copy’). Ensure that the atom and bond types are correct and modify as necessary. Select the atoms of the methyl groups attached to the 2- and 3-position oxygen atoms and delete them (‘Edit > Delete > Selected’) and then add hydrogens. Select the newly added hydrogen atom corresponding to the C4 oxygen atom and change its atom type to O.3. Add hydrogen atoms again to add the missing hydrogen. Name the substructure ‘ZbB’ and optionally name the atoms as described previously. Export the structure as ‘ZbB.mol2’. For D-glucuronic acid 2-sulfate (GlcA(2S); Zb2B), follow the same procedure except instead of removing both methyl groups at the 2- and 3-positions, remove only the one at the 3-position. Modify the 2-position methyl group carbon atom to type S.o2 and the methyl group hydrogens to type O.co2. Modify the S–O bond types to ‘ar’ if necessary and then add hydrogens.

Building α -L-iduronic acid (IdoA) and its derivatives: IdoA is capable of adopting equilibrium between the chair (1C_4) and skew-boat (2S_0) forms, the two most prevalent conformations known in H/HS oligosaccharides. The H/HS library takes into account both

conformations explicitly. For the 1C_4 conformation of IdoA, import the 1HPN structure from the PDB (model number 2) into SYBYL. Select an IdoA monomer and copy it to an empty molecule area. Continue using process similar to that described above for the GlcA-based structures to generate the uaA and ua2A monosaccharides. To generate the monomer in 2S_O conformation, import model number 1 from 1HPN, Select an IdoA monomer and copy it to an empty molecule area. Proceed using the same procedure as described for the 1C_4 conformation to generate the ucA and uc2A monosaccharides.

Building α -D-glucosamine (GlcN) and its derivatives: To build the α -anomer of D-glucosamine 2,3,6-sulfate (GlcNS(3S,6S); Yb236A) in the 4C_1 conformation perform the following steps: From the 1TB6 crystal structure, select the substructure named GU6_446 corresponding to the 'F' residue of the H5 sequence and copy it to an empty molecule area ('Edit > Copy'). Modify the atom and bond types as described in section 2.2.4 and then add hydrogens. Select the newly added hydrogen atom corresponding to the C4 oxygen atom and change its atom type to O.3. Add hydrogen atoms again to add the missing hydrogen. Name the substructure 'Yb236A' and optionally name the atoms as described previously. Export the structure as 'Yb236A.mol2'. Continue modifying this structure and exporting for each of the eight remaining GlcN-based monosaccharides listed in Table 1.

When complete, the library should consist of all the monomers listed in Table 1 with appropriate molecule name, and exported as separate SYBYL mol2 files. It is critical when generating the H/HS 'building blocks' that the configuration and conformation of the mono- and disaccharide units are correct. Whichever method is employed to generate the monosaccharide structures, ensure that the substitution pattern, configuration and conformation of the structures are correct either by visual inspection or (preferably) using SPL scripts that will facilitate this process with greater simplicity and reliability.

2.2.7. Combining Monosaccharides to Form Disaccharides—A virtual library of the twenty-three HS disaccharides observed in nature [48] can now be created. The coordinates for the disaccharide sequences can be generated by using the JOIN command in SYBYL where the reducing end OH group attached to the C1 atom of the first monomer (a uronic acid derivative) is attached to the non-reducing end OH group attached to the C4 atom of the second monomer (a glucosamine derivative), thereby forming the 1 $\rightarrow$ 4 glycosidic linkage and removing H₂O in the process. Use the MODIFY TORSION command in SYBYL to assign the average backbone values to the glycosidic linkage torsions as shown in Table 2. Name the disaccharides so that the names reflect the sequence (e.g. the disaccharide consisting of uc2A and Yb23A would be named uc2A-Yb23A) and save them as separate mol2 files. The completed disaccharide 'building block' library should contain all of the naturally-occurring disaccharide units as listed in Table 3.

2.2.8. Building Longer Sequences—The disaccharide library has 36 unique sequences [48]. These 36 sequences will be used to generate a combinatorial virtual library of H/HS hexasaccharides. The pseudocode routine shown in Fig. 6 outlines the steps required to create the library and will generate hexasaccharide H/HS sequences for all permutations of disaccharides **r**, **s** and **t** using steps similar to those described previously through the use of an SPL script. The individual disaccharide units are joined and the glycosidic torsion angles

are set to the ‘average backbone’ values using a procedure analogous to that described in Section 2.2.7. The SLN query in Fig. 5B is then used to identify the atoms of the hexasaccharide backbone and the backbone atoms are renumbered as shown in Fig. 5A, with the atoms of the glycosidic linkages being numbered sequentially. Next, the hexasaccharides are energy-minimized using a two-stage minimization process. In the first stage, the atoms of the backbone are kept rigid while the pendant groups are allowed move in order to alleviate major repulsive steric interactions among them. Minimization parameters are as described in Section 2.2.6. In the second stage, the minimization procedure is repeated but this time all atoms are allowed to move, with only a small torsional constraint of 0.01 kcal-mol⁻¹-deg⁻² applied to the glycosidic torsion angles. Finally, the structures are named and exported to a database as mol2 files. The virtual library of hexasaccharide sequences combinatorially built from 36 disaccharide building blocks (Table 3) should now include $36 \times 36 \times 36 = 46,656$ total sequences.

