NMR METHODOLOGY FOR THE CHARACTERIZATION OF PROTEIN-GLYCOSAMINOGLYCAN INTERACTIONS

by

QI GAO

(Under the Directions of James H. Prestegard and Joshua S. Sharp)

ABSTRACT

This dissertation focuses on the characterization of interactions of glycoproteins with glycosaminoglycans (GAGs) using primarily nuclear magnetic resonance (NMR) methodology. Glycoproteins are proteins carrying covalently linked glycans; many glycoproteins play crucial roles in human physiology and disease. Many function by interacting with other glycans, including the highly sulfated and structurally diverse glycans found in the extracellular matrix (GAGs). The characterization of these systems is best performed on properly glycosylated forms produced by the expression of the proteins in mammalian cell culture. However, mammalian protein characterization by traditional NMR methodology is challenging since the uniform isotopic labeling with isotopes needed for NMR observation (^{13}C, ^{15}N and ^{2}H) becomes extraordinarily expensive and deuteration is very detrimental to cell growth. I describe an alternative methodology based on sparse labeling with single isotopically enriched amino acids. The primary limitation of sparse labeling is that the connectivities between isotopically labeled residues are lost. As a result, traditional triple resonance assignment approaches are no longer applicable. To overcome this obstacle, a new strategy is developed to assign the crosspeaks in a heteronuclear single quantum coherence (HSQC) spectrum of a sparsely labeled protein sample.

This strategy uses a genetic algorithm to search for an optimal pairing of HSQC crosspeaks with labeled sites based on the experimental and predicted values of chemical shifts, nuclear Overhauser effects and residual dipolar couplings. This methodology has been validated on a set of previously assigned proteins and a sparsely labeled two-domain construct from a glycosylated signaling protein, Robo1-Ig1-2. Using available NMR assignments, I have characterized Robo1-Ig1-2 interacting with two heparan sulfate tetramers and an octamer using a series of high structural content NMR experiments. A model of this complex has been generated and used to rationalize how heparan sulfate may modulate interaction with Robol's signaling partner, Slit2. This methodology was then applied to study the interaction between glycoprotein LAR and heparan sulfate, another glycoprotein-GAG interaction that is important for signal transduction. I present a model for LAR-Ig1-2 interacting with a particular heparan sulfate pentasaccharide, and use this to assess heparan sulfate modifications that may lead to enhanced binding and induced LAR dimerization.

INDEX WORDS: NMR, Glycosylated mammalian protein, Sparse labeling, NMR assignment, Chemical shift prediction, Nuclear Overhauser effects (NOEs), Residual dipolar couplings (RDCs), Chemical shift perturbation (CSP), Saturation transfer difference(STD), Transferred NOE (trNOE), Lanthanide binding peptide, Pseudocontact shift (PCS), Paramagnetic relaxation enhancement (PRE), Molecular docking, Molecular dynamics (MD) simulation, Roundabout 1 (Robo1), Leukocyte common antigen-related protein (LAR), Glycosaminoglycan (GAG), Heparan sulfate (HS).

NMR METHODOLOGY FOR THE CHARACTERIZATION OF PROTEIN-GLYCOSAMINOGLYCAN INTERACTIONS

by

QI GAO

Bachelor of Science, Northwest University, China, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

NMR METHODOLOGY FOR THE CHARACTERIZATION OF PROTEIN-GLYCOSAMINOGLYCAN INTERACTIONS

by

QI GAO

Major Professor:	James H. Prestegard
Co-Advisor:	Joshua S. Sharp
Committee:	Ron Orlando
	Jeffery Urbauer

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

DEDICATION

To my dear mother, Yang Mu, and my dear father, Liancheng Gao, for giving me life and endless love.

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the help of many people. It is my great pleasure to express my sincere appreciation and acknowledge them for giving me guidance, courage, and support.

I would like to express my deepest gratitude to my advisor, Professor James Prestegard for being my mentor and guiding me to the field of NMR science. The past six years of study under his direction not only armed me as an analytical scientist with diverse skill sets but also built my systematic concepts to solve research problems. I am very grateful to Dr. Prestegard for sharing his vast knowledge of biology and chemistry and broadening my scientific horizons by providing me opportunities to work with excellent researchers on highly collaborative projects and communicate with talented scientists in different capacities. His careful attitude, sense of responsibility and enthusiasm and persistence for science have influenced me during my research process and will continue to inspire me in the rest of my life.

I would also like to thank Dr. Joshua Sharp for teaching me mass spectrometry which is a powerful analytical technology. I am grateful to him for giving me valuable guidance, providing great opportunities to know the mass spectrometry community and training me to become a comprehensive thinker and multi-tasker.

Thanks must also go to Dr. Jeffery Urbauer and Dr. Ron Orlando for serving as my committee members and sharing ideas that are of great help towards my study and research. I also would like to thank Dr. Urbauer and Ramona Urbauer for generously providing me access to instrumentation in their lab and providing intensive training.

I appreciate the time and efforts from all of my wonderful collaborators: Dr. Kelly Moremen's lab (Shuo Wang, Jeong-Yeh Yang, Annapoorani Ramiah, Pradeep Prabhakar), Gordon Chalmers, Dr. Geert-Jan Boons's lab (Chengli Zong, Maria Moure), Dr. Robert Woods's lab (David Thieker, Arunima Singh), Dr. Daniel Häussinger and Dr. Gottfried Otting. I thank them for their thoughts and materials that made projects progress nicely and efficiently.

Special thanks go to Dr. John Glushka for explaining basic NMR concepts and the design of pulse sequences, training me in NMR application and providing continual technical support. I especially thank John for sharing his thoughts and experiences in both science and music. I also want to thank Iris for being a lab mom who always comforts me, relieved my homesickness and supported my life during my graduate studies. Thanks are also due to Laura Morris for helping me solve all the computer problems I encountered and teaching me MD simulations. I want to thank current and previous lab members: David, Joy, Cheng-yu, Kari, Younghee, Alex, Chuck and Monique, for always creating a wonderful working environment where we can have great discussions, exchange ideas and progress together. I have learned a lot from every member.

I want to thank my best friends at UGA for their companionship, sharing my happiness and sorrows. Special thanks go to Joy and Yiwen, who supported and encouraged me throughout my dissertation writing. The journey here is enjoyable and unforgettable with all of you around. Thank you Fengtao, my boyfriend, for traveling overseas and being by my side. Last, I am extremely grateful to my mother and father. Thank you so much for providing me the best education. I feel so blessed to be your child and have grown up in a family full of happiness and love.

Thank you all for making the past six years the best time in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background

In order to properly develop, differentiate and function as tissues, cells detect and respond actively to different signaling transduction processes through a series of molecular events[1]. In mammals, these events are activated and mediated by a group of cell-cell adhesion, communication and signaling proteins[2-3]. Many of these are glycoproteins that can form diverse types of complexes and junctions to join cells and the extracellular matrix. Many glycoproteins share a similar structural topology: the extracellular region begins with one or more immunoglobulin-like (Ig-like) domains which are highly glycosylated, followed by zero to several fibronectin type III (FN3) repeats, a transmembrane domain, and a cytoplasmic domain which conveys the signal to intracellular systems. The majority of signaling pathways begin with interactions between the glycosylated terminal domains of these cell surface proteins and other molecules, including free-floating ligands and domains of other membrane proteins on the same cell (cis-interactions) or proximate cells (trans-interactions). A molecular understanding of the interactions between these glycosylated proteins and various ligands can reveal the complex mechanisms responsible for signaling in these pathways. This understanding can, in turn, facilitate intervention in these processes for biomedical purposes.

Glycosylation is the most common, yet most complex, protein post-translational modification. It affects protein function in many ways[4-5], including facilitating protein folding and complex assembly, guiding protein trafficking, aiding ligand recognition and triggering cell

signaling and immunogenicity. The common glycans found in mammals include hexoses (glucose, galactose and mannose), hexosamines, deoxyhexoses (fucose), uronic acids (iduronic acid, glucuronic acid) and sialic acids. These monosaccharide units are linked together by flexible glycosidic bonds to form various oligosaccharide chains. They are linked to the protein through N- or O- linkages to specific asparagine, serine or threonine residues. In general, for the pyranoses mentioned above, the conformation of the monosaccharide is fairly rigid except for IdoA which, in solution, has two different well-populated conformations ($^1C_4$-chair and $^2S_0$-skew boat). The motional flexibility of glycans is primarily the result of variations in glycosidic torsion angles which are defined as Φ and Ψ angles (in the α(1,4) linkages these are defined as O5-C1-O1-C4 and C1-O1-C4-C3 respectively). There are extra degrees of freedom when glycans are linked through exocyclic hydroxymethylenes (variation in the angle ω for 1-6 linked glycans). There are many challenges in glycosylated protein characterization, most of which are caused by the heterogeneity, complexity and flexibility of native glycans.

In order to trigger certain signaling pathways that regulate cell properties and function, glycoproteins on the cell surface commonly interact with other extracellular matrix (ECM) components. Among the different types of ECM constituents, proteoglycans, present in almost all tissues, are the most abundant. Proteoglycans consist of a core protein covalently linked to glycosaminoglycans (GAGs). GAGs are linear polysaccharides built from repeating dissaccharide units of an N-acetylated or N-sulfated hexoamine and either a uronic acid (glucuronic in chondroitin sulfate and glucuronic or iduronic in heparan sulfate) or galactose (keratin sulfate)[6]. Except for hyaluronan (a polymer of glucuronic acid and N-acetylglucosamine), all of the GAGs contain sulfates at different positions. GAGs have a significant degree of internal mobility, both in terms of variations in glycosidic bond torsion

angles and, for heparan sulfate, alterations in iduronic acid ring forms. GAGs typically interact with cell-surface glycoproteins through an interaction between the negatively charged sulfates and carboxylates and positively charged amino acids in the protein (lysine or arginine). The specificity and affinity of these interactions is believed to depend critically on the structure and sulfate distributions on the sugar chains. Protein-GAG complexes have important roles and functions in cell migration, proliferation and, by extension, cancer progression[7-8]. In the research presented here, the interaction of two glycoproteins with GAGs are completely characterized by solution-based NMR spectroscopy methodology.

1.2 NMR Methodology

One methodology that has the capacity to deal with the complexity and flexibility of glycans, as well as the structure of proteins to which they are attached, is nuclear magnetic resonance (NMR). Over the past half century, *de novo* protein structure determination by NMR has rapidly advanced: from single dimension to multiple dimension experiments and from small protein characterization (1-6K in size) to macromolecule complex determination (for example, those involving G-protein coupled receptors) [9-10]. High-resolution NMR spectroscopy has benefitted tremendously from developments and improvements in modern NMR cryogenic probes[11], high field superconducting magnets[12] and console electronics. In addition, there have been methodological developments such as multiple isotopic labeling strategies[13-14] , innovative pulse-sequence design[15-16], utilization of new data types[17-19], as well as coupling NMR with other structural[20-21] and simulation techniques[22-23].

Compared with other structure biology methods, such as X-ray crystallography, which is by far the dominant tool for high-resolution protein structure determination[24], and cryo-electron microscopy, which has been developing dramatically in the past 10 years [25], NMR still holds

advantages for dealing with certain types of biomolecules, for example, glycoproteins, and systems with high degrees of internal motions, for example glycans[26-28] and intrinsically disordered proteins [29].

Preparing samples with a homogeneous glycan composition has been difficult, and even when successful, the conformational heterogeneity of flexible glycans inhibits crystallization[30]. As a result, the mutation of specific glycosylation sites to eliminate glycosylation or the use of bacterial-hosts for expression without glycosylation have become alternatives for the crystallographic studies of glycoproteins. There are certainly concerns about the accuracy of functional interpretations based on structures of these non-glycosylated forms. This issue becomes far more serious when considering the fact that more than half of eukaryotic proteins are, in fact, glycoproteins[31]. In order to address some of these concerns, I attempted to develop and apply new combinations of NMR methodology to better describe the structure and function of glycosylated mammalian proteins and their ligands. These methods are briefly outlined below.

1.2.1 Chemical Shift Perturbation

Chemical shift perturbation, or CSP, is a widely used NMR experimental variable for studying protein-ligand binding[32]. CSP is very useful since both disassociation constants (*Kd*s) and binding site location can be extracted from the same set of measurements[33]. Experimentally, a non-labeled ligand, which can be a small molecule or a macromolecule, is titrated into a ^{15}N labeled protein. The titration process is monitored at each stage by acquiring a 2D heteronuclear single quantum coherence (HSQC) spectrum. Direct effects on shift due to the close approach of the ligand, or slight changes in protein conformation induced by ligand binding can alter chemical shift. The resonances with the most chemical shift changes are more likely to come

from residues that are involved in the binding process. In cases of the rapid exchange of ligand on and off of the binding site, the dissociation constant of the ligand *Kd* can be obtained by fitting the change in chemical shift as a function of ligand concentration, to an expression for fractional population of a binding site as a function of the dissociation constant and protein and ligand concentrations.

1.2.2 Saturation Transfer Difference

Saturation transfer difference (STD) experiments are used to identify important binding epitopes on the ligand[34]. STD experiments are most applicable if the ligand spends significant time on the protein but is not tightly attached to it. In other words, the off rate is less than typical spin-lattice relaxation times. Most glycan-protein interactions fall into this category. In fact, the first STD NMR experiment was conducted to study the interaction between GlcNAc and wheat germ agglutinin[35]. STD experiments complement the chemical shift perturbation experiments nicely in that they can be used for epitope mapping on the ligand, as opposed to the protein. Experimentally, the sample contains both the ligand and the receptor, with approximately 100 fold excess of ligand. An 'on-resonance' one dimensional ^{1}H NMR spectrum is recorded with irradiation on some region with protein proton signals but no ligand proton signals (e.g. -1.5 ppm or 8 ppm). Protein resonances will be saturated and some of this saturation will be transferred to ligand resonances by distance dependence spin-spin relaxation mechanisms. Similarly, an 'off-resonance' spectrum is recorded using an irradiation frequency set to a value that is significantly different from both the protein and the ligand frequency region (e.g. 30 ppm). Subtraction of the second spectrum from the former yields a difference spectrum, containing primarily signals that result from the saturation transfer from protein to ligand resonances.

1.2.3 Transferred Nuclear Overhauser effects

The conformation of the ligand when bound to the protein receptor can be determined through the use of transferred Nuclear Overhauser Effects (trNOEs)[36]. Transfers of perturbed magnetization from one proton in a ligand to another occur in a $1/r^6$ dependent fashion leading to a crosspeak in a typical two dimensional NOESY spectrum. These effects scale up approximately in proportion to the rotational correlation time for the complex. Observations on ligands exchanging between bound and free states, even when the ligand is in excess, reflect primarily the geometry of the ligand in the bound state. NOEs taken at short mixing times can be converted to distances using the $1/r^6$ dependence and used to determine the ligand conformation in the bound state.

1.2.4 Residual Dipolar Couplings

Residual dipolar couplings (RDCs) arise when a molecule in solution partially orients in a magnetic field[17]. This results in incomplete averaging of anisotropic magnetic interactions, including dipole-dipole interactions between a pair of NMR active spins. The latter is represented by the term Dij, which can be expressed in the equation below, where r is the distance between a specific pair of nuclei, $\gamma_{i,j}$ are the magnetogyric ratios for the nuclei, μ_0 is the permittivity of space, h is Planck's constant, and θ is the angle between the measured internuclear vector and the magnetic field. Site specific RDCs contain rich structural, dynamic and orientation information for proteins. They can provide long-range constraints on structures in situations in which distance-dependent NOEs cannot be observed.

$$D_{ij} = -\frac{\mu_0 \gamma_i \gamma_j h}{(2\pi r)^3} \left\langle \frac{3cos^2\theta - 1}{2} \right\rangle$$

Unfortunately the extraction of this information is a little complicated. The brackets in the above expression denote averaging that can, in principle, be established by defining a principal alignment frame relative to a molecular coordinate system and then determining the principal order parameter and asymmetry parameter to describe motion about this frame. This means that at least five independent RDCs have to be collected for each semi-rigid molecular fragment.

Many glycoproteins contain more than one extracellular immunoglobulin-like domain. There are often crystal or NMR structures of single or pairs of domains and the internal geometries of domains are well preserved regardless of the extents of glycosylation or the presence or absence of other domains; these domains can serve as the semi-rigid fragments needed for RDC analysis.

The relative domain-domain orientations in solution is one of our primary questions. When the protein used to determine an X-ray crystal structure lacks glycosylation, questions of whether or not the existence of glycosylation can affect the relative orientation of each domain remains. By using RDCs these questions can be answered. If the relative orientation of domains is reasonably well-defined, the resulting orientation frames and degrees of order for properly positioned rigid domains should coincide. If there is some dynamic motion between domains, the principal order parameters for the two alignment frames will be different[37]. Thus, residual dipolar couplings can provide long-range information about the existence of motion and relative domain orientation in multidomain proteins.

1.2.5 Paramagnetic effects

Paramagnetic lanthanide ions offer unique opportunities for structural investigations by NMR spectroscopy. Typically, these ions are incorporated into the protein by chelating to a covalently linked peptide fragment, referred to as a ‘tag’. Among the paramagnetic effects

produced by lanthanide ions, paramagnetic relaxation enhancements (PREs) and pseudocontact shifts (PCS) stand out for providing long-range (10–40 Å) distance constrains and orientational information, which is complementary to the RDC alignment [38-39]. The geometric sensitivity of the PCS is described by the equation below:

$$\Delta\delta(PCS) = \frac{1}{12\pi r^3}[\Delta\chi_{ax}(3cos^2\theta - 1) + \frac{3}{2}\Delta\chi_{rh}sin^2\theta cos2\varphi]$$

where $\Delta\delta(PCS)$ denotes the difference in chemical shifts measured between diamagnetic and paramagnetic samples, r is the distance between the metal ion and the nuclear spin, $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$are the axial and rhombic components describing the anisotropy of the tag's magnetic susceptibility tensor (χ), and the angles θ and φ describe the position of the nuclear spin with respect to the principal axes of the χ tensor. Among different paramagnetic metal ions, Tb^{3+}, Dy^{3+} and Tm^{3+} ,Yb^{3+} are most commonly used because they have the largest anisotropy, Δχ. Gd^{3+} can only cause PRE effects which is solely distance dependent because of its isotropic susceptibility tensor.

1.2.6 Combining sparse-label NMR data for molecular structure determination

Molecular docking is a computational technique aimed at accurately predicting the configuration of a protein-ligand complex. It often uses the experimental constraints that define a receptor binding site and ligand geometry along with estimates of the corresponding interaction energy [40-41]. It has been widely used in modern drug design[42]. Following docking, the complex configurations are analyzed and scored using molecular simulation and data mining methods. Docking can complement different biophysical experiments and provide invaluable information on protein-ligand binding modes in a highly efficient manner[40].

HADDOCK (High Ambiguity Driven biomolecular DOCKing)[43] is a docking program based on a simulated annealing method. It uses diverse biochemical and biophysical interaction data to predict a near-native complex conformation. The restrains used in HADDOCK may derive from NMR, mass spectrometry, chemical cross-linking, cryo-EM, SAXS, fluorescence and so on. When high resolution data are available, such as distance restraints from trNOESY experiments or pseudocontact shift restraints from paramagnetic titration, the resulting complex structures are more reliable. To date, there are 140 complex structures determined by NMR with HADDOCK and deposited in the protein data bank[44].

HADDOCK has been used to derive the models of protein-GAG complexes in this thesis. Restraints involving residues of the protein or parts of the ligand identified as being involved in an interaction by chemical shift perturbation or STD intensity were entered as ambiguous interaction restraints. Interproton distances derived from trNOE data connecting specific pairs of protons were converted to distance constraints. PCSs data for both the protein and ligand were implemented using the XPCS restraints as defined in the program. The detailed docking procedure is discussed in chapters 2 and 4.

1.3 Sparse labeling and NMR assignments

Most structural characterization of proteins has depended on nuclear Overhauser effects (NOEs). NOEs provide short range data (usually < 4Å), and rely on having numerous proton-proton NOEs to yield total protein structures from the derived short distance constraints. The methods described above are distinct in that they use longer range data and can provide structure swith a relatively small number of constraints. This goes hand-in-hand with a sparse labeling strategy in which NMR detectable isotopes are introduced at only a subset of amino acid sites. Triple resonance approaches based on heteronuclear magnetization transfer along the

polypeptide backbone were breakthroughs in the biomolecular NMR field in late1980s[45-46] and are still widely employed in NMR structural studies today. A series of three dimensional triple resonance experiments enable both backbone and side chain resonance assignments. Triple resonance approaches usually depend on uniform $^{15}N/^{13}C$ isotope labeling technologies using *E. coli* as a host for over-expression. By combining $^{15}N/^{13}C$ isotope labeling with extensive ^{2}H labeling the relaxation times of carbon nuclei and amide protons (reintroduced by exchange from protonated water) can be lengthened and the NMR resolution improved. The combination of these methods has allowed for the assignment and structural characterization of proteins up to approximately 60 KDa in molecular weight. While *E. coli* systems for uniformly labeling proteins are well developed, inexpensive and easily implemented [13], they cannot be used to express natively glycosylated proteins, since most bacteria lack the ability to make complex glycans and post-translationally modify proteins. Mammalian cells, such as Human Epithelial Kidney cells (HEK) for example, produce complex oligosaccharides, which fulfills the demand for proper glycosylation. These non-bacterial hosts are also able to regulate disulfide bond formation during protein folding. However, uniform labeling with ^{13}C and ^{15}N can be very costly in mammalian cell lines and ^{2}H labeling using high concentration of deuterium is detrimental to cells.

Sparse labeling, where a single or a small subset of isotopically labeled amino acids are used to introduce ^{13}C and /or ^{15}N labels, is a good alternative which can be directly employed with Eukaryotic hosts[47]. There are several advantages to sparse labeling. First, it is an economical strategy. Many isotopically labeled amino acids are quite inexpensive, which makes sparse labeling versatile. Since only a certain number of protein residues are labeled, the

complexity of the resulting NMR spectra decreases. Therefore, even without employing of deuteration, the resolution of the HSQC experiment can be greatly enhanced.

The primary limitation of sparse labeling is that the one-bond connectivity between isotopically labeled sites is lost. Therefore, traditional triple resonance approaches are not applicable and a new strategy is needed to assign the crosspeaks in an HSQC spectrum of a sparsely labeled protein sample. This can be built on information available from some of the same experiments described above for structural determination, providing suitable models for a protein structure is available. Chemical shifts of labeled ^{15}N or ^{13}C and protons can be measured from ^{15}N-^{1}H or ^{13}C-^{1}H HSQC spectra. NOEs can be obtained from a ^{15}N or ^{13}C edited NOESY spectrum. Other long range orientation and geometry restraints such as residual dipolar couplings (RDCs) or pseudocontact shifts (PCSs) can also be collected from sparsely labeled proteins. Once these observables are combined and compared with the predicted values from available protein structures (or domain structures), each crosspeak is able to be uniquely assigned. This work will be described in detail in chapter 3. Once the crosspeaks are assigned, structural information obtained from the NMR measurements described above can be utilized.

1.4 Targeting protein-GAG complexes

1.4.1 Robo1-HS interaction

This thesis focuses on two specific proteins that utilize GAGs to modulate signaling; roundabout proteins 1, (Robo1) and leucocyte common antigen receptor (LAR). Robos, or Roundabouts, are single-pass transmembrane cell adhesion molecules that are highly conserved across many animals. [48] Mammals have four roundabout receptors (Robo1-4). Robo proteins share a similar structure, consisting of five immunoglobulin-like domains, three fibronectin type III (FN3) repeats, a transmembrane domain, and a cytoplasmic domain with up to four conserved

motifs (CC0-3)[49]. Signaling by Robos involves a second protein, a Slit protein secreted by midline glia cells. In mammals, Slit-Robo signaling is required for the proper development of the CNS, lung, kidney and endothelial cell migration [50-52]. For example, the binding of Slit2 to Robo1 triggers cytoskeletal rearrangements within the axon growth cone, resulting in axon repulsion. This fundamental function of the Slit-Robo interaction is conserved between invertebrates and vertebrates. Biochemical and genetic experiments have shown that heparan sulfate (HS) is required for Slit-Robo signaling in most systems[53-54]. Previous research has also shown that the Slit-Robo interaction is mediated by the D2 domain of Slit and Ig1-2 domains of Robo. The existence of HS promotes the formation of a Slit-Robo-HS signaling complex [55]. There is evidence that the interactions with HS vary depending on particular sulfation patterns and other structural details of this polymeric ligand [56]. There is a substantial amount of previous structural information, including crystal structures of two domain constructs from both human and drosophila homologs [54-55], but these have used no-glycosylated proteins. The drosophila structure shows a piece of heparin sandwiched between two Robo1 molecules. There is also a crystal structure of the Robo1-D1 domain in complex with the D2 domain of Slit2 [54].Our studies of Robo1 interacting with HS use this prior structural information as a starting point and provide a useful model for how a well-defined HS oligomer may modulate Slit2-Robo1 signaling in the presence of native glycosylation. These studies are described in chapter 2.

1.4.2 LAR-HS interaction

LAR, or leukocyte common antigen-related protein, is one of the type IIa receptor protein tyrosine phosphatases (RPTPs). Structurally, type IIa RPTPs, including LAR, share a very similar domain architecture; they contain three immunoglobulin-like (Ig) domains, followed by nine fibronectin type II (FN) units, a single transmembrane helix and two intracellular

phosphotyrosine-specific phosphatase domains. Different from other RPTPs most of which remain as orphan receptors[57], the type IIa RPTPs have been shown to bind a variety of cell surface proteins or soluble ligands and are believed to be highly involved in cell-cell or cell-matrix contacts[58-61]. LAR regulates diverse biological events, such as axonal guidance and outgrowth during neural development [62], synaptic organization[63], cell proliferation[64] and immune response[65]. Heparan sulfate proteoglycans (HSPGs) and chondroitin sulfate proteoglycans (CSPGs) can both interact with LAR and modulate RPTP signaling at neuronal growth cones while giving opposite results [61]. HSPGs complexing with LAR result in proteins clustering and promote neuron extension to the post synapse and interaction with postsynaptic proteins such as TrkC receptor protein tyrosine kinase[61, 66]. On the other hand, CSPGs complexing with LAR disrupt protein clustering, and lead to an inhibition of neural growth and regeneration[59, 61]. Crystallography and site-direct mutagenesis studies have suggested that the first Ig domain (Ig-1) is structurally crucial for glycosaminoglycan binding and the first two Ig domains (Ig 1-2) are a minimum structural requirement for interaction with the postsynaptic ligand TrkC[60]. Based on previous structural information, we aimed to generate a structure for a specific interaction between LAR and a well-defined HS, and provide an improved understanding of RPTP protein clustering and signaling pathways. The results of this research will be presented in chapter 4.

The results described in both chapters 2 and 4 depend, to some extent, on paramagnetic perturbations introduced by inserting a lanthanide-binding peptide into the native protein structure. This procedure has some limitations in that proteins stability and function are easily perturbed. In chapter 5, we describe some efforts to explore some alternative options for introducing paramagnetic metals into proteins and measuring PREs and PCSs. These methods include attaching metal ion binding chelates that carry a sulfhydryl group to a cysteine in the

protein via a disulfide bond. We also attempted to improve the performance of lanthanide binding peptides by screening various constructs using MD simulation. The detailed methods and results of these approaches are described in chapter 5.

Together, these studies demonstrate a novel approach to the assignment of NMR resonances in glycosylated mammalian proteins and provide significant insights into how GAGs regulate signaling events for a pair of related protein systems. The methods presented here will pave the way for the structural characterization of other glycosylated proteins and their GAG binding properties.

1.5 References

1. Berg, J. M.; Tymoczko, J. L.; Stryer, L., Signal-transduction pathways: an introduction to information metabolism. **2002**.

2. Braga, V. M., Cell–cell adhesion and signalling. *Current opinion in cell biology* **2002,** *14* (5), 546-556.

3. Hynes, R. O., Integrins: versatility, modulation, and signaling in cell adhesion. *Cell* **1992,** *69* (1), 11-25.

4. Walsh, G., Post-translational modifications of protein biopharmaceuticals. *Drug discovery today* **2010,** *15* (17-18), 773-780.

5. Endo, T., New Era of Glycoscience: Intrinsic and Extrinsic Functions Performed by Glycans Foreword. *Biol Pharm Bull* **2009,** *32* (5), 765-766.

6. In *Essentials of Glycobiology*, 3rd ed.; Varki, A.; Cummings, R. D.; Esko, J. D.; Stanley, P.; Hart, G.; Aebi, M.; Darvill, A.; Kinoshita, T.; Packer, N. H.; Prestegard, J. J.; Schnaar, R. L.; Seeberger, P. H., Eds. Cold Spring Harbor (NY), 2015.

7. Esko, J. D.; Linhardt, R. J., Proteins that Bind Sulfated Glycosaminoglycans. In *Essentials of Glycobiology*, 2nd ed.; Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E., Eds. Cold Spring Harbor (NY), 2009.

