Modeling large protein–glycosaminoglycan complexes using a fragment-based approach

Isaure Chauvot de Beauchêne

Modeling large protein-glycosaminoglycan complexes using a fragment-based approach

Sergey A. Samsonov [1], Martin Zacharias [2], Isaure Chauvot de Beauchene [3]

1. Faculty of Chemistry, University of Gdańsk. ul. Wita Stwosza 63, 80-308 Gdańsk, Poland.
2. Physics Dep., Technical University of Munich. James-Franck Strasse 1, 85748 Garching, Germany.
3. CNRS, LORIA (CNRS, Inria NGE, Université de Lorraine). Campus Scientifique, 615 rue du Jardin Botanique, Vandœuvre-lès-Nancy, F-54506, France.

Correspondence to: Isaure Chauvot de Beauchene (E-mail: isaure.chauvot-de-beauchene@loria.fr), Sergey A. Samsonov (E-mail: sergey.samsonov@ug.edu.pl)

Abstract

Glycosaminoglycans (GAGs), a major constituant of the extracellular matrix, participate in cell-signaling by binding specific proteins. Structural data on protein-GAG interactions is crucial to understand and modulate these signaling processes, with potential applications in regenerative medicine. However, experimental and theoretical approaches used to study GAG-protein systems are challenged by GAGs high flexibility limiting the conformational sampling above a certain size, and by the scarcity of GAG-specific computational tools. We present for the first-time an automated fragment-based method for docking GAGs on a protein binding site. In this approach, trimeric GAG fragments are flexibly docked to the protein, assembled based on their spacial overlap, and refined by molecular dynamics. The method appeared more successful than the classical full-ligand approach for most of 13 tested complexes with known structure. The approach is particularly promising for docking of long GAG chains, which represents a bottleneck for classical docking approaches applied to these systems.

Keywords: Glycosaminoglycans docking, fragment-based docking, Glycosaminoglycans-protein complex, glycans modeling.

Introduction

Protein-GAG structures

Glycosaminoglycans (GAGs) play essential roles in many physiological processes: cell proliferation, migration and differentiation, inter-cellular communication, blood coagulation, viral invasion and others.[1,2] These linear negatively charged hetero-polysaccharides consist of repetitive disaccharide units containing an uronic acid and an amino sugar, and display diverse patterns of sulfation known as 'sulfation code' in the field of protein-GAG interaction studies .[3] Although intuitively this code could be associated with a particular sulfation pattern of the GAG region directly interacting with a protein and contributing to its specificity,[4] there is no strict definition for such a ‘code’, which renders it important to use this term very carefully. All GAGs, with the exception of hyaluronic acid (HA), are covalently bound to proteins to form proteoglycans, which are a major constituent of the plasma membrane and extra-cellular matrix (ECM). Natural and artificially sulfated GAGs are very promising targets for the design of novel bioinspired functional biomaterials with potential medical applications in the field of bone and skin regeneration.[5] The functions of GAGs are mainly mediated by interactions with their target proteins. For instance, heparin (HE) binds and regulates the function of stem cells growth factors,[6,7] extracellular matrix proteins and surface proteins of pathogens;[8,9] chondroitin sulfates (CS) have been shown to stimulate the outgrowth of embryonic hippocampal neurons by recruiting growth factors to the cell surface;[10–12] hyaluorinic acid (HA) has been shown to be involved in the CD44 receptor related molecular mechanisms.[13] However, the experimental determination of high quality structures for such protein-GAG complexes is difficult to obtain by e.g. NMR or X-ray crystallography, which can be attributed to periodicity and high flexibility of GAGs as well as their high negative charge that renders it challenging to yield pure homogeneous samples. Therefore, computational approaches can be very useful to complement experiments aimed to study protein-GAG complexes structures.

Protein-GAG docking

GAGs bear a high negative charge, and their binding to protein surfaces is mainly guided by electrostatic interactions.[14] Therefore, simple calculation of electrostatic potential isosurfaces for protein molecules is a powerful approach for the prediction of GAG binding regions.[15] In addition, rigid or multiple docking of a GAG or its fragments was shown to assist the determination of its binding site. [12,16–18] This allows computational docking methods to focus the initial search on the approximately known binding site (local docking approach). When the putative binding site of a GAG on a protein is known, it is still challenging to predict a binding pose for even short GAGs. Recent evaluation of six widely-used docking programs in terms of their general applicability for protein-GAG systems for local docking performance showed that only a free docking software AutoDock 3 (AD3) [19] and the commercial program Glide showed reasonable performance for docking protein-GAG complexes.[15] Their best poses yielded RMSD values about 3.5 Å

with a reference to the experimental structures, and top poses were of significantly lower quality for the studied GAGs with the length dp2-dp6. This suggests that when docking GAGs, clustering and experimental data should be applied to effectively predict a protein-GAG binding pose.[20] Moreover, even when those fully flexible docking methods perform well for the GAG of dp2-dp6 with a limited number of degrees of freedom (up to 30), the challenge arises when GAGs are already longer than dp6, and the corresponding number of degrees of freedom is higher. Such poor performance is indicated by the correlation between the length of the GAGs docked and the structural difference of the obtained binding poses to the ligand in an experimental structure.[15] In other studies, conventional docking tools have been applied with limited success to isolated protein-GAG systems containing GAGs with degree of polymerisation (dp) typically up to dp4-dp6.[21–24] Recently, DarwinDock docking method was adapted to the protein-GAG systems: first, “a coarse docking” is used to predict a GAG binding site, and then a“fine grained” approached is employed to predict strong binding poses that are further minimized. This approach used flexible whole-ligand docking and was benchmarked starting from the bound conformation. A dynamic molecular docking method that is based on a steered MD procedure works better for longer GAGs than other known docking tools but has a disadvantage of being computationally far more expensive.[26] Therefore, new computational tools are needed to effectively model longer GAGs (dp >= 6) bound to a protein.

Fragment-based docking

Fragment-based approaches permit to deal with the challenge of a large ligand with a high number of degrees of freedom, by applying successive or parallel docking runs for smaller parts of the ligand on known binding sites.[27–29] The focus of the positional search and the reduction of the size of the ligand make the exhaustive conformational sampling feasible. Current fragment-based docking methods for GAGs apply incremental construction of the ligand from a seed docking pose of one fragment. The success of this type of approach strongly depends on the accuracy of the seed pose, and is mainly applied to small ligands binding in a well known and delimited binding site (e.g. the active site of an enzyme). Fragments docking has also been used as a binding site prediction tool, by counting the number of contacts with docking poses made by the protein residues.[30,31] Such method proved to have predictive value on one tested case, but does not take into account the constrain of connectivity of the fragments in the case of a long multi-fragments ligand. We recently presented a new method expanding the application of such approaches to docking long linear ssRNA on a protein surface, when no binding site for any of the monomers is known, i.e. when an incremental construction from a seed fragment cannot be applied.[32,33]
We present here a novel fragment-based method with a fully flexible ligand, which docks long GAGs on a coarsely known binding site. We validate the method on a benchmark of 13 protein-GAG structures, providing the first reported fragment-based docking approach that successfully reproduces experimental

binding poses of different GAGs (HE, CS, HE) of length dp5-dp7. This is the first time near-native models were generated for long dp7 GAGs.

Methods

Construction of the benchmark

We tested the method on all non-redundant protein-GAG complexes from the PDB that contain at least 5 successive monomers bound to the protein. A monomer with at least two atoms within the distant cut-off of 5 Å from at least one atom of the protein was considered as bound. To avoid redundancy, only a complex with a higher resolution was taken for analysis if several complexes with the same protein counterpart were available. The structure 4c4n displays two binding modes, corresponding to two GAG chains of same sequence bound to two protein chains of same sequence. For each docking pose, we computed the RMSD toward both bound ligands (after superposing the two protein chains with the crystal structure) and considered the lowest one for further analysis. For unbound docking, we modeled each target from its closest homologue structure in PDB (S1 Table). For 1rid, the closest homologous structure had 38% sequence identity (PDB code 2xrb), leading to a model at more than 10 A interface RMSD from the bound form. This model was therefore not considered further for unbound docking. For 1fq9, we used the structure of 1bfc as “unbound” structure (see Table 4).