2.3. Docking of H/HS Sequences

The typical automated molecular docking study consists of the following main steps: (a) preparing the ligands (here the combinatorial virtual library) (b) preparing the protein receptor and setting up the simulation system, which involves identifying the binding site of interest (c) performing the actual docking simulations and (d) post-processing and analysis of the results. The steps involved in combinatorial virtual library generation have been discussed in detail in the previous section. The next important step is protein preparation.

Protein preparation and identification of GAG binding sites—In this example the crystal structure of the ternary complex of antithrombin–thrombin–heparin (PDB ID: 1TB6) is used. To prepare the protein, extract the antithrombin coordinates from this structure using the SYBYL Biopolymer Structure Preparation Tool. Add hydrogen atoms and check for amino acid side chains that need repair. It is not uncommon for the surface-exposed lysine and arginine side chains that are critical for GAG recognition to be incompletely resolved in the crystal structure, resulting in missing atoms. The SYBYL Mutate Residues feature allows coordinates to be assigned to such missing atoms. For a reasonable docking study it is important to have the residues in the binding site to be in their correct protonation state; this can be done by “Set Protonation State” in the SYBYL structure preparation tool. Then minimize the protein structure with fixed heavy-atom coordinates using the TFF for 1000 iterations subject to a termination gradient of 0.05 kcal/(mol Å).

In general, the GAG binding site information can be obtained from the co-crystallized protein–GAG complex deposited in the PDB; if this structural information is not available then it may be inferred by site-directed mutagenesis or through computational prediction of GAG binding sites. Numerous studies have been reported in identifying the common structural features in H/HS binding site of proteins, by both experimental studies (X-ray and NMR) and through molecular modeling approaches [49,50]. The binding site in antithrombin was defined to include all atoms within 14 Å of the K125 N^ζ (NZ) atom. This binding site definition covers all basic residues K11, R13, R46, R47, K114, K125, R129 and R132 in the AT pentasaccharide binding site (PBS).

Docking Algorithm—GOLD is a molecular docking program that uses a GA search strategy that ‘evolves’ docked solutions starting from a set of random ligand conformations, and includes rotational flexibility for selected receptor residues along with full ligand flexibility. The algorithm that GOLD uses to generate docked solutions involves several discrete steps and is outlined in Fig. 7A. The structural state of the ligand (and any flexible receptor side chains) is encoded by a ‘chromosome’, representing its conformation and hydrogen bonding interactions (Fig. 7B). The ‘chromosome’ is a binary string in which every torsional angle is encoded by a ‘gene’. To generate different ligand conformations, each rotatable torsion angle is allowed to vary between -180° and 180° . A library of preferred conformations is used to assign specific values to the torsion angles. The GA uses two genetic operators: crossover and mutation, as illustrated in Fig. 7B. A GA operator is applied to parent chromosomes that are randomly chosen from the existing population with a bias towards the fittest members (*i.e.* stronger members are more likely to be parents). Two integer strings encode possible hydrogen bonds between the protein and the ligand. For a given GA iteration, the ligand–receptor hydrogen bonds are subsequently matched using a least-squares fitting protocol to maximize the number of inter-molecular hydrogen bonds (*i.e.* hydrogen bonds between the ligand and the receptor), thus placing the ligand in the binding site and generating a ligand ‘pose’. As a consequence the GA pose generation is biased towards inter-molecular hydrogen bonds. The GOLDScore fitness function that evaluates the quality of the docked ligand poses is comprised of three terms: a hydrogen bond term; a 4–8 intermolecular dispersion potential and a 6–12 intramolecular dispersion potential energy for the internal energy of the ligand. Although this scoring function in general correlates modestly with the observed free energy of binding, a modified form of the scoring function has been found to be more reliable [12]. This modified GOLD score, which utilizes hydrogen-bonding and van der Waals interactions as described above (Eq. 3), can also be used to rank the final docked solutions.

$$\text{GOLDScore} = \text{HB}_{\text{EXT}} + 1.375 \times \text{VDW}_{\text{EXT}} \quad (3)$$

Here, HBEXT and VDWE XT are the “external” (*i.e.* nonbonded interactions taking place between the ligand and receptor) hydrogen bonding and van der Waals terms, respectively. The GOLDScore energy function is designed such that larger (more positive) values are more favorable than negative scores.