8. Afratis, N.; Gialeli, C.; Nikitovic, D.; Tsegenidis, T.; Karousou, E.; Theocharis, A. D.; Pavao, M. S.; Tzanakakis, G. N.; Karamanos, N. K., Glycosaminoglycans: key players in cancer cell biology and treatment. *Febs J* **2012,** *279* (7), 1177-1197.

9. Liang, B. Y.; Tamm, L. K., NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat Struct Mol Biol* **2016,** *23* (6), 468-474.

10. Kay, L. E.; Gardner, K. H., Solution NMR spectroscopy beyond 25 kDa. *Current opinion in structural biology* **1997,** *7* (5), 722-731.

11. Kovacs, H.; Moskau, D.; Spraul, M., Cryogenically cooled probes - a leap in NMR technology. *Prog Nucl Mag Res Sp* **2005,** *46* (2-3), 131-155.

12. Weijers, H. W.; Markiewicz, W. D.; Gavrilin, A. V.; Voran, A. J.; Viouchkov, Y. L.; Gundlach, S. R.; Noyes, P. D.; Abraimov, D. V.; Bai, H.; Hannahs, S. T., Progress in the Development and Construction of a 32-T Superconducting Magnet. *IEEE Transactions on Applied Superconductivity* **2016,** *26* (4), 1-7.

13. Tugarinov, V.; Kanelis, V.; Kay, L. E., Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nature protocols* **2006,** *1* (2), 749-54.

14. Zhang, H. Y.; van Ingen, H., Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. *Current opinion in structural biology* **2016,** *38*, 75-82.

15. Roche, J.; Ying, J. F.; Shen, Y.; Torchia, D. A.; Bax, A., ARTSY-J: Convenient and precise measurement of (3)J(HNH alpha) couplings in medium-size proteins from TROSY-HSQC spectra. *J Magn Reson* **2016,** *268*, 73-81.

16. Solyom, Z.; Schwarten, M.; Geist, L.; Konrat, R.; Willbold, D.; Brutscher, B., BEST-TROSY experiments for time-efficient sequential resonance assignment of large disordered proteins. *J Biomol Nmr* **2013,** *55* (4), 311-321.

17. Prestegard, J. H.; Bougault, C. M.; Kishore, A. I., Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* **2004,** *104* (8), 3519-3540.

18. Chen, K.; Tjandra, N., The Use of Residual Dipolar Coupling in Studying Proteins by NMR. *Top Curr Chem* **2012,** *326*, 47-67.

19. Sanchez-Medina, C.; Sekhar, A.; Vallurupalli, P.; Cerminara, M.; Munoz, V.; Kay, L. E., Probing the Free Energy Landscape of the Fast-Folding gpW Protein by Relaxation Dispersion NMR. *J Am Chem Soc* **2014,** *136* (20), 7444-7451.

20. Aznauryan, M.; Delgado, L.; Soranno, A.; Nettels, D.; Huang, J. R.; Labhardt, A. M.; Grzesiek, S.; Schuler, B., Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *P Natl Acad Sci USA* **2016,** *113* (37), E5389-E5398.

21. Wang, X.; Sharp, J. S.; Handel, T. M.; Prestegard, J. H., Chemokine Oligomerization in Cell Signaling and Migration. *Prog Mol Biol Transl* **2013,** *117*, 531-578.

22. Roche, J.; Louis, J. M.; Bax, A.; Best, R. B., Pressure-induced structural transition of mature HIV-1 protease from a combined NMR/MD simulation approach. *Proteins-Structure Function and Bioinformatics* **2015,** *83* (12), 2117-2123.

23. van Zundert, G. C. P.; Rodrigues, J. P. G. L. M.; Trellet, M.; Schmitz, C.; Kastritis, P. L.; Karaca, E.; Melquiond, A. S. J.; van Dijk, M.; de Vries, S. J.; Bonvin, A. M. J. J., The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **2016,** *428* (4), 720-725.

24. Garman, E. F., Developments in X-ray Crystallographic Structure Determination of Biological Macromolecules. *Science* **2014,** *343* (6175), 1102-1108.

25. Callaway, E., The Revolution Will Not Be Crystallized. *Nature* **2015,** *525* (7568), 172-174.

26. Barb, A. W.; Meng, L.; Gao, Z. W.; Johnson, R. W.; Moremen, K. W.; Prestegard, J. H., NMR Characterization of Immunoglobulin G Fc Glycan Motion on Enzymatic Sialylation. *Biochemistry-Us* **2012,** *51* (22), 4618-4626.

27. Mulloy, B.; Hart, G. W.; Stanley, P., Structural Analysis of Glycans. In *Essentials of Glycobiology*, 2nd ed.; Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E., Eds. Cold Spring Harbor (NY), 2009.

28. Battistel, M. D.; Azurmendi, H. F.; Yu, B.; Freedberg, D. I., NMR of glycans: Shedding new light on old problems. *Prog Nucl Mag Res Sp* **2014,** *79*, 48-68.

29. Jensen, M. R.; Ruigrok, R. W. H.; Blackledge, M., Describing intrinsically disordered proteins at atomic resolution by NMR. *Current opinion in structural biology* **2013,** *23* (3), 426-435.

30. Kwong, P. D.; Wyatt, R.; Desjardins, E.; Robinson, J.; Culp, J. S.; Hellmig, B. D.; Sweet, R. W.; Sodroski, J.; Hendrickson, W. A., Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1). *Journal of Biological Chemistry* **1999,** *274* (7), 4115-4123.

31. Lepenies, B.; Seeberger, P. H., Simply better glycoproteins. *Nature biotechnology* **2014,** *32* (5), 443-445.

32. Zuiderweg, E. R., Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **2002,** *41* (1), 1-7.

33. McCoy, M. A.; Wyss, D. F., Spatial localization of ligand binding sites from electron current density surfaces calculated from NMR chemical shift perturbations. *Journal of the American Chemical Society* **2002,** *124* (39), 11758-63.

34. Bhunia, A.; Bhattacharjya, S.; Chatterjee, S., Applications of saturation transfer difference NMR in biological systems. *Drug discovery today* **2012,** *17* (9-10), 505-13.

35. Mayer, M.; Meyer, B., Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew Chem Int Edit* **1999,** *38* (12), 1784-1788.

36. Post, C. B., Exchange-transferred NOE spectroscopy and bound ligand structure determination. *Current opinion in structural biology* **2003,** *13* (5), 581-8.

37. Fischer, M. W.; Losonczi, J. A.; Weaver, J. L.; Prestegard, J. H., Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. *Biochemistry* **1999,** *38* (28), 9013-22.

38. Otting, G., Prospects for lanthanides in structural biology by NMR. *Journal of biomolecular NMR* **2008,** *42* (1), 1-9.

39. Nitsche, C.; Otting, G., Pseudocontact shifts in biomolecular NMR using paramagnetic metal tags. *Prog Nucl Mag Res Sp* **2016**.

40. Sturlese, M.; Bellanda, M.; Moro, S., NMR-Assisted Molecular Docking Methodologies. *Mol Inform* **2015,** *34* (8), 513-525.

41. Smith, G. R.; Sternberg, M. J., Prediction of protein–protein interactions by docking methods. *Current opinion in structural biology* **2002,** *12* (1), 28-35.

42. Ferreira, L. G.; dos Santos, R. N.; Oliva, G.; Andricopulo, A. D., Molecular docking and structure-based drug design strategies. *Molecules* **2015,** *20* (7), 13384-13421.

43. Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J., HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **2003,** *125* (7), 1731-1737.

44. The Protein Data Bank. *Comput Sci Eng* **2010,** *12* (5), 10-10.

45. Kay, L. E.; Ikura, M.; Tschudin, R.; Bax, A., Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *Journal of Magnetic Resonance (1969)* **1990,** *89* (3), 496-514.

46. Grzesiek, S.; Bax, A., Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J Am Chem Soc* **1992,** *114* (16), 6291-6293.

47. Prestegard, J. H.; Agard, D. A.; Moremen, K. W.; Lavery, L. A.; Morris, L. C.; Pederson, K., Sparse labeling of proteins: Structural characterization from long range constraints. *J Magn Reson* **2014,** *241*, 32-40.

48. Evans, T. A.; Bashaw, G. J., Slit/Robo-mediated axon guidance in Tribolium and Drosophila: divergent genetic programs build insect nervous systems. *Developmental biology* **2012,** *363* (1), 266-78.

49. Dickinson, R. E.; Duncan, W. C., The SLIT-ROBO pathway: a regulator of cell function with implications for the reproductive system. *Reproduction* **2010,** *139* (4), 697-704.

50. Grieshammer, U.; Le, M.; Plump, A. S.; Wang, F.; Tessier-Lavigne, M.; Martin, G. R., SLIT2-mediated ROBO2 signaling restricts kidney induction to a single site. *Developmental cell* **2004,** *6* (5), 709-17.

51. Xian, J.; Clark, K. J.; Fordham, R.; Pannell, R.; Rabbitts, T. H.; Rabbitts, P. H., Inadequate lung development and bronchial hyperplasia in mice with a targeted deletion in the Dutt1/Robo1 gene. *Proceedings of the National Academy of Sciences of the United States of America* **2001,** *98* (26), 15062-6.

52. Dickson, B. J.; Gilestro, G. F., Regulation of commissural axon pathfinding by slit and its Robo receptors. *Annual review of cell and developmental biology* **2006,** *22*, 651-75.

53. Johnson, K. G.; Ghose, A.; Epstein, E.; Lincecum, J.; O'Connor, M. B.; Van Vactor, D., Axonal heparan sulfate proteoglycans regulate the distribution and efficiency of the repellent slit during midline axon guidance. *Current biology : CB* **2004,** *14* (6), 499-504.

54. Morlot, C.; Thielens, N. M.; Ravelli, R. B. G.; Hemrika, W.; Romijn, R. A.; Gros, P.; Cusack, S.; McCarthy, A. A., Structural insights into the Slit-Robo complex. *Proceedings of the National Academy of Sciences of the United States of America* **2007,** *104* (38), 14923-14928.

55. Fukuhara, N.; Howitt, J. A.; Hussain, S. A.; Hohenester, E., Structural and functional analysis of slit and heparin binding to immunoglobulin-like domains 1 and 2 of Drosophila Robo. *The Journal of biological chemistry* **2008,** *283* (23), 16226-34.

56. Zhang, F. M.; Moniz, H. A.; Walcott, B.; Moremen, K. W.; Linhardt, R. J.; Wang, L. C., Characterization of the interaction between Robo1 and heparin and other glycosaminoglycans. *Biochimie* **2013,** *95* (12), 2345-2353.

57. Mohebiany, A. N.; Nikolaienko, R. M.; Bouyain, S.; Harroch, S., Receptor-type tyrosine phosphatase ligands: looking for the needle in the haystack. *Febs J* **2013,** *280* (2), 388-400.

58. Coles, C. H.; Jones, E. Y.; Aricescu, A. R., Extracellular regulation of type IIa receptor protein tyrosine phosphatases: mechanistic insights from structural analyses. *Semin Cell Dev Biol* **2015,** *37*, 98-107.

59. Fisher, D.; Xing, B.; Dill, J.; Li, H.; Hoang, H. H.; Zhao, Z. Z.; Yang, X. L.; Bachoo, R.; Cannon, S.; Longo, F. M.; Sheng, M.; Silver, J.; Li, S. X., Leukocyte Common Antigen-Related Phosphatase Is a Functional Receptor for Chondroitin Sulfate Proteoglycan Axon Growth Inhibitors. *J Neurosci* **2011,** *31* (40), 14051-14066.

60. Coles, C. H.; Mitakidis, N.; Zhang, P.; Elegheert, J.; Lu, W. X.; Stoker, A. W.; Nakagawa, T.; Craig, A. M.; Jones, E. Y.; Aricescu, A. R., Structural basis for extracellular cis and trans RPTP sigma signal competition in synaptogenesis. *Nat Commun* **2014,** *5*.

61. Coles, C. H.; Shen, Y.; Tenney, A. P.; Siebold, C.; Sutton, G. C.; Lu, W.; Gallagher, J. T.; Jones, E. Y.; Flanagan, J. G.; Aricescu, A. R., Proteoglycan-specific molecular switch for RPTPsigma clustering and neuronal extension. *Science* **2011,** *332* (6028), 484-8.

62. Dunah, A. W.; Hueske, E.; Wyszynski, M.; Hoogenraad, C. C.; Jaworski, J.; Pak, D. T.; Simonetta, A.; Liu, G.; Sheng, M., LAR receptor protein tyrosine phosphatases in the development and maintenance of excitatory synapses. *Nat Neurosci* **2005,** *8* (4), 458-467.

63. Takahashi, H.; Craig, A. M., Protein tyrosine phosphatases PTPδ, PTPσ, and LAR: presynaptic hubs for synapse organization. *Trends in neurosciences* **2013,** *36* (9), 522-534.

64. Tonks, N. K., Protein tyrosine phosphatases: from genes, to function, to disease. *Nat Rev Mol Cell Bio* **2006,** *7* (11), 833-846.

65. Mustelin, T.; Vang, T.; Bottini, N., Protein tyrosine phosphatases and the immune response. *Nature Reviews Immunology* **2005,** *5* (1), 43-57.

66. Takahashi, H.; Arstikaitis, P.; Prasad, T.; Bartlett, T. E.; Wang, Y. T.; Murphy, T. H.; Craig, A. M., Postsynaptic TrkC and Presynaptic PTP sigma Function as a Bidirectional Excitatory Synaptic Organizing Complex. *Neuron* **2011,** *69* (2), 287-303.

CHAPTER 2

STRUCTURAL ASPECTS OF HEPARAN SULFATE BINDING TO ROBO1-IG1-2[1]

[1]. Reproduced in part with permission from [Gao, Q.; Chen, C. Y.; Zong, C.; Wang, S.; Ramiah, A.; Prabhakar, P.; Morris, L. C.; Boons, G. J.; Moremen, K. W.; Prestegard, J. H., Structural Aspects of Heparan Sulfate Binding to Robo1-Ig1-2. *ACS Chemical Biology* **2016,** *11* (11), 3106-3113.] Copyright [2016] American Chemical Society.

2.1 Acknowledgement

The following project involved a collaboration with Dr. Kelley Moremen's lab and Dr. Geert-Jan Boons lab from Complex Carbohydrate Research Center CCRC. The protein constructs were designed by Kelley Moremen. All sparsely labeled Robo1 protein samples were expressed by Shuo Wang. The haparan sulfates were synthesized by Chengli Zong. Laura Morris from the Prestegard lab, David Thieker and Arunima Singh from the Woods lab gave valuable advice in setting up MD simulations and post energy analysis. Roberto Sonon from the analytical services center of the Complex Carbohydrate Research Center provided kind assistance in N-glycosylation profiling of Robo1.

2.2 Abstract

Roundabout 1, or Robo1, is a cell surface signaling molecule important in axon guidance during mammalian development. Its interaction with heparan sulfate (HS) chains and members of the Slit protein family is essential to its activity, making characterization of these interactions by structural methods such as NMR highly desirable. However, the fact that Robo1 is a glycosylated protein prevents employment of commonly used bacterial hosts for expression of properly glycosylated forms with the uniform ^{15}N, ^{13}C, and ^{2}H labeling needed for NMR studies. Here, we demonstrate an alternative methodology based on sparse labeling with single isotopically enriched amino acids, combined with high structural content NMR experiments to characterize the binding geometries of a two domain construct of glycosylated Robo1 (Robo1-Ig1-2) interacting with two different synthetic HS tetramers (IdoA-GlcNS6S-IdoA2S-GlcNS6S-$(CH_2)_5NH_2$ and IdoA-GlcNS6S-IdoA-GlcNS6S-$(CH_2)_5NH_2$) as well as an octamer (GlcA-GlcNS6S-IdoA-GlcNS-IdoA2S-GlcNS6S-IdoA-GlcNAc6S-$(CH_2)_5NH_2$). Significant chemical shift perturbation of crosspeaks from K81 on titration of Robo1-Ig1-2 with the HS tetramer

containing 2-sulfation of an iduronic acid residue provides initial evidence for the location of a binding site, and a disassociation constant of 255 μM. The second tetramer lacking this sulfation is shown to occupy a similar site, but has enhanced binding affinity, as does the octamer which carries additional N and 6-O-sulaftion. The binding epitopes and bound conformations of the HS tetramers have been further characterized by saturation transfer difference (STD), transferred nuclear Overhauser effect (trNOE) and paramagnetic perturbations. A model of the complex with the 2-sulfated tetramer has been generated using the docking program, HADDOCK, and constraints derived from the various NMR experiments. A post-processing energetic analysis provides a rationale for the lower binding affinity of the 2-sulfated tetramer, despite favorable electrostatic interactions between the sulfate and the positively charged K81 residue, and examination of the binding site in comparison to previously identified Robo-Slit interactions provides a rationale for modulation of Robo-Slit interactions by HS.

2.3 Introduction

Robo1 (roundabout receptor 1) is one of four members of the human ROBO family; all are developmentally important cell-surface signaling molecules most recognized for their role in axon guidance [1], but also for their role in angiogenesis [2], and the development of many internal organs [3], including those of the reproductive system[4]. Robo1 is also involved in tumorigenesis, cancer progression and metastasis, possibly through its regulation of growth factors or chemokines in the tumor microenvironment [5]. Its signaling is regulated by interactions with Slit2, one of three members of a family of very large secreted glycoproteins. Interactions between Robos and Slits, including Robo1 and Slit2, are modulated by interactions with heparn sulfate (HS) [1, 6] (Figure 2.1A). There is evidence that the interactions with HS vary depending on particular sulfation patterns and other structural details of this polymeric ligand [7]. Producing

a structure illustrating specific interactions between protein and ligand for a well-defined segment of HS would provide a basis for understanding this specificity and using this understanding in the design of molecules that could compete in modulating important physiological processes. Here we present a model for the interaction of a particular 2-sulfated HS tetramer (IdoA-GlcNS6S-IdoA2S-GlcNS6S-$(CH_2)_5NH_2$, ligand 1 in Figure 2.1B) with the terminal two domains of Robo1, and use that model to rationalize the higher affinity of an analog lacking 2-sulfation (IdoA-GlcNS6S-IdoA-GlcNS6S-$(CH_2)_5NH_2$, ligand 2 in Figure 2.1B) and an octamer in which ligand 1 is extended by two residues (IdoA-GlcNS6S) at each end . The model is based on a combination of NMR cross-relaxation data that define bound ligand geometry, saturation transfer difference (STD) data that identify binding epitopes of the ligand, and paramagnetic perturbation and ligand induced chemical shift data that allow placement of the ligand in a binding site. The resulting ligand geometry and binding site location are found to complement previously proposed Robo1-Slit2 interactions.

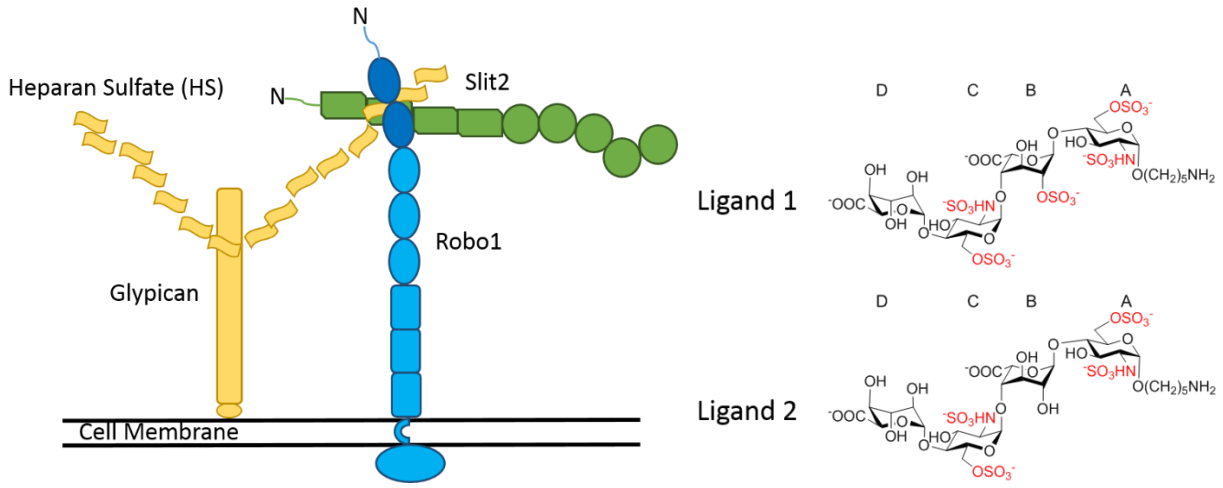


Figure 2.1. (A) Cartoon representation of the Robo1-Slit2-HS interaction and domain organization. Robo1 is shown in blue, with the Ig1-2 domains labeled in dark blue. Slit2 is in

green and the heparan sulfate chain attached to proteoglycan, glypican in yellow. (B) Structures of heparan sulfate ligands 1 and 2 used in this study. All of the sulfate groups are labeled in red.

Previous work on Robo1 – HS complexes have in general used natural isolates, often from depolymerized heparin rather than HS, or similar isolates in which levels of sulfation have been chemically or enzymatically modified [11-12]. These are in general not homogeneous preparations, making structural characterization difficult. Here we capitalize on a synthetic strategy directed at the preparation of tetramers and octamers with specific sulfation patterns [11]. Previous work has suggested that both 6-sulfation and N-sulfation are important [7]. We selected the pair, IdoA-GlcNS6S-IdoA2S-GlcNS6S-$(CH_2)_5NH_2$), and IdoA-GlcNS6S-IdoA-GlcNS6S-$(CH_2)_5NH_2$) for study and provide a structural model for the former. The model allows us to rationalize these differences.

NMR can provide a number of different data types that build on existing structural data to provide details of ligand-protein interactions. Transferred nuclear Overhauser effect (trNOE) and saturation transfer difference (STD) experiments provide information about bound ligand geometry and interaction epitopes of the ligand respectively. Neither require isotope labeling of the ligand, but they do require assignment of proton resonances from the ligand. Those assignments are provided.

Among the most widely used indicator of interacting groups on the protein surface is chemical shift perturbation of specific protein resonances upon ligand binding. While there is not a strict dependence on proximity of the ligand to a residue showing perturbed resonances, residues showing perturbations are very likely to be in the binding site. Data acquisition usually employs a two dimensional ^{15}N-^{1}H heteronuclear single quantum coherence (HSQC) experiment, which provides an observable crosspeak (juncture of ^{1}H and ^{15}N resonances) for nearly every

amino acid in a protein, provided protein can be uniformly isotopically labeled with ^{15}N. We use chemical shift perturbation data in the current study, however, Robo1 is a glycoprotein, as well as a protein with two conserved disulfide bonds, that has resisted expression in the bacterial hosts normally used for uniform isotopic labeling. Here we use a sparse labeling strategy that can be applied in mammalian (HEK293) cells [13-14]. Proteins can be labeled by supplementation with a single isotopically labeled type of amino acid or by supplementation with complementary groups of isotopically labeled amino acids. Here we select labeling with ^{15}N-labeled lysine because of the frequent involvement of this positively charged amino acid in interaction with negative sulfates and carboxylates on HS. We also label separately with ^{15}N-labeled phenylalanine as this amino acid is sometimes found in glycan binding sites.

Crosspeaks in spectra from sparsely labeled proteins are typically well resolved and selective perturbation of any peak provides both an indication of binding site location and a means of determining binding constants though the concentration dependence of chemical shift changes. However, a structural interpretation requires sequence specific assignment of perturbed crosspeaks. Here we document an approach to the assignment of sparsely labeled proteins that combines chemical shift prediction with data from residual dipolar couplings (RDCs) and ^{15}N-edited NOEs to make assignments. The strategy is successful and provides important experimental information on the location of the HS binding site in Robo1.

Because chemical shift perturbation provides only qualitative structural information, we also provide data coming from paramagnetic perturbations of ligand chemical shifts by a lanthanide ion bound to a site in the protein. Lanthanide ions such as Dy^{3+} and Tm^{3+} cause both paramagnetic relaxation enhancement (PRE) which broadens and decreases intensity of HSQC crosspeaks with an inverse sixth power distance dependence and pseudo contact shifts (PCSs)

which move resonances with a dependence on the inverse third power of the distance from the ion as well as the orientation of the vector connecting the ion and observed site. The Robo1 construct does not have a native site capable of binding a lanthanide ion. To introduce a site a lanthanide binding loop has been engineered into the Robo1 construct [15-17]. This has provided more quantitative information on the location of the bound 2-sulfated ligand, as well as a comparison of how the ligands with and without 2-sulfation bind.

All of the NMR data have been combined in a constrained docking approach using the program HADDOCK [18]. The resulting structure shows an interaction between lysine 81 of Robo1 and the 2-sulfate of the internal iduronic acid in ligand 1, as well as the participation of arginine residues previously identified by mutational studies [9]. Despite the involvement of a favorable lysine 81- 2-sulfate interaction, the ligand without 2-sulfation appears to bind in the same site with an even higher affinity. We are able to rationalize the higher affinity based on a high penalty for desolvation of the extra 2-sulfate. This is confirmed by calculating the solvation energy using an MM-PB/GBSA method, followed by per-residue energy decomposition analysis [19-21]. The sites used by these ligands lie adjacent to a site previously determined for binding of the D2-domain of Slit2 [10]. The docked structures provide a useful model for how HS may modulate Slit2-Robo1 signaling.

2.4 Materials and Methods

2.4.1 Materials

^{15}N-Phe, ^{15}N- Lys, and deuterium oxide were purchased from Cambridge Isotope Laboratories. Pf1 Phage was purchased from ASLA biotech. All other chemicals were purchased from Sigma-Aldrich unless otherwise stated. Heparan sulfate teramer 1 (IdoA-GlcNS6S-IdoA2S-GlcNS6S-$(CH_2)_5NH_2$), tetramer 2 (IdoA-GlcNS6S-IdoA-GlcNS6S-$(CH_2)_5NH_2$) and octamer

(GlcA-GlcNS6S-IdoA-GlcNS-IdoA2S-GlcNS6S-IdoA-GlcNAc6S-$(CH_2)_5NH_2$) were synthesized by Chengli Zong in the Geert-Jan Boons laboratory.