Docking with AutoDock 3

Each GAG ligand is made of repeated dp2 units (dp stays for degree of polymerization) that we denote as (AB)n. For each complex, we docked the two possible dimers (AB, BA) and the two possible trimers (ABA, BAB) on the receptor. Through the whole manuscript, "fragment" refers to a fragment of the GAG ligand sequence, and "pose" refers to a docking solution for a particular fragment. Docking calculations for fragments were performed with AutoDock 3 (AD3) [19] with a spacing grid of 0.375 Å. We also tested to use a spacing grid of 0.27 Å, which AD3 did not improve the docking results and was not further applied. AGs were treated completely flexible and the protein receptor rigid. GAG charges were obtained from the GLYCAM06 force field [34] and from the literature [35] for sulfate groups, the charges of for the protein were assigned by AD3. Receptor structures were extracted from crystallographic complexes. We used roughly half of the protein surface for docking, taking into account GAG binding sites known from the corresponding experimental structure and centering a grid box on them. Such a bias towards a known binding site could be justified by our previous findings, which clearly demonstrate that a GAG binding site could be effectively predicted by applying of electrostatic potential calculations for the GAG protein target [15, 26]. The Lamarckian genetic algorithm with an initial population size of 300 and a termination condition of 10^5 generations or 9995×10^5 energy evaluations were used. A total of 10^3 independent runs were performed for each docking experiment. For a "taboo search", the grid box was designed to exclude

the regions where we aimed to avoid sampling. 1000 top scored solutions were taken for further analysis using fragment assembling approach. Afterward, AD3 scoring function was considered not accurate enough to distinguish between those 1000 solutions, and scores and ranks were not considered in the chain assembly process (all poses in the top-ranked 1000 were considered equiprobable *a priori*). For the whole ligand docking with AD3, we used the same parameters as for fragments docking described above. 50 top ranked solutions were then considered for further analysis as it was done previously in the study evaluating performance of different docking software.[15]

Chain assembly

To accelerate the assembly process, we used a coarse grained (CG) representation of the GAG, consisting in keeping only the O, N and S atoms. If several chemically equivalent atoms are connected to the same atom (e.g. oxygen atoms of a sulfate group), this last atom is kept instead (e.g. sulfur atom in a sulfate group). This is done in order to avoid artificial distinction between chemically equivalent atoms. The docking poses were assembled into chains of compatible poses, based on an overlap RMSD criteria. For each pair of poses of different types, we measured the CG RMSD between the overlapping parts of the two poses (e.g. BA for A<u>BA</u> vs <u>BA</u>B). The overlap cut-off was progressively increased by 0.1 Å steps until at least 1000 or 10.000 chains were found for dp5 or dp6/7 respectively, or until a maximum cut-off of 3 Å was reached. Only for complex 2axm, as no chains were found at 3.0 Å overlap cut-off, the cut-off was increased to 4.0 Å. We repeated this "chain assembling" procedure for (sub)-chains of different length (starting from dp5) and for chains (AB)n and (BA)n when relevant. For instance, for a ligand ABABAB, we assembled along the three assembling modes: {ABA + BAB + ABA => ABABA}, {BAB + ABA + BAB => BABAB} and {ABA + BAB + ABA + BAB => ABABAB}.

Poses filtering

For each fragment in terms of its position in the sequence, the docking poses were sorted by their connectivity, i.e. their occurrence in the assembled chains. We then pooled together the poses obtained from all assembling modes (longest chain and two sub-chains) and from all fragments of the same type (e.g. poses ABA from fragments 1 and 3 or poses BAB from fragment 2 and 4). The redundant poses were removed, and a maximum of 100 poses with the highest connectivity were retained. Each pool of retained poses was compared (by RMSD computation) to each bound fragment of the corresponding type in the experimental structure. For each fragment, we compared the number of good (RMSD < 3 Å) or acceptable (RMSD < 5 Å) poses in the highly connected poses and in the same number of top-ranked poses by AD3 scoring.

Comparison of rigid *versus* flexible docking

To estimate the gain obtained by flexible docking over rigid docking (e.g. with a fragment library), we compared to each bound fragment the conformations of poses obtained by flexible docking of the same fragment type on the corresponding protein ("bound docking") or on the other proteins ("cross docking"). On each bound fragment of each complex structure, we fitted all poses from all docking runs of the same fragment type. We then computed the RMSD to the bound fragment of the best-fitted pose among the poses obtained by bound docking or by cross docking. Only fragments of the 11 complexes containing HE were tested, as CS and HA were each present in only one complex of the benchmark.

Refinement

The assembled chains of poses are discontinuous, as the overlapping between poses is not perfect (Fig 1). To reconnect the poses, we first transformed chains of poses into chains of monomers by averaging the atomic positions of the overlapping parts of the poses.[32]According to our previous findings, many docking programs are capable of producing GAG docking poses of high quality but fail to score them properly.15 Therefore, clustering of docking poses based on pairwise RMSD criteria to eliminate redundancies increases the chances to get a correct pose in the top-ranked poses (ranked by score) compared to non clustered top-ranked docking poses obtained by imperfect scoring alone. The chains were then clustered with a 0.5 Å cut-off to remove redundancy. The averaging of monomer coordinates produces monomer conformations with incorrect geometry. To correct those conformational inconsistencies, we created a monomer library for each GAG residue in the benchmark, by pooling together the docking poses obtained for all complexes and extracting the considered residue. Each averaged monomer was replaced by the best-fitting monomer in the library in all-atom representation. Each chain of all-atoms monomers in complex with the protein was then refined by fully flexible minimization. The structures of the protein-GAG complexes obtained from the docking calculations were used for MD simulations carried out with AMBER 14 3535. Parameters from the ff14SB and GLYCAM-06j [34] force fields were used for proteins and GAGs, respectively. The complexes were solvated in a TIP3P octahedral periodic box with a minimal distance to the periodic box border of 8 Å and counter ions. Two energy-minimization steps were carried out: first 0.5×10^3 steepest descent cycles and 10^3 conjugate gradient cycles with harmonic force restraints on solute atoms, and then 3×10^3 steepest descent cycles and 3×10^3 conjugate gradient cycles without constraints. Afterwards, the system was heated up to 300 K for 10 ps, equilibrated for 100 ps at 300 K and 10^6 Pa in isothermal isobaric ensemble (NPT). Finally, an another two-step minimization of 3×10^3 steepest descent cycles and 3×10^3 conjugate gradient cycles without constraints was carried out. The SHAKE algorithm, 2 fs time integrations 8 Å cutoff for non-bonded interactions and the Particle Mesh Ewald method were used. For each GAG from the complex, pyranose rings were harmonically restrained in 4C1 (for IdoA2S in 1C4) conformations. It is widely known that IdoA2S ring could be in other conformations, which could

significantly affect the performance of heparin molecular modeling approaches including molecular docking. [37] Ideally, all possible combinations of heparin trisaccharides with both major IdoA2S ring conformations ($^{1}C_{4}$ and $^{2}S_{O}$) should be considered. In this study, we used the ring conformations corresponding to the experimental structures of the benchmarking complexes from the PDB. Moreover, the ring conformation would not affect the glycosidic linkages, which conformations are substantial for the assembling procedure, and, therefore, the shape of a longer GAG built by the fragment-based approach. [38,39]

The final models were clustered at 0.5 Å and evaluated by computing the RMSD on all heavy atoms of the GAG compared to the experimental structure after superposing the protein.