Setting up a GOLD docking run—GOLD docking runs can be set up by using the Hermes visualizer, the graphical front-end of GOLD. Using Hermes, the input files for docking are prepared in an interactive manner to define the binding site, specify constraints, and set up other parameters for the docking run. In Hermes, the top-level menu has an icon called GOLD; click this icon and there are two options:

1. Setup and Run a Docking
2. Wizard

For a new user it's highly recommended to use the **Wizard** icon, as this will walk the user through the entire process of setting up the GOLD docking run. The usual necessary steps in the wizard are the following:

1. **Proteins.** Choose the protein molecule to be used for docking. The protein molecule that was saved after processing through the protein preparation tool in SYBYL is used here.
2. **Define Binding Site.** There are multiple ways to define the binding site, the simplest being to select an atom and an associated radius. Select an atom in the protein; in this case LYS125 NZ from the 'I' chain of 1TB6. Specify a radius of 14 Å to include all the atoms within 14 Å of the selected atom.
3. **Select Ligands.** Select the library of mol2 files in the SYBYL database, or concatenate all of the sequences in the library into a single multi-mol2 file. Specify the number of GA runs. For the first CVLS filter (*vide infra*), set the value to 10; for the second filter set it to 100.
4. **Ligand Flexibility.** To keep the glycosidic linkages fixed during the docking runs, specify the atom numbers of the two atoms joined by each glycosidic bond. Due to the consistent numbering scheme, these will be the same for all of the sequences in the library (see Fig. 5A). Check the option to fix ligand rotatable bonds, select 'fix specific' and then 'Specify Bonds' to enter the atom numbers and 'Close' when done.
5. **Fitness & Search Options.** Select GOLDScore.
6. **GA Settings.** Check the 'Preset' option. For the first CVLS filter, set the value to 10,000 operations; for the second filter set it to 100,000 operations.
7. **Output Options.** As a very large number of files will be created, select the option to create output sub-directories for each ligand.
8. **Finish.** Save the configuration file (*gold.conf* by default) for further use and to run GOLD outside of the Hermes interface.

Evaluating the docked poses and optimization of docking conditions—To ensure that the GAG orientation and the position obtained from the docking studies were likely to represent valid and reasonable binding modes, the GOLD program docking parameters were first used to dock the heparin pentasaccharide (H5) sequence in the crystal structure (native) on AT. The H5 sequence is constrained only at their glycosidic linkage.

Parameters were optimized by performing multiple docking runs and changing various parameters such as the number of GA runs (10 to 200), the maximum number of iterations and radius to cover all the atoms in the binding site. The docking procedure was repeated three times, and the results were analyzed and compared with the crystal structure of this complex [36]. The results of the docking study using the optimal parameters indicate that the highest-ranked solution is spatially close to the native structure (H5 analog co-crystallized in PDB = 1TB6). Because the crystal structure for this particular GAG–protein interaction is known, a good measure of the reliability and specificity of docking is through an RMSD

calculation using the backbone atoms, which for the best examined parameter set is 0.96 Å. In addition, the hydrogen bond and van der Waals interactions were found to be identical, thereby suggesting the docking procedure to be reliable to obtain a specific sequence. These parameters are subsequently used in the second filter of the CVLS algorithm.

2.4. Reproducibility of Observed H5 Binding Mode

To evaluate the ability of the GOLD algorithm to reproduce the observed binding mode of H5 (Fig. 8) with AT using the ‘average backbone’ approximation with fixed glycosidic linkage torsion angles, a small library consisting of only H5-containing sequences found in the entire 46,656 sequence hexasaccharide library was constructed and docked to the AT PBS using the 1TB6 crystal structure parameters corresponding to the second CVLS filter. This H5-containing library contained 44 sequences of the form X-Yb[C,2]6A-ZbB-Yb236A-u[a,a2,c,c2]A-Yb26A, where X represents any of the six uronic acid derivatives. GOLD reproduced the observed binding mode for each of the 22 sequences containing iduronic acid in the ²S_O conformation at monosaccharide position 5. In addition, 21 of these had GOLDScores that placed them in the top 50% of the database (see Fig. 9).

2.5. Application of the CVLS Approach

The previous section outlined how to set up the molecular docking simulation for each of the sequences in the H/HS library for both the first and second CVLS filters. The same GOLD configuration file is used for docking the entire database (46,656 hexasaccharide sequences) against the AT PBS for a given CVLS filter. The GOLD configuration file can be set up in such a way to read every sequence from a single mol2 file or to read the entire set of sequences from a single multi-mol2 file. Taking into account of the ‘average backbone’ approximation, the glycosidic torsions are kept fixed in the GOLD configuration file. The atom numbers of the atoms contributing to the glycosidic linkage in different sequences would normally vary from sequence to sequence, which would complicate the identification and setting up of fixed torsions for all sequences. However, the atom numbering convention discussed above was used to number the atoms in the glycosidic linkage in a consistent manner. This provides glycosidic linkage atom numbers that are common for all 46,656 sequences in the library. The CVLS docking calculations were performed using an in-house Linux cluster utilizing 120 concurrent processes; the time required to finish the entire hexasaccharide library was two days for the first filter and two additional days for the second filter.