2.4.2 Protein expression and purification

The protein sequence of Robo1 was analyzed for domain boundaries using the UniProt database and a truncated protein sequence comprised of Ig domains 1 and 2 (plus flanking regions) was chosen for gene synthesis. For the Robo1 construct containing a lanthanide binding loop a site between β-strands C and D of Ig domain 1 was chosen based on the similarity between separation of strand ends and loop ends in a previous insert in an α-helical protein [15]. The lanthanide binding loop differed from the original construct from the Imperiali lab by the replacement of a tryptophan with an alanine, addition of a serine at the N terminus and replacement of a leucine and alanine with a serine and glycine at the C terminus [17]. The final sequence with the inserted tag in bold is: GSRLRQEDFPPRIVEHPSDLIVSKGEPATLNCK AEGRPTPTIEWYKG**SYIDTNNDGAYEGDELSG**GERVETDKDDPRSHRMLLPSGSLFFLR IVHGRKSRPDEGVYVCVARNYLGEAVSHNASLEVAILRDDFRQNPSDVMVAVGEPAVME CQPPRGHPEPTISWKKDGSPLDDKDERITIRGGKLMITYTRKSDAGKYVCVGTNMVGER ESEVAELTVLERPSFVK. Numbering is from 58 to 266 based on the Uniprot sequence. DNAs corresponding to this sequence was synthesized by GenScript (Piscataway, NJ) with codons optimized for utilization in mammalian cell expression. DNA fragments for the native Robo1-Ig1-2 sequence, the sequence with the loop insert, a R136AK137A mutant that showed diminished binding capability, and a separate Robo1-D1 sequence were cloned into the mammalian expression vector containing an excretion signal, a His-tag, a GFP super-folder sequence, and a TEV cleavge site (pGEn2) using restriction digestion and ligation into sites designed into the vector. Large scale DNA preparations were prepared and transiently transfected

into HEK293S suspension culture cells in FreeStyle 293 media (Thermo Fisher Scientific, Waltham MA). The cell media was exchanged to Freestyle dropout media (missing Lys or Phe amino acids) supplemented with 150 mg/L isotopically labeled (or unlabeled) Phe or Lys on the second day of transfection. The recombinant protein was harvested from culture supernant after 6 days of growth using Ni^{2+} -NTA chromatography and concentrated to ~1 mg/mL. The resulting protein preparation was digested with recombinant TEV to cleave between Robo1 and GFP and then subjected to Ni^{2+}-NTA chromatography a second time to remove GFP. It was subsequently purified by size exclusion chromatography. The average protein yields were 10 mg/L. Examination of trypsinized fragments containing lysine and phenylalanine indicate that ^{15}N labeling is about 75%. N-glycan profiling was based on release of N-linked oligosaccharides from Robo1-Ig1-2 by treatment with PNGase F. The resulting glycan mixture was analyzed by MALDI-TOF/MS (AB Sciex 5800, Applied Biosystems) and ESI-MS/MS (LTQ-Orbitrap, Thermo Scientific). Glycans were identified by comparison of masses to those expected for N-glycans commonly found in mammalian systems. The released glycans are very heterogeneous with most major peaks belonging to biantennary structures having core fucosylation (see Figure 2.2)

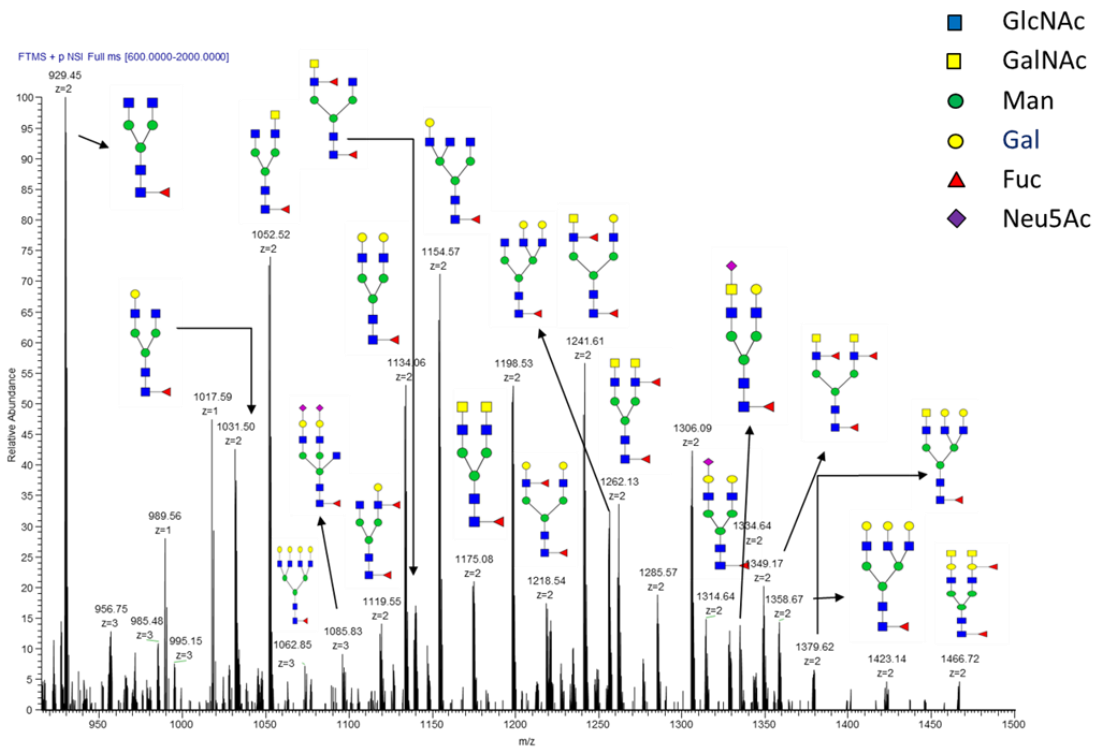


Figure 2.2. N-glycans profile of Robo1-Ig1-2, the possible structure are listed on the top of each m/z.

2.4.3 NMR spectroscopy

All the NMR spectroscopy was performed on Varian/Agilent instruments with DD2 (21.1 T and 18.8 T) consoles and 5 mm cryogenically cooled triple resonance probes. NMR protein samples were 150 µM in 10% D_2O buffer containing 25 mM Tris and 100 mM potassium chloride at pH 7.0 for the two-dimensional ^{15}N-HSQC titration experiments, three dimensional ^{15}N-filtered NOE experiments and RDC experiments. Robo1 loop samples for measurement of paramagnetic perturbations contained lanthanides (Dy^{3+}, Tm^{3+}, Tb^{3+} or Gd^{3+}, as well as Lu^{3+} for a diamagnetic control) at lanthanide to protein ratios slightly less than 1:1 under the same buffer conditions. Pseudo contact shifts (PCSs) of proteins and ligands, as well as RDCs for the protein, were determined with heparan sulfate tetramer 1 and 2 from standard ^{15}N-HSQC spectra with a

1:2 protein/ligand ratios. Samples for the STD and trNOE experiments were 15 μM in protein and 900 μM in ligand, all in 100% D_2O buffer containing 20 mM phosphate, 100 mM potassium chloride and pH 7.0. All samples contained Dimethyl-2-silapentane-5-sulfonate (DSS) as an internal reference.

For titration experiments, increasing concentrations of heparan sulfate tetramer 1 and 2 (0 μM to 560 μM in steps of 70 μM) were added to the Robo1-Ig1-2 sample (150 μM) and binding was monitored by ^{15}N HSQC. Each ligand was added from a highly concentrated solution of ligand such that the addition of ligand causes effectively no dilution of protein.

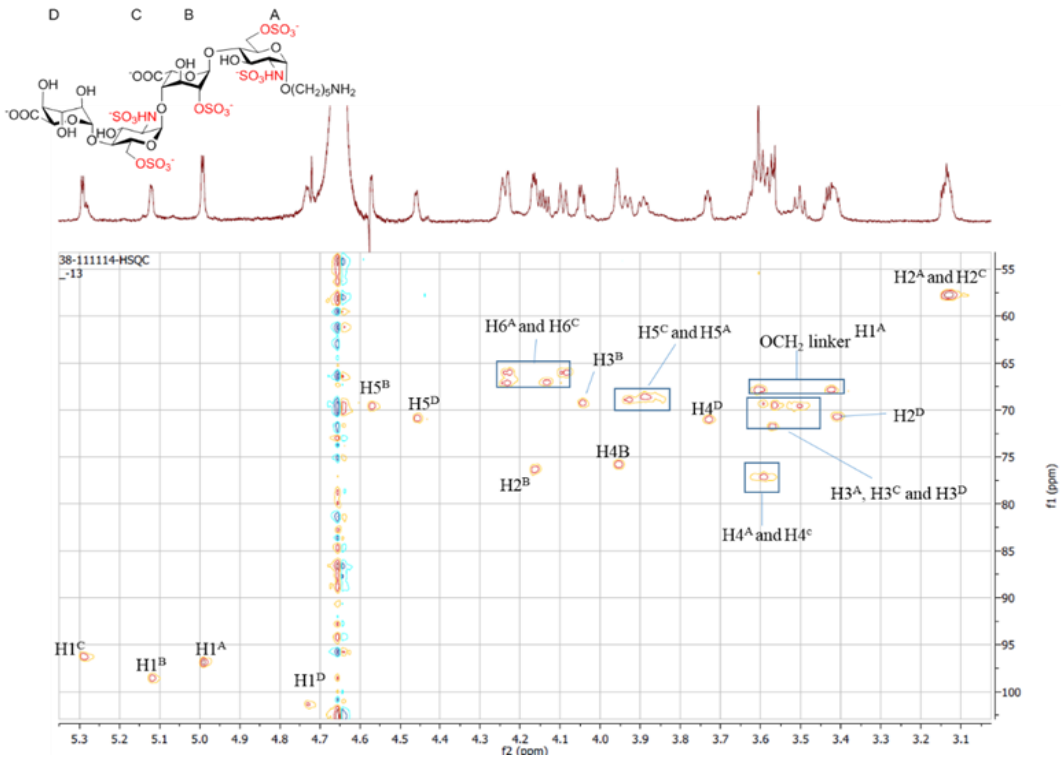
^{15}N-filtered NOEs were collected using a standard three-dimensional NOESY-HSQC sequence from the Agilent/Varian Biopack using a mixing time of 150 ms.

Residual dipolar couplings (RDCs) data were measured on a protein sample containing 12.5mg/mL Pf1 phage. The $^{1}J_{N\text{-}H}$ couplings were measured, in isotropic solution and in anisotropic Pf1 phage using a pulse sequence in which cross-peaks in HSQC spectra are modulated by J+D coupling in the ^{15}N dimension [22]. The modulation delays varied from 0.5 to 14 ms in 8 steps for all experiments.

Both STD and trNOE experiments were standard Biopack experiments. STD on Robo1-Ig1-2 with heparan sulfate tetramer 1 and 2 were carried out separately with a 15 μM protein at a 1:60 protein/ligand ratio and a 15 μM protein sample in the absence of ligand as control. Both the control and experimental samples were irradiated at -1.5 ppm, and saturation times were increased from 1 to 4 s in steps of 1 s. To obtain the final spectrum, the STD NMR spectrum of Robo1-Ig1-2 in the absence of ligands was subtracted from the STD NMR spectrum in the presence of ligands. The trNOE experiments on Robo1-Ig1-2 with heparan sulfate tetramer 1

and 2 were performed using the same samples as the STD experiments with a mixing time of 150 ms.

To allow interpretation of STD and trNOE experiments, spectra of the ligand must be assigned. This was accomplished by acquiring ^{1}H proton, ^{1}H-^{1}H TOCSY, ^{1}H-^{1}H NOESY, ^{13}C-^{1}H HSQC and ^{13}C-^{1}H HMQC. A ^{13}C-^{1}H HSQC spectrum with complete assignments is included in Figure 2.3.



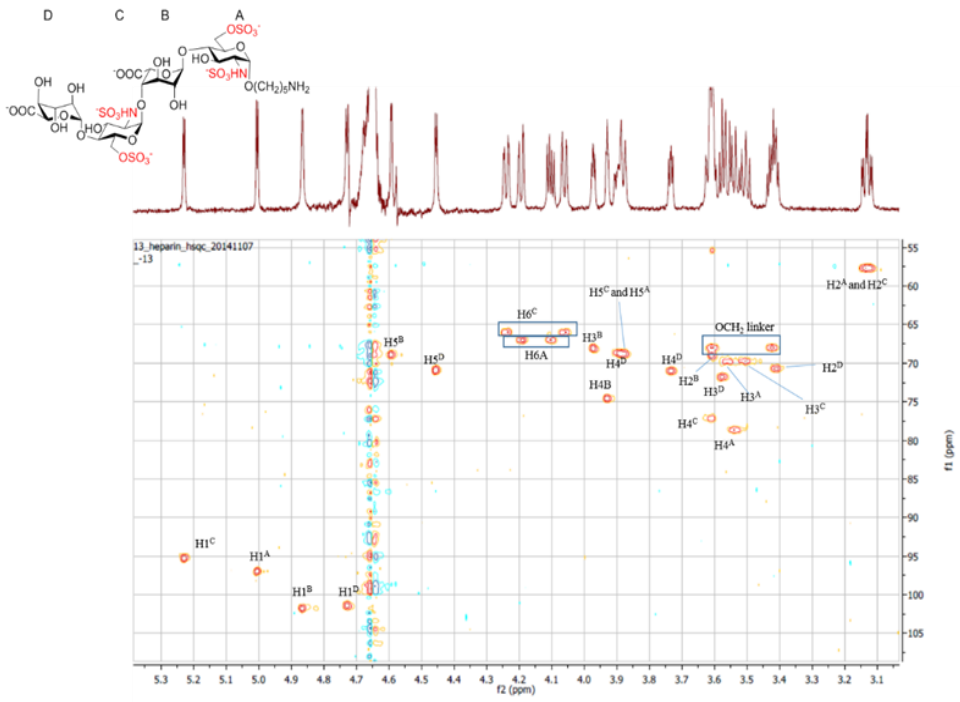


Figure 2.3. An illustration of 13C-1H HSQC spectrum of ligand 1 and 2 with 1D proton trace on the top. The assignment of each proton is labeled by the side. The full assignments for both ligand 1 and 2 are given in the follow.

Ligand 1: 1H NMR (500 MHz, Deuterium Oxide) δ 5.32 (d, 1H, J = 3.4 Hz, H1C), 5.15 (d, 1H, J = 3.1 Hz, H1B), 5.03 (d, 1H, J = 3.6 Hz, H1A), 4.78 (d, 1H, J = 4.5 Hz, H1D), 4.63 (d, 1H, J = 2.0 Hz, H5B), 4.48 (d, 1H, J = 2.4 Hz, H5D), 4.29 – 3.38 (m, 18H, H3A, H4A, H5A, H6A, H2B, H3B, H4B, H3C, H4C, H5C, H6C, H2D, H3D, H4D, OCH2 Linker), 3.20 – 3.15(m, 2H, H2A, H2C), 2.92 (t, 2H, J = 7.4 Hz, NCH2 Linker), 1.71 – 1.28 (m, 6H, 3*CH2 Linker).

Ligand 2: 1H NMR (800 MHz, D2O): δ 5.33 (d, 1H, J= 3.7 Hz, H1C), 5.11 (d, 1H, J= 3.7 Hz, H1A), 4.97 (d, 1H, J = 2.0 Hz, H1B), 4.85 (d, 1H, J = 4.9 Hz, H1D), 4.71 (d, 1H, J = 2.2 Hz, H5B), 4.57 (d, 1H, J = 3.9 Hz, H5D), 4.35 (dd, 1H, J = 2.2 and 11.2 Hz, H6aC), 4.29 (dd, 1H, J = 1.7 and 11 Hz, H6aA), 4.23 (dd, 1H, J = 5.4 and 11.2 Hz, H6bA), 4.17 (dd, 1H, J =1.7 and 11 Hz, H6bC), 4.08 (t, 1H, J = 4.0 Hz, H3B), 4.04 (t, 1H, J = 2.6 Hz, H4B), 4.00 – 3.98 (bm, 2H, H5A,

H5C), 3.85 (dd, 1H, J = 3.9 Hz, J = 5.9 Hz, H4D), 3.74 – 3.61 (m, 7H, OCHH Linker, H2B, H4C, H3D, H4A, H3A, H3C), 3.55 – 3.52 (m, 2H, OCHH Linker, H2D), 3.25 (dd, 1H, J = 3.9 and 10.3 Hz, H2A), 3.23 (dd, 1H, J = 3.7 and 10.8 Hz, H2C), 2.98 (t, 2H, J = 7.5 Hz, NCH2 Linker), 1.69 – 1.44 (m, 6H, 3*CH2 Linker).

2.4.4 Data processing and analysis

All the NMR data were processed with NMRPipe/NMRDraw [23] and SPARKY [24]. Chemical shift changes for residues showing substantial changes in HSQC spectra were combined using the formula $\sqrt{\frac{1}{2}[\delta_H^2 + (0.14\delta_N)^2]}$ [25]. The resulting curves as a function of ligand concentration were fit to equation 1 to extract dissociation constants for the heparan sulfate tetramers [25].

$$\Delta\delta_{obs} = \Delta\delta_{max}\{([P]_t + [L]_t + K_d) - [([P]_t + [L]_t + K_d)^2 - 4[P]_t[L]_t]\}/2[P]_t \quad (1)$$

All the trNOEs were based on peak pick volumes from SPARKY. These were converted to distances using a $1/r^6$ distance dependence and a 2.5 Å separation for the GlcNAc H1-H2 pair as a calibration distance.

Assignment of crosspeaks to specific protein sites was initially accomplished by comparing a variety of experimental data to data based on predictions using the coordinate file, 2v9r. Data included ^{1}H and ^{15}N chemical shifts, ^{15}N-filtered NOEs, and RDCs, Predictions of chemical shifts were done with PPM_one [26], NOEs were predicted from the crystal structure using a $1/r^3$ distance dependence, and RDCs were calculated using REDCAT [27]. All of the information has more recently been combined in a genetic algorithm approach within MATLAB scripts (MathWorks, Inc.) [28] to confirm assignments.

2.4.5 Susceptibility tensor determination

To extract susceptibility tensors needed in calculating PCS derived constraints on the ligand, experimental RDCs and PCSs for ^{15}N labeled protein sites were used. To combine RDC and PCS data we took advantage of the fact that the same anisotropic magnetic susceptibilities of lanthanides that produces PCSs also induces the alignment needed for RDC measurement. RDCs produced by field –induced alignment of molecules with anisotropic susceptibilities rise with magnetic field squared and are significant above 14T. RDCs were extracted for the sites in domain D1 of Robo1-Ig1-2 using the formula in equation 2 and data collected at field strengths corresponding to proton observation at 600 and 900 MHz.

$$D_{NH,\exp(900)} = \frac{900^2}{900^2-600^2}[(J+D)_{HN\,(900)} - (J+D)_{HN\,(600)}] \qquad (2)$$

The PCSs and RDCs (6 PCSs and 3 RDCs) were then combined in the program REDCAT to determine the susceptibility tensor (order tensor). This requires appropriate scaling using different RDCmax and PCSmax constants (24350 Hz for ^{15}N-^{1}H RDCs and 18.54×10^{6} ppm for PCSs at 900 MHz) and the extraction of inter-nuclear and ion-nuclear vectors from an appropriate pdb file.

2.4.6 MD simulation of the Robo1-Ig1-2-loop construct

To provide an appropriate pdb file for loop containing Robo1-Ig1-2 a 1 μs molecular dynamics (MD) trajectory was produced. The molecular dynamics simulation was carried out using the AMBER 14 package [30] and the ff14SB force field [31] with the SANDER module. The GLYCAM_06j-1 force field [32] was adapted to the ff14SB force field for carbohydrate simulation. The crystal structure of Robo1-Ig1-2 (pdb: 2v9r) was used to obtain the initial atomic coordinates and the lanthanide-binding loop was modeled in using tools in CHIMERA [29]. A

cubic box of TIP3P [33] water molecule was used to solvate the protein, maintaining a distance 8.0 Å between face of the box and the solute. The system was first energy minimized by 2000 steps of minimization including steepest descents and conjugate gradients. Then the system was heated gradually to 300 K at 2 fs stepwise for 400 ps. The MD simulation last for 1 μs with a 2 fs time step with the pressure and temperature maintained. Frames from 401 ns to 1000 ns were used to find the average ion and isotopically labeled site positions required by REDCAT. The resulting order tensor elements were Sxx=-8.22×10^{-5}, Syy=-3.31×10^{-4} and Szz=4.13×10^{-4}. The Q factor for back-calcualted data was 0.22. This order tensor and the PCSmax were then used to predict PCSs for the ligands to be used in scoring of docked poses as described in the following section.

2.4.7 Robo1- HS complex assembly by HADDOCK

Models of Robo1- HS complex were generated using the docking program HADDOCK [18]. The average Robo1-Ig1-2-loop structure generated from MD simulation was used as the input protein structure. Ligands structures were produced using the GLYCAM web server [30]. Restraints involving residues of the protein or parts of the ligand identified as being involved in an interaction by chemical shift perturbation or STD intensity were entered as ambiguous interaction restraints. Interproton distances derived from trNOE data connecting specific pairs of protons were converted to upper and lower bounds for distance constraints by adding or subtracting 0.6 Å and 0.3 Å. PCSs data for both the protein and ligand were implemented using the XPCS restraints as defined in HADDOCK. The ligand was set to be fully flexible and the loops of the protein containing the residues having the most perturbed chemical shift on ligand addition or identified as interacting residues in mutational studies were specified as semi-flexible. The docking began with rigid-body energy minimization followed by semi-flexible

refinement using simulated annealing and ended with water refinement using default force field parameters. 200 refined models ranked by the weighted sum of electrostatic and van der Waals energies were obtained and the twenty top scoring models were submitted to further analysis.

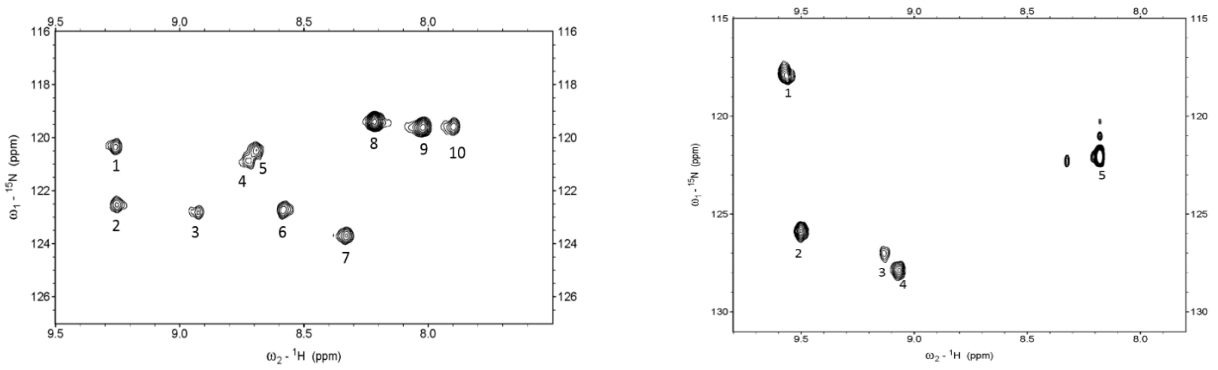
2.4.8 Solvation energy calculations

Similar MD trajectories were initiated as described for the Robo1-Ig1-2-loop construct, but now with the 2-sulfated HS tetrasaccharide docked into the Robo1-Ig1-2-loop construct as in the top five results from HADDOCK docking. In each simulation, the solvated complex was equilibrated by carrying out 50 ps of minimization, heating and density equilibration followed by 2200 ps of constant pressure equilibration at 300 K. Production runs of 50 ns were then initiated. Various properties including density, temperature, pressure and potential energies were monitored to ensure that the equilibration had been achieved and was well maintained. In order to calculate the free binding and solvation energy of the protein-ligand interactions, the molecular mechanics generalized Born surface area (MM-GBSA) method followed by per-residue decomposition analysis was conducted. The full length MD (50ns) was taken for the post-processing free binding energetic analysis with igb = 5 for each 10th frame (500 frames in total) [20].

2.5 Results

2.5.1 Chemical shift perturbation of sparsely labeled Robo1-Ig1-2

Robo1 is a highly glycosylated protein which prevents employment of bacterial hosts commonly used for protein expression. Instead, sparse labeling of Robo1-Ig1-2 from HEK293 cell with lysine and phenylalanine has been used. There are 12 lysines and 5 phenylalanines in the Robo1-Ig1-2 construct. The 2D ^{15}N-^{1}H HSQC spectra of lysine labeled and phenylalanine labeled Robo1-Ig1-2 are shown in Figures 2.4A and 2.4B respectively.



A B

Figure 2.4. 2D ^{15}N-^{1}H HSQC HSQC spectra of (A) ^{15}N Lys labeled Robo1-Ig1-2 and (B) ^{15}N Phe labeled Robo1-Ig1-2. Each peak is labeled with a peak number as a reference for the following assignment.

Crosspeaks for all phenylalanine residues are observed, but two crosspeaks are missing from the HSQC of lysine labeled Robo1-Ig1-2. It is common to find missing HSQC crosspeaks due to high rates of amide proton exchange in solvent exposed regions or in regions that are dynamic and suffer line broadening due to modulation of chemical shifts on timescales near the reciprocal of shift changes in Hz. Comparing the HSQC with that of a mutant, R136AK137A, that was found to lack binding activity [31], one finds a superimposable spectrum with no additional crosspeaks missing. This shows that one missing crosspeak belongs to K137. The other missing crosspeak belongs to K266. Mass spectral analysis of the intact protein shows a mass deficiency equivalent to that of a lysine and a trypsin digest of Robo1-Ig1-2 detects only C-terminal peptides missing K266. It is likely that some proteolysis has occurred during expression and isolation.

Chemical shift perturbation is a qualitative method for studying protein-ligand binding and both dissociation constants, K_d,and binding site location can be determined from the same

set of experiments. The ^{15}N-^{1}H HSQC spectra for the lysine labeled Robo1-Ig1-2 in the presence of different concentration of HS tetramer 1 (IdoA-GlcNS6S-IdoA2S-GlcNS6S-$(CH_2)_5NH_2$) and 2 (IdoA-GlcNS6S-IdoA-GlcNS6S-$(CH_2)_5NH_2$)) are overlaid in Figure 2.5A and 2.5B. One lysine residue shows significant chemical shift perturbation (> 0.05) when titrated with ligand 1 and moderate perturbation with ligand 2. No phenylalanine residues show any significant perturbation during titration, which indicates that phenyalanines are not highly involved in the binding process. But even perturbation of a single peak can allow determination of a binding constant. Binding affinities have been extracted by fitting curves for chemical shift as a function of concentration to the binding equation given in methods (see Figure 2.6). These are of 255 ± 30 µM and 45 ± 30 µM for ligands 1 and 2 respectively. A similar titration with an HS octamer containing the IdoA2S at the third position from the non-reducing end (GlcA-GlcNS6S-IdoA-GlcNS-IdoA2S-GlcNS6S-IdoA-GlcNAc6S-$(CH_2)_5NH_2$) shows a K_d of 86 ± 11 µM with K81 showing a very similar perturbation (0.06 ppm).

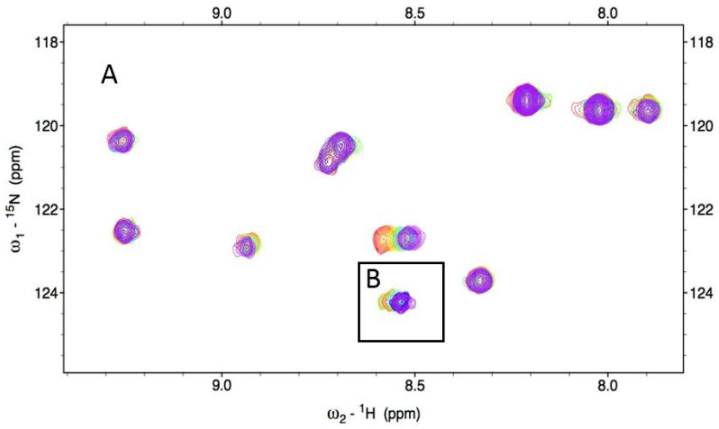
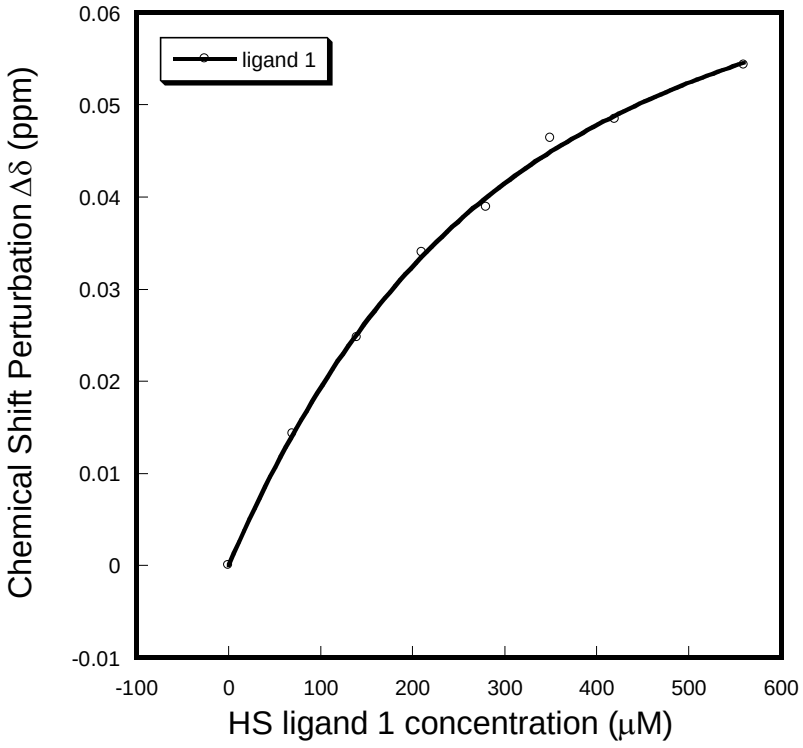
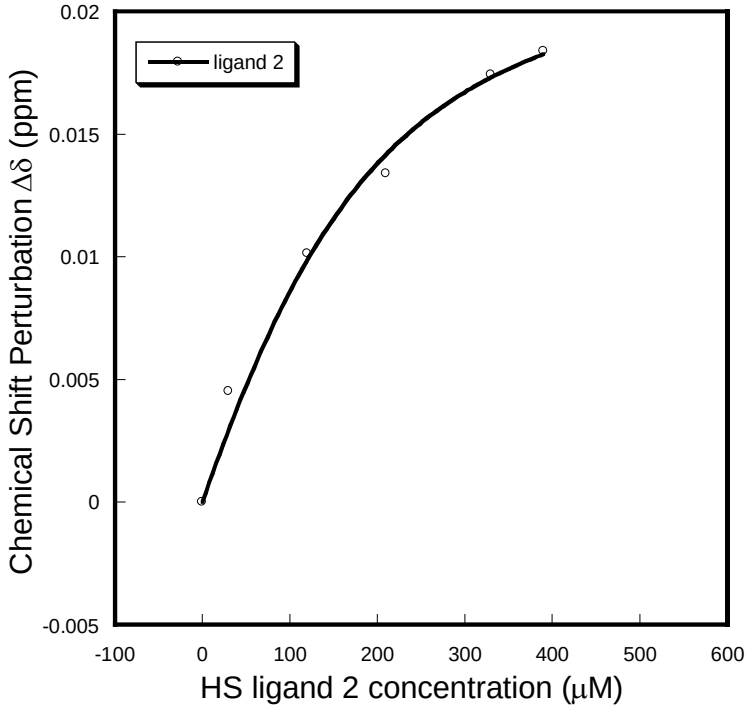


Figure 2.5. (A) Overlaid $^{1}H–^{15}N$ HSQC spectra for Robo1-Ig1-2 with ligand 1. Shifted resonances are presumed to be close to the binding site. A dissociation constant of ~255 µM is obtained by fitting titration data. (B) A portion of the $^{1}H–^{15}N$ HSQC spectrum for Robo1-Ig1-2

with ligand 2 is shown. The same resonance shows less perturbation but a lower a dissociation constant of ~45 μM is extracted. The overlaid titration spectra are in rainbow color coding with increased concentration of HS tetramer (0 μM of ligand one in red and 560 μM in purple with a stepwise increase of 70 μM; 0 μM of ligand 2 in red and 390 μM in purple with a first step of 30 μM and later steps of 60 μM).