Results

In each complex of the benchmark, the GAG, made of repeats of a dp2 unit [AB]n is cut into two types of fragments (ABA and BAB) (Fig 1). Each fragment type is docked once on the protein, and the poses from each docking run are compared to each of the bound fragments of the same type. All retained poses are considered equiprobable at this stage (e.g. the AD3 scores were not taken into account). Poses for the 1st and 3rd fragments are always obtained from one docking run, and poses for the 2nd and 4th fragment (if existing) from another docking run. The poses are then assembled into chains to reconstitute the whole ligand, based on an overlap RMSD criteria for the overlapping parts of the two fragments (e.g. A<u>BA</u> + <u>BA</u>B, see Methods). The method was tested on a benchmark of 13 X-ray structures of GAG-protein complexes (S1 Table). This includes 12 different proteins from 11 different families bound to GAGs of three different types, heparin (HE), chondroitin sulfate (CS) and hyaluronic acid (HA), with length ranging from dp5 to dp7. The quality of docking prediction was assessed by computing the RMSD with respect to the bound form. A cutoff of 2.0 Å RMSD is typically used as acceptance criterion for drug-like small molecules (MW < 500). Given the larger size of our dp5-dp7 ligands and the notorious challenge that their flexibility and periodicity represent for docking,[40] we defined adapted criteria of 3.0 – 5.0 Å RMSD for good and acceptable solution, respectively, based on previously obtained results for dp6-7 docking reported in the literature.[15]

AutoDock 3 results

We used Autodock 3 because this program so far performed the best among other programs benchmarked for protein-GAG complexes in terms of both docking pose generation and scoring.[15] Despite its simple scoring function, this software seems to be powerful to account for electrostatics-driven nature of protein-GAG interactions. For all complexes, bound docking with AD3 sampled good (RMSD < 3 Å) solutions for one or more fragments. However, it sampled correctly all fragments for only 2 out of the 13 complexes (1gmn and 3ina). For docking performance evaluation, the number of good solutions per

fragment is more important than the best RMSD of poses per fragment, as it has a higher impact on the probability to be able to construct an acceptable full-ligand after assembly. The percentage of good solutions found by AD3 is reported in Table 1. We also performed bound docking of dp2 GAGs, which resulted in poses that were highly non-specific. This confirms previous observations that only GAG fragments of length dp 3 or more can establish specific binding with a protein.[21]

Table 1. Quality of the docking poses sampled by AD3

Complex (PDB ID)	Best RMSD in Å (% acceptable poses)									
	bound docking						unbound docking			
	Frag 1	Frag 2	Frag 3	Frag 4	Frag 5	Frag 1	Frag 2	Frag 3	Frag 4	Frag5
1bfc	4.6 (0)	4.5 (0)	2.2 (12)	1.5 (38)	-	7.8 (0)	**4.2** (0)	3.0 (8)	3.2 (16)	-
1fq9	2.1 (4)	1.7 (2)	2.2 (3)	4.9 (0)	-	-	-	-	-	-
1xmn	4.0 (2)	2.5 (4)	2.6 (18)	2.9 (4)	-	**3.8** (**3**)	3.2 (3)	3.4 (6)	4.2 (1)	-
2axm	2.8 (4)	3.4 (11)	1.8 (11)	3.7 (1)	-	**2.4** (3)	**2.6** (**15**)	4.2 (0)	7.1 (0)	-
1rid	2.5 (0)	3.6 (1)	3.8 (0)	3.7 (0)	6.0 (0)	-	-	-	-	6.5 (0)
1gmn	1.9 (4)	1.6 (8)	2.8 (3)	-	-	3.3 (0)	1.7 (3)	3.0 (**6**)	-	-
2jcq	1.9 (2)	1.1 (6)	2.9 (1)	5.2 (0)	6.4 (0)	3.9 (**1**)	1.1 (1)	4.4 (1)	**4.6** (0)	**2.3 (1)**
2hyv	4.1 (1)	2.6 (1)	4.6 (0)	-	-	4.5 (0)	2.9 (1)	**4.5** (0)	-	-
3ina	3.2 (1)	1.8 (4)	2.4 (0)	2.4 (0)	3.1 (0)	4.0 (0)	3.6 (1)	2.4 (0)	6.5 (0)	-
3mpk	5.0 (0)	2.6 (2)	1.5 (7)	2.7 (11)	-	6.1 (0)	3.0 (2)	1.8 (5)	2.7 (**12**)	-
3c9e	1.1 (1)	2.9 (1)	5.0 (0)	3.6 (1)	-	1.9 (1)	**1.9** (4)	5.3 (0)	4.2 (0)	-
4ak2	2.2 (0)	2.4 (2)	3.3 (2)	4.8 (1)	-	2.5 (**1**)	3.2 (2)	**3.0** (0)	**4.0** (1)	-
4c4n	4.6 (1)	2.8 (1)	4.0 (1)	3.4 (1)	-	3.9 (1)	4.1 (1)	4.5 ()0	4.8 (0)	-

Bold: fragments equally well or better sampled by unbound than bound docking. Underlined: fragments much less well sampled by unbound than bound docking.

Fragments assembly and filtering

We compared the results of pose selection by chain-assembly and by AD3 scoring in order to identify the best procedure for filtering acceptable poses. We compared the number of acceptable solutions for each fragment *(i)* in the most-connected assembled poses and *(ii)* in the same number of top-ranked poses in AD3 scoring. As some of the GAG terminal fragments can be badly sampled due to higher flexibility in the complex, we assembled not only the whole ligand but also sub-chains shortened by one or two monomers at either end of the ligand, and we pooled together the most-connected poses from all (sub-)chains. For a qualitative evaluation, we considered primarily the number of fragments for which at least one correct pose was found, and secondarily the total number of correct poses per fragment. The assembly procedures proved much more efficient in filtering good fragment solutions than AD3 scoring, regarding both criteria, for 10/11 complexes (Fig 2, S2 Table). The only four fragments for which poses were better filtered by AD3 are fragments located in the center of the chain (Fig 2).

Chain building.

The assembled chains of poses are discontinuous, as the spatial overlap between poses is not perfect (Fig 1). Therefore, atom coordinates in the overlapping sections were averaged over the overlapping poses, and redundant chains were removed. For a dp7 ligand, represented by a chain of 5 dp3 fragments, assembling monomers 1 to 5 (fragments 1 to 3) or 3 to 7 (fragments 3 to 5) is equivalent, as poses for fragments 1, 3 and 5 are the same, and poses for fragments 2 and 4 are the same. Therefore, the results of sub-chains assembly are presented in Table 2 for each of the two assembly modes.

We could sample acceptable (RMSD ≤ 5 Å) dp5 GAG chains for 11/13 complexes and good (RMSD ≤ 3 Å) chains for 8/13 complexes. When assembling longer chains, those ratios only slightly diminished to 8/11 and 5/11 for dp6, and to 1/3 and 1/3 for dp7. Despite the high number of chains obtained (dozen to thousands), we still get more than 1% of acceptable solutions in 10/13 complexes for dp5 and 7/11 complexes for dp6. Note that the dp7 GAG (3ina: HE bound to Heparinase) showed a very good prediction accuracy when considering the length of the ligand: a best-RMSD solution at 2.1 Å RMSD, and 13% of acceptable solutions. This outstanding performance is consistent with the general observation that all docking approaches perform significantly better on the complexes formed with GAG-specific enzymes: the GAG recognition in a well-defined cavity with specific complementarity between the enzyme and the substrate is easier to predict than the recognition of the GAG on the protein surface lacking a defined cavity. [15]

The relatively low performance of dp6 assembling for complex 1bfc is probably due to the quite bad sampling by AD3 of fragments 1 and 2, that results from a shift of the poses toward the binding sites of fragments 3 and 4 (see section AutoDock 3 results). This substantially decreased the number of compatible pairs of acceptable poses for fragments 1 and 2. For complex 1rid, the low performance of the assembly of short chains (dp5) is less expected, given the rather good sampling of its fragments from AD3 (comparable to 2axm). Given the better performance of dp7 assembly for 1rid, this might indicate that for this complex the interactions of the GAG within the whole binding site are essential for establishing the specificity of binding. Finally, for complexes 3c9e and 2jcq, the bad sampling of fragment 3 and 6 respectively prevented the assembling of dp6 chains.