Analysis of the Results and Pharmacophore Identification—Two functions (filters) to evaluate the best sequence from the virtual screening experiment were combined; affinity in the form of GOLD score and specificity in the form of consistency of docking, which represents as a powerful tool to deduce high-affinity GAG sequences that bind proteins with high specificity. In the first CVLS filter, the entire library of 46,656 sequences was screened for ‘affinity’ using ten GA runs for each sequence with 10,000 GA iterations per GA run. The best-ranked solution for each of the 46,656 sequences was selected, and the scores were sorted based on GOLDScore. The sequences corresponding to the top 0.1% of the database (47 sequences) were selected for further analysis to advance to the second CVLS filter. Of the 47 sequences that passed the first filter, 36 were placed at the same

location in the PBS as the H5-like sequence in the 1TB6 crystal structure; the remaining 11 sequences overlapped with the 1TB6 H5-like sequence but were usually shifted by one or two monosaccharides residues toward the extended heparin binding site (EHBS). Of the 36 sequences that overlapped the 1TB6 H5-like sequence, 33 contained H5 or H5-like sequences. The elements of the classical H5 AT PBS recognition motif (i.e. pharmacophore) were the following: Each sequence contained either ZbB or Zb2B in the third ('E') position and either ucA or uc2A in the fifth ('G') position (see Fig. 10A).

In the second CVLS filter, the 47 sequences that passed the first filter were evaluated for 'specificity', as measured by docking consistency: Instead of performing a single docking experiment (see note #11), the docking experiment was performed in triplicate for each sequence under more rigorous docking conditions (100 GA runs, 100,000 GA iterations per run). This procedure also favors reproducibility and reduces the possibility of false positives. The two highest-scoring solutions from each of the three docking experiments (i.e. six docked solutions) were compared and those sequences whose six docked solutions exhibited a backbone atom RMSD of ≤ 2.5 Å were considered for further analysis. Of the 47 sequences that passed the first filter, twenty went on to pass the second filter. Of these twenty, nine reproduced the binding mode exhibited by the 1TB6 H5-like sequence, with the remaining 11 extending into the EHBS. All nine 1TB6-matching sequences shared additional H5 pharmacophoric elements above and beyond those from the first filter: Each contains either Yb26A or YbC6A in the second ('D') position, ZbB or Zb2B in the third ('E') position, either Yb23A or Yb236A in the fourth ('F') position and either ucA or uc2A in the fifth ('G') position (see Fig. 10B).

These results show that the CVLS algorithm may be used to identify pharmacophoric patterns in GAG sequences that specifically recognize GAG binding sites on proteins. Such information will be invaluable not only in advancing our understanding of biological processes involving GAGs at the molecular level but also in the design of novel therapeutics targeted toward the circumvention of GAG-associated disease states.

3. Conclusions

While GAG-related glycobiology is of immense physiological relevance, it has not caught favor amongst scientists due to the inherent problems associated with the field. Current experimental methodologies are unable to purify homogenous samples of GAGs, with few exceptions. Synthesis is also difficult due to structural complexity. Under these circumstances, exploratory searches conducted *in silico* can be of immense use as a preliminary information mining resource to guide experimental efforts. We have presented methodology above that allows such applications, including studying interactions between a specific GAG sequence with a protein, or alternatively identification of specific binders from a combinatorial library of GAG sequences.

There is much room for improvement in the field of computational GAG-related glycoscience. For instance, docking a relatively small conformational library of approximately 100,000 GAG structures takes approximately 2 weeks on a reasonably sized compute-cluster. This, too, only samples a very small portion of conformational space due to

the excess of rotatable bonds. Luckily, current information seems to suggest that there are some stable conformations adopted by GAG oligosaccharides, which makes the currently presented methodology relevant. However, it is not guaranteed that the same conformations of the same oligosaccharide will interact with every protein. Oligosaccharides exist in multiple interchangeable ensembles of conformations in solution. Thus, first and foremost, the problem of conformational sampling must be solved. If done in the trial and error or brute force way, this will put immense pressure on currently available computational resources and consequently will also increase simulation time by several orders of magnitude. Hence, new technology must be developed to satisfy this need—particularly the development of artificial intelligence-based methodologies must be pursued.

Furthermore, much work is needed to develop force fields and scoring functions that are able to accurately assess the affinity of protein–GAG interactions. While docking and scoring have been useful tools in the case of drug-like molecules, the average success rate is usually on the order of 5–10%. This is because current scoring functions are unable to distinguish between binders and non-binders unless there is approximately a 1000-fold difference in affinities. With protein–GAG systems, the accuracy of available force fields falls even further due to the lack of parameterization for sulfated GAGs. Recently, a series of perspectives has suggested rigorous work is needed for improvement of scoring functions and force fields across the next 25 years, particularly with the aim of identifying drug-like molecules [51]. Perhaps this will serve as impetus in computational glycosciences as well. In the meantime, it is expected that docking and scoring using protocols such as CVLS will contribute towards our understanding of glycobiology.