A



B

Figure 2.6. Binding affinity of Robo1 with (A) ligand 1 and (B) ligand 2.

2.5.2 Assignment of sparsely labeled Robo1-Ig1-2

The fact that residues whose chemical shifts are perturbed on ligand addition are likely involved in ligand binding makes their resonance assignment a high priority. Triple resonance approaches are widely employed in most NMR crosspeak assignments of proteins. However, the requirement for uniform isotopic enrichment in ^{13}C, ^{15}N and frequently ^{2}H makes this approach impractical for proteins that must be expressed in mammalian cells. A series of new resonance assignment strategies applicable to sparsely labeled proteins is under development [14, 32]. Here we combine a number of measurements that can be made on sparsely ^{15}N labeled proteins with

predictions based on known domain structures to achieve these assignments for Robo1-Ig1-2. The measured NMR observables include backbone chemical shifts, RDCs and NOEs.

Since the individual measurement types seldom give unambiguous assignments one would ideally examine predictions for all measurement types using an entire set of permuted assignments to make a decision. Even for the 10 unassigned lysine crosspeaks this is an arduous task (10! is 3.6288 million possible assignments). Instead we have sequentially applied each measurement type retaining only assignments that agree within generous error limits at each step. First, amide chemical shifts were predicted by PPM_one [26], which incorporates effects from both backbone geometry and remote groups as seen in the crystal structure, and compared with each of the crosspeak positions using a 2 sigma standard deviation. Next, NOEs for protons within 4 Å of each labeled amide were predicted using distances extracted from 2v9r using Chimera [32] and shift predictions from PPM_one to position NOEs in the proton dimension of a hypothetical 3-dimensional spectrum. They were compared to the experimental NOE measurements from a 15N-edited HSQCNOESY, again considering a shift difference of 2 sigma (0.34 ppm) and observation of 2/3 of the expected peaks as acceptable. Last, RDCs were measured for each labeled site and compared with predictions using REDCAT [27]. ^{1}H-^{15}N RDCs are inherently dependent on the protein structure at each site because they reflect the average of $(1-3\cos^2\theta)$ where θ is the angle of an H-N bond vector relative to the magnetic field. However, their measurement requires partial alignment in a suitable medium (here a dilute bacteriophage solution), and defining this alignment requires determination of five alignment parameters. 10 lysine RDCs are adequate for both alignment parameter determination and assignment screening, providing we can assume the two Robo1-Ig1-2 domains are rigidly oriented with respect to one another. This is not strictly correct, as we shall discuss below, but we decided to proceed using

the crystal structure, 2v9r, which proves to have domain orientations near the dominant structures seen in a long (1 μs) MD trajectory. Solutions with RDC Q factors less than 0.3 were regarded as acceptable. The experimental and predicate data are summarized in Table 2.1 for crosspeaks from domain Ig1-2. At each step predictions are included only for possibilities resulting from the precious step. Two additional pieces of information were used to confirm assignments and resolve ambiguities arising from the presence of crosspeaks from two domains. First, consistency with distance information from a Robo1-Ig1-2 loop construct with Gd^{3+} complexed was examined (see section on paramagnetic perturbations below). Second, a construct containing only the first domain was labeled and the correct association of crosspeaks with each of the two domains was confirmed.

Most important among these assignments is that of crosspeak 6 to K81. This is the peak that shifts with addition of ligand, particularly HS ligand 1. While there is no specific relationship between shifts and the separation of the shifted residue and the ligand, it is highly likely that the distance of separation is short. Moreover, the fact that ligand 1, which has a sulfate on the 2 position of the internal iduronic acid produces a significantly larger shift than ligand 2, suggests that K81 is involved in an ion pair interaction with this sulfate group. This, along with abolition of binding in the R136A, K137A mutant, identifies a potential binding site on the protein.

Table 2.1. Experimental (Exp.) and predicted (Pred.) data leading to assignment of labeled sites in the Robo1-Ig1-2.

Peak number	15N,1H shift (ppm)		NOE (ppm)		RDC (Hz)	Final Assignment
	Exp.	Pred.	Exp.	Pred.	Exp./Pred.	
Domain 1 Lys						
Lys -3	122.8,8.9	K81(123.9,8.6) K90(123.4,8.4), K103(124.2,8.9),	0.80,0.95,4.83, 5.02,6.70	K103(0.92,1.29,1.92,2.65,4.92,8.43) K90(0.18,1.70,1.79,2.87,4.58,7.89, 9.10)	-11.3±4.9/K103(-7.5)	K103
Lys - 6	122.7,8.6	K81(123.9,8.6), K90(123.4,8.4),	0.57,1.44,1.73, 1.80,3.75,4.22, 4.66	K81(0.56,0.65,1.78,3.78,4.66) K90(0.18,1.70,1.79,2.87,4.58,7.89, 9.10)	3.3±3.2/K81(3.4)	K81
Lys - 7	123.7,8.3	K81(123.9,8.6), K90(123.4,8.4),	0.70,0.80,1.31, 1.41,1.61,1.64, 1.73,4.50.4.59, 4.78	K81(0.56,0.65,1.78,3.78,4.66) K90(0.18,1.70,1.79,2.87,4.58,7.89, 9.10)	3.0±0.7/K90(3.0)	K90
Lys - 10	119.6,7.9	K90(123.4,8.4), K112(117.9,7.8)	1.54,1.67,1.73, 1.79,4.12,4.78 7.93	K112(1.49,1.79,1.59,4.33,4.36,7.69,8.17, 8.27,8.54)	-1.5±2.2/ K112(-1.5)	K112

Peak number	15N,1H shift (ppm)		NOE (ppm)		RDC (Hz)	Final Assignment
	Exp.	Pred.	Exp.	Pred.	Exp./Pred.	
Domain 2 Lys						
Lys - 1	120.1,9.2	K205(120.3,9.3), K232(122.5,8.7), K237(119.3,8.1)	0.72,0.67,0.59, 0.97,0.91,1.84, 1.89,1.72,1.65, 1.58,1.41,1.23, 3.05	K205(0.65,0.68,0.73,1.52,3.12,4.13,5.51, 9.12) K232(0.73,0.94,1.72,1.87,1.49,1.65,1.72, 1.87,4.70,8.27)	-15.6±3.4/K205(-16.3)	K205
Lys - 2	122.4,9.3	K205(120.3,9.3) K206(124.9,8.3) K232(122.5,8.7),	0.70,0.68,0.59, 0.97,0.93,1.84, 1.89,1.72,1.65, 1.58,1.41,1.23, 3.05	K205(0.65,0.68,0.73,1.52,3.12,4.13,5.51, 9.12) K232(0.73,0.94,1.72,1.87,1.49,1.65,1.72, 1.87,4.70,8.27)	6.7±2.7/K232(6.9)	K232

Lys - 4	121.0,8.8	K205(120.3,9.3), K224(121.8,7.8), K237(119.3,8.1)	0.96,0.86,1.45, 1.30,1.24,4.78, 4.60	K205(0.65,0.68,0.73,1.52,3.12,4.13,5.51, 9.12) K237(0.77,1.49,1.70,4.25,4.20,9.29)	-13.8±3.7/K237(-13.3), K205(-16.3)	K237
Lys - 5	120.5,8.7	K205(120.3,9.3) K214(117.8,7.9), K224(121.8,7.8), K237(119.3,8.1)	0.92,0.82,0.95, 1.60,1.54,1.47, 1.32,1.24,4.55	K214(1.87,2.64,2.67,7.66) K224(1.46,1.67,4.02,4.28,8.73) K237(0.77,1.49,1.70,4.25,4.20,9.29)	1.1±0.9/K214(1.2), K224(1.3)	K214 or K224
Lys - 8	119.4,8.2	K214(117.8,7.9), K224(121.8,7.8), K237(119.3,8.1)	1.44,1.91,1.68, 1.53,3.04,2.90, 2.52,4.64,4.54, 7.68,	K214(1.87,2.64,2.67,7.66) K224(1.46,1.67,4.02,4.28,8.73)	7.7±1.6/K224(7.3), K214(7.7)	K224 or K214
Lys - 9	120.2,8.0	K205(120.3,9.3) K206(124.9,8.3) K237(119.3,8.1) K224(121.8,7.8),	0.85,0.80,0.72, 1.89,1.73,1.64, 1.47,1.24,1.06, 4.71,4.78,4.01, 5.13,8.98	K205(0.65,0.68,0.73,1.52,3.12,4.13,5.51, 9.12) K206(0.73,1.15,1.32,4.93,7.6)	-15.2±1.7/K205(-16.3), K206(-14.6)	K206

Peak number	15N,1H shift (ppm)		NOE (ppm)		RDC (Hz)	Final Assignment
	Exp.	Pred.	Exp.	Pred.	Exp./Pred.	
Phe						
Phe - 1	117.9,9.6	F66(119.0,9.1)	0.77,1.01,1.61, 2.59	F66(0.46,1.56,2.69,2.37,2.96,6.93,7.11,6. 85)	0.5±2.5/F66(1.8)	F66
Phe - 2	125.8,9.5	F128(127.8,9.3), F129(126.9,9.6)	2.61,3.18,4.83	F128(0.35,0.45,1.60,1.98,2.88,4.65,4.80,8 .98) F129(1.04,2.66,5.12,7.03,7.15)	15.6±5.0/F128(15.6), F66(12.2)	F128
Phe - 3	127.1,9.1	F128(127.8,9.3), F129(126.9,9.6), F172(123.3,8.6)	1.01,4.53,6.89	F128(0.35,0.45,1.60,1.98,2.88,4.65,4.80,8 .98) F129(1.04,2.66,5.12,7.03,7.15)	16±5.0/F129(11)	F129
Phe - 4	127.7,9.0	F128(127.9,8.9), F172(123.3,8.6)	0.77,1.05,2.86, 5.34	F128(0.35,0.45,1.60,1.98,2.88,4.65,4.80,8 .98) F172(1.43,2.57,2.92,4.76,7.09)	12.1±5.2/F172(9.4)	F172
Phe - 5	122.1,8.2	F264(124.9,8.3)*	2.25,4.35,4.75	F264(1.57,3.05,4.37,4.10,7.96,8.20)*	-6.8±0.8/NA*	F264

*Assignment of Peaks 5 and 8 to K214 and K224 remains ambiguous

2.5.3 Saturation transfer difference NMR

Saturation transfer difference (STD) experiments complement the chemical shift perturbation experiments nicely in that they identify potential interaction epitopes on the ligands. They are applicable when exchange of ligands on and off proteins occurs on a timescale less than typical NMR spin relaxation times. The ligands studied here easily fall into that category. The mechanism is very much like that of an NOE experiment; saturation of magnetization of a particular proton on the protein is transferred to a proton on the ligand in a $\frac{1}{r^6}$ dependent way, where r is the distance between two protons, and a ligand resonance with a close approach to a protein proton will decrease in intensity, identifying it as a part of the binding epitope on the ligand. However, saturation of protein protons is not specific; spin exchange among protons in a large molecule like a protein is so efficient that saturation of one set of protons (those on methyl groups near -1.5 ppm in our case) quickly spreads to protons in the proteins. Hence, epitope identification on the ligand is qualitative, much like the identification of binding site residues in the proteins by chemical shift perturbation. There are some complexities. The changes in intensity are sometimes small and there are changes in the protein spectrum as well. Taking a difference between spectra with and without protein saturation, and filtering out the much broader resonances of the protein usually leaves a spectrum dominated by resonances for protons in close contact with the protein. For the protein studied here, a glycoprotein, there are additional complications in that many of the resonances from the attached glycan are not broad and persist in the difference spectrum at positions that often overlap with those of our ligand. Hence, a double difference spectrum was produced using an equivalent STD spectrum acquired on the protein in the absence of ligand.

STD experiments were acquired with a series of saturation times ranging from1 to 4 s (see Figure 2.7); the build-up rates are useful in comparing to simulated STD spectra once a model of the complex is obtained [33]. However, in Figure 2.8 we report data only using the longest saturation time (4 s). Quantification of the signal has been made by dividing the STD signal intensity by original ligand proton spectrum intensity and then scaling all intensities relative to that with the highest level. For ligand 1, the largest signals arise from H2C (or H2A which is overlapped), H2B, and H4D. H2B is on the internal IdoA residue, supporting a direct interaction of this residue with the protein surface near K81. Most other protons show a significant signal reduction suggesting that significant spin diffusion among ligand protons may challenge our ability to interpret these signals as specific interactions at the protein surface. Ligand 2 shows STD signals for most ligand protons as well. H2C and H4D are among those most perturbed as with ligand 1 and there is a significant signal for H2B, which is strongly perturbed in ligand 1. A strong signals is also seen for H6A in ligand 2 which is only moderately in ligand 1. The similarities support the suggestion that the two ligands occupy a similar site, but with some significant differences in detailed contacts. In any event, STD spectra provide another useful source of information for molecular docking.

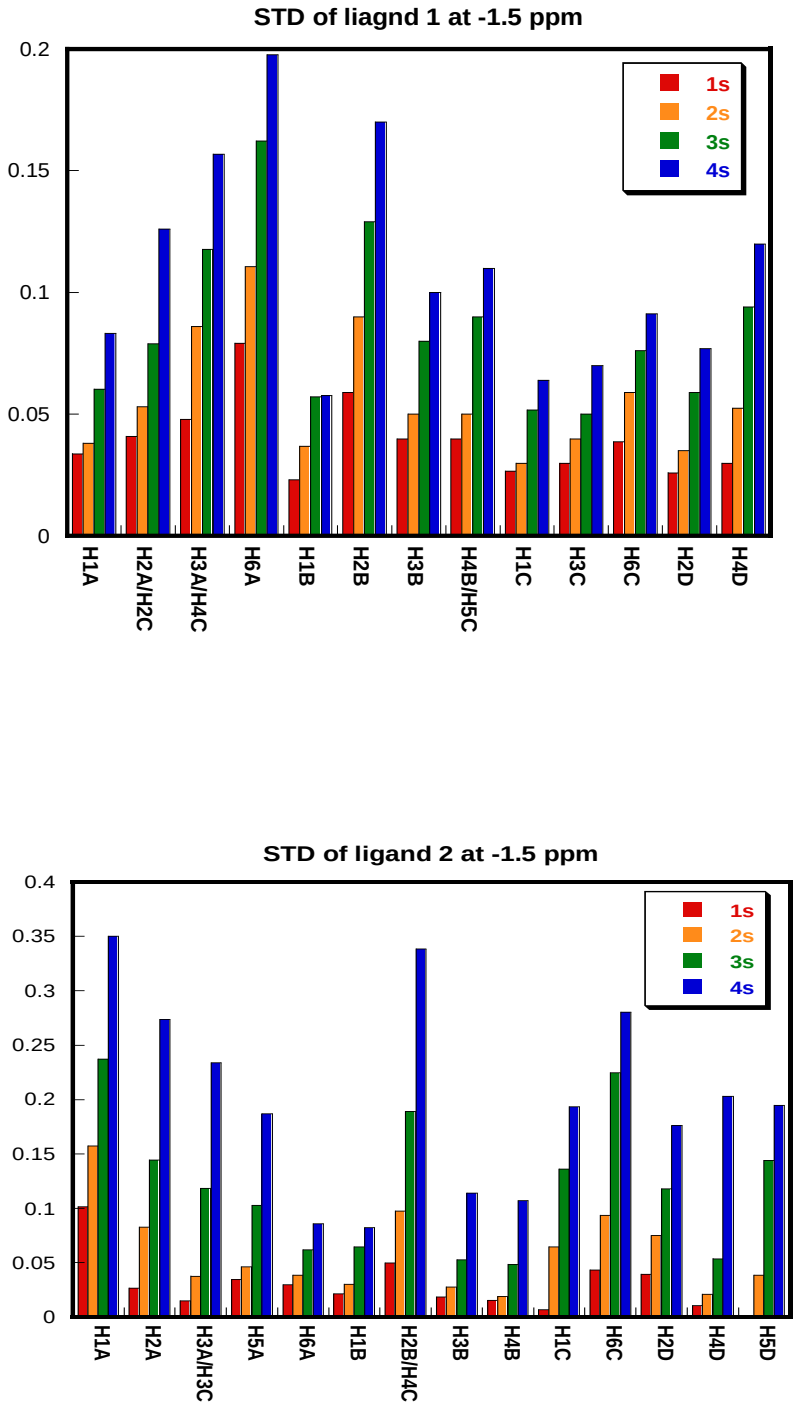


Figure 2.7. Experimental STD build up curves of Robo1 with ligand 1 (A) and ligand 2 (B) for saturation at -1.5 ppm with saturation times from 1 to 4 s.

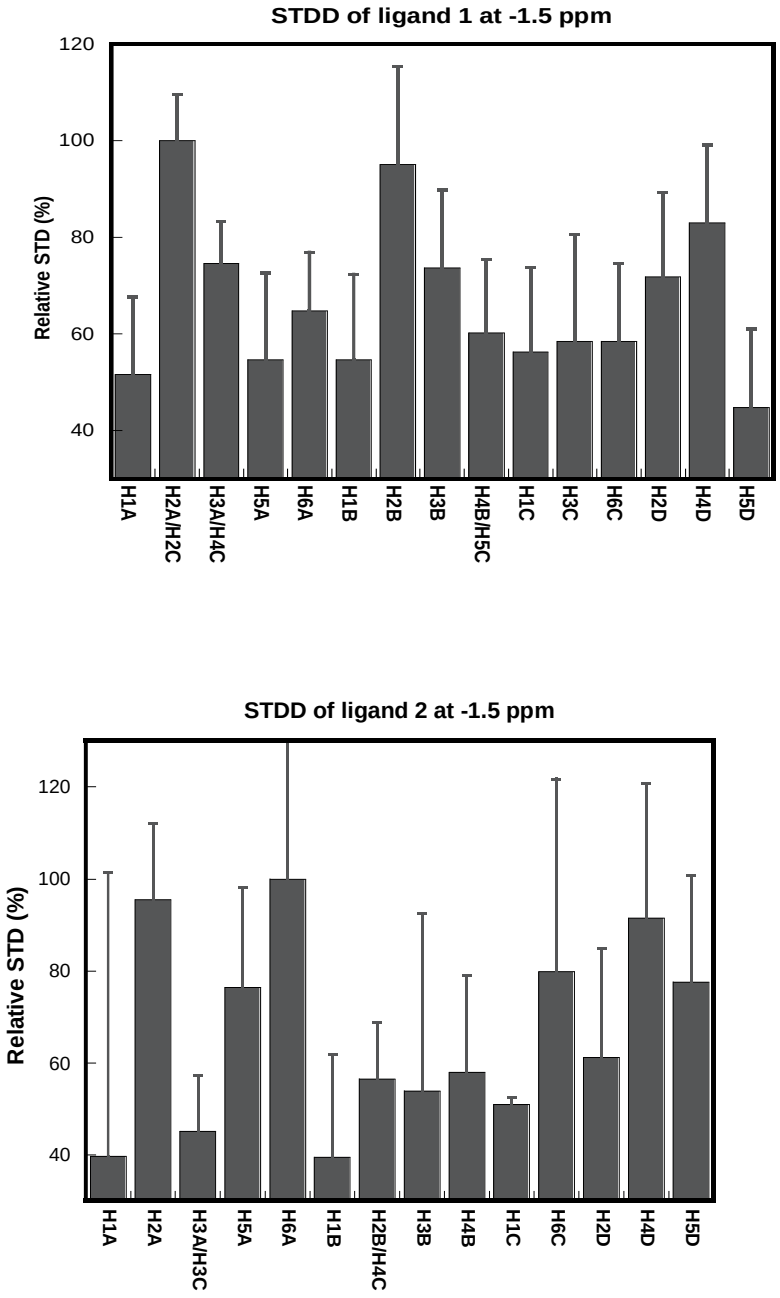


Figure 2.8. Quantification of experimental saturation transfer double difference data on various resonances with a saturation time of 4 s at -1.5 ppm of (A) ligand 1 and (B) ligand 2. Errors are derived based on RMS noise limits.

2.5.4 Transferred nuclear Overhauser effect

Heparan sulfates (HS) are a group of glycans with a significant degree of internal mobility, both in terms of variations in glycosidc bond torsion angles and iduronic acid ring forms. A subset of these conformers are likely to be selected on binding to Robo1-Ig1-2. Nuclear Overhauser Effects (NOEs) provide insight into conformations sampled through their $\frac{1}{r^6}$ dependence on interproton distances between protons. Transferred Nuclear Overhauser Effects (trNOE) report rather specifically on bound conformations because of the enhanced transfers of magnetization in large molecular assemblies (effects scale up approximately in proportion to the rotational correlation time for the complex). This allows contributions to observed ligand NOEs from bound ligands to dominate over contributions from ligands in solution even at ligand to protein ratios of 20:1. NOEs taken at short mixing times can be converted to distances using the $1/r^6$ dependence and a reference NOE with a known distance, in our case that for the GlcNAc C H2 and H4 pair at 2.5Å. Distances converted in this way are are weighted averages of sampled conformations. However, the number of conformers sampled in the bound state tends to be small and derived distances in this case should be close to those in the minimum energy bound conformer. The derived distances between pairs of nuclei on opposite sides of the glycosidic bonds in bound conformers of ligands 1 and 2 are listed in Table 2.2.

Table 2.2. Transglycosidic distances of ligand 1 and 2 measured in bound and free state from trNOEs. Errors are derived based on RMS noise limits in reference and sample spectra.

Ligand 1	Sugar Ring 1	Sugar Ring 2	Atom 1	Atom 2	Bound Ligand (Å)	Free ligand (Å)
	Linker	GlcNAc A	methylene	H1	2.27 ± 0.04	2.74 ± 0.04
	GlcNAc A	IdoA B	H4	H1	1.90 ± 0.04	2.38 ± 0.05
	IdoA B	GlcNAc C	H3	H1	2.47 ± 0.04	2.65 ± 0.04
	IdoA B	GlcNAc C	H4	H1	2.25 ± 0.04	2.45 ± 0.05

Ligand 2	Sugar Ring1	Sugar Ring 2	Atom 1	Atom2	Bound Ligand (Å)	Free ligand (Å)
	Linker	GlcNAc A	methylene	H1	2.38 ± 0.06	2.60 ± 0.05
	GlcNAc A	IdoA B	H4	H1	3.02 ± 0.05	2.77 ± 0.07
	GlcNAc A	IdoA B	H6	H1	3.10 ± 0.05	2.98 ± 0.06
	IdoA B	GlcNAc C	H3	H1	2.48 ± 0.13	2.37 ± 0.06
	IdoA B	GlcNAc C	H4	H1	2.69 ± 0.07	2.57 ± 0.07
	GlcNAc C	IdoA D	H6	H1	2.95 ± 0.04	2.26 ± 0.06

For comparison, distances derived from NOEs for the same pairs in the absence of protein are also listed. There are some moderate differences between the bound and free state. For example, for ligand 1, distance between GlcNAc A H4 and IdoA B H1 in the bound state is found to be 1.90 ± 0.04 while that found in the free state is marginally larger (2.38 ± 0.05 Å). For ligand 2, the distance between GlcNAc C H6 and IdoA D H1 is found to be 2.95 ± 0.04 Å while that in the free state is significantly shorter (2.26 ± 0.06 Å). Note that these deviations involve the terminal residues of the tetrasaccharide which may have more motional freedom in solution; no significant differences were observed for the central portion of the tetrasaccharide suggesting that something close to the minimum energy conformer found in solution is selected for the bound state. The bound state data provide additional distance restraints when implementing docking.

2.5.5 Pseudo contact shifts (PCSs) of Robo1-Ig1-2 and the HS complex

Paramagnetic effects caused by lanthanide ions offer unique opportunities to more quantitatively position ligands in protein-ligand complexes [34]. Pseudo contact shifts (PCSs) are changes in chemical shift caused by an average magnetic field from an induced dipole moment centered on the unpaired electron distribution of the lanthanide. It depends not only on the distance between a nucleus and the metal ion (decreases with r^{-3}) but also on the orientation and

magnitude of the anisotropic part of the ion's susceptibility tensor. Once the susceptibility tensor is known, precise long-range distance and orientational constraints can be deduced from PCSs. To provide a site capable of binding a paramagnetic ion in Robo1-Ig1-2, a short polypeptide (SYIDTNNDGAYEGDELSG) has been engineered into the Robo1-Ig1-2 construct between strands C and D of the D1 domain. Luminescence data based on a tryptophan to Tb^{3+} energy transfer show the site to have a binding affinity of 62 nM (see Figure 2.9); affinities for other lanthanides are expected to be similar. The superimposed spectra of lysine and phenylalanine labeled Robo1-Ig1-2-Loop with Tm^{3+} (a paramagnetic lanthanide) and Lu^{3+} (diamagnetic lanthanide) are shown in Figure 2.10. The unique diagonal shifts in peak positions are used to pair the resonances in each spectrum (since PCSs are independent of the nucleus, near identical chemical shifts are observed in both ^{1}H and ^{15}N dimensions). Similar measurements can be made on at least well resolved resonances from ligands in 1D proton experiments. In these cases an average of the resonance position for the uncomplexed ligand and the complexed ligand is measured (our ligands are in fast exchange) and shifts have to be scaled by the percentage bound (data shown in Table 2.3). The qualitative similarity of shifts for ligands 1 and 2 indicate that they adopt a similar pose when binding with Robo1-Ig1-2.

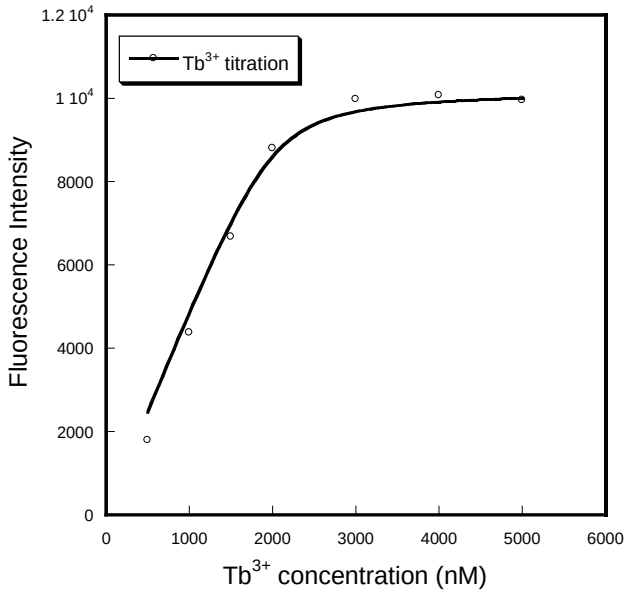
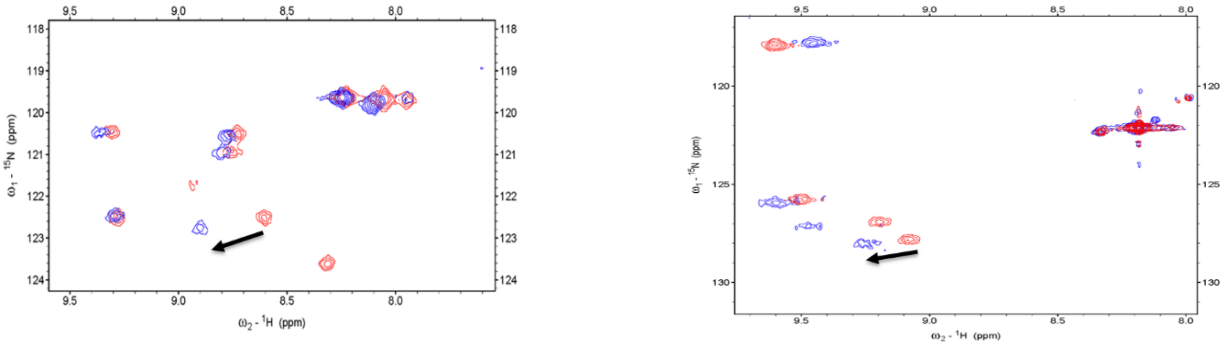


Figure 2.9. Tb^{3+} binding affinity of Robo1 LBT construct.