Table 2. Docking results after fragments assembly.

length	Complex (PDB ID)	Chain	bound docking			unbound docking		
			Best RMSD (Å)	% acceptable	Number of chains	Best RMSD (Å)	% acceptable	Number of chains
dp5	1bfc	1 - 5	6.9	0	64	6.2	0	198
		2 - 6	3.1	80	15	4.9	1	75
	1fq9	1 - 5	1.8	45	487	-	-	-
		2 - 6	1.7	74	288	-	-	-
	1xmn	1 - 5	2.9	31	216	3.6	11	985
		2 - 6	2.9	22	158	3.5	3	1540

	2axm	1 - 5	2.8	7	218	3.8	5	230
		2 - 6	4.2	6	93	5.6	0	449
	1rid	1 - 5	8.6	0	343	-	-	-
		2 - 6	6.7	0	125	-	-	-
		3 - 7	9.0	0	343	-	-	-
	1gmn	1 - 5	2.5	69	155	2.9	8	486
	2jcq	1 - 5	1.6	3	153	5.8	0	612
		2 - 6	3.6	1	69	5.8	0	522
		3 - 7	6.1	0	153	5.9	0	612
	2hyv	1 - 5	3.9	1	248	5.4	0	562
	3ina	1 - 5	2.3	1	129	9.1	0	418
		2 - 6	2.1	3	206	6.3	0	294
		3 - 7	2.4	1	129	5.5	0	418
	3mpk	1 - 5	7.7	0	49	6.3	0	312
		2 - 6	2.1	40	68	2.7	39	238
	3c9e	1 - 5	5.5	0	37	4.3	6	121
		2 - 6	7.7	0	51	4.7	1	249
	4ak2	1 - 5	2.4	6	71	2.7	2	772
		2 - 6	5.9	0	86	3.9	3	442
	4c4n	1 - 5	4.7	0	233	4.2	0	924
		2 - 6	5.6	0	248	5.8	0	1483
dp6	1bfc	1 - 6	4.4	19	32	5.9	0	717
	1fq9	1 - 6	1.4	86	1211	-	-	-
	1xmn	1 - 6	2.6	50	980	3.6	7	1626
	2axm	1 - 6	2.9	14	367	4.8	0	1054
	1rid	1 - 6	5.8	0	1060	-	-	-
		2 - 7	4.9	0	275	-	-	-
	2jcq	1 - 6	3.1	2	565	5.5	0	1122
		2 - 7	6.0	0	505	5.9	0	853
	3ina	1 - 6	2.0	6	390	6.1	0	1179
		2 - 7	2.0	2	393	9.3	0	557
	3mpk	1 - 6	7.0	0	73	6.6	0	506
	3c9e	1 - 6	10	0	85	6.3	0	578
	4ak2	1 - 6	3.0	3	264	3.4	3	825
	4c4n	1 - 6	4.5	0	1073	6.8	0	968
dp7	1rid	1 - 7	5.4	0	583	-	-	-
	2jcq	1 - 7	5.3	0	422	5.5	0	1053

Fragment-based versus whole ligand docking.

We compared results obtained by fragment assembly to previous results obtained with AD3 alone (Table 3).[15] AD3 failed to find acceptable solutions for 7/13 complexes. Particularly, no solutions within 10 Å RMSD were found for the three dp7 ligands. In contrast, our fragment assembly method failed only in 4/13 cases and could find solutions within 5.3 Å RMSD for all these three dp7 ligands. However, only 50 solutions were obtained by AD3 in each case, versus 32 to 1211 solutions by fragments assembly. To compare similar numbers of solutions, we clustered our assembly solutions at 3.0 Å and retained the centers of the 50 most-populated clusters (except for 1bfc for which we had obtained only 32 solutions). The clustering lead to only a slight decrease in docking performance in terms of best-RMSD solution, the number of complexes with acceptable solutions decreasing from 9/13 to 8/13. The percentage of

acceptable solutions remained significantly higher than with AD3 for most cases. The decrease in precision is particularly significant for dp7 GAGs, indicating that a higher number of solutions should be considered for such long ligands.

Table 3. Docking results for the whole ligand, obtained either by AD3, or by chain assembly and clustering at 3.0 Å (bound docking).

Complex (PDB ID)	length	Best RMSD[c] (Å)		% acceptable solutions		Number of solutions		Qualitative comparison [a]
		assembly	AD3	assembly	AD3	assembly	AD3	
1bfc [b]	dp6	4.4	3.7	19	10	32	50	-
1fq9	dp6	1.6	9.4	82	0	50	50	++
1xmn	dp6	3.2	8.7	40	0	50	50	++
2axm	dp6	3.9	7.3	18	0	50	50	++
1rid	dp7	8.7	17	0	0	50	50	++
1gmn	dp5	2.5	1.6	63	23	48	50	~
2jcq	dp7	7.0	11	0	0	50	50	++
2hyv	dp5	4.6	5.8	2	0	50	50	+
3ina	dp7	2.4	17	14	0	50	50	++
3mpk	dp6	7.3	4.4	0	4	50	50	-
3c9e	dp6	>10	1.7	0	4	40	50	--
4ak2	dp6	4.1	3.8	2	4	50	50	~
4c4n	dp6	7.8	5.0	0	2	46	50	~

[a] [- - / - / ~ / + / ++] : assembly [much less / less / equivalently / more / much more] effective than AD3 docking of whole ligand.
[b] No clustering was applied, as the initial number of chains was small.
[c] Those results were published for AD3 with another metric (RMSatd) accounting for local quality of the pose.[15] We give here the RMSD as a global quality metric.

Chain refinement

Distorted conformations of glycosidic linkages known to be produced by AD,[41] and those produced by averaging the overlapping parts in the chain, were corrected by an all-atom refinement procedure. The chains of monomers were converted into all-atom representation, a minimization and short MD equilibration of each chain-protein complex with AMBER resolved clashes due to this conversion, and reconnected the GAG monomers in low-energy conformations. The best-RMSD structures obtained after refinement for each complex are presented on Fig S1. Those refinement steps improved the best-RMSD solution in most cases by up to 3.4 Å, but did not significantly improve the percentage of acceptable solutions in most cases (S3 Table, Fig 3).

Effect of protein and ligand flexibility

To estimate the gain in performance obtained by flexible docking over rigid docking of the fragments,[27] for each bound fragment of a given complex, we compared the conformations of poses

obtained either by cognate docking (on the protein structure of that complex) or by cross-docking (on other protein structures). Among the 44 HE fragments, only one displayed more than 0.2 Å RMSD improvement in the best-fitted conformation among poses from cognate docking compared to cross-docking (Table 4). Moreover, two complexes in the benchmark, 1bfc and 1fq9, consist of the same GAG type binding to the same protein FGF-2 but either monomeric or bound to FGFR-1. The superposition of the binding site (residues within 5 Å from GAG) of the two complexes displays a similar backbone (0.4 Å RMSD), and a 1.5 Å RMSD of the side-chains. When comparing poses obtained by AD3 docking to one protein with the bound ligand of the other complex (cross-docking), the quality of sampling diminished significantly, as expected (Table 4). Yet poses at 5 Å RMSD from the cross-reference were still found for half of the fragments, and docking to 1bfc allowed to model an acceptable dp5 GAG when compared to 1fq9. The docking results obtained on the protein from 1fq9 complex in comparison to the experimental 1bfc complex are of apparently lower quality than for the reverse cross-docking, as all docking solutions were above 5 Å RMSD. This is explained by the fact that the first GAG units of 1bfc produce clashes with FGFR-1 (which is co-crystallized in case of 1fq9) when superimposing the two complexes on FGF-2.

Table 4. Best RMSD (Å) of fragment poses and chain poses by bound-docking and cross-docking.

	Docking on 1bfc		Docking on 1fq9	
	poses	chains	poses	chains
Comparison to 1bfc	frag1 4.6 frag2 4.6 frag3 2.8 frag4 1.8	dp5 3.8 dp6 5.3	frag1 8.5 frag2 7.0 frag3 6.0 frag4 5.1	dp5 6.9 dp6 6.8
Comparison to 1fq9	frag1 4.9 frag2 4.6 frag3 5.0 frag4 6.9	dp5 4.8 dp6 6.1	frag1 2.5 frag2 1.7 frag3 2.4 frag4 4.9	dp5 1.8 dp6 2.0

In gray/white background: bound/cross docking

Unbound docking.