4. Notes

1. The parameters (and filtering steps) given in Fig. 2 can be adjusted to suit the user's computing resources. For example, if the user's computational resources are sufficiently great, the more rigorous GA settings used in the second filter may be applied in the first filter to reduce the number of false negatives.
2. Alternate SLN numbering schemes other than that described in Fig. 5 could also be used.
3. The energy minimization process can be lengthy for larger libraries; however, it only needs to be performed once.
4. Although the staged energy minimization described above prevents distortion of the pyranose ring and glycosidic linkage conformations, it can occasionally result in the distortion of the sulfate and/or other non-backbone groups due to steric clashes in the conformation of the initially-built oligosaccharide. This situation may be remedied by identifying distorted functional groups, re-adjusting their geometry to an 'ideal' state and re-minimizing the sequence.
5. GAGs are highly complex molecules, and thus analysis tools to assess conformational and configurational integrity are essential in assessing the integrity of the GAG sequences.

6. The mol2 files containing the mono- and disaccharide structures may need to be edited manually in order to, for example, have the substructure label displayed consistently at the same atom position. In our libraries the 'ROOT' atom of the substructure (where the substructure label is displayed) corresponds to the anomeric C1 carbon.
7. When building individual GAG sequences, values for the glycosidic linkage torsion angles may also be assigned values corresponding to other defined macroscopic conformations rather than those of the 'average backbone'. For example, a regular linear H/HS sequence may be generated by assigning values to ϕ and ψ corresponding to those found in PDB ID = 1HPN [52].
8. Naming of substructures in SYBYL can result in names in which all of the letters are capitalized. Because the naming convention employed here is case-sensitive, modification of the saved mol2 files may be necessary to correct this.
9. When docking the H/HS sequences in GOLD, ensure that the 'ring corner flipping' option is disabled. Otherwise, the predefined pyranose ring conformations will be modified during the docking process. Additionally, enable the 'flip pyrimidal nitrogen' option to allow *N*-sulfate nitrogen atoms to change conformation.
10. The side chains of the protein may also be made flexible during the docking runs. However, doing so exponentially increases the size of the conformational search space, resulting in an exponential increase in the number of iterations (and time) required by the docking program to find good docked solutions. Thus, it is not generally recommended for virtual screening applications like CVLS.
11. GOLD starts with a population of 100 arbitrarily ligand conformations, docks and 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, the 'best' docked poses and their associated scores generated during individual GA runs can be different; because of this several such GA runs are required to more reliably predict correct bound conformations. In the second CVLS filter, 100 GA runs are performed with the GOLDScore as the "fitness" function. Collectively, these 100 GA runs are referred to as one docking experiment.

Acknowledgments

This work was supported by the grants HL090586 and HL107152 from the National Institutes of Health and by Award Number S10RR027411 from the National Center For Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center For Research Resources or the National Institutes of Health.

References

1. Esko, JD.; Kimata, K.; Lindahl, U. Proteoglycans and sulfated glycosaminoglycans. In: Varki, A.; Cummings, RD.; Esko, JD., editors. *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press; Cold Spring Harbor, New York: 2009. p. 229-248.
2. Blow N. A spoonful of sugar. *Nature (London)*. 2009; 457:617–620. [PubMed: 19177129]