A B

Figure 2.10. Superposition of ^{15}N-^{1}H HSQC spectra of (A) ^{15}N-Lys labeled and (B) ^{15}N-Phe labeled Robo1-Ig1-2, engineered with lanthanide binding peptide loaded with Lu^{3+} (red) or Tm^{3+} (blue).

Table 2.3A. Limiting PCS of ligand 1

Resonance	PCS (ppm)
H1A	0.109
H2A/C	0.186
H3A/H4C	N/A
H6A/H2B	0.125
H1B	0.130
H3B	0.167
H1C	0.154
H3C	N/A
H5C/H4B	0.128
H6C	N/A
H1D	0.128
H2D	0.101
H4D	0.127
H5D	0.147

Table 2.3B. Limiting PCS of ligand 2

Resonance	PCS (ppm)
H1A	0.060
H2A	N/A
H3A/H3C	N/A
H5A	0.084
H6A	N/A
H1B	0.071
H2B/H4C	N/A
H3B	0.070
H4B	0.054
H1C	0.071
H6C	N/A
H1D	0.058
H4D	0.060
H5D	0.030

2.5.6 Location and tensor alignment for the Ln-binding loop

Before the observed PCSs for the ligand can be converted to useful constraints on ligand poses in the binding site, the position of the lanthanide ion and the anisotropic part of its susceptibility tensor must be determined. Similar to the case with RDCs five independent elements of a tensor (now a susceptibility tensor) must be determined, plus additional translational coordinates must be specified to properly place the ion in the coordinate frame of the protein. Placement of the ion was done by averaging positions found in an extensive MD simulation (see methods). For the tensor elements it would be tempting to use PCSs for the 10 observable sites in the HSQC spectrum of ^{15}N-lysine labeled Robo1-Ig1-2 and 5 observable sites in the HSQC spectrum of ^{15}N-phenylalanine labeled Robo1-Ig1-2. However, we would again need to assume a rigid Robo1-Ig1-2 structure, and previous literature do indicate significant flexibility between Robo1 D1 and D2 domains [9-10]. Using data for just domain D1, where the loop is attached, there are just 3 lysines and 3 phenylalanine making the number of data points for tensor determination marginal. Hence, we used the fact that the same anisotropic part of the susceptibility tensor is responsible for boththe PCSs and field induced RDCs. Only three filed-induced RDCs could be measured, but this raised the number of data points to 9. All data used for tensor determination are listed in Table 2.4.

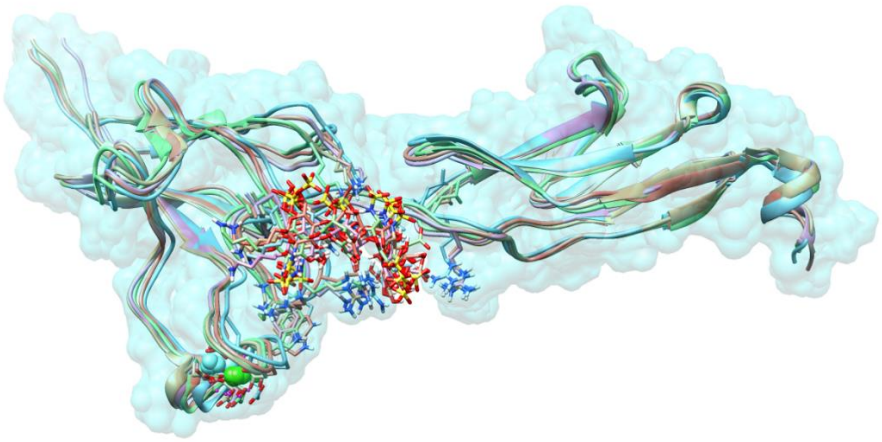
Table 2.4. Measured pseudo contact shifts and field-induced RDCs of Robo1-LBP loaded with Tm^{3+}. The values in bold are from D1 domain and used for tensor determination.

HSQC peak number	H PCS (ppm)	Field Induced RDC (Hz)
Lys NH1	0.073	-2.9
Lys NH2	0.027	2.9
Lys NH3	Broad away	Broad away
Lys NH4	0.064	-2.5

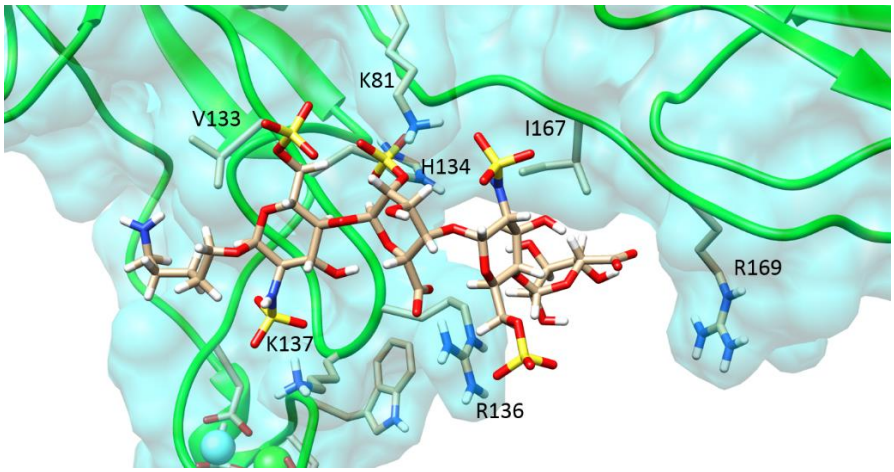
Lys NH5	0.062	-1.2
Lys NH6	**0.304**	**7.0**
Lys NH7	**-0.062**	**-14.7**
Lys NH8	0.036	-3.3
Lys NH9	0.057	-13.4
Lys NH10	**0.172**	**7.8**
Phe NH1	**-0.151**	**NA**
Phe NH2	**0.088**	**NA**
Phe NH3	**0.283**	**NA**
Phe NH4	0.182	NA
Phe NH5	0.004	NA

2.5.7 Computational docking

High ambiguity driven biomolecular docking (HADDOCK) was used to combine all of the structural constrains to determine the structure of a Robo1-Ig1-2 HS complex. Haddock makes use of a variety biochemical and biophysical data to characterize protein ligand complexes. More qualitative information, such as those coming from chemical shift perturbations and STD experiments, are represented as ambiguous restraints and the more quantitative ones, such as those coming from trNOE and PCSs, are treated explicitly in error functions that that compare experimental data and predictions calculated from various models. The energetically minimized coordinates of HS tetramers 1 and 2 were generated using GLYCAM [30, 35], and the docking process was as described in the Materials and Methods. In the end, the top 5 HADDOCK structures with the lowest energy and score as well as no distance restraint violations greater than 0.5 Å were obtained for pose evaluation. Figure 2.11A shows that the top 5 structures form a single cluster. A single average representation of the protein is shown in Figure 2.11B in a ribbon format with residues making frequent contacts with the ligand shown as ball and stick entities.



A



B

Figure 2.11. (A) Overlaid top 5 HADDOCK structures of Robo1-Ig1-2-HS with highest scores and lowest energy. (B) Expanded view of the binding pocket for the best HADDOCK structure with interacting residues within 1 Å of van der Waals contact presented in a stick representation.

2.6 Discussion

The top 5 structures show a well-clustered binding location as well as well-defined ligand conformation. The structure with the highest docking score and lowest energy is used as an

illustration in Figure 2.11B: the residues within 1 Å of van der Waals contact of the ligand include K81, V133, H134, G135, R136, K137, I167, and R169. Five of these are positively charged residues, a number close to the number of negatively charged residues in ligands 1 and 2, namely 7 and 6 respectively. Among the listed residues, K81, R136 and K137 have been confirmed by site-directed mutagenesis to participate in the interaction with HS [6, 9]. More specifically R169 is close enough for a strong electrostatic interaction with IdoA D (4.1 Å). The sidechain of R136 is close enough to interact with the 6-sulfate of GlcNAc C (3.0 Å) as well as the carboxylate of the 2-sulfated IdoA B (1.6 Å). The side chain of K81 is close enough to interact with the 2-sulfate of IdoA B (1.8 Å) as well as the N-sulfate of GlcNAc C (1.8 Å). The side chain of H134 is close enough to interact with the 2-sulfate of IdoA B (2.9 Å). K137 is close enough to interact with the N-sulfate of GlcNAc A (2.0 Å). There are also strong van der Waals interactions that may explain some of the stronger STD signals. The epsilon methylene protons of K18 are close enough to H2 of IdoA B for at least transicent interactions (2.7 Å). The gamma methyl protons of I167 are in van der Waals contact with H4 of IdoA D and the methyl protons of V133 are in van der Waals contact with the reducing terminus extension on GlcNAc A.

Given the number of potential of favorable interactions with the 2-sulfate of IdoA B, including K81, it may seem strange that removal of this sulfate in ligand 2, which appears to occupy essential the same binding site, actually leads to a higher affinity. However, note that K81 does interact with other electronegative groups (the N-sulfate of GlcNAc C) which may become stronger if not shared with a 2-sulfate on IdoA B, and the carboxyl group of IdoA B favorably interacts with R136 that may also be stronger if another positive residue is not in play. In rationalizing binding affinities, it is also important to remember that these relate not just to

favorable interactions with the protein, but to the difference in free energies between the ligand-protein complex in solvent and the separated ligand and protein in solvent. To properly consider these effects molecular mechanics-generalized Born surface area (MM-GBSA) calculations were utilized. In order to specifically study the role the 2-SO3 of IdoA B we treated it as a separate residue in a per-residue energy decomposition analysis. The top 5 docked structures were chosen for energy calculation. The average electrostatic interactions with K81 are indeed favorable (-49.34 kcal/mol). However, the desolvation penalty in moving this sulfate from water to the protein is actually larger (52.33 kcal/mol). Hence, according to this analysis, the 2-sulfate does not produce a net favorable reduction in binding free energy, despite the favorable interactions between ligand and residues seen in the model for the binding site. The decrease in binding affinity is less than a factor of 6 and does not correspond to a large free energy difference (~ 1 kcal/mol). The net free energy changes suggested by on our model are not highly precise, but are well within this range. Based on our calculations, the N-sulfates appear to reduce the net free energy of binding more significantly and their positioning in the site may contribute more to specificity for segments of HS with N-sulfated GlcNAcs separated by IdoA residues.

Based on the data from trNOE (Table 2.2) as well as the final docked structure, the overall conformation, as defined by inter-residue glycosidic torsion angles (Table 2.5), of the bound tetramer is not significantly different from the conformations dominating the free state. All of the IdoA residues also prefer a $^{1}C_{4}$ chair conformation. Starting structures with ligand residues in both the chair $^{4}C_{1}$ and the skew-boat $^{2}S_{0}$ conformation were tested. Neither gave clusters with competitive scores or energies. The $^{1}C_{4}$ chair conformation is also known to be more energetically favorable in solution especially when it is a non-reducing terminal [36]. Most of the previous studies on ligand interactions of Robo1 use depolymerized heparin rather than

HS as ligands. Homogeneous HS oligomers with specific sulfation patterns make it possible to directly demonstrate the importance of each sulfate group, as illustrated here for the 2-sulfated group on the internal IdoA of our tetramers.

Table 2.5. Glycosidic torsion angles of ligand 1 extracted from docked structure and crystal structure created by GLYCAM. The Φ and Ψ angles in the α(1,4) linkages are defined as O5-C1-O1-C4 and C1-O1-C4-C3 respectively.

Sugar ring	Φ angle (docked)	Φ angle (free)	Ψ angle (docked)	Ψ angle (free)
GlcNAc	94.8°	72.3°	89.0°	61.9°
IdoA B	-86.8°	-63.7°	103.6°	107.9°
GlcNAc	79.6°	72.7°	119.6°	96.5°
IdoA D	-43.4°	-65.0°	127.0°	112.8°

Comparison to crystal structures: To date, there have been several structural characterizations of Robo1-ligand interactions using different methodologies [7, 9-10, 31]. It is of particular interest to compare our results on docking of a heparin sulfate ligand to the crystal structure of drosophila Robo1-D1,D2 in which a heparin tetramer has been modeled. The protein sequence is 53% identical, and the HS tetramers are identical except that the heparin fragment has both IdoA residues 2-sulfated. In the crystal structure, the heparin fragment is sandwiched between two Robo1-Ig1-2 monomers. The residue corresponding to K81 in our structure in both monomers is involved with binding. On one side this contacts the 2-sulfate of the internal IdoA, as in our model, but then the tetramer turns away from this monomer to make contacts between the corresponding lysine of the other monomer and the 6-sulfate of the penultimate GlcNAc and the 2-sulfate of the non-reducing terminal IdoA. In our model the tetramer continues along the surface of the D1 domain making contact with K137 and R136.

There is no evidence for dimer formation under the conditions of our experiment. An average correlation time of 13 ns measured from cross-correlation experiments of Robo1 alone and the Robo1-Ig1-2-HS tetramer complex reveals that the protein remains monomeric before and after interacting with HS tetramer in solution (see Figure 2.12). Moreover, there is a single glycosylation site near the C-terminus of domain 1 that may well influence dimerization and inter-domain geometry. Most crystal structures have employed material lacking this glycosylation.

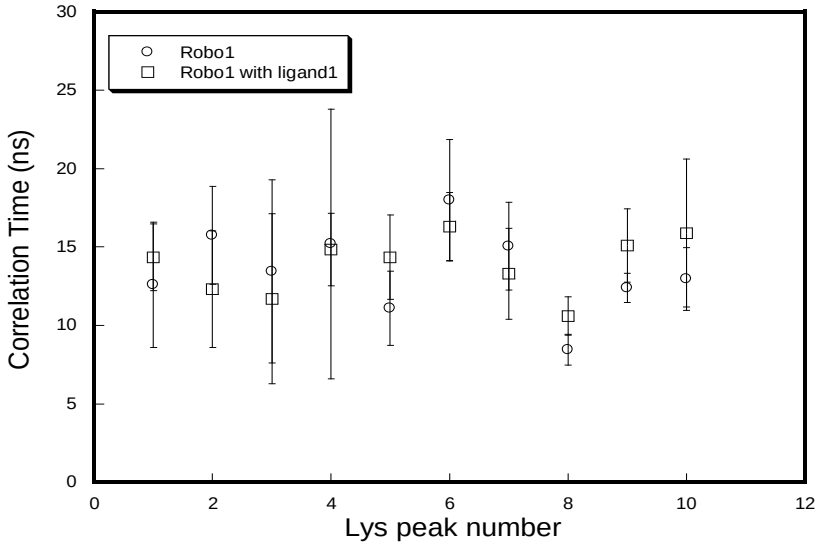


Figure 2.12. Rotational correlation time τc of Lys residues in Robo1 in the presence and absence of ligand 1.

Crystal structures of Robo1-Ig1-2 show different positions of domain D2 relative to D1. For example, structures in the 2v9q structure differ from that in 2v9r by a bend between D1 and D2 domains of 35 degree. From our MD simulation, the protein tends to adopt the more bent

conformation after the first 400 ns of stabilization. The average conformation between 400 ns to 1000 ns also shows a curved structure which indicates that this conformation tends to be more energetically favorable in the solution state. This curved structure of Robo1-Ig1-2 is able to make more efficient contact between the interacting amino acid residues and each glycan ring compared with the straight form, particularly I167 and R169.

One further point of note is that the HS octamer containing the IdoA2S in the middle (GlcA-GlcNS6S-IdoA-GlcNS-IdoA2S-GlcNS6S-IdoA-GlcNAc6S-$(CH_2)_5$NH2) binds to Robo1-Ig1-2 with an affinity higher than the corresponding tetramer 1 and a very similar perturbation of K81 (0.06 ppm when extrapolated to its limiting value (see Figure 2.13). This suggests that the ligand binds the central four residues in a very similar manner to the tetramer with the extra residues extending in both directions and finding additional favorable interactions. Examination of the model finds additional positively charged residues in both directions, R173 and R195 for the non-reducing end and R116, R119 and R131 for the reducing end.

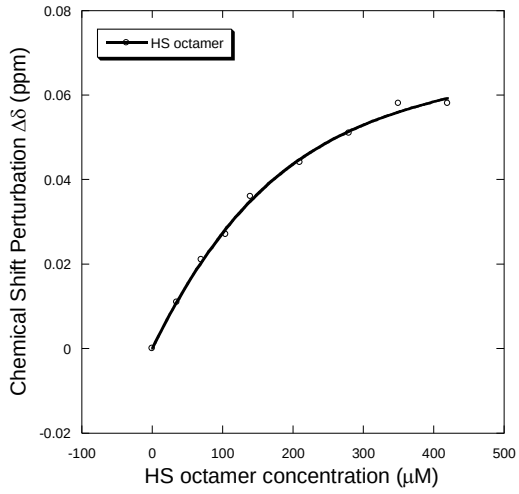


Figure 2.13. Binding affinity of Robo1 with HS octamer with a kd of 86 ± 11 μM.

Robo1 signaling is initiated by interactions with Slit2 [4], and this interaction is known to be facilitated by interactions with HS [37]. There is no structure showing the trimeric Robo1, Slit2, HS complex, but there is a crystal structure of a Robo1-D1-Slit2 D2 complex. Having a model for HS interacting with Robo1-D1-2, we can build a model for the trimeric complex by superimposing Robo1-D1 in our model with Robo1-D1 in the crystal structure. This model is presented in Figure 2.14. The HS tetramer (in beige) sits well in the grove between the two proteins and some of the positively charged residues highlighted in red located on the suggested extended binding surface of Robo1-Ig1-2, and the surface of the docked Slit2-D2, show potential interaction sites by which by which a longer HS segment could stabilize the trimeric complex. Production of a suitable Slit2 construct to explore this possibility is under consideration.

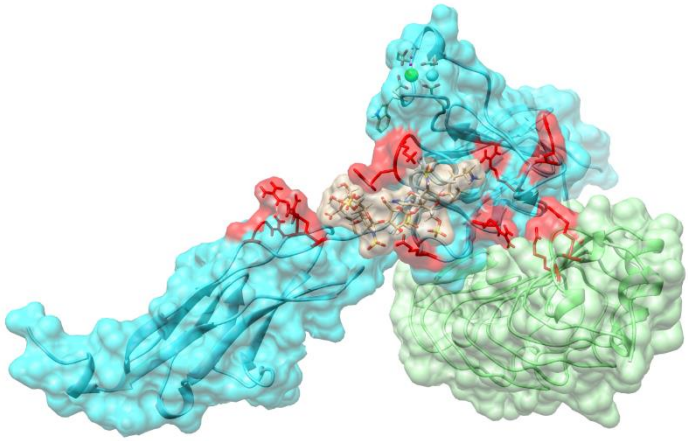


Figure 2.14. Model of trimeric Robo1-Ig1-2 (blue) –HS (beige) –Slit (green) with positive residues labeled in red.

2.7 Conclusion

A detailed model for the interaction of a HS tetramer with a two domain fragment of Robo1 has been determined. The HS tetramer has a specific sulfation pattern that includes 2-sulfation of an internal iduronic acid residue. A second tetramer lacking this sulfation is shown

to occupy a very similar site, but have enhanced binding affinity. A structural rationalization for this difference is suggested. The model also leads to a plausible explanation for how HS facilitates the interaction between Robo1 and its signaling partner, Slit2, providing a guide for further studies using longer HS oligomers and complexes involving both Robo1 and Slit2. The methods used in the current study also set a precedent useful in studies of other complexes of glycosylated proteins. The methods exploit a number of NMR experiments that can be applied to glycosylated proteins sparsely labeled with NMR active isotopes. These methods should be applicable to the large number of other systems found on the surfaces of mammalian cells.

2.8 References

1. Kastenhuber, E.; Kern, U.; Bonkowsky, J. L.; Chien, C. B.; Driever, W.; Schweitzer, J., Netrin-DCC, Robo-Slit, and Heparan Sulfate Proteoglycans Coordinate Lateral Positioning of Longitudinal Dopaminergic Diencephalospinal Axons. *J Neurosci* **2009,** *29* (28), 8914-8926.

2. Andrews, W.; Liapi, A.; Plachez, C.; Camurri, L.; Zhang, J. Y.; Mori, S.; Murakami, F.; Parnavelas, J. G.; Sundaresan, V.; Richards, L. J., Robo1 regulates the development of major axon tracts and interneuron migration in the forebrain. *Development* **2006,** *133* (11), 2243-2252.

3. Domyan, E. T.; Branchfield, K.; Gibson, D. A.; Naiche, L. A.; Lewandoski, M.; Tessier-Lavigne, M.; Ma, L.; Sun, X., Roundabout Receptors Are Critical for Foregut Separation from the Body Wall. *Developmental cell* **2013,** *24* (1), 52-63.

4. Dickinson, R. E.; Duncan, W. C., The SLIT-ROBO pathway: a regulator of cell function with implications for the reproductive system. *Reproduction* **2010,** *139* (4), 697-704.

5. Gara, R. K.; Kumari, S.; Ganju, A.; Yallapu, M. M.; Jaggi, M.; Chauhan, S. C., Slit/Robo pathway: a promising therapeutic target for cancer. *Drug discovery today* **2015,** *20* (1), 156-164.

6. Hussain, S. A.; Piper, M.; Fukuhara, N.; Strochlic, L.; Cho, G.; Howitt, J. A.; Ahmed, Y.; Powell, A. K.; Turnbull, J. E.; Holt, C. E.; Hohenester, E., A molecular mechanism for the heparan sulfate dependence of Slit-Robo signaling. *Journal of Biological Chemistry* **2006,** *281* (51), 39693-39698.

7. Zhang, F. M.; Moniz, H. A.; Walcott, B.; Moremen, K. W.; Linhardt, R. J.; Wang, L. C., Characterization of the interaction between Robo1 and heparin and other glycosaminoglycans. *Biochimie* **2013,** *95* (12), 2345-2353.

8. Dickson, B. J.; Gilestro, G. F., Regulation of commissural axon pathfinding by slit and its Robo receptors. *Annual review of cell and developmental biology* **2006,** *22*, 651-75.

9. Fukuhara, N.; Howitt, J. A.; Hussain, S. A.; Hohenester, E., Structural and functional analysis of slit and heparin binding to immunoglobulin-like domains 1 and 2 of Drosophila Robo. *The Journal of biological chemistry* **2008,** *283* (23), 16226-34.

10. Morlot, C.; Thielens, N. M.; Ravelli, R. B. G.; Hemrika, W.; Romijn, R. A.; Gros, P.; Cusack, S.; McCarthy, A. A., Structural insights into the Slit-Robo complex. *P Natl Acad Sci USA* **2007,** *104* (38), 14923-14928.

11. Zong, C. L.; Venot, A.; Dhamale, O.; Boons, G. J., Fluorous Supported Modular Synthesis of Heparan Sulfate Oligosaccharides. *Org Lett* **2013,** *15* (2), 342-345.

12. Chappell, E. P.; Liu, J., Use of biosynthetic enzymes in heparin and heparan sulfate synthesis. *Bioorgan Med Chem* **2013,** *21* (16), 4786-4792.

13. Barb, A. W.; Meng, L.; Gao, Z. W.; Johnson, R. W.; Moremen, K. W.; Prestegard, J. H., NMR Characterization of Immunoglobulin G Fc Glycan Motion on Enzymatic Sialylation. *Biochemistry-Us* **2012,** *51* (22), 4618-4626.

14. Prestegard, J. H.; Agard, D. A.; Moremen, K. W.; Lavery, L. A.; Morris, L. C.; Pederson, K., Sparse labeling of proteins: Structural characterization from long range constraints. *J Magn Reson* **2014,** *241*, 32-40.

15. Barb, A. W.; Ho, T. G.; Flanagan-Steet, H.; Prestegard, J. H., Lanthanide binding and IgG affinity construct: Potential applications in solution NMR, MRI, and luminescence microscopy. *Protein Sci* **2012,** *21* (10), 1456-1466.

16. Zhuang, T. D.; Lee, H. S.; Imperiali, B.; Prestegard, J. H., Structure determination of a Galectin-3-carbohydrate complex using paramagnetism-based NMR constraints. *Protein Sci* **2008,** *17* (7), 1220-1231.

17. Barthelmes, K.; Reynolds, A. M.; Peisach, E.; Jonker, H. R. A.; DeNunzio, N. J.; Allen, K. N.; Imperiali, B.; Schwalbe, H., Engineering Encodable Lanthanide-Binding Tags into Loop Regions of Proteins. *J Am Chem Soc* **2011,** *133* (4), 808-819.

18. Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J., HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **2003,** *125* (7), 1731-1737.

19. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts Chem Res* **2000,** *33* (12), 889-897.

20. Gandhi, N. S.; Mancera, R. L., Free energy calculations of glycosaminoglycan-protein interactions. *Glycobiology* **2009,** *19* (10), 1103-1115.

21. Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E., MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *J Chem Theory Comput* **2012,** *8* (9), 3314-3321.

22. Tjandra, N.; Grzesiek, S.; Bax, A., Magnetic field dependence of nitrogen-proton J splittings in N-15-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J Am Chem Soc* **1996,** *118* (26), 6264-6272.

23. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *J Biomol Nmr* **1995,** *6* (3), 277-293.

24. Goddard, T.; Kneller, D., SPARKY 3. *University of California, San Francisco* **2004,** *15*.

25. Williamson, M. P., Using chemical shift perturbation to characterise ligand binding (vol 73, pg 1, 2013). *Prog Nucl Mag Res Sp* **2014,** *80*, 64-64.

26. Li, D. W.; Bruschweiler, R., PPM_One: a static protein structure based chemical shift predictor. *J Biomol Nmr* **2015,** *62* (3), 403-409.

27. Valafar, H.; Prestegard, J. H., REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **2004,** *167* (2), 228-241.

28. Sumathi, S.; Surekha, P., MATLAB-Based Genetic Algorithm. *Computational Intelligence Paradigms: Theory and Applications Using Matlab* **2010**, 547-589.

29. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-1612.

30. Woods, R., glycam Web. *Complex Carbohydrate Research Center. Athens, GA: University of Georgia* **2005**.

31. Li, Z. X.; Moniz, H.; Wang, S.; Ramiah, A.; Zhang, F. M.; Moremen, K. W.; Linhardt, R. J.; Sharp, J. S., High Structural Resolution Hydroxyl Radical Protein Footprinting Reveals an Extended Robo1-Heparin Binding Interface. *Journal of Biological Chemistry* **2015,** *290* (17), 10729-10740.

32. Prestegard, J. H.; Sahu, S. C.; Nkari, W. K.; Morris, L. C.; Live, D.; Gruta, C., Chemical shift prediction for denatured proteins. *J Biomol Nmr* **2013,** *55* (2), 201-9.

33. Pederson, K.; Mitchell, D. A.; Prestegard, J. H., Structural Characterization of the DC-SIGN-Lewis(X) Complex. *Biochemistry-Us* **2014,** *53* (35), 5700-5709.

34. Otting, G., Protein NMR Using Paramagnetic Ions. *Annual Review of Biophysics, Vol 39* **2010,** *39*, 387-405.

35. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: A generalizable Biomolecular force field. Carbohydrates. *J Comput Chem* **2008,** *29* (4), 622-655.

36. Ferro, D. R.; Provasoli, A.; Ragazzi, M.; Casu, B.; Torri, G.; Bossennec, V.; Perly, B.; Sinay, P.; Petitou, M.; Choay, J., Conformer Populations of L-Iduronic Acid Residues in Glycosaminoglycan Sequences. *Carbohyd Res* **1990,** *195* (2), 157-167.

37. Johnson, K. G.; Ghose, A.; Epstein, E.; Lincecum, J.; O'Connor, M. B.; Van Vactor, D., Axonal heparan sulfate proteoglycans regulate the distribution and efficiency of the repellent slit during midline axon guidance. *Current biology : CB* **2004,** *14* (6), 499-504.

CHAPTER 3

NMR ASSIGNMENTS OF SPARSELY LABELED PROTEINS USING A GENETIC ALGORITHM[2]

[2] Submitted to the *Journal of Biomolecular NMR*,11/30/2016

3.1 Acknowledgement

The following project involved a collaboration with Gordon R. Chalmers from the department of Computer Science at UGA. I was responsible for collecting all the experimental data, searching and preparing test cases and designing a matlab script for making NOE vectors. The software was coded and tested by Gordon Chalmers. We contributed equally to the project. We would like to thank Kari Pederson and Alex Eletsky for their constructive comments during the course of this project.