The performance of the unbound docking by AD3 is in general worse compared to bound docking, yet 9 out of 44 fragments are better sampled by unbound than by bound docking (Table 1). In 3ina and 2axm, the misplacement of the side chain of K255 and R122 respectively, which establish polar contacts with the 5th GAG monomer in 3ina and the 5th and 6th in 2axm, might partially explain the worse sampling of fragment 4 (which has the 5th monomer at his center) and of fragments 5 and 6 respectively. In 1bfc, the misplaced side chain of R121 clashes with the bound position of the 1st and 2nd monomer, impairing the correct placement of fragments 1 and 2. After assembly, we could sample acceptable dp5 chains (RMSD ≤ 5 Å) solutions for 8 of the 11 complexes, and good dp5 chains (RMSD ≤ 3 Å) for 3 of them (Table 2).

We sampled acceptable dp6 chains for 3 of 9 complexes, and one dp7 chain at 5.5 A RMSD for one of the two dp7 complexes. Among 75 to 1626 sampled chains, we get 1% or more acceptable solutions in 7/11 complexes for dp5 and 2/9 complexes for dp6. We did not find a direct correlation between the interface RMSD of the protein model toward the protein bound structure and the quality of the docking results. The sampling of acceptable solutions for the two dp7 complexes, 2jcq and 3ina, is impaired by the bad sampling by AD3 of fragments 5 and 4 respectively (Table 1). Despite the above mentioned misplacement of a critical side chain in 1bfc and 2axm, one acceptable dp6 chain could be sampled for each of those complexes. The refinement of the best chain for each case by applying a minimization and a short MD simulation did not change the obtained results significantly in terms of the RMSD to the experimental structure. Among the 14 cases with acceptable best models, the RMSD increased by 0.5 – 1.9 Å in 8 cases and decreased by 0.1 – 0.9 Å in 6 cases (S3 Table).

Discussion

Progress in protein-GAG docking

Our new non-incremental fragment-based approach allowed to dock dp5 – 6 GAGs with an accuracy of 5 Å for 11/13 – 8/11 of bound cases and 8/11 – 3/9 of unbound cases, and we could sample a dp7 GAG at 3 – 5.5 Å RMSD for 1 out of 3 bound – unbound docking cases. Moreover, several percentage of acceptable solutions could be obtained for almost all successful cases (Table 2). While this performance can be regarded as low compared to current standards in small-ligand docking, they do constitute a significant improvement in the state-of-the-art for dp5-7 GAGs. Both sampling and scoring are challenging for long GAGs docking, and in this study we concentrated on the first challenge: our method brings an essential progress in terms of placement for long GAGs in comparison to all other available software. The fact that GAGs in ECM are long heterogeneous periodic polymers hinders the application of many classical computational approaches developed for short peptides or small molecules. Those limitation were particularly pointed by the CAPRI experiment ("Critical Assessment of PRediction of Interactions"): despite being provided structures of the bacterial surface protein Bt4661 and its ligand heparin very close to the bound structures (0.78 and 0.23 Å respectively), the entire modeling community could produce only 5 medium-quality models (lRMSD < 5 Å) of their protein-GAG complex, with a best model at 3.2 Å RMSD, among 256 submitted models. In this context, our fragment-based approach achieved a major advance in the field of GAG-protein docking, providing at least 1% of medium-quality models for most test cases. Regarding the scoring problem, the choice of a procedure to assemble AD3 scores of fragments into a full-chain score would be not trivial. A better choice would probably be to train a dedicated scoring function for samples obtained by fragment-based docking, [42] which will be a subject of our further work.

We present in the rest of the discussion some considerations such as requirements and applicability of such a non-incremental fragment-based approach in general, and for GAGs in particular.

Non-incremental fragment-based docking.

In a classical incremental fragment-based method, at least one fragment, usually *a priori* known, must be docked correctly. Else, starting from an only acceptable pose to construct the next fragments will add a bias that would propagate along the chain of fragments. In the absence of knowledge on which fragment could provide accurate and precise docking poses, considering all possibilities in many incremental docking experiment is inefficient for large ligands. Then, all fragments should be considered as equivalent and assembled simultaneously. This is particularly the case for complexes where the determinants for binding affinity and specificity are evenly shared among the fragments (i.e. there is not one particular "hot spot" in the fragment) as in periodic GAG ligands.[27]

In this study, we removed from consideration the terminal GAG units that do not bind the protein in the experimental structure. In a real case, this information would not be known a priori, while fragment docking will force the binding of all assembled units. Further longer MD simulation could be performed to take into account this effect, which could be overlooked when docking poses are analyzed after a short MD refinement. Therefore, the results obtained by docking should be further analyzed by MD if, for example, free energy calculations are needed to be performed to characterize the obtained binding poses more rigorously or to define a minimal GAG binding unit.

Fragments sampling.

For a simultaneous assembly to succeed, each fragment must have been correctly sampled. Here, most of the badly sampled fragments by AD3 correspond to terminal parts of the GAGs, thus allowing in principle the assembling of some contiguous correct poses. For all but one docking run, acceptable poses are found in at least one of the correct binding sites of the docked fragment type. This might be explained by the fact that interactions within one site are significantly stronger than within the others binding sites, according to AD3 scoring. This would drag the poses to this particular binding site and deplete the other binding sites. To check this hypothesis, we computed the ratio between the number of poses with smaller deviation toward one or another bound fragment of the same type for each docking run (S4 Table). In complexes 1bfc and 3mpk, the badly sampled fragments do correspond to a strong shift of the docking poses toward one of their correct binding sites. For 1bfc, this is consistent with the experimental data showing that a HE dp4 represents an energetically essential binding motif to occupy FGF2 high affinity HE binding site.[43] In contrast, in 2hyv, none of the two binding sites for fragments 1 and 3 provided interactions strong enough according to AD3 scoring, resulting in a bad sampling for both fragments. This

might indicate that fragment 2 could potentially represent a key binding motif of GAG in terms of binding affinity for the 2hyv complex.

Fragment libraries

A main advantage of fully flexible docking is its ability to account for the induced fit mechanism of binding. Docking flexible fragments on the bound form of the protein should favour the sampling of conformations specifically induced by the conformation of their binding site (especially by the side-chains orientations of protein residues). Our comparison of pose conformations docked on the different bound proteins suggest that flexible ligand docking does not induce substantial changes in the total pool of ligand conformations. This suggests that the GAG bound conformations are not significantly specific for the binding site, but can be “picked up” in a large enough ensemble of conformations. This ensemble could be obtained from a reasonably exhaustive sampling without using the bound protein structure, such as docking to other GAG-binding proteins or MD simulations of the unbound fragments. Therefore, rigid docking of fragment libraries, which is substantially less computationally expensive, could be an alternative to fully flexible docking and will be tested in our future research.

Fragment assembly

The second requirement for the methodology to be successful is a criteria for selecting compatible poses that should be tight enough to discard wrong poses but loose enough to account for inaccuracies even in the best-RMSD fragment poses. The choice of a fixed RMSD overlap cut-off is not suitable, as the RMSD would strongly depend on the size of the overlapping part from one GAG to another, and on the spatial distribution of the atoms retained in the coarse grained representation. Instead, we chose to apply the smallest cut-off that retains at least 10^3 or 10^4 chains for dp5 and dp6/7 for each complex, respectively. Our results show that his procedure is more successful in filtering out incorrect solutions than selection by AD3 scoring. As expected, chain-assembly was particularly suited to retain correct poses of terminal fragments, which AD3 tends to rank worse than poses of the same fragment type located at the binding site of a central fragment.

GAGs heterogeneity

Here we have approximated each GAG as a regular periodic polysaccharide.In nature, long GAG chains do not have any regular structure in terms of disaccharide units composition. Theyrepresent highly heterogeneous samples in comparison to the homogeneous samples that could be obtained by chemical synthesis. Here, the experimental structures from the PDB used for verification of our methodology contain mostly stereo-regular GAGs, justifying the proposed approximation to consider the GAG as homogeneous a priori. An exception is complex 3ina, the 5th SGN residue of the heparin being replaced by a SUS. This did not prevent our method to find near-native solutions by bound docking of regular

heparin, and decent solutions for residues 3-to-7 by unbound docking (Table 2). But in the general case, when a protein-GAG complex structure is unknown, and the molecular docking technique is applied, this approximation could in principle not always be correct. The heterogeneity of GAGs represents a general challenge in computational analysis of GAG interactions, and evaluating the impact of such heterogeneity on the docking quality and the homogeneity of the binding site could be the subject of a further study.