3. Dementiev A, Petitou M, Herbert JM, Gettins PGW. The ternary complex of antithrombin–anhydrothrombin–heparin reveals the basis of inhibitor selectivity. *Nat Struct Mol Biol.* 2004; 11:863–867. [PubMed: 15311268]
4. Olson ST, Björk I. Predominant contribution of surface approximation to the mechanism of heparin acceleration of the antithrombin–thrombin reaction. Elucidation from salt concentration effects. *J Biol Chem.* 1991; 266:6353–6364. [PubMed: 2007588]
5. Cardin AD, Weintraub HJR. Molecular modeling of protein-glycosaminoglycan interactions. *Arterioscl Thromb Vasc Biol.* 1989; 9:21–32.
6. Hileman RE, Fromm JR, Weiler JM, Linhardt RJ. Glycosaminoglycan–protein interactions: Definition of consensus sites in glycosaminoglycan binding proteins. *Bioessays.* 1998; 20:156–167. [PubMed: 9631661]
7. Margalit H, Fischer N, Ben-Sasson SA. Comparative analysis of structurally defined heparin binding sequences reveals a distinct spatial distribution of basic residues. *J Biol Chem.* 1993; 268:19228–19231. [PubMed: 8366075]
8. Sobel M, Soler DF, Kermonde JC, Harris RB. Localization and characterization of a heparin binding domain peptide of human von Willebrand factor. *J Biol Chem.* 1992; 267:8857–8862. [PubMed: 1577724]
9. Bitomsky W, Wade RC. Docking of glycosaminoglycans to heparin-binding proteins: Validation for aFGF, bFGF, and antithrombin and application to IL-8. *J Am Chem Soc.* 1999; 121:3004–3013.
10. Gandhi NS, Coombe DR, Mancera RL. Platelet endothelial cell adhesion molecule 1 (PECAM-1) and its interactions with glycosaminoglycans: 1. Molecular modeling studies. *Biochemistry.* 2008; 47:4851–4862. [PubMed: 18393439]
11. Krieger E, Geretti E, Brandner B, Goger B, Wells TN, Kungl AJ. A structural and dynamic model for the interaction of interleukin-8 and glycosaminoglycans: Support from isothermal fluorescence Titrations. *Proteins.* 2004; 54:768–775. [PubMed: 14997572]
12. Raghuraman A, Mosier PD, Desai UR. 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.* 2006; 49:3553–3562. [PubMed: 16759098]
13. Raghuraman A, Mosier PD, Desai UR. Understanding dermatan sulfate–heparin co-factor II interaction through virtual library screening. *ACS Med Chem Lett.* 2010; 1:281–285. [PubMed: 20835364]
14. Rogers CJ, Clark PM, Tully SE, Abrol R, Garcia C, Goddard WA III, Hsieh-Wilson LC. Elucidating glycosaminoglycan–protein–protein interactions using carbohydrate microarray and computational approaches. *Proc Natl Acad Sci USA.* 2011; 108:9747–9752. [PubMed: 21628576]
15. Cole, JC.; Nissink, JWM.; Taylor, R. Protein–Ligand docking and virtual screening with GOLD. In: Alvarez, J.; Shoichet, B., editors. *Virtual Screening in Drug Discovery.* Taylor & Francis CRC Press; Boca Raton, FL: 2005. p. 379–416.
16. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem.* 2009; 16:2785–2791. [PubMed: 19399780]
17. Dominguez C, Boelens R, Bonvin AMMJ. HADDOCK: A protein– protein docking approach based on biochemical or biophysical information. *J Am Chem Soc.* 2003; 125:1731–1737. [PubMed: 12580598]
18. Sapay N, Nurisso A, Imberty A. Simulation of carbohydrates, from molecular docking to dynamics in water. *Methods Mol Biol.* 2013; 924:469–483. [PubMed: 23034760]
19. Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ. GLYCAM06: A generalizable biomolecular force field. *Carbohydrates.* *J Comput Chem.* 2008; 29:622–655. [PubMed: 17849372]
20. Woods RJ, Dwek RA, Edge CE. Molecular mechanical and molecular dynamical simulations of glycoproteins and oligosaccharides. 1. GLYCAM_93 parameter development. *J Phys Chem.* 1995; 99:3832–3846.
21. Guvench O, Hatcher E, Venable RM, Pastor RW, MacKerell AD Jr. CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. *J Chem Theory Comput.* 2009; 5:2353–2370. [PubMed: 20161005]

22. Huige CJM, Altona C. Force field parameters for sulfates and sulfamates based on *ab initio* calculations: Extensions of AMBER and CHARMM fields. *J Comput Chem*. 1995; 16:56–79.
23. Mallajosyula SS, Guvench O, Hatcher E, MacKerell AD Jr. CHARMM additive all-atom force field for phosphate and sulfate linked to carbohydrates. *J Chem Theory Comput*. 2012; 8:759–776. [PubMed: 22685386]
24. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Krüger P, van Gunsteren WF. The GROMOS biomolecular simulation program package. *J Phys Chem A*. 1999; 103:3596–3607.
25. Jin L, Barran PE, Deakin JA, Lyon M, Uhrin D. Conformation of glycosaminoglycans by ion mobility mass spectrometry and molecular modelling. *Phys Chem Chem Phys*. 2005; 7:3464–3471. [PubMed: 16273147]
26. Mikhailov D, Linhardt RJ, Mayo KH. NMR solution conformation of heparin-derived hexasaccharide. *Biochem J*. 1997; 328:51–61. [PubMed: 9359833]
27. Mikhailov D, Mayo KH, Vlahov IR, Toida T, Pervin A, Linhardt RJ. NMR solution conformation of heparin-derived tetrasaccharide. *Biochem J*. 1996; 318:93–102. [PubMed: 8761457]
28. Murphy KJ, McLay N, Pye DA. Structural studies of heparan sulfate hexasaccharides: New insights into iduronate conformational behavior. *J Am Chem Soc*. 2008; 130:12435–12444. [PubMed: 18722428]
29. Pol-Fachin L, Becker CF, Guimarães JA, Verli H. Effects of glycosylation on heparin binding and antithrombin activation by heparin. *Proteins*. 2011; 79:2735–2745. [PubMed: 21769943]
30. Zhang Z, McCallum SA, Xie J, Nieto L, Corzana F, Jiménez-Barbero J, Chen M, Liu J, Linhardt RJ. Solution structures of chemoenzymatically synthesized heparin and its precursors. *J Am Chem Soc*. 2008; 130:12998–13007. [PubMed: 18767845]
31. Gandhi NS, Mancera RL. Free energy calculations of glycosaminoglycan–protein interactions. *Glycobiology*. 2009; 19:1103–1115. [PubMed: 19643843]
32. Pichert A, Samsonov SA, Theisgen S, Thomas L, Baumann L, Schiller J, Beck-Sickinger AG, Huster D, Pisabarro MT. Characterization of the interaction of interleukin-8 with hyaluronan, chondroitin sulfate, dermatan sulfate, and their sulfated derivatives by spectroscopy and molecular modelling. *Glycobiology*. 2012; 22:134–145. [PubMed: 21873605]
33. Gilson MK, Honig B. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins*. 1988; 4:7–18. [PubMed: 3186692]
34. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. *Science*. 1995; 268:1144–1149. [PubMed: 7761829]
35. Tsui V, Case DA. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers*. 2001; 56:275–291. [PubMed: 11754341]
36. Li W, Johnson DJD, Esmon CT, Huntington JA. Structure of the antithrombin–thrombin–heparin ternary complex reveals the antithrombotic mechanism of heparin. *Nat Struct Mol Biol*. 2004; 11:857–862. [PubMed: 15311269]
37. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*. 1997; 267:727–748. [PubMed: 9126849]
38. Gandhi NS, Mancera RL. The structure of glycosaminoglycans and their interactions with proteins. *Chem Biol Drug Des*. 2008; 72:455–482. [PubMed: 19090915]
39. Mulloy B, Forster MJ. Conformation and dynamics of heparin and heparan sulfate. *Glycobiology*. 2000; 10:1147–1156. [PubMed: 11087707]
40. Cremer D, Pople JA. A general definition of ring puckering. *J Am Chem Soc*. 1975; 97:1354–1358.
41. Forster MJ, Mulloy B. Molecular dynamics study of iduronate ring conformations. *Biopolymers*. 1993; 33:575–588.
42. Rao, VSR.; Qasba, PK.; Balaji, PV.; Chandrasekaran, R. Conformation of carbohydrates. Harwood Academic Publishers; Amsterdam: 1998.
43. Pol-Fachin L, Verli H. Depiction of the forces participating in the 2-*O*-sulfo- α -l-iduronic acid conformational preference in heparin sequences in aqueous solutions. *Carbohydr Res*. 2008; 343:1435–1445. [PubMed: 18452898]

44. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inform Comput Sci.* 1988; 28:31–36.
45. Jin L, Abrahams JP, Skinner R, Petitou M, Pike RN, Carrell RW. The anticoagulant activation of antithrombin by heparin. *Proc Natl Acad Sci USA.* 1997; 94:14683–14688. [PubMed: 9405673]
46. Johnson DJD, Li W, Adams TE, Huntington JA. Antithrombin– S195A factor Xa–heparin structure reveals the allosteric mechanism of antithrombin activation. *EMBO J.* 2006; 25:2029–2037. [PubMed: 16619025]
47. McCoy AJ, Pei XY, Skinner R, Abrahams JP, Carrell RW. Structure of β -antithrombin and the effect of glycosylation on antithrombin's heparin affinity and activity. *J Mol Biol.* 2003; 326:823–833. [PubMed: 12581643]
48. Esko JD, Selleck SB. Order out of chaos: Assembly of ligand binding sites in heparan sulfate. *Annu Rev Biochem.* 2002; 71:435–471. [PubMed: 12045103]
49. Carter WJ, Cama E, Huntington JA. Crystal structure of thrombin bound to heparin. *J Biol Chem.* 2005; 280:2745–2749. [PubMed: 15548541]
50. Faham S, Hileman RE, Fromm JR, Linhardt RJ, Rees DC. Heparin structure and interactions with basic fibroblast growth factor. *Science.* 1996; 271:1116–1120. [PubMed: 8599088]
51. Stouch TR. Looking forward into the Next 25 Years: The 25th anniversary issue of the Journal of Computer-aided Molecular Design. *J Comp Aided Mol Design.* 2012; 26:1.
52. Mulloy B, Forster MJ, Jones C, Davies DB. N.M.R. and molecular-modelling studies of the solution conformation of heparin. *Biochem J.* 1993; 293:849–858. [PubMed: 8352752]

Fig. 1.
Bridging mechanism of T inhibition by AT and H/HS (PDB ID: 1TB6).

Fig. 2.
Overview of the dual-filter CVLS algorithm.

Fig. 3. Monosaccharide constituents of H/HS. *A.* β -D-Glucuronic acid (GlcA), 4C_1 conformation. *B.* α -D-Glucosamine (GlcN), 4C_1 conformation. *C.* α -L-Iduronic acid (IdoA), 1C_4 conformation. *D.* α -L-Iduronic acid (IdoA), 2S_0 conformation. Atom numbers are shown in red.

Fig. 4. Inter-glycosidic linkage definitions. A. Location of glycosidic torsion angles ϕ (ϕ) and ψ (ψ) within a typical H/HS disaccharide, GlcA(2S)-GlcNS(6S). Atom numbers are shown in red. B. Newman projection representations of ϕ and ψ showing the names of the atoms involved in their definition.

Fig. 5.

Representation of the H/HS hexasaccharide backbone atoms. *A.* Numbering scheme. The reducing (R) and non-reducing (NR) ends are labeled. *B.* SYBYL Line Notation (SLN). See section 2.2.2 for details.

```
for r=1,N                                # r is first disaccharide
  for s=1,N                              # s is second disaccharide
    for t=1,N                            # t is third disaccharide
      Join r + s = rs                    # tetramer formation
      Join rs + t = rst                  # hexamer formation
      Renumber                           # renumber glycosidic linkage atoms
      O5-C1-O1-C4' =  $\varphi$               #  $\varphi$  and  $\psi$  modification
      C1-O1-C4'-C5' =  $\psi$ 
      Minimize energy                    # relax structures
      Save mol2 file                     # finished hexasaccharide sequence
    endfor
  endfor
endfor
```

Fig. 6.