3.2 Abstract

Sparse isotopic labeling of proteins for NMR studies using single types of amino acid (^{15}N or ^{13}C enriched) has several advantages. Resolution is enhanced by reducing numbers of resonances for large proteins, and isotopic labeling becomes economically feasible for glycoproteins that must be expressed in mammalian cells. However, without access to the traditional triple resonance strategies that require uniform isotopic labeling, NMR assignment of crosspeaks in heteronuclear single quantum coherence (HSQC) spectra is challenging. We present an alternative strategy which combines readily accessible NMR data with known protein domain structures. Based on the structures, chemical shifts are predicted, NOE cross-peak lists are generated, and residual dipolar couplings (RDCs) are calculated for each labeled site. Simulated data are then compared to measured values for a trial set of assignments and scored. A genetic algorithm uses the scores to search for an optimal pairing of HSQC crosspeaks with labeled sites. While none of the individual data types can give a definitive assignment for a particular site, their combination can in most cases, and a completely correct assignment is typically found near the top of an ordered list of possibilities. Four test proteins previously assigned using triple resonance methods and a sparsely labeled glycosylated protein, Robo1,

previously assigned by manual analysis, are used to illustrate the method and develop a criterion for identifying sites assigned with high confidence.

3.3 Introduction

In the structural biology field, NMR is most noted for its ability to produce *de novo* structures of small proteins in solution. Its contributions are significant in that the structures are produced in solution and often involve proteins that have resisted crystallization. Nevertheless, interest of the structural biology community is shifting to larger proteins and multi-protein complexes. Some of the proteins of interest are also glycoproteins. Here, applications are limited, because most current resonance assignment methodology depends on a set of triple resonance experiments that require uniform enrichment of proteins in ^{15}N and ^{13}C, and for larger proteins, perdeuteration is required as well. For glycoproteins native glycosylation is often important, and expression in mammalian cells is preferred. The need to supplement media for these cells with labeled amino acids makes uniform labeling extraordinarily expensive and the cells do not tolerate high deuterium content. However, sparse labeling with single or small subsets of isotopically labeled amino acids is still possible. Certain subsets of isotopically labeled amino acids are relatively inexpensive, making application to glycoproteins economical, and the reduction in numbers of resonances increases resolution for larger proteins [1-3]. Even with fewer labeled sites chemical shift perturbation can provide information on ligand binding or protein-protein association [4], and residual dipolar couplings (RDCs) can constrain relative orientation of structural units in multi-domain proteins or protein-protein complexes [5-6]. The only prerequisite is the replacement of triple resonance methods with an alternative assignment strategy. Here we present a strategy based on collection of data readily obtained on sparsely

labeled sites and describe a convenient software implementation that uses a genetic algorithm to search for an optimal set of assignments.

The types of data that can be acquired on sparsely labeled proteins include heteronuclear single quantum coherence (HSQC) spectra from which chemical shifts of both protons and heteronuclei (^{15}N or^{13}C) can be measured. HSQC experiments also provide the basis for collection of ^{15}N- or ^{13}C-edited nuclear Overhauser effects (NOEs) [7] and one-bond residual dipolar couplings (RDCs) [5]. When structures are available for at least the domains comprising multi-domain proteins, or the proteins comprising multi-protein complexes, it is possible to predict each of these data types in a site specific manner. There are now several different chemical shift prediction programs [8-11] that are based on the existence of three dimensional structures for protein domains. The NOE vectors associated with a given crosspeak in the ν2-ν3 plane of a 3D NOESY-HSQC can also be predicted from a three dimensional structure, assuming a $1/r^6$ dependence of NOE intensity on interproton distances, and RDCs can be predicted using programs such as REDCAT [12] or PALES [13], provided a sufficient number of RDCs are available to simultaneously evaluate the level and direction of partial orientation.

It is clear that comparison of any of the above data types with predictions can be used to facilitate assignments. The automated assignment of some NOE peaks in the course of structure determination is quite common [14-16], and there are several examples of the use of NOEs for structural characterization without explicit NOE crosspeak assignments [17-18]. There is also increasing use of both NOE and chemical shift data to validate or extend assignments made by traditional triple resonance methods [19-21]. However, most of these procedures rely to some extent on uniform isotope labeling. For sparsely labeled systems there has been some prior effort at resonance assignment. We have attempted to use amide exchange rate correlations in NMR and

MS data to achieve assignments of ^{15}N labeled sites [22-23]. There have also been approaches that label proteins with combinations of amino acids to make assignments based on connectivities similar to those seen in triple resonance experiments [24-26], and of course, assignment in sparsely labeled proteins can be facilitated by mutating residues to remove crosspeaks from labeled sites one at a time [27]. However, these approaches are labor intensive. ^{13}C-methy labeling of large proteins can also be seen as a type of sparse labeling, and this has fostered the development of assignment strategies that depend on data that can be collected via these sites, primarily methyl-methyl NOEs [28] and paramagnetic effects [29]. It has also stimulated exploration of probabilistic approaches to assigning using these and other types of data [30-31].

Recently, we utilized the data types described above, namely chemical shifts, RDCs and NOEs, to assign a glycoprotein sparsely labeled with ^{15}N enriched amino acids. Resonance assignment used a largely manual approach in which each type of prediction is sequentially used to exclude possible assignments [32]. However, it is difficult to set strict exclusion limits in a sequential strategy. In principle, it is better to assign a score based on agreement between measurement and prediction of all data types and use these scores in a search over all possible pairings of labeled sites with measured crosspeaks. A systematic search over all possibilities would be feasible with modern computers and a small number of sites, but this quickly becomes unmanageable. Even at 12 sparsely labeled sites, the number of possible assignments is enormous; 12! = 470,001,600. Instead we use a procedure based on evolutionary algorithms, more specifically a genetic algorithm [33]. Genetic algorithms have been used previously to facilitate NMR assignments [34-35], but not for the assignment of sparsely labeled proteins. We test our implementation of a genetic algorithm search for an optimal assignment on a set of proteins for which the required data and assignments have been deposited. Four of the test proteins come

from a recent summary of X-ray/NMR structure pairs produced by the Northeast Structural Genomics group (NESG) [36]. The existing assignments in this case are by traditional triple resonance methods on uniformly labeled proteins, but we mimic a sparsely labeled set by selecting resonances for a subset of amino acids. We also test the procedure on a sparsely labeled two domain fragment of the Robo1 protein whose assignment is achieved by a sequential approach [32]. Robo1 is a glycosylated protein whose activity in developmentally related cell-signaling is regulated by interaction with certain heparan sulfate (HS) epitopes. Assignment was critical to locating the HS binding site and modeling an HS-Robo1 complex.

The new procedure proves to be quite successful. It is relatively fast, requiring from several minutes to a few hours of computational time. In three of the five cases the completely correct assignment is found among the top five scoring solutions (chromosomes), and the correct assignment is always found in the set of solutions having scores less than a score corresponding to measurements falling within one or two standard deviations of predictions. Many crosspeaks are also assigned consistently to correct sites within most of this solution set. Based on this consistency we suggest a criterion for identification of crosspeaks which can be assigned with high confidence in the absence of known assignments. The procedure, therefore, provides a robust means of assigning NMR spectra and sets the stage for answering many more questions involving ligand binding and protein-ligand complex assembly for some of the more challenging structural biology systems.

3.4 Materials and Methods

3.4.1 Test set selection

Four test proteins were chosen from the 40 pairs of NMR-X-ray structures produced by the Northeast Structural Genomics group, imposing the additional requirement that the resolution

of the X-ray structures be below 2 Å [36]. The NMR data for Robo1-Ig1-2 are from our paper reporting its interaction with heparan sulfate [32]. There are several X-ray structures for the Ig1-2 construct, but these show significant differences in inter-domain orientation. For the purpose of this application domain motions were simulated in a long MD trajectory (1μs) [32] and an x-ray structure (PDB 2V9R) that closely approximated the domain orientation in the most highly populated state of this trajectory was selected. The PDB accession codes for the structures and a summary of experimental data used for all test proteins are summarized in Table 3.1.

Table 3.1. Structure and experimental information on chosen test proteins and a glycoprotein.

PDB		Total Residues	Labeled Sites	Available RDC Data	Inter-residue NOE crosspeaks per residue
X-ray (resolution Å)	NMR				
3C4S (1.70)	2JZ2	58*	Ala 4, Val 8	12	4.8
3CWI (1.90)	2K5P	70	Ala 7, Val 9	11	7.9
3LMO (2.00)	2KW2	93	Ala 12, Lys 6	17	6.4
3FIA (1.45)	2KHN	111	Ala 8, Lys 6	8	8.2
2V9R (2.00)	NA	212	Lys 12, Phe 5	17	7.4

*3C4S is a dimer. Monomer A with 58 residues was used.

3.4.2 Experimental and predicted data

^{15}N-^{1}H HSQC chemical shifts. Because RDC analysis requires at least 5 measurements to fit order parameters before data can be used to assess quality of alignment or crosspeak assignment, we strove to have at least 10 measurements. Based on the expression construct for Robo1-Ig1-2, lysine and phenylalanine would yield 17 measurements, but two lysine crosspeaks were not observed. One N-terminal site is missing due to proteolysis, as confirmed by mass spectral analysis. The other unobserved site was identified as K137 by a selective mutation that produced no change in the number of HSQC crosspeaks. The two sites were thus eliminated

from assignment consideration. Data for one phenylalanine was also not used because it exhibited high levels of internal motion, consistent with its being very near the end of the non-structured C-terminus (absent in some crystal structures). Hence, 14 of the 17 sites were subject to assignment by our methodology. To mimic a similar level of sparse labeling in the 4 uniformly labeled test proteins, we selected data from alanine and valine or alanine and lysine. The specific numbers of sites are listed in Table 3.1. Experimental errors for chemical shifts are all small compared to prediction errors and will, therefore, be neglected.

There are several program options for the prediction of chemical shift data [8-10], and they have very similar estimated precision for amide N and H shifts. Here, we chose PPM_one [8] to perform chemical shift prediction. Both backbone amide nitrogen and proton chemical shifts of the labeled sites were predicted using the crystal structures listed in Table 3.1. Errors for predicted ^{1}H and ^{15}N chemical shifts were taken to be 0.17 ppm and 1 ppm respectively, numbers consistent with limits suggested by the authors of prediction programs [8-9].

Nuclear Overhauser Effects (NOEs). Experimental NOE data to be used for ^{1}H-^{15}N crosspeak assignment are most useful when both chemical shifts and intensities of NOE crosspeaks are available. For most deposited NOE data only peak lists containing chemical shifts and constraint files containing upper and lower distance limits for proton pairs are available. However, there is no reliable way to work back from distance limits to a crosspeak intensity. Therefore, we chose cases where the original ^{15}N-filtered NOE peak lists included the intensity of each crosspeak. Because the crosspeaks from intra-residue contacts are less useful in making assignments, we removed these by considering crosspeaks from a ^{15}N-edited TOCSY spectrum. NOE crosspeaks at corresponding chemical shifts were eliminated from the list. Experimental amide proton NOE vectors containing only inter-residue data were then

constructed by spreading intensity over a range of chemical shift equal to the estimated accuracy of predicted shifts. Since the vectors mimic columns emanating from H-N crosspeaks in 3D NOESY-HSQC data sets, it is also possible to take vectors directly from the columns in the 3D data sets when these are available.

For predicted NOE data the intensities for peaks in amide proton NOE vectors were predicted using a $\frac{1}{r^6}$ dependence on interproton distances derived from crystal structures. This would be correct for a rigid spherical protein with NOEs measured from initial slopes. We cannot assume the selected data meet these conditions, but since we are not seeking perfect scores, just best scores, we believe the treatment is adequate. The intensities were placed in predicted amide proton inter-residue NOE vectors at the chemical shift positions predicted by PPM_1 and spread over a range equal to the estimated chemical shift accuracy.

Residual Dipolar Couplings (RDCs). ^{15}N-^{1}H Residual dipolar couplings (RDCs) reflect the orientation of each H-N bond vector relative to the magnetic field; hence, they are very dependent on the structure of each site. Experimental RDCs (8-17 in number) were obtained along with their estimated errors (typically 1Hz for the 4 uniformly labeled test proteins, and 2-5 Hz for the Robo1 set). Unlike the previous data types, predicted values could not be obtained independently for each site, because the data must be used to determine the five order parameters in addition to the RDCs. Therefore, for each possible set of assignments, all RDCs were used simultaneously, in combination with coordinates for nitrogen and proton pairs at potential sites, to calculate order parameters by singular value decomposition. A set of predicted RDCs were then back-calculated using the parameters and compared to measurements.

Missing and substandard data. It is not always possible to have complete data sets or data sets with uniform high quality. Sometimes crosspeaks are not observed in HSQC spectra, because of motional contributions to line widths, interference from solvent and other contaminants, or accidental overlap of a pair of crosspeaks. Sometimes there is supplementary information that makes the interpretation of certain measurements suspect, and one would choose to disregard that data. For example, spin relaxation data may suggest a high level of internal motion that would compromise the interpretation of RDCs. As the number of measurements of a given type must always equal the number of sites, the software package described below handles missing measurements by entering 999 as a default measurement. The package also handles cases where the number of measurements exceeds the number of sites by entering 999 for null sites.

3.4.3 Program development

Our implementation of a genetic algorithm search for correct assignments of HSQC crosspeaks is based on routines available in MATLAB [37-38]. The complete package, "ASSIGNments for Sparsely Labeled Proteins", ASSIGN_SLP, can be downloaded from the internet site http://tesla.ccrc.uga.edu/software/. The program executing the genetic algorithm search is designated ASSIGN_SLP.mat.

Implementation of the genetic algorithm. An initial population of chromosomes (assignments) is constructed randomly, usually a number approaching 10,000; this value may seem large for a genetic algorithm, however the size of the search space and the complexity of the objective function require it. With respect to the number of possible assignments, this is still a relatively small sampling (about 0.3 % for 10 sites). For each chromosome, sites remain in

fixed order and crosspeaks (with their associated data) are randomly assigned to each site. The only restriction is that each crosspeak is used only once in an assignment.

Selection and mutation of chromosomes. Chromosomes are selected from the entire pool with a frequency biased toward high scores, which means farther from the best individual score, and these are subjected to modification. There are two general processes used, a permutation mutation and a pairwise crossover process, as illustrated in the flow chart. The frequencies with which the processes are used is governed by rate constants which are normally tuned for a particular application. As we intend our program to be applicable to a range of different data types we attempted to eliminate the normal tuning process by looping through a combination of rates (0.2, 0.4, 0.6 and 0.8 for both mutation and crossover). An initial set of rates is selected, a new set of chromosomes is generated, the new set is re-scored and it is subjected to another cycle. This cycle is continued for a given crossover and mutation rate until a convergence criterion is reached. In our case, the convergence criterion is that the lowest score does not change after 100 iterations or that a set maximum number of 500 iterations is reached. The program then selects another set of mutation and crossover rates, generates another set of random assignments and starts the process over. When all pairs of rates have been used the program ends.

Output of results. There are two post-processing programs which are used. The first is designated "OutputAnalysis"; it is used to convert the cell array MATLAB file to a text file, with all duplicates removed, and sorted from lowest to highest score. The text file has a header with all the information necessary to reproduce the search. Examination of the ordered list may suggest adjustment of weighting factors for different data types or extraction of a few top scoring assignments for validation against unused data. The next program, "StatisticalAnalysis", is used

to generate a heatmap and histogram from the text file output of the previous program. The output of "StatisticalAnalysis" is two MATLAB figures, a heatmap and a histogram, which are saved in a user specified location.

The software package contains documentation and examples. The documentation explains how the programs are to be used. There are 5 examples in sub-directories which contain all the input files and the output files needed to reproduce the examples. In order to make the software user-friendly, a preparation file is given with the commands used to generate the output of all the examples.

The work flow through the primary program is depicted in Figure 3.1. Input includes: the pdb file, a list of sites to be assigned (a-h in the flow chart), output from chemical shift programs, output from the NOE vector script and other experimental data with estimates of errors. Coordinates are extracted by the provided script from the pdb file for the calculation of predicted RDCs and NOEs, and predicted chemical shifts for the relevant sites are extracted from output files of PPM_one or SHIFTX2.

One of the most important aspects of the search program is the objective function, in our case a sum of scores for different data types. It must minimize at a global optimal solution, weight each data type appropriately and provide a useful limit on what we regard as an acceptable solution. For most data types (RDCs and chemical shifts) our scores are based on a root mean square deviation (RMSD) between the predicted value for each site (pred(i)) and the experimental value being assigned to that site, exp(q(i)). See Equation 1.

$$\frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(\exp(q(i))-pred(i)))^2 \Big/ (error(q(i)))^2} \qquad (1)$$

The error estimate, error(q(i)), is used to scale each score contribution by a number that reflects the information content of the measurement type. For example, in the case of chemical shifts, the error is from the estimated precision of predictions supplied by the authors of the prediction programs. If the range of measurements (largest deviation of prediction and measurement) is divided by the error, we have an estimate of the information content. As measurements assigned to particular sites begin to approximate predicted values, their contribution to the score is reduced. When the differences between experiment and prediction are at the estimated error, the score for each measurements type would equal one, and any total score less than one or two times the number of measurements types should be considered acceptable. Hence, the scaling provides both a weighting by information content and a convenient cut off for acceptable solutions.

In calculating an RMSD the number of comparisons contributing to the mean for a given measurement type (n) is normally the total number of independent measurements. For chemical shifts this is just the number of measurements, but for RDCs the number of independent measurements is less than the number of experimental data by five because of the order parameters that must be determined. In the case of missing data there should also be an additional subtraction for the number of data points entered as 999. This would appropriately make the errors seem somewhat larger. However, we must also consider information content. This is particularly important for RDCs. As the number of RDCs approaches 5, the five parameters will, in most cases, allow a perfect match of predicted to experimental data, and the information content for assignment will actually approach zero. Hence, we have added an additional scaling factor of (n-5)/n to the RDC part of the scoring function.

There are some limitations to the above RMSD-based procedure. First, it is hard to define an RMSD for some types of measurement, NOEs, for example. For NOEs we calculate a Pearson correlation coefficient that compares the predicted and experimental NOE vectors. This has the advantage of making the absolute intensity of the vectors irrelevant; only NOE patterns matter. The coefficient, R, is 1 for a perfect match between experimental and predicted NOE vectors and zero for no correlation. To mimic an RMSD that goes to zero for a perfect match we use the square of (1-R) averaged over all NOE vectors. To replace the estimate of error we use data from the four uniformly labeled test proteins, where we know correct assignments, to estimate NOE information content. To do this we calculate heat maps (see below). When predicted and experimental NOEs are listed in the correct order the scores on the diagonal represent correct pairings and the scores off the diagonal represent those for incorrect pairings. The ratios of the averages on and off the diagonal are about 5 for the test proteins and this is what we used for a weighting factor (equivalent to using an error estimate of about 0.2).

The above procedures, particularly those that simply use error estimates as opposed to a heat map analysis, are not perfect. Improper weighting relative to information content can still occur. This is often recognizable in an ordered list of acceptable solutions that includes contributions to total scores from individual data types. A contribution from one particular data type may consistently have an unusually small or large contribution, or its contributions may fall continuously through the list indicating that is dominating the algorithm. It is possible in these cases to further adjust weighting factors to eliminate this behavior.

Scores can also be adjusted to take into account additional knowledge about pairings. Since sparse labeling often includes data from samples labeled with different amino acids, we know that a specific set of crosspeaks must be associated with sites having a particular amino

acid type. Adding 100 to experimental and predicted chemical shifts for one amino acid type forces incorrect associations to have unacceptable scores. Other types of knowledge, for example, knowledge about surface versus interior protein location can be introduced in a similar fashion.

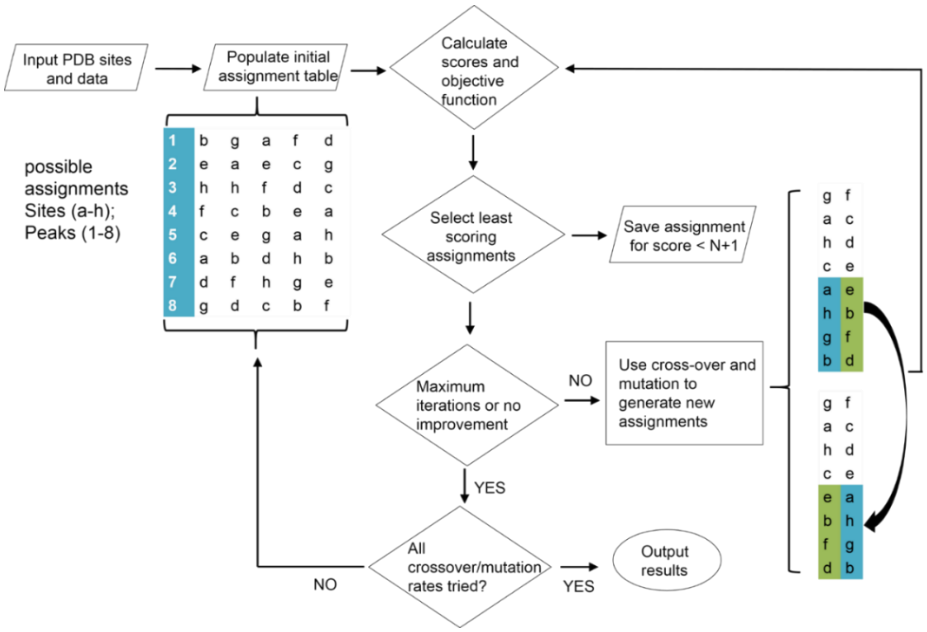


Figure 3.1. Work flow of the assignment strategy using a genetic algorithm. The sparsely labeled sites are represented by letters a-h and the HSQC crosspeaks are numbered 1-8. The region where certain mutations happened is labeled by blue and green.

The objective function is evaluated for each of the trial assignments (chromosomes). Chromosomes with a total score less than a user specified value is saved in an output file; this maximum score usually is set to two to three times the number of measurement types. Detailed information of chromosomes selection and mutation rates determination are described in the Supplemental Materials. Chromosomes are then selected from the entire pool with a frequency biased toward high scores, which means farther from the best individual score, and these are subjected to modification. There are two general processes used, a permutation mutation and a

pairwise crossover process, as illustrated in the flow chart. The frequencies with which the processes are used is governed by rate constants which are normally tuned for a particular application. As we intend our program to be applicable to a range of different data types we attempted to eliminate the normal tuning process by looping through a combination of rates (0.2, 0.4, 0.6 and 0.8 for both mutation and crossover). An initial set of rates is selected, a new set of chromosomes is generated, the new set is re-scored and it is subjected to another cycle. This cycle is continued for a given crossover and mutation rate until a convergence criterion is reached. In our case, the convergence criterion is that the lowest score does not change after 100 iterations or that a set maximum number of 500 iterations is reached. The program then selects another set of mutation and crossover rates, generates another set of random assignments and starts the process over. When all pairs of rates have been used the program ends.

The raw output is a MATLAB file in a "cell array," which is a Java data class used by MATLAB. The output contains possible solutions generated in each of the mutation/crossover cycles that have scores less than a user specification. In addition to detailing the match of crosspeaks to sites, contributions of each data type to the total score are given along with information that allows a direct comparison of predicted to experimental values.

3.5 Results

Assignments for our four uniformly labeled test proteins and one glycoprotein have been produced using the programs introduced above. The four uniformly labeled test proteins range in size from 55 to 212 amino acids. Different mixes of secondary structures are represented, including those rich in alpha-helix, rich in beta-sheet and a combination of both. There are instances of missing data and different levels of internal motion. For the two domain construct from Robo1, our example of a sparsely labeled glycoprotein, ^{15}N-^{1}H HSQC spectra for the lysine

and phenylalanine labeled versions are shown in chapter 2. These spectra give examples of the resolution that can be expected for a sparsely labeled, non-deuterated 23 kDa protein.

Working with the first four proteins, for which assignments are well documented by traditional methods, provides an opportunity to evaluate the degree to which each measurement type contributes to the assignment process. Their contributions can be visualized in the heatmaps similar to those generated by the auxiliary statistical analysis program. The examples shown in Figure 3.2 use the X-ray structure, 2K5P, for prediction and the deposited information for the NMR structure, 3CWI, for experimental data. Experimental assignments are listed on the y-axis and predicted assignments are listed on the x-axis, both ordered with respect to increasing residue numbers. Correct assignments fall on the diagonal. The contributions to the total score from each data type have been generated using an in-house MATLAB script (available at the ASSIGN_SLP download site). The values are represented on the plots in gray-scale, with black representing zero (best score) and white a normalized score of 1. The amino acids represented are 7 alanines and 9 valines. Since we do not allow cross-assignments between the amino acid types, white regions exist for coordinates1-7, 8-16 and 8-16, 1-7. The first 3 panels are heatmaps of the scores for individual data types; chemical shifts, NOEs and RDCs. From these heatmaps, it is clear that NOEs are the most informative since the darkest spots for most possible assignments fall on the diagonal. However, it is also obvious that there are cases with little distinction between pairs of possible assignments (scores for peaks 1, 6, and 13), and an incorrect assignment would be indicated for peaks 9 and 12. Data for RDCs and chemical shifts are typically less definitive, but still useful. Adding all the scores together produces a plot in which the diagonal box is darkest for all but one possible site. The heatmaps have already

been used to extract an error estimate for NOEs, but they could be used to evaluate the proper weighting for all data types. We will examine this possibility in the future.

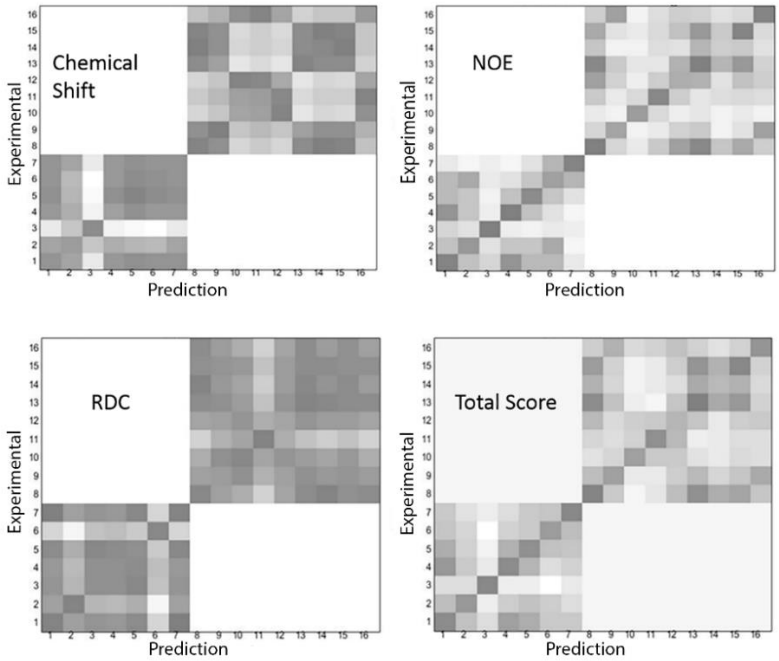


Figure 3.2. Heatmaps comparing predicted and experimental values of each type of measurement (chemical shift, NOEs and RDCs) and total score contribution. Each number on both X- and Y- axes represent one labeled residue. The amino acid type is assumed to be known for the two sets of crosspeaks.

An example of the output of our assignment program for the 3CWI-2K5P protein is shown in Table 3.2. The output contains not only all the possible assignments but also the solution rank and the score contributions from each type of measurement. There is a comparison of experimental and predicted RDC data for each site. For chemical shifts the experimental data are given for each site; the predicted data are given in the output header. For NOEs individual score contributions in terms of $(1\text{-}R)^2$ are given. In the example presented, the first rank solution is a single-swap of two residue assignments (peak 3 should assigned to 48 and peak 4 should

assigned to 32); peak 3 has no RDC data, making assignments for this pair somewhat ambiguous.

The results of application of our assignment program to all four uniformly labeled test cases are summarized in Table 3.3. In all cases ^{1}H and ^{15}N chemical shifts, NOE peak lists for HSQC crosspeaks, and a single set of RDCs were available. The top score solutions contain at least 70% of correct assignments (case 3FIA) and can reach 100% of correct assignments (case 3C4S). The correct solution is always found near the top of the list; the worst case is number 11 out of 7493 solutions for 3FIA which has 6 missing RDCs.

The application to Robo1-Ig1-2 deserves a separate discussion. Robo1-Ig1-2 is both larger than the other test proteins, (212 residues), it has the potential complication of internal motion between domains, and it is a glycoprotein where sparse labeling with individual amino acids is necessary. The top ranked assignment from an initial run (having a score of 4.09) contains 10 correct assignments and the completely correct solution was solution number 366. However, the Robo1 protein is a good example of using some intelligence in changing the weights of the contributions in the objective function to improve performance. The calculation using initial error estimates had high chemical shift contributions to the scores and several of the RDC's did not agree with the back-calculation of individual contributions. By increasing the errors of the chemical shift terms to lessen their importance in the objective function, the correct solution moved from a rank of 366 to a rank of 18 in a list of 354 acceptable solutions with scores less than 5.09. The top ranked solution still had 10 correct assignments.