Strengths, limitations and perspectives

The fragment-based approach allowed to obtain a higher percentage of acceptable solutions and/or a better best-RMSD solution than whole-ligand docking with AD3, especially for longer GAGs (dp7). The decrease we observe in the percentage of successful cases when assembling longer chains can be interpreted as the increase in the probability that at least one fragment is incorrectly sampled by AD3, or that two fragments have too few acceptable poses to find a correspondingly connected pair. The lower performance for a long ligand is therefore not systematic, which partially explains the excellent results obtained for dp7 in the case of 3ina. Due to its fragment-based nature, this docking approach is in principle applicable to any protein-GAG system without limitations related to GAG length. Yet this absence of technical limit does not ensure that we would reach the desired accuracy for longer GAGs. This should be verified on experimental structures of complexes with longer bound GAG, which are not available at the moment. Nevertheless, experimental data originating from NMR, MS or SAXS could be very useful to guide the selection of docking solutions for further analysis.

In addition to ligand size and flexibility, we evaluated the influence of the protein conformation on the docking and assembly results. Unbound docking and docking on homology models yielded quantitatively worse results, which is a common feature in the molecular docking field, but qualitatively comparable ones to the results from the bound docking: The bound docking results we obtain are within the accuracy of 5 Å RMSD to the experimental structure for all complexes and of 3 Å RMSD for half of them; The unbound docking results are within the accuracy of 5 Å for most complexes, and even using a homology model at 3.8 Å i-RMSD from the bound form allowed to retrieve a dp5 pose at less than 3 Å RMSD.

Conclusion

In this study, we developed the first automated fragment-based method to dock GAGs locally. The method combines flexible docking of dp3 GAG fragments by AutoDock 3 with combinatorial assembly of the compatible poses into GAGs chains, followed by fully flexible refinement. The method was successfully applied to a benchmark of 13 protein-GAG complexes containing GAGs of different types (heparin, chondroitin sulfate and hyaluronic acid) with the length of dp5-dp7. This is the first reported assembly method to dock diverse dp5-7 GAGs with an accuracy below 5 Å RMSD in most cases. In addition, we observe that the conformations of the GAG docking poses are not significantly specific for

different binding sites, suggesting that rigid docking of dp3 fragment libraries could be an alternative to fully flexible docking, being significantly less computationally expensive. In summary, our novel fragment assembly method specifically developed to treat the complexes of proteins with long GAG can provide a higher level of structural details that should improve our understanding of the molecular basis of the interactions in those challenging systems.

Acknowledgment

The authors thank Sjoerd J. de Vries for useful discussions and critical reading of the manuscript.

Supporting information

S1 Table. Protein-GAG benchmark

S2 Table. Comparison of pose filtering by assembly or by AD3 scoring (bound docking).

S3 Table. Comparison of solutions obtained by poses assembly before and after refinement (bound docking).

S4 Table. Rratio of poses toward two bound fragments.

S1 Fig. Chain with best RMSD obtained for each complex, after refinement by all-atom minimization (bound docking). Same representation code as Figure 3. Each complex is showed in two orthogonal views. For some complexes, part of the protein was removed to improve the visibility of the ligand. The pictures were created using PyMOL.

References

1. Perrimon, N., Bernfield, M. *Nature*, 2000, **404**, 725–728.
2. Kreuger, J., Spillmann, D., Li, J., Lindahl, U. *J. Cell Biol.*, 2006, **174**, 323–327.
3. Habuchi, H., Habuchi, O., Kimata, K. *Glycoconj. J.*, 2004, **21**, 47–52.
4. Jin, L., Abrahams, J. P., Skinner, R., Petitou, M., Pike, R. N., Carrell, R. W. *Proc. Natl. Acad. Sci. U. S. A.*, 1997, **94**, 14683–8.
5. Scharnweber, D., Hübner, L., Rother, S., Hempel, U., Anderegg, U., Samsonov, S. A., Pisabarro, M. T., Hofbauer, L., Schnabelrauch, M., Franz, S., Simon, J., Hintze, V. *J. Mater. Sci. Mater. Med.*, 2015, **26**, 232.
6. Forsberg, M., Holmborn, K., Kundu, S., Dagälv, A., Kjellén, L., Forsberg-Nilsson, K. *J. Biol. Chem.*, 2012, **287**, 10853–10862.
7. Kraushaar, D. C., Rai, S., Condac, E., Nairn, A., Zhang, S., Yamaguchi, Y., Moremen, K., Dalton, S., Wang, L. *J. Biol. Chem.*, 2012, **287**, 22691–700.
8. Bernfield, M., Götte, M., Park, P. W., Reizes, O., Fitzgerald, M. L., Lincecum, J., Zako, M. *Annu. Rev. Biochem.*, 1999, **68**, 729–77.
9. Rostand, K. S., Esko, J. D. *Infect. Immun.*, 1997, **65**, 1–8.
10. Clement, A. M., Sugahara, K., Faissner, A. *Neurosci. Lett.*, 1999, **269**, 125–8.

11. Gama, C. I., Tully, S. E., Sotogaku, N., Clark, P. M., Rawat, M., Vaidehi, N., Goddard, W. A., Nishi, A., Hsieh-Wilson, L. C. *Nat. Chem. Biol.*, 2006, **2**, 467–73.
12. Rogers, C. J., Clark, P. M., Tully, S. E., Abrol, R., Garcia, K. C., Goddard, W. A., Hsieh-Wilson, L. C. *Proc. Natl. Acad. Sci. U. S. A.*, 2011, **108**, 9747–52.
13. Peach, R. J., Hollenbaugh, D., Stamenkovic, I., Aruffo, A. *J. Cell Biol.*, 1993, **122**, 257–64.
14. Imberty, A., Lortat-Jacob, H., Pérez, S. *Carbohydr. Res.*, 2007, **342**, 430–9.
15. Samsonov, S. A., Pisabarro, M. T. *Glycobiology*, 2016, **26**, 1–12.
16. Gandhi, N. S., Freeman, C., Parish, C. R., Mancera, R. L. *Glycobiology*, 2012, **22**, 35–55.
17. Agostino, M., Mancera, R. L., Ramsland, P. A., Yuriev, E. *J. Mol. Graph. Model.*, 2013, **40**, 80–90.
18. Agostino, M., Sandrin, M. S., Thompson, P. E., Yuriev, E., Ramsland, P. A. *Mol. Immunol.*, 2009, **47**, 233–46.
19. Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., Olson, A. J. *J. Comput. Chem.*, 1998, **19**, 1639–1662.
20. Hofmann, T., Samsonov, S. A., Pichert, A., Lemmnitzer, K., Schiller, J., Huster, D., Pisabarro, M. T., von Bergen, M., Kalkhof, S. *Methods*, 2015, **89**, 45–53.
21. Pichert, A., Samsonov, S. A., Theisgen, S., Thomas, L., Baumann, L., Schiller, J. J., Beck-Sickinger, A. G., Huster, D., Pisabarro, M. T. *Glycobiology*, 2012, **22**, 134–145.
22. Hintze, V., Samsonov, S. A., Anselmi, M., Moeller, S., Becher, J., Schnabelrauch, M., Scharnweber, D., Pisabarro, M. T. *Biomacromolecules*, 2014, **15**, 3083–3092.
23. Gandhi, N. S., Mancera, R. L. *J. Chem. Inf. Model.*, 2011, **51**, 335–358.
24. Bitomsky, W., Wade, R. C. *J. Am. Chem. Soc.*, 1999, **121**, 3004–3013.
25. Griffith, A. R., Rogers, C. J., Miller, G. M., Abrol, R., Hsieh-Wilson, L. C., Goddard, W. A. *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 13697–13702.
26. Samsonov, S. A., Gehrcke, J. P., Pisabarro, M. T. *J. Chem. Inf. Model.*, 2014, **54**, 582–592.
27. Brenke, R., Kozakov, D., Chuang, G.-Y., Beglov, D., Hall, D., Landon, M. R., Mattos, C., Vajda, S. *Bioinformatics*, 2009, **25**, 621–7.
28. Imai, T., Oda, K., Kovalenko, A., Hirata, F., Kidera, A. *J. Am. Chem. Soc.*, 2009, **131**, 12430–40.
29. Sadjad, B., Zsoldos, Z. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, 2011, **8**, 1120–1133.
30. Agostino, M., Gandhi, N. S., Mancera, R. L. *Glycobiology*.
31. Babik, S., Samsonov, S. A., Pisabarro, M. T. *Glycoconj. J.*, 2017, **34**, 427–440.
32. Beauchene, I. C. De, Vries, S. J. De, Zacharias, M. 2016, **44**, 4565–4580.
33. Beauchene, I. C. De, Vries, S. J. De, Zacharias, M. 2016, 1–21.
34. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., Woods, R. J. *J. Comput. Chem.*, 2008, **29**, 622–55.
35. Huige, C. J. M., Altona, C. *J. Comput. Chem.*, 1995, **16**, 56–79.
36. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., Woods, R. J. *J. Comput. Chem.*, 2005, **26**, 1668–1688.
37. Samsonov, S. A., Pisabarro, M. T. *Carbohydr. Res.*, 2013, **381**, 133–137.
38. Samsonov, S. A., Bichmann, L., Pisabarro, M. T. *J. Chem. Inf. Model.*, 2015, **55**, 114–124.
39. Sattelle, B. M., Shakeri, J., Almond, A. *Biomacromolecules*, 2013, **14**, 1149–59.
40. Raghuraman, A., Mosier, P. D., Desai, U. R. *J. Med. Chem.*, 2006, **49**, 3553–62.
41. Nivedha, A., Thieker, D., Wood, robert *J Chem Theory Comput*, 2016, **8**, 583–592.
42. Vajda, S., Kozakov, D. 2013, **37**, 62–70.
43. Faham, S., Hileman, R. E., Fromm, J. R., Linhardt, R. J., Rees, D. C. *Science*, 1996, **271**, 1116–20.