Pseudocode outlining the steps involved in construction of an H/HS hexasaccharide library.

Fig. 7.

The genetic algorithm (GA). **A.** Flowchart outlining the major steps involved. **B.** Application of the GA to a typical H/HS disaccharide, YbC6A-ua2A. A simplified 'chromosome' encoding its 12 rotatable bond torsion angles is shown, with the fixed backbone torsion angles labeled in red and the invariant values ('genes') shown in grey cells. The evolutionary processes of mating and mutation are also illustrated.

Fig. 8.
Structure and names of the monosaccharide constituents of the AT-specific H5 pentasaccharide sequence.

Fig. 9.
H5-containing hexasaccharide sequences that reproduced the binding mode found in the PBS of the 1TB6 crystal structure.

Fig. 10.

A. CVLS hits from the first filter. Thirty-three hexasaccharides containing H5 and H5-like hexasaccharides (capped sticks; green carbon atoms) successfully reproduced the bound conformation observed in the 1TB6 crystal structure (ball-and-stick; orange carbon atoms). The N-terminus (Nt) and helices involved in the recognition of the oligosaccharides are labeled, as are GAG monosaccharide units corresponding to residues D–H in the H5 pentasaccharide sequence. **B.** CVLS hits from the second filter. The nine hexasaccharides (capped sticks; cyan carbon atoms) reproducing the 1TB6 H5-like sequence are shown.

Table 1

Naming convention for the H/HS monosaccharide building blocks.

CVLS Library Name ^a	Conventional Name
ZbB	GlcA (⁴ C ₁ conformation, β-anomer)
Zb2B	GlcA(2S) (⁴ C ₁ conformation, β-anomer)
uaA	IdoA (¹ C ₄ conformation, α-anomer)
ua2A	IdoA(2S) (¹ C ₄ conformation, α-anomer)
ucA	IdoA (² S _O conformation, α-anomer)
uc2A	IdoA(2S) (² S _O conformation, α-anomer)
Yb2A	GlcNS (⁴ C ₁ conformation, α-anomer)
Yb23A	GlcNS(3S) (⁴ C ₁ conformation, α-anomer)
Yb26A	GlcNS(6S) (⁴ C ₁ conformation, α-anomer)
Yb236A	GlcNS(3S,6S) (⁴ C ₁ conformation, α-anomer)
YbCA	GlcNAc (⁴ C ₁ conformation, α-anomer)
YbC6A	GlcNAc(6S) (⁴ C ₁ conformation, α-anomer)
YbHA	GlcN (⁴ C ₁ conformation, α-anomer)
YbH3A	GlcN(3S) (⁴ C ₁ conformation, α-anomer)
YbH36A	GlcN(3S,6S) (⁴ C ₁ conformation, α-anomer)

^aSymbols used: Z = D-glucuronic acid, u = L-iduronic acid, Y = D-glucosamine. Ring conformations: a = ¹C₄; b = ⁴C₁; c = ²S_O. Glucosamine substituents: H = unsubstituted 2-position (NH₃⁺), C = N-acetyl, 2 = N-sulfate, 3 = 3-O-sulfate, 6 = 6-O-sulfate. Uronic acid substituents: 2 = 2-O-sulfate. Anomeric carbon configuration: A = α, B = β.

Table 2

Values used in the ‘average backbone’ approximation.

Disaccharide	ϕ (O5–C1–O1–C4')	ψ (C1–O1–C4'–C5')
GlcA(β 1→4)GlcN	–81.8	–114.0
IdoA(α 1→4)GlcN	–87.7	–128.3
GlcN(α 1→4)GlcA	91.1	–151.6
GlcN(α 1→4)IdoA	87.4	–132.3

Table 3

The 36 disaccharide building blocks used to construct virtual H/HS oligosaccharide libraries.

GlcA-containing	IdoA-containing		
ZbB-YbCA	uaA-YbCA	ua2A-Yb2A	uaA-Yb236A
ZbB-YbC6A	ucA-YbCA	uc2A-Yb2A	ucA-Yb236A
ZbB-Yb2A	ua2A-YbCA	uaA-Yb26A	ua2A-YbH3A
Zb2B-Yb2A	uc2A-YbCA	ucA-Yb26A	uc2A-YbH3A
ZbB-Yb26A	uaA-YbC6A	ua2A-Yb26A	ua2A-YbH36A
Zb2B-Yb26A	ucA-YbC6A	uc2A-Yb26A	uc2A-YbH36A
ZbB-Yb23A	ua2A-YbC6A	uaA-Yb23A	
Zb2B-Yb23A	uc2A-YbC6A	ucA-Yb23A	
ZbB-Yb236A	uaA-Yb2A	ua2A-Yb23A	
ZbB-YbHA	ucA-Yb2A	uc2A-Yb23A	