For Robo1-Ig1-2 it is possible to see some of the reason for the four missed assignments in the top ranked solution. The RDC degeneracy makes it hard to distinguish peak 4 from 9, peak 6 from 7 and peak 13 from 14. Therefore, swaps between assignments for these pairs might

have been expected. 10 correct out of 14 is in fact not a bad result. We might also have expected the RDC data to be compromised in the Robo1-Ig1-2 case by the existence of inter-domain motion. This could have led to different alignment tensors for the two domains and completely incorrect RDC predictions when assuming a rigid structure and extracting a single set of alignment parameters. The fact that RDCs fit reasonably well may mean that motions are fairly restricted in the presence of the large attached glycan. The crystal structures showing large variations in inter-domain geometry were all produced on non-glycosylated material.

It may seem convenient to focus on top-ranked solutions, however, this is not particularly valuable for a protein for which there is not prior knowledge of the correct assignment. Identifying sites which are assigned with high confidence is actually more important than obtaining a complete assignment. For example, in applications to ligand binding by chemical shift perturbation, one only needs to know the assignment of the perturbed peak, and for domain orientation using RDC measurements, one only needs an adequate number of confident assignments to use RDCs.

Table 3.2. First rank solution for assignments of 3CWI-2K5P.

Solution Rank	1															
Peak number	2	5	1	3	4	7	6	12	15	9	14	11	8	13	16	10
Residue number	15	26	30	32	48	51	59	5	12	20	29	35	37	43	54	60
Exp.RDC	2.69	7.24	0.96	999	0.74	-9.87	3.33	-3.24	999	3.2	9.5	999	-0.17	999	-1.28	999
Calculated RDC	3.04	8.41	-0.14	0	0.42	-8.94	1.1	-1.92	0	3.5	9.6	0	0.15	0	0.11	0
Exp. shift (N)	121.2	126.1	131.1	120.2	119.9	119.5	121.1	226*	225*	217.9*	219.2*	221.2*	227.6*	226.8*	226.2*	221.8*
Exp. shift (H)	7.36	8.71	8.54	8.01	8.03	7.54	8.33	108.56 *	108.56 *	107.81 *	107.58 *	107.33 *	109.62 *	108.87 *	109.22 *	109.06 *
NOE score	0.01	0.2	0	0.11	0.23	0.1	0.01	0.05	0.05	0.15	0.03	0.2	0	0.43	0	0
Data type	RDC	N	H	NOE	Sum/Score											
Total score	1.48	0.99	1.19	1.52	5.19											

*100 is automatically added to the chemical shift for the second type of amino acid so that different types of amino acid will not be cross-assigned. 999 is used to indicate data that are not available. The incorrect assignment is colored in gray; 3 and 4 should be interchanged.

Table 3.3 Assignment summary of four test protein cases.

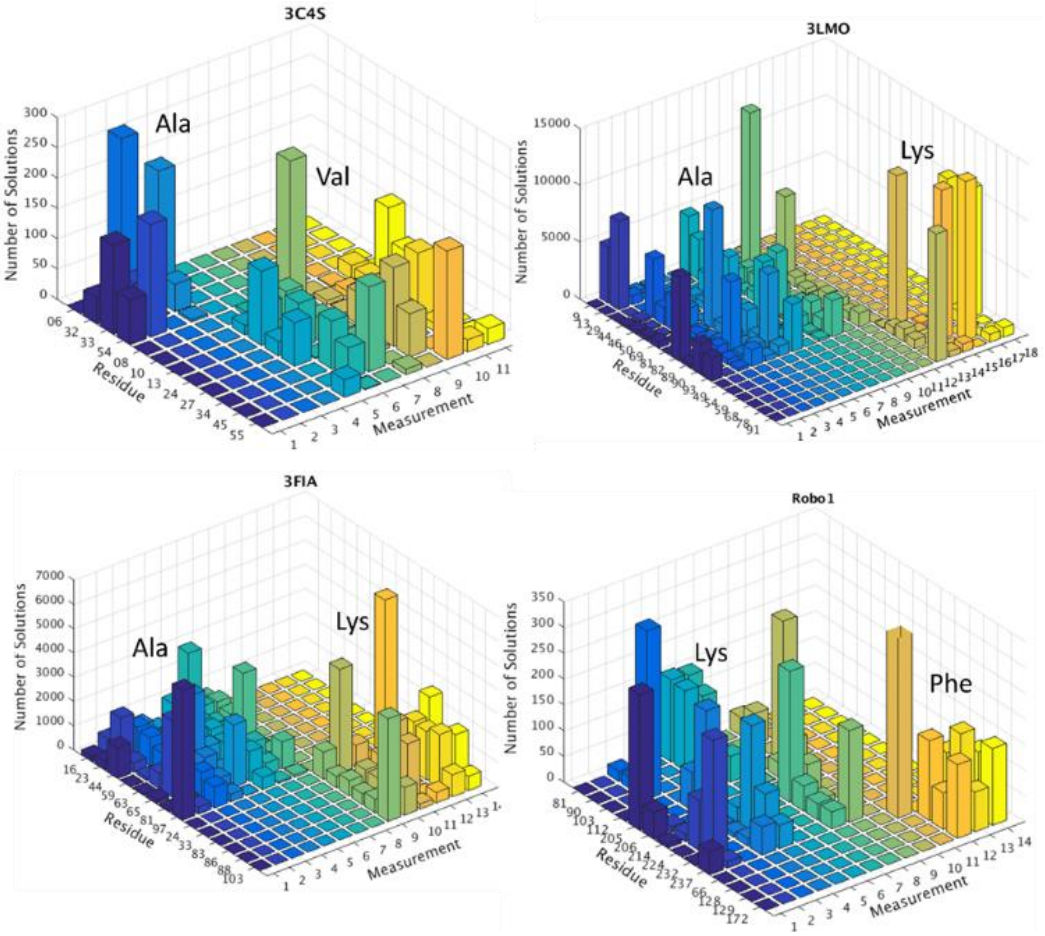
PDB	3C4S	3CWI	3LMO	3FIA
Labeled Sites and Number	Ala 4, Val 8	Ala 7, Val 9	Ala 12, Lys	Ala 8, Lys 6
Number of Acceptable Solutions*	260	1376	14006	7493
Top Score Solution (correct/total)	12/12	14/16	16/18	10/14
Correct Solution Rank	1	4	2	11
Consistently Assigned Crosspeaks	7/12	10/16	15/18	10/14
Missing Data	0	5 RDCs	1 RDCs	6 RDCs

*If the lowest score for an application is below 4, all the solutions with a score under 5 are collected. If the lowest score is above four, all the solutions with a score more than the lowest score plus 1.0 are collected.

One approach to assessing the probability of a correct assignment is to look at the frequency with which a crosspeak is assigned to the same site in solutions which fall within a standard deviation or so of satisfying experimental data. Since we have tried to scale scores relative to estimated error for each data type, the cut-off for solutions to examine should be roughly equal to the number of data types used. Our test cases had 4 data types. We would expect to see a significant number of solutions with scores below 4. Two of the test proteins fall in this class, 3CWI had a best solution score of 3.66 and 3FIA had a best solution score of 3.87. The other 3 had best scores of 4.45, 5.38, and 4.09. The higher scores could represent an underestimate of error, a systematic deviation in some data due to internal motion, or minor differences in structure between solution and crystal. To get an adequate sampling of solutions we will examine solutions with scores less than 5.0 if the minimum score is less than 4 and one plus the minimum when the minimum score is larger or equal to 4. We consider these to be acceptable assignments.

A visual way of presenting this analysis is shown in Figure 3.3. Histograms show the number of times a crosspeak is assigned to each site. If we take consistency to be assignment of

the same residue to a given crosspeak in more than 50% of the acceptable assignments, we find the following: of the 60 assignments which we can compare to the results of traditional triple resonance assignments, we would assign with confidence 35 peaks or about 60% of them. We find that among these 35 we would make one mistake. This would correspond to being correct 97% of the time, something close to a 95% confidence limit. The Robo1 system is a little different because we have good reason to believe that the structural model may be inadequate. Nevertheless, applying the same criteria we find that we can assign 7 of the 14 peaks with confidence and all of these agree with our manual assignment.



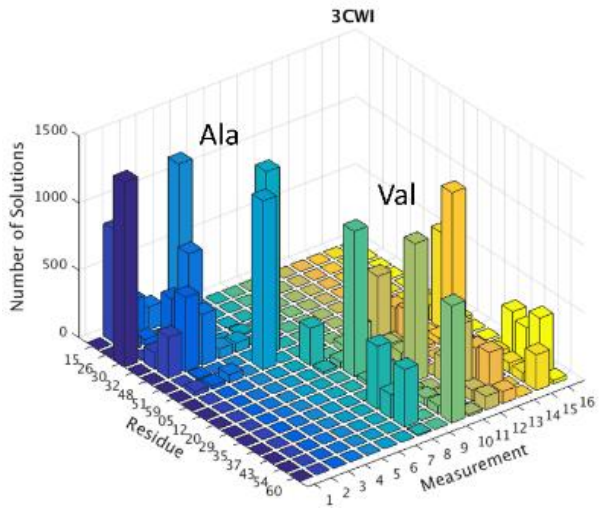


Figure 3.3. Histogram showing the frequency with which each crosspeak (measurement) is assigned to each site (residue) for test proteins and Robo1.

3.6 Discussion

Thus, we have clearly demonstrated an alternate procedure for assignments of HSQC crosspeaks from sparsely labeled proteins. This is particularly useful for glycoproteins that may have to be expressed in mammalian cell culture. Certain isotopically labeled amino acids are only moderately expensive. Drop-out media for mammalian cell culture is available and procedures for expression using 100-300 mg/L of labeled amino acid have been described [39]. The basic experiments for data collection are straightforward and not extremely time consuming. For Robo1-Ig1-2 only a single protein sample was needed for a complete set of experiments on each amino acid type; this is less than that typically required for a complete set of traditional 3D NMR experiments. The computation times are also modest, and more importantly, do not require personnel time. For the proteins studied, each required 5 hours or less to cover all cross-over mutation rates (0.1, 0.2, 0.4, 0.6 and 0.8) on a Xeon E5-2640 CPU.

We have emphasized the use of ^{15}N-enriched amino acids, but methods are equally applicable to ^{13}C-enriched amino acids. Metabolic labeling in methyl groups of isoleucine, leucine and valine, combined with perdeuteration, has become a popular approach to NMR-based structural work on large proteins [40]. Some alternate assignment strategies have also been developed for these systems [28-31], but we believe the approach described here could offer some advantages. Labeling with ^{13}C-methyl alanine provides the same excellent sensitivity and resolution, but it also provides RDCs that are backbone centered, much like ^{15}N-^{1}H amide RDCs. These can be used in our assignment strategy. Complete deuteration of large proteins would have to be sacrificed to collect the type of NOE data we use, but there is precedent for collection of NOEs on partially deuterated proteins that give well resolved spectra [41 42].

The procedure we describe does require structures for at least the individual domains comprising a protein, or proteins comprising a multi-protein complex. Structures for many of the proteins of interest today have been produced by X-ray crystallography or NMR and are available through readily accessible databanks. In principle, many other proteins can be modeled from homologous proteins in databanks [43-44]. We have not examined the use of modeled protein structures. However, it is likely that high quality structures will be required, as RDCs, NOEs and chemical shifts are all highly sensitive to three dimensional structure. Some limitations may also arise, because, NOEs are not sensitive just to an average structure or minimum energy structure, but to all structures sampled on the timescale of an NMR measurement. Accounting for conformational sampling using molecular dynamics trajectories remains one of the most promising options for improving predictions of chemical shifts, RDCs and NOEs. There have already been attempts to improve both chemical shift predictions and NOE predictions using these trajectories.

We have used just four types of data in this presentation. However, addition of other data types is relatively straightforward. Pseudo-contact shifts have the same geometry dependence as RDCs and can be predicted with the same single-value-decomposition algorithms used for RDCs. A precedent for use of these in resonance assignment has been established[29,45]. Paramagnetic relaxation enhancements (PREs) also have the same dependence as an NOE. There are other types of readily accessible data that would require design of distinctly different scoring functions. Amide protein exchange rates, for example, are easily measured from ^{15}N-^{1}H HSQCs, and there have been some attempts at making predictions based on structure [46]. These additions promise significant improvements in the applicability of our sparse-label assignment strategy in the future.

In summary, we have successfully demonstrated an NMR resonance assignment strategy that does not rely on triple resonance experiments and is applicable to proteins that benefit from sparse isotope labeling as opposed to uniform isotopic labeling. A program, "ASSIGNments for Sparsely Labeled Proteins", that uses a genetic algorithm to search for the best match of readily accessible experimental data to data predicted from known domain structures, has been developed. While a set of relatively small, previously assigned, proteins has been used to validate methods, the approach is applicable to larger proteins and a growing number of glycoproteins that are proving important in the study of human physiology and disease.

3.7 References

1. Prestegard, J. H.; Agard, D. A.; Moremen, K. W.; Lavery, L. A.; Morris, L. C.; Pederson, K., Sparse labeling of proteins: Structural characterization from long range constraints. *J Magn Reson* **2014,** *241*, 32-40.

2. Goto, N. K.; Kay, L. E., New developments in isotope labeling strategies for protein solution NMR spectroscopy. *Current opinion in structural biology* **2000,** *10* (5), 585-592.

3. Kainosho, M.; Torizawa, T.; Iwashita, Y.; Terauchi, T.; Ono, A. M.; Guntert, P., Optimal isotope labelling for NMR protein structure determinations. *Nature* **2006,** *440* (7080), 52-57.

4. Williamson, M. P., Using chemical shift perturbation to characterise ligand binding (vol 73, pg 1, 2013). *Prog Nucl Mag Res Sp* **2014,** *80*, 64-64.

5. Lipsitz, R. S.; Tjandra, N., Residual dipolar couplings in NMR structure analysis. *Annu Rev Bioph Biom* **2004,** *33*, 387-413.

6. Chen, K.; Tjandra, N., The Use of Residual Dipolar Coupling in Studying Proteins by NMR. *Top Curr Chem* **2012,** *326*, 47-67.

7. Ikura, M.; Kay, L. E.; Tschudin, R.; Bax, A., Three-dimensional NOESY-HMQC spectroscopy of a 13C-labeled protein. *Journal of Magnetic Resonance (1969)* **1990,** *86* (1), 204-209.

8. Li, D. W.; Bruschweiler, R., PPM_One: a static protein structure based chemical shift predictor. *J Biomol Nmr* **2015,** *62* (3), 403-409.

9. Han, B.; Liu, Y. F.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol Nmr* **2011,** *50* (1), 43-57.

10. Shen, Y.; Bax, A., SPARTA plus : a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol Nmr* **2010,** *48* (1), 13-22.

11. Sahakyan, A. B.; Vranken, W. F.; Cavalli, A.; Vendruscolo, M., Structure-based prediction of methyl chemical shifts in proteins. *J Biomol Nmr* **2011,** *50* (4), 331-346.

12. Valafar, H.; Prestegard, J. H., REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **2004,** *167* (2), 228-241.

13. Zweckstetter, M., NMR: prediction of molecular alignment from structure using the PALES software. *Nature protocols* **2008,** *3* (4), 679-690.

14. Lange, O. F.; Rossi, P.; Sgourakis, N. G.; Song, Y. F.; Lee, H. W.; Aramini, J. M.; Ertekin, A.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D., Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *P Natl Acad Sci USA* **2012,** *109* (27), 10873-10878.

15. Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M., ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **2003,** *19* (2), 315-316.

16. Buchner, L.; Guntert, P., Systematic evaluation of combined automated NOE assignment and structure calculation with CYANA. *J Biomol Nmr* **2015,** *62* (1), 81-95.

17. Orts, J.; Walti, M. A.; Marsh, M.; Vera, L.; Gossert, A. D.; Guntert, P.; Riek, R., NMR-Based Determination of the 3D Structure of the Ligand-Protein Interaction Site without Protein Resonance Assignment. *J Am Chem Soc* **2016,** *138* (13), 4393-4400.

18. Grishaev, A.; Steren, C. A.; Wu, B.; Pineda-Lucena, A.; Arrowsmith, C.; Llinas, M., ABACUS, a direct method for protein NMR structure computation via assembly of fragments. *Proteins-Structure Function and Bioinformatics* **2005,** *61* (1), 36-43.

19. Dashti, H.; Tonelli, M.; Lee, W.; Westler, W. M.; Cornilescu, G.; Ulrich, E. L.; Markley, J. L., Probabilistic validation of protein NMR chemical shift assignments. *J Biomol Nmr* **2016,** *64* (1), 17-25.

20. Huang, Y. J.; Mao, B.; Xu, F.; Montelione, G. T., Guiding automated NMR structure determination using a global optimization metric, the NMR DP score. *J Biomol Nmr* **2015,** *62* (4), 439-451.

21. Piai, A.; Gonnelli, L.; Felli, I. C.; Pierattelli, R.; Kazimierczuk, K.; Grudziaz, K.; Kozminski, W.; Zawadzka-Kazimierczuk, A., Amino acid recognition for automatic resonance assignment of intrinsically disordered proteins. *J Biomol Nmr* **2016,** *64* (3), 239-253.

22. Nkari, W. K.; Prestegard, J. H., NMR Resonance Assignments of Sparsely Labeled Proteins: Amide Proton Exchange Correlations in Native and Denatured States. *J Am Chem Soc* **2009,** *131* (14), 5344-5349.

23. Feng, L.; Lee, H. S.; Prestegard, J. H., NMR resonance assignments for sparsely 15N labeled proteins. *J Biomol Nmr* **2007,** *38* (3), 213-9.

24. Lohr, F.; Tumulka, F.; Bock, C.; Abele, R.; Dotsch, V., An extended combinatorial N-15, C-13(alpha), and C-13 ' labeling approach to protein backbone resonance assignment. *J Biomol Nmr* **2015,** *62* (3), 263-279.

25. Maslennikov, I.; Choe, S., Advances in NMR structures of integral membrane proteins. *Curr Opin Struc Biol* **2013,** *23* (4), 555-562.

26. Kato, K.; Yamaguchi, Y.; Arata, Y., Stable-isotope-assisted NMR approaches to glycoproteins using immunoglobulin G as a model system. *Prog Nucl Mag Res Sp* **2010,** *56* (4), 346-359.

27. Tzakos, A. G.; Grace, C. R. R.; Lukavsky, P. J.; Riek, R., NMR techniques for very large proteins and RNAs in solution. *Annu Rev Bioph Biom* **2006,** *35*, 319-342.

28. Xiao, Y.; Warner, L. R.; Latham, M. P.; Ahn, N. G.; Pardi, A., Structure-Based Assignment of Ile, Leu, and Val Methyl Groups in the Active and Inactive Forms of the

Mitogen-Activated Protein Kinase Extracellular Signal-Regulated Kinase. *Biochemistry-Us* **2015,** *54* (28), 4307-4319.

29. John, M.; Schmitz, C.; Park, A. Y.; Dixon, N. E.; Huber, T.; Otting, G., Sequence-specific and stereospecific assignment of methyl groups using paramagnetic lanthanides. *J Am Chem Soc* **2007,** *129* (44), 13749-13757.

30. Mishra, S. H.; Frueh, D. P., Assignment of methyl NMR resonances of a 52 kDa protein with residue-specific 4D correlation maps. *J Biomol Nmr* **2015,** *62* (3), 281-290.

31. Chao, F. A.; Kim, J. G.; Xia, Y. L.; Milligan, M.; Rowe, N.; Veglia, G., FLAMEnGO 2.0: An enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J Magn Reson* **2014,** *245*, 17-23.

32. Gao, Q.; Chen, C. Y.; Zong, C.; Wang, S.; Ramiah, A.; Prabhakar, P.; Morris, L. C.; Boons, G. J.; Moremen, K. W.; Prestegard, J. H., Structural Aspects of Heparan Sulfate Binding to Robo1-Ig1-2. *ACS Chem Biol* **2016,** (DOI: 10.1021/acschembio.6b00692).

33. Schmitt, L. M., Theory of genetic algorithms. *Theoretical Computer Science* **2001,** *259* (1), 1-61.

34. Yang, Y.; Fritzsching, K. J.; Hong, M., Resonance assignment of the NMR spectra of disordered proteins using a multi-objective non-dominated sorting genetic algorithm. *J Biomol Nmr* **2013,** *57* (3), 281-296.

35. Lin, H. N.; Wu, K. P.; Chang, J. M.; Sung, T. Y.; Hsu, W. L., GANA - a genetic algorithm for NMR backbone resonance assignment. *Nucleic Acids Research* **2005,** *33* (14), 4593-4601.

36. Everett, J. K.; Tejero, R.; Murthy, S. B.; Acton, T. B.; Aramini, J. M.; Baran, M. C.; Benach, J.; Cort, J. R.; Eletsky, A.; Forouhar, F.; Guan, R.; Kuzin, A. P.; Lee, H. W.; Liu, G.;

Mani, R.; Mao, B.; Mills, J. L.; Montelione, A. F.; Pederson, K.; Powers, R.; Ramelot, T.; Rossi, P.; Seetharaman, J.; Snyder, D.; Swapna, G. V.; Vorobiev, S. M.; Wu, Y.; Xiao, R.; Yang, Y.; Arrowsmith, C. H.; Hunt, J. F.; Kennedy, M. A.; Prestegard, J. H.; Szyperski, T.; Tong, L.; Montelione, G. T., A community resource of experimental data for NMR / X-ray crystal structure pairs. *Protein Sci* **2016,** *25* (1), 30-45.

37. Chipperfield, A.; Fleming, P. In *The MATLAB genetic algorithm toolbox*, Applied control techniques using MATLAB, IEE Colloquium on, IET: 1995; pp 10/1-10/4.

38. Sumathi, S.; Surekha, P., MATLAB-Based Genetic Algorithm. *Computational Intelligence Paradigms: Theory and Applications Using Matlab* **2010**, 547-589.

39. Barb, A. W.; Meng, L.; Gao, Z. W.; Johnson, R. W.; Moremen, K. W.; Prestegard, J. H., NMR Characterization of Immunoglobulin G Fc Glycan Motion on Enzymatic Sialylation. *Biochemistry-Us* **2012,** *51* (22), 4618-4626.

40. Tugarinov, V.; Kanelis, V.; Kay, L. E., Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nature protocols* **2006,** *1* (2), 749-54.

41. Kay, L. E.; Gardner, K. H., Solution NMR spectroscopy beyond 25 kDa. *Current opinion in structural biology* **1997,** *7* (5), 722-731.

42. Nietlispach, D.; Clowes, R. T.; Broadhurst, R. W.; Ito, Y.; Keeler, J.; Kelly, M.; Ashurst, J.; Oschkinat, H.; Domaille, P. J.; Laue, E. D., An approach to the structure determination of larger proteins using triple resonance NMR experiments in conjunction with random fractional deuteration. *J Am Chem Soc* **1996,** *118* (2), 407-415.

43. Bordoli, L.; Kiefer, F.; Arnold, K.; Benkert, P.; Battey, J.; Schwede, T., Protein structure homology modeling using SWISS-MODEL workspace. *Nature protocols* **2009,** *4* (1), 1-13.

44. Webb, B.; Sali, A., Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **2014**, 5.6. 1-5.6. 32.

45. Skinner, S. P.; Moshev, M.; Hass, M. A. S.; Ubbink, M., PARAssign-paramagnetic NMR assignments of protein nuclei on the basis of pseudocontact shifts. *J Biomol Nmr* **2013,** *55* (4), 379-389.

46. McAllister, R. G.; Konermann, L., Challenges in the Interpretation of Protein H/D Exchange Data: A Molecular Dynamics Simulation Perspective. *Biochemistry-Us* **2015,** *54* (16), 2683-2692.

CHAPTER 4

STRUCTURAL CHARATERIZATION OF HEPARAN SULFATE INTERACTING WITH LAR-IG1-2[3]

[3] To be submitted to the *Journal of Biological Chemistry*.

4.1 Acknowledgement

The following project involved a collaboration with Dr. Kelley Moremen's lab from Complex Carbohydrate Research Center CCRC. The protein constructs were designed by Kelley Moremen. All sparsely labeled LAR protein samples were expressed by Jeong-Yeh Yang. Laura Morris from the Prestegard lab and David Thieker from the Woods lab gave valuable advice in setting up MD simulations and post energy analysis. We thank them for their great efforts in this project.

4.2 Abstract

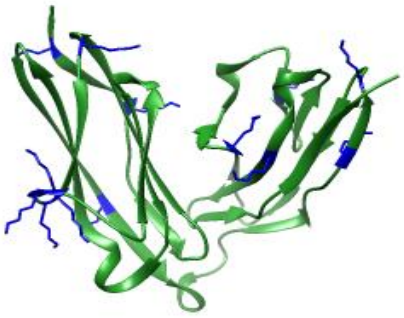
Leukocyte common antigen-related (LAR) protein is one of the type IIa receptor protein tyrosine phosphatases (RPTPs) which are important for signal transduction at the axon surfaces. Heparan sulfate chains play essential roles in the modulation of LAR signaling. Here, we report the structural characterization of the first two immunoglobulin domains (Ig1-2) of LAR interacting with a heparan sulfate pentasaccharide (GlcNS6S-GlcA-GlcNS3,6S-IdoA2S-GlcNS6S-OME, trade name fondaparinux) using multiple solution-based NMR methods. Because LAR is natively glycosylated we chose to express the protein in mammalian cells. Perdeuteration is not an option under these conditions and uniform labeling with ^{15}N and ^{13}C can be very expensive. Therefore, to maintain resolution and reduce expense we applied a sparse labeling strategy in which supplementation with a single type of isotopically enriched amino acids is used (in this case ^{15}N-enriched lysine). The assignments of labeled crosspeaks have been achieved using the software package, ASSIGN_SLP, and the aid of a lanthanide binding peptide construct to decrease crosspeak overlap and take advantage of distance dependent paramagnetic relaxation enhancement. Titration of LAR with fondaparinux reveals significant perturbations of crosspeaks assigned to Lys 68, 69, 71 and 72, allowing these residues to be associated with the

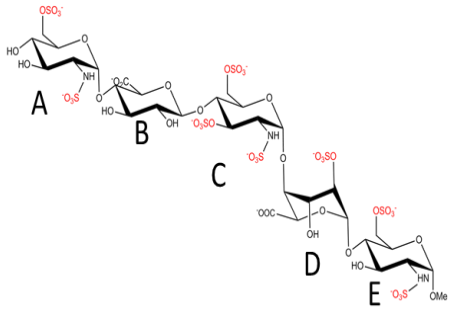
binding site and a disassociation constant of 60 μM to be obtained. Saturation transfer difference (STD) and transferred nuclear Overhauser effect (trNOE) experiments have identified binding epitopes and bound conformations of fondaparinux. NMR restraints derived from STD, trNOE, and shift perturbation data were combined in the docking program, HADDOCK, to generate models for the LAR-fondaparinux complex. The modeled complex is further analyzed by post-processing energetic analysis to identify key residues involved in the binding process.

4.3 Introduction

LAR, or leukocyte common antigen-related protein, is one of the type IIa receptor protein tyrosine phosphatases (RPTPs). Different from other RPTPs, most of which remain as orphan receptors[1], the type IIa RPTPs have been shown to bind a variety of cell surface proteins or soluble ligands and are believed to be highly involved in cell-cell or cell-matrix contacts[2-5]. LAR regulates diverse biological events, such as axonal guidance and outgrowth during neural development [6], synaptic organization[7], cell proliferation[8] and immune response[9]. Among the ligands that can interact with LAR and modulate RPTP signaling at neuronal growth cones are heparan sulfate proteoglycans (HSPGs) and chondroitin sulfate proteoglycans (CSPGs), [5]. Extensive previous studies have shown that the HSPGs and CSPGs oppositely regulate synaptic function upon binding with LAR. HSPGs complexing with LAR result in proteins clustering and promote neuron extension to the post synapse and interaction with postsynaptic proteins such as TrkC receptor protein tyrosine kinase[5, 10]. Contrarily, CSPGs complexing with LAR disrupt protein clustering, and lead to an inhibition of neural growth and regeneration[3, 5]. Protein-glycosaminoglycan interactions are heavily dependent on the structural characteristics of the glycan, such as length of the sugar chain and sulfation patterns[11-12]. For example, among different chondroitin sulphate isoforms (e.g. CS-A, CS-C and CS-E), CS-E shows a preference in

the interaction with RPTP sigma and inhibits the signaling pathways[13]. However, no similar specificity characterization or structural detail has been revealed for the heparan sulfate LAR interaction. Therefore, generating a structure of specific interaction between LAR and a well-defined HS would provide structural insight and an improved understanding of protein clustering and signaling pathways. In this paper, we present a solution model for the first two extracellular immunoglobulin-like domains of LAR (LAR-Ig1-2) (crystal structure shown in Figure 4.1. A) and their interaction with a specific HS pentasaccharide chain (GlcNS6S-GlcA-GlcNS3, 6S-IdoA2S-GlcNS6S-OME, trade name fondaparinux). The structure of fondaparinux, with our residue naming convention is shown in Figure 4.1.B).