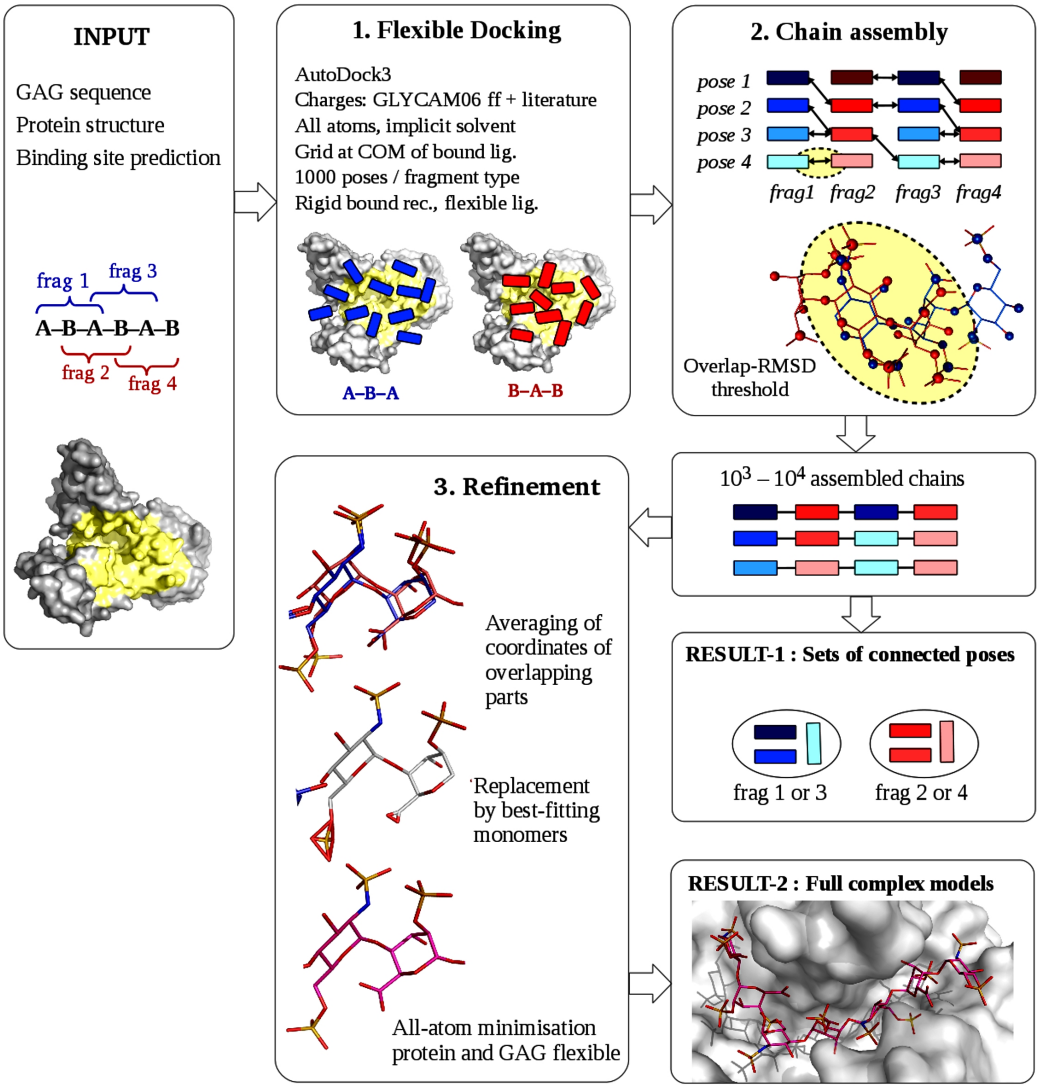

Fig 1. Flowchart illustrating the fragment-based docking approach.

The GAG sequence, the protein structure and the binding area are considered as known. The GAG sequence is cut in overlapping dp3 fragments. Each fragment is docked in a fully flexible mode on the protein by AD3, ussing a docking grid centered on the binding area. The docking poses are then assembled into connected chains, pose-pose connectivity being determined by an RMSD criterion of the overlapping parts. The percentages of acceptable poses (e.g. close to any fragment of the bound GAG in the experimental structure) is computed either among the poses participating into chains or among all the docking poses, in order to evaluate the enrichment in acceptable solution provided by chain assembling. Then, chains of connected poses are converted into GAG structures with correct geometry, refined by all-atom energy minimization, and compared to the bound GAG.

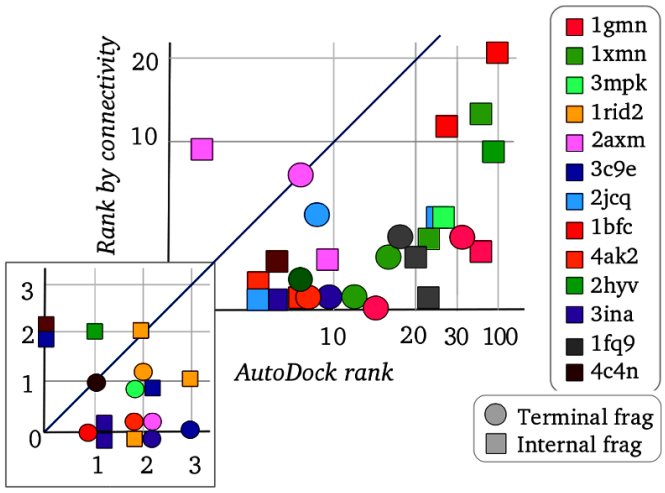

Fig 2. Poses filtering by chain assembly *vs* AD3 scoring.

The ranking of acceptable solutions by connectivity or by AD scoring are compared for each fragment of each complex. Terminal fragments in the whole ligand are represented as discs, the others as squares. Full view (upper graph) in logarithmic scale, and zoom (lower graph) in linear scale. For ligands with same values, the coordinates of one of the corresponding points are shifted to avoid full superposition.

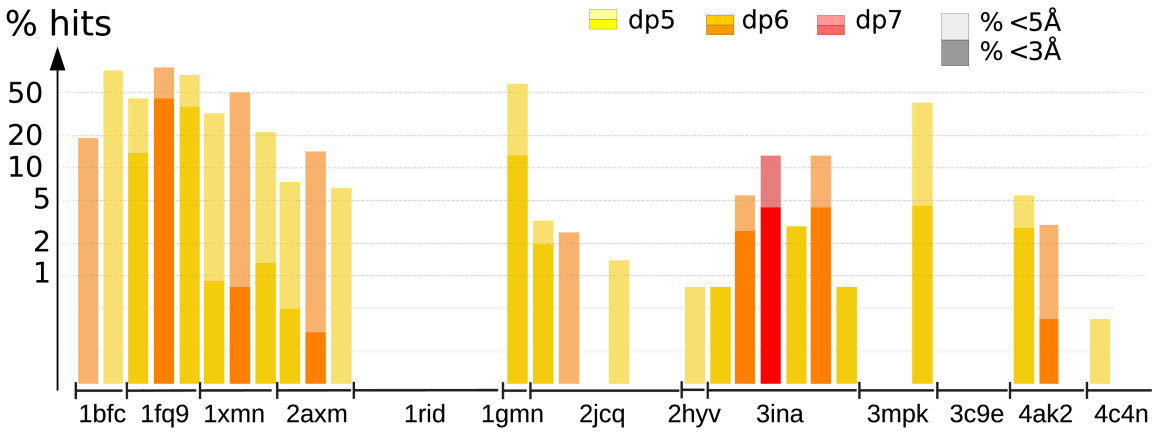

Fig 3. Quality of docking solutions obtained by assembly and refinement.

One bar is drawn per assembly mode, colored according to the chain length (dp5 in yellow, dp6 in orange and dp7 in red). The total height of the bar represents the percentage of acceptable solutions, the darker part of the bar represents the percentage of good solutions. We use a logarithmic scale for more clear representation of the data. The corresponding values are all presented in S3 Table.

S1 Table. Protein-GAG benchmark.

PDB ID	Protein	Ligand	Length	Resolution(Å)	Unbound PDB ID	% identity	Interface RMSD* (Å)
2axm	FGF-1	HE	dp6	2.2 [46]	1rg8	100	1.4
1bfc	FGF-2	HE	dp6	2.9 [47]	1rg8	56	1.8
1fq9	FGF-2/FGFR-1	HE	dp6	3.0 [48]	-	-	-
1gmn	NK1 [HGF)	HE	dp5	2.3 [49]	3sp8	81	3.8
1rid	VCP	HE	dp7	2.1 [50]	-	-	-
1xmn	Thrombin	HE	dp6	1.9 [51]	2bdy, 3nxp	98, 98	3.4
2hyv	Annexin 2A	HE	dp5	2.3 [52]	1w3w	46	2.5
2jcq	CD44	HA	dp6	1.3 [53]	4pz3	86	1.7
3c9e	Cathepsin K	CS-4	dp6	1.8 [54]	4x6h	100	0.9
3ina	Heparinase I mut	HE	dp7	1.9 [55]	3ikw	98	1.6
3mkp	VFT2	HE	dp5	2.8 [56]	2qj2	99	1.4
4ak2	BT4661 [Suse-like)	HE	dp6	1.4 [57]	4ak1	100	0.9
4c4n	Hedgehog morphogen	HE	dp6	2.4 [58]	3k7i	91	1.2

** all heavy atoms*

S2 Table. Comparison of pose filtering by assembly or by AD3 scoring (bound docking).

Complex (PDB ID)	Nb successful fragments [1]		Nb near-native solutions	
	Assembly	AD3	Assembly	AD3
1gmn	3	2	94	7
2hyv	2	2	8	4
2axm	4	3	19	18
1bfc	3	2	125	37
1fq9	4	3	62	8
1xmn	4	4	94	21
3c9e	2	2	5	3
4ak2	3	3	19	4
4c4n	2	3	7	6
3mpk	3	3	88	15
2jcq	3	3	37	12
3ina	5	2	19	2
1rid	4	3	9	4

[1] fragments for which at least one near-native solution was selected.

S3 Table. Results obtained by poses assembly before and after refinement (bound docking). Cases for which no near-native solution was found neither before nor after refinement are not shown.

GAG length	complex	chain	before refinement		Nb of chains	after refinement				
			Best RMSD (Å)	% near-native		Best RMSD (Å)	% near-native	RMSD top 1 (Å)	Best RMSD in top 10 (Å)	Rank of best chain
dp5	1gmn	1 - 5	2.3	63	155	2.5	61	3.2	3.2	82
	2hyv	1 - 5	4.8	1	248	3.9	1	3.9	3.9	1
	2axm	1 - 5	3.3	7	218	2.8	7	>10	4.3	31
		2 - 6	3.9	6	93	4.2	6	5.0	5.0	88
	1bfc	2 - 6	3.7	80	15	3.1	80	4.4	3.1	4
	1xmn	1 - 5	3.6	30	216	2.9	31	5.2	4.2	59
		2 - 6	3.2	16	158	2.9	22	6.5	3.7	14
	1fq9	1 - 5	1.9	45	487	1.8	45	>10	1.9	83
		2 - 6	1.6	73	288	1.7	74	6.2	1.7	2
	3c9e	1 - 5	4.2	5	37	5.5	0	>10	5.7	36
	3mpk	2 - 6	2.7	40	68	2.0	40	>10	9.4	32
	4ak2	1 - 5	2.8	6	71	2.4	6	>10	5.8	39
		1 - 5	4.9	1	233	4.7	0	9.0	6.6	55
	2jcq	1 - 5	2.3	4	153	1.6	3	>10	6.0	105
		2 - 6	4.5	1	69	3.6	1	>10	8.5	47
	3ina	1 - 5	2.7	2	204	2.3	1	>10	8.1	84
		2 - 6	2.0	3	206	2.1	3	>10	>10	82
		3 - 7	3.0	0	204	2.4	1	>10	>10	120
dp6	2axm	1 - 6	3.6	11	367	2.9	14	7.7	5.5	349
	1bfc	1 - 6	5.2	0	32	4.4	19	4.4	4.4	1
	1xmn	1 - 6	3.4	47	980	2.6	50	4.4	3.8	789
	1fq9	1 - 6	1.8	84	1211	1.4	86	2.3	2.0	1027
	4ak2	1 - 6	3.6	3	264	3.0	3	>10	>10	3.0
	4c4n	1 - 6	5.1	0	1073	4.5	0	>10	>10	566
	2jcq	1 - 6	4.2	2	565	3.1	2	>10	5.5	113
	3ina	1 - 6	2.4	6	390	2.0	6	>10	>10	147
		2 - 7	2.3	2	393	2.0	2	>10	2.0	10
	1rid	1 - 6	5.0	0	1060	5.8	0	>10	7.2	91
		2 - 7	4.8	0	275	4.9	0	>10	8.8	993
dp7	3ina	1 - 7	2.4	13	162	2.1	13	>10	>10	10

S4 Table. Ratio of bound docking poses toward two bound fragments.

Complex (PDB ID)	fragments	ratio
2axm	frag1 / frag3	0.42
	frag2 / frag4	0.21
1bfc	frag1 / frag3	0.10
	frag2 / frag4	0.16
1fq9	frag1 / frag3	0.24
	frag2 / frag4	0.12
1gmn	frag1 / frag3	0.86
1rid	frag1 / frag3	0.25
	frag2 / frag4	0.49
	frag3 / frag5	0.81
2jcq	frag1 / frag3	0.31
	frag2 / frag4	0.68
	frag3 / frag5	0.47
2hyv	frag1 / frag3	0.53

complex	fragments	ratio
3ina	frag1 / frag3	0.11
	frag2 / frag4	0.08
	frag3 / frag5	0.08
3mpk	frag1 / frag3	0.15
	frag2 / frag4	0.38
4c4n	frag1 / frag3	0.48
	frag2 / frag4	0.41
4ak2	frag1 / frag3	0.45
	frag2 / frag4	0.42
3c9e	frag1 / frag3	0.41
	frag2 / frag4	0.87
1xmn	frag1 / frag3	0.62
	frag2 / frag4	0.37