GlcNS6S-GlcA-GlcNS3, 6S-IdoA2S-GlcNS6S-OME

Figure 4.1. (A) The LAR-Ig1-2 structure is shown as a ribbon diagram with Lys residues labeled in blue. (B) Structures of the heparan sulfate pentasaccharide used in this study with sulfate groups labeled in red.

Structurally, type IIa RPTPs, including LAR, share a very similar domain architecture, containing three immunoglobulin-like (Ig) domains, followed by nine fibronectin type II (FN) units, a single transmembrane helix and two intracellular phosphotyrosine-specific phosphatase domains[2]. Earlier crystallography and site-direct mutagenesis studies have suggested that the first Ig domain (Ig-1) is structurally crucial for the glycosaminoglycans binding and the first two Ig

domains (Ig 1-2) are a minimum structural requirement for interaction with the postsynaptic ligand TrkC[4]. Therefore, we selected the first two Ig domains for this study. The availability of a Ig1-2 crystal structure[5] will not only facilitate the protein assignment process but also allow a direct comparison with structures from our study. Compared with the remaining Ig domains, Ig 1-2 are believed to display the least inter-domain orientation flexibility[4]. This may be a valuable characteristic for synaptic signaling, and verification of retention of this structure in solution could be important.

To produce an acceptable sample, expression in a mammalian host was chosen (HEK 293 cells). LAR has a single N-glycosylation site in the first two domains, more specifically at N117 on the Ig1 domain. Normally the glycans would be complex and heterogeneous. To make the glycosylation homogenous, we expressed LAR-Ig1-2 in a cell line deficient in GnT1, stopping synthesis at a high mannose type; we then cleaved these high mannose forms to a single GlcNAc. This single sugar, covalently attached to the protein, has been shown maintain structural stability and some level of function in many glycoproteins[14]. We believe it will do so for LAR. To allow study by NMR a sparse labeling strategy was chosen[15]. ^{15}N-enriched lysines were chosen for selective labeling because of the likely involvement of lysines in GAG-related binding processes. The Ig1-2 construct has 13 lysines well dispersed throughout the structure, and they are expected to give well dispersed crosspeaks in two-dimensional ^{15}N-^{1}H heteronuclear single quantum coherence (HSQC) spectra.

Previously we developed strategies to make crosspeak assignments for a sparsely labeled protein (see chapters 2 and 3 of this thesis).[16-19]. Here we have applied ASSIGN_SLP[20], a software package designed to use a genetic algorithm to search for an optimal assignment of HSQC crosspeaks to specific lysine sites using predicted and experimental values for chemical

shifts, RDCs and NOEs. These procedures are complemented by the use of paramagnetic effects from a lanthanide binding peptide introduced into the protein construct[21]. The distance dependence of these effects allowed separation of crosspeaks into subgroups which could then be matched with a smaller set of sites. These methods proved quite efficient and successful. The crosspeak assignments provide a basis for functional interpretation and structural analysis.

To date, the achieved structural characteristics of LAR-HS complexes are based on using either long-chain heparin[22] or depolymerized heparan sulfates (i.e., dp 4, 6, 8 or 10), both of which are highly heterogeneous, or a HS mimic (sucrose octasaccharide) [5] which lacks the structural and functional features of HS. In this study, we use a structurally well-defined and homogenous HS pentassacharide, GlcNS6S-GlcA-GlcNS3,6S-IdoA2S-GlcNS6S-OME (trade name fondaparinux), which is of reasonable length, to generate a more realistic model for HS binding to LAR. A series of NMR techniques are used to generate data, including chemical shift perturbation (CSP) which defines the protein binding pocket, saturation transfer differences (STD) and transferred nuclear Overhausser effects (trNOE) which identify the ligand binding epitope and geometry. The model is generated by the docking program, HADDOCK[23], using the data as restraints. The binding geometry is rationalized by binding free energy analysis and per-residue decomposition. The resulting structure shows an extended binding cleft which encompasses the binding site for sucrose octasulfate (an HS mimic) identified in an existing crystal structure[5]. This structural characterization begins to build a solid foundation for understanding LAR-HS signaling.

4.4 Materials and Methods

4.4.1 Materials

^{15}N- Lys and deuterium oxide were purchased from Cambridge Isotope Laboratories. All other chemicals including HS fondaparinux, sodium salt, were purchased from Sigma-Aldrich unless otherwise stated.

4.4.2 Protein expression and purification

The detailed expression and purification procedure was described previously[24]. Briefly, the genes for LAR-Ig1-2 and a construct containing a lanthanide-binding peptide loop, LAR-loop-Ig1-2, were synthesized by GenScript (Piscataway, NJ) with codons optimized for utilization in mammalian cell expression. The DNA fragments for both sequences were cloned into the mammalian expression vector, pGEn2, which includes codes for an export signal, His-tag, AviTag, GFP-superfolder, and TEV cleavage site, using restriction digestion and ligation into a site following the TEV sequence (sequences given in Figure 4.2).

Large scale DNA preparations were prepared and transiently transfected into Lec1 (GnT1-) suspension culture cells in FreeStyle 293 media (Thermo Fisher Scientific, Waltham MA). The cell medium was exchanged to Freestyle dropout medium (Lys, Phe, Tyr, and Val amino acids) supplemented with 150 mg/L isotopically labeled Lys and non-isotopically labeled Phe, Tyr and Val in the second day of transfection. The recombinant protein was harvested after 6 days of growth and was purified from the culture supernatant using Ni^{2+} -NTA chromatography and concentrated to ~1 mg/mL. The resulting protein preparation was digested with recombinant TEV to cleave between LAR and GFP and then subjected to Ni^{2+}-NTA chromatography a second time to remove GFP. The N-glycan was trimmed off by Endo-H cleavage and one GlcNAc was left attached to the protein. The protein was subsequently purified by size exclusion

chromatography. The final protein yield was 12 mg/L. An average ^{15}N labeling efficiency for lysines was 77%, as determined by analyzing the isotopic envelope of the tryptic peptides containing lysine residues by mass spectrometry.

>LAR_GFP_fusion_construct
MAPEPAPGRTMVPLVPALVMLGLVAGAHGDSKPVFIKVPEDQTGLSGGVASFVCQATG
EPKPRITWMKKGKKVSSQRFEVIEFDDGAGSVLRIQPLRVQRDEAIYECTATNSLGEINTS
AKLSVLEEEQLPPGFPSIDMGPQLKVVEKARTATMLCAAGGNPDPEISWFKDFLPVDPA
TSNGRIKQLRSGALQIESSEESDQGKYECVATNSAGTRYSAPANLYVRVRRVAEFENLYFQ
GMSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPT
LVTTLTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDT
LVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQL
ADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGITHGMSGL
NDIFEAQKIEWHEHHHHHHHH
>LARLoop_GFP_fusion_construct
MAPEPAPGRTMVPLVPALVMLGLVAGDSKPVFIKVPEDQTGLSGGVASFVCQATGEPK
PRITWMKKGSYIDTNNDGAYEGDELSGKKVSSQRFEVIEFDDGAGSVLRIQPLRVQRDE
AIYECTATNSLGEINTSAKLSVLEEEQLPPGFPSIDMGPQLKVVEKARTATMLCAAGGNP
DPEISWFKDFLPVDPATSNGRIKQLRSGALQIESSEESDQGKYECVATNSAGTRYSAPANL
YVRVRRVAEFENLYFQGMSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKL
TLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKD
DGTYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKA
NFKIRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLL
EFVTAAGITHGMSGLNDIFEAQKIEWHEHHHHHHHH
Color code: Signal sequence Human LAR (LBT-LOOP) TEV cleavage site
Superfolder GFP AviTag His tag

Figure 4.2. Construct sequences of LAR and LAR_loop.

4.4.3 NMR spectroscopy

All the NMR spectroscopy was carried out on Agilent instruments with DD2 (18.8 T and 14.0 T) consoles and 5 mm cryogenically cooled triple resonance probes. NMR protein samples were 220 µM in 10% D_2O buffer containing 10 mM MES and 100 mM NaCl at pH 6.0 for ^{15}N HSQC titration, 3D ^{15}N-filtered NOE, and RDC experiments. LAR loop samples contained lanthanides at lanthanide to protein ratios slightly less than 1:1. The protein-ligand complex samples for STD and trNOE experiments were in 100% D_2O buffer containing 20 mM sodium

phosphate and 100 mM NaCl at pH 6.5, at a protein ligand ratio of 1:30 with a ligand concentration of 1.2 mM.

NMR experiments were standard Biopack experiments conducted at 25 °C. A mixing time of 40 ms and 90 ms was used in trNOE and HSQC-NOESY experiments (pulse sequence: gnoesyNfhsqcA). Two sets of RDCs were measured on protein samples containing either 12.5 mg/mL Pf1 phage (ASLA biotech) or 4.2 % PEG (C12E5/hexanol) bicelle using a pulse sequence in which cross-peaks in HSQC spectra are modulated by J+D coupling in the ^{15}N dimension[25]. LAR-Ig1-2 (170 µM) was titrated with increasing concentration of fondaparinux from 50 µM to 550 µM in steps of 50 µM from 0 µM to 150 µM and in steps of 100 µM from 150 µM to 550 µM, and recorded by ^{1}H-^{15}N HSQC (pulse sequence: gNfhsqc). The disassociation constant was determined by fitting the chemical shift perturbation as a function of concentration for each sparsely labeled residue as described before[16]. The STD (pulse sequence: dpfgse_satxfer) samples were irradiated at both -2.5 and 9.5 ppm with interleaved irradiation at 30 ppm to create difference spectra, and saturation times were increased from 1 to 4 s in steps of 1 s. The residue specific rotational correlation times for the protein were obtained from an SCT–CCR experiment using the same sample from the titration experiment[26]. The ligand proton resonances were assigned by acquiring and analyzing ^{1}H proton, ^{1}H-^{1}H TOCSY, ^{1}H-^{1}H NOESY, ^{13}C-^{1}H HSQC and ^{13}C-^{1}H HMBC spectra. All the NMR data were processed with NMRPipe[27] and analyzed with SPARKY[28].

4.4.4 Assignment of sparsely labeled LAR

The assignment of each labeled Lys residues in LAR-Ig1-2 was achieved using the program Assign_SLP[20]. The input experimental data include amide ^{1}H and ^{15}N chemical shifts, two sets of RDCs and ^{15}N-filtered NOEs. NOEs were converted to vectors using a script

contained in the program package. In order to improve crosspeak assignments, LAR with the lanthanide binding loop loaded with Gd^{3+} was used to separate labeled sites into a nearby group and a distant group based on paramagnetic broadening and loss of crosspeak intensity. There is less potential degeneracy of data when groups are smaller, making the assignment more robust. Different snapshots (600 ns, 800 ns and 1000 ns) from an MD simulation of LAR-Ig1-2 containing the lanthanide binding loop were utilized to identify 5 sites closest in distance to the metal. Assignment to these groups was enforced by adding penalties to the genetic algorithm objective function in proportioninter to elements of a user-specified constraint matrix in which elements denoting assignment to a correct group have value zero and elements denoting assignment to and incorrect group have value one (proportionality constant 10 in this case). In principle, the efficiency of the calculation could also be improved by running the ASSIGN_SLP separately on the two groups. The number of assignments to be screened is given by N! where N is the number of sites to be assigned. Dividing the calculation into two steps, one of size n and one of size (N-n), has a complexity of (N-n)! x n!, which is much less than N!. However, the use of RDC data requires the determination of 5 order parameters and at least 6 pieces of data for any calculation to be effective. The size of one of our groups is too small to take this route. Other details of the genetic algorithm implementation have been described previously[20]. Briefly, a combination of mutation and crossover rates (0.2, 0.4, 0.6 and 0.8 for both mutation and crossover) was used to allow a thorough search of assignment space, and all the assignments with a score under 8 were collected as possible solutions and saved for analysis.

4.4.5 Computational Docking

The docking program HADDOCK[23] was used to generate the molecular models of the LAR-fondaparinux complex. The crystal structure of human LAR-Ig1-2 (PDB 2YD5) was used

as the input protein structure. Two PDB structures of fondaparinux with the 2S IdoA group in either $^{1}C_{4}$ or $^{2}S_{0}$ ring forms were generated using the GLYCAM[29] web server. Ambiguous interaction restraints were set for the protein residues identified as involved in binding by chemical shift perturbation in the NMR titration experiments and similar restraints were set for residues in the ligand based on STD data. Interproton distance restraints within the ligand were set based on distances calculated from trNOE data using a $1/r^{6}$ dependence of NOE intensity and a reference distance from the intra-residue ^{1}H-^{2}H pair of the GlcNAc residue. The upper and lower limits were set by adding or subtracting 0.3 Å to the calculated distance. During the flexible docking part of the routine, the ligand was set to be fully flexible and strands of the protein containing direct binding sites were specified as semi-flexible. The detailed docking process has been described previously [16]. In the end, 20 top scoring models with the lowest rms NOE deviation and lowest energy were obtained.

4.4.6 MD simulation, free binding energy calculation and per-residue decomposition

In order to provide pdb structures of loop containing LAR-Ig1-2, a molecular dynamics (MD) trajectory was carried out using the SANDER module of AMBER 14[30] and the ff14SB[31] force field. The atomic coordinates were initially obtained from pdb 2YD5 with the lanthanide binding loop modeled in using CHIMERA[32]. The protein was solvated by a cubic box of TIP3P water. The MD simulation lasted for 1 μs after 2000 steps of minimization followed by 400 ps of heating. Similar MD (MD) trajectories were initiated with the top 4 output HADDOCK structures using the AMBER 14 package. The GLYCAM_06j-1[33] force field was used for carbohydrate simulation. For free energy calculation and per-residue decomposition, the same protein and ligand structures produced by HADDOCK were used to extract the initial atom coordinates. This MD simulation lasted for 50 ns after minimization, heating and density

equilibration. The molecular mechanics generalized Born surface area (MM-GBSA)[34] method followed by per-residue decomposition was applied to calculate the free binding energy of the bound state and the solvation energy of both the protein and ligand in solution. Detailed parameterization was described elsewhere[16].

4.5 Results

4.5.1 Assignment of LAR HSQC crosspeaks

Figure 4.3 shows superimposed ^{15}N-^{1}H HSQC spectra of the ^{15}N-Lys labeled two domain constructs of LAR, one containing a lanthanide binding loop carrying a diamagnetic ion (blue), and one without that loop (red). Several options for loop insertion had been explored, however, insertion in the loop normally connecting β-strands E and F (between residues G70 and K71) was the only position giving acceptable levels of expression, retention of LAR structure and useful lanthanide affinity. 12 and 13 crosspeaks are observed in the respective spectra; there are 13 lysine sites in the construct and we would expect 13 peaks. Peaks that are absent or of reduced intensity often reflect exchange broadening due to internal motion near the corresponding sites. Preservation of crosspeak positions on comparing the non-loop and a loop sample with a diamagnetic ion verify the retention of basic structural features of LAR, except for regions corresponding to crosspeaks 5, 10 and 11.

There are three lysines within two residues of the loop insertion, and it is not surprising that crosspeaks from these sites would shift. The affinity for fondaparinux also proves to be greatly reduced in the loop construct. This is also not surprising as the three sequentially proximate lysines are among the residues previously suggested to be involved in HS binding. Hence, we will use the loop construct for crosspeak assignment purposes and for verification of

the in-solution inter-domain structure of the LAR two domain construct. We will return to the non-loop structure for determination of a fondaparinux-LAR complex.

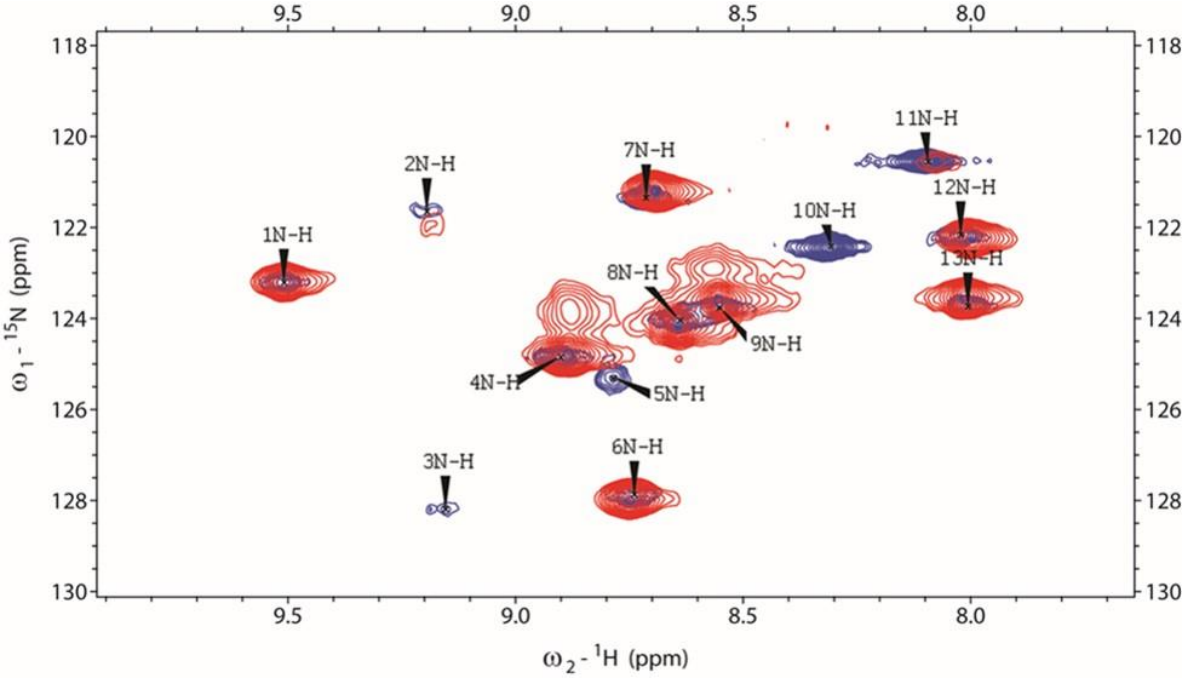


Figure 4.3. 2D ^{15}N-^{1}H HSQC spectra of (red) ^{15}N Lys labeled LAR and (blue) ^{15}N Lys labeled LAR-loop. Each peak in the LAR-loop spectrum is labeled with a peak number as a reference for the following assignment.

The LAR loop construct was selected to conduct initial resonance assignments partly considering the sharper lines and the more dispersed chemical shifts, especially in the direct dimension. This dispersion is particularly critical for collection of subsequent ^{15}N-filtered NOESY spectra, which were collected in a pseudo 2D fashion relying on dispersion in the proton dimension. To carry out assignments we used the program ASSIGN_SPL which requires multiple NMR observables including amide chemical shifts, ^{15}N-filtered NOEs and RDCs, as well as corresponding predicted values, using domain structures available from the crystal structure[5]. The predicted proton and nitrogen chemical shifts were generated using the PPM_one

software[35]. NOEs were predicted using a simple $1/r^6$ dependence of NOE intensity on interproton distance, and these intensities were used to generate a predicted vector in similar format to the experimental values. Predicted RDCs have to be determined for each possible assignment because of the need to determine five order parameters along with the back-calculated RDC set. So, this is done internally in ASSIG_SLP. Predicted and experimental values for chemical shifts and NOEs are given in Table 4.1 and Table 4.2. Experimental RDCs are given in Table 4.3.

Table 4.1. Experimental and predicted chemical shifts of Lys labeled ^{15}N-HSQC. Predicted chemical shifts are based on the 1000th ns snapshot from the LAR-loop MD run. Residue numbers are from the MD run, those in parentheses are from the crystal structure.

Crosspeak	Experimental		Lys residue	Predicted	
	N (ppm)	H (ppm)	MD (crystal structure)	N(ppm)	H (ppm)
1N-H	123.113	9.462	3(32)	126.975	8.427
2N-H	121.556	9.148	8(37)	118.833	8.123
3N-H	128.096	9.107	32(61)	121.242	8.294
4N-H	124.751	8.854	39(68)	121.665	8.535
5N-H	125.224	8.738	40(69)	127.394	7.845
6N-H	127.772	8.692	61(71)	121.681	8.325
7N-H	121.254	8.666	62(72)	120.668	8.528
8N-H	123.945	8.592	111(121)	120.616	8.688
9N-H	123.664	8.505	134(144)	124.628	8.701
10N-H	122.317	8.26	138(148)	123.97	8.229
11N-H	120.449	8.047	160(170)	118.78	8.671
12N-H	122.068	7.975	175(185)	125.576	8.702
13N-H	123.632	7.959	194(204)	120.347	8.348

Table 4.2. Experimental and predicted NOEs within 4 Å. Predicted chemical shifts are based on the 1000th ns snapshot of LAR-loop.

Experimental NOEs		Predicted NOEs				
HSQC crosspeak number	^{1}H (ppm)	Lys residues (MD naming)	Predicted contact residue name, number and atom			chemical shift (ppm)
1	1.221	3	SER	2	HA	4.539
	0.664	3	LYS	3	HB3	1.77
	0.306	3	LYS	3	HG3	1.381
	0.125	3	LYS	3	HA	4.359
	-0.069	3	LYS	3	HG2	1.381
	9.462	3	LYS	3	HB2	1.77
	5.004	8	ILE	7	H	8.676
	7.95	8	PHE	6	HB3	2.837
	4.639	8	GLN	26	H	8.608
	9.478	8	GLN	26	HB2	2.005
		8	LYS	8	HA	4.477
2	4.67	8	ILE	7	HB	1.722
	1.137	8	LYS	8	HB3	1.755
	1.491	8	LYS	8	HB2	1.755
	9.157	8	PHE	6	HA	5.259
		8	ILE	7	HA	4.001
3	4.425	8	ALA	27	HA	5.508
	1.503	8	PHE	6	HB2	2.837
	1.133	8	VAL	9	HG12	0.387
	9.132	32	LYS	32	HG2	1.551
		32	PRO	31	HA	4.513
4	9.246	32	LYS	32	HB2	1.713
	4.762	32	LYS	32	HA	4.13
	1.403	32	PRO	31	HB3	2.096
	1.157	32	LYS	32	HG3	1.551
	0.629	32	GLU	30	HA	4.627
	0.453	32	PRO	31	HB2	2.096
	0.154	32	LYS	32	HB3	1.713
	-0.069	39	MET	38	HA	5.012
	4.506	39	MET	38	HB3	1.84
	8.849	39	LYS	39	HB2	1.584
		39	LYS	39	HB3	1.584
5	4.839	39	LYS	39	HA	4.144
	2.364	39	MET	38	HB2	1.84
	1.884	39	LYS	62	HA	4.501
	1.62	39	VAL	63	HG13	1.045
	1.045	40	LYS	39	HA	4.144
	0.811	40	TYR	95	HA	5.135

	8.756	40	LYS	40	HB2	1.597
		40	LYS	40	HB3	1.597
6	4.651	40	LYS	40	HA	4.105
	1.878	40	GLU	96	H	8.834
	1.579	40	GLU	96	HG2	2.143
	1.033	40	LYS	39	HG3	0.811
	0.84	40	LYS	39	HG2	0.811
	0.594	61	GLY	60	HA3	4.018
	0.324	61	LYS	61	HG2	1.376
	-0.063	61	LYS	61	HA	4.687
	9.248	61	GLY	60	HA2	4.018
	8.709	61	LYS	61	HG3	1.376
		61	LYS	39	HB3	1.584
7	3.83	61	LYS	39	H	8.535
	1.508	61	LYS	61	HB3	1.721
	1.033	61	MET	38	HB3	1.84
	1.209	61	LYS	61	HB2	1.721
	4.651	62	LYS	62	HB3	1.868
	4.33	62	LYS	61	HA	4.687
	8.72	62	LYS	61	HB2	1.721
		62	LYS	62	HG3	1.496
8	8.76	62	LYS	62	HA	4.501
	4.604	62	LYS	61	HB3	1.721
	4.421	62	LYS	62	HB2	1.868
	4.163	62	MET	38	HG2	2.306
	2.446	62	LYS	62	HG2	1.496
	1.776	111	ALA	110	HA	4.644
	1.594	111	ALA	110	HB3	1.244
	1.244	111	LYS	111	HB3	1.822
	8.596	111	LYS	111	HB2	1.822
		111	ASP	12	HA	4.736
9	4.526	111	LYS	111	HA	5.137
	3.765	111	ASP	12	HB3	2.78
	1.536	111	ALA	110	HB1	1.244
	1.305	111	ALA	110	HB2	1.244
	1.165	111	ILE	94	HG12	1.381
	0.717	134	LEU	133	HA	4.661
	0.506	134	LEU	133	HG	1.863
	8.52	134	LYS	134	HA	4.637
		134	LYS	134	HG3	1.346
10	4.733	134	MET	144	HE3	1.807
	4.276	134	LEU	213	HA	4.952
	1.643	134	LEU	133	HD22	0.722

	1.315	134	LYS	134	HG2	1.346
	8.26	134	LYS	134	HB3	1.836
		134	TYR	214	HD2	7.105
11	8.233	134	TYR	214	H	8.642
	8.345	138	GLU	137	HA	4.558
	4.716	138	LYS	138	HB3	1.785
	4.264	138	LYS	138	HB2	1.785
	3.931	138	LYS	138	HA	3.94
	1.649	138	VAL	217	HA	4.33
	1.325	138	GLU	137	HB3	2.009
	8.05	138	VAL	215	HG12	0.711
		138	ARG	216	H	8.421
12	7.201	138	GLU	137	HG3	2.234
	9.266	160	PHE	159	HA	5.05
	4.667	160	PHE	159	HB3	2.825
	1.678	160	LYS	160	HA	4.257
	1.372	160	LYS	160	HB2	1.429
	0.671	160	VAL	165	HG23	0.691
	0.383	160	PHE	159	HD2	7.059
	1.203	160	LYS	160	HB3	1.429
	7.978	160	LEU	163	H	7.942
		160	PHE	159	HB2	2.825
13	4.757	160	PRO	164	HA	4.434
	1.362	175	ILE	174	HA	4.584
	0.828	175	LYS	175	HB2	1.616
	0.377	175	ILE	174	HG23	0.771
	0.23	175	GLN	183	H	8.76
	7.996	175	LYS	175	HA	4.455
		175	ILE	174	HG22	0.771
		175	LYS	175	HB3	1.616
		175	LEU	182	HG	1.158
		175	GLN	183	HB3	1.841
		175	ILE	184	HA	4.601
		194	LYS	194	HB2	1.653
		194	LYS	194	HG2	1.365
		194	GLY	193	HA3	4.099
		194	GLY	193	HA2	4.099
		194	LYS	194	HA	4.296
		194	LYS	194	HG3	1.365
		194	LYS	194	HB3	1.653
		194	ASN	212	HD22	6.994

Table 4.3. Experimental RDCs of LAR_loop using phage and PEG as alignment media.

Crosspeak	Phage RDC(Hz)	Error (Hz)	PEG RDC(Hz)	Error (Hz)
1	-3.037	1.76	-1.491	1.366
2	-4.746	5.457	-4.771	1.89
3	-0.971	9.244	0.941	7.01
4	-3.514	1.201	-5.516	1.123
5	-4.677	5.721	-9.298	4.052
6	-7.213	0.982	3.811	1.652
7	-9.676	1.578	3.776	1.557
8	7.576	1.417	4.244	1.464
9	-2.98	1.278	-9.005	1.242
10	-1.019	1.104	-1.06	0.548
11	-0.477	0.636	-0.631	0.299
12	-8.8	1.946	-1.722	1.917
13	-0.222	0.778	-14.583	1.935

The assignment of crosspeaks to lysine sites was further constrained by correlating pramagnetically perturbed crosspeaks with sites close to the ion carried by the lanthanide binding loop in the LAR-Loop construct. The superimposed spectra of lysine labeled LAR-Loop with and without Gd^{3+} (a paramagnetic lanthanide) are shown in Figure 4.4. From the overlaid spectra we can tell that the intensities of crosspeaks 2, 3 5, 10 and 11 decrease the most; this indicates these residues to be the closest in distance to the lanthanide. By comparing the distances from the MD simulation, the 5 residues nearest the ion belong to K68, K69, K71, K72 and K121. ASSIGN_SLP assigns penalties to any assignment which does not pair one of the five sites with one of the 5 perturbed crosspeaks.

The statistical results of assignment using 3 different MD frames as input protein structures are summarized in Table 4.4, and the histograms are shown in Figure 4.5. Previous work (see chapter 3[20]) has suggested a validation criterion based on the frequency of assignment of a crosspeak to the same site (more than 50% of the time corresponds to a 95% confidence limit). These assignments are indicated with an asterisk in the table. Other assignments

represent the most frequent assignment. Multiple residues are listed when the frequencies of assignment to multiple residues are similar. Assignments using the 1000ns frame are most definitive, yielding six definitive assignments. Confident assignments from the 600 ns frame were used to confirm two additional assignments. If we accept the two additional singly most probable assignments from the 1000 ns frame, the remaining three crosspeaks can be assigned by elimination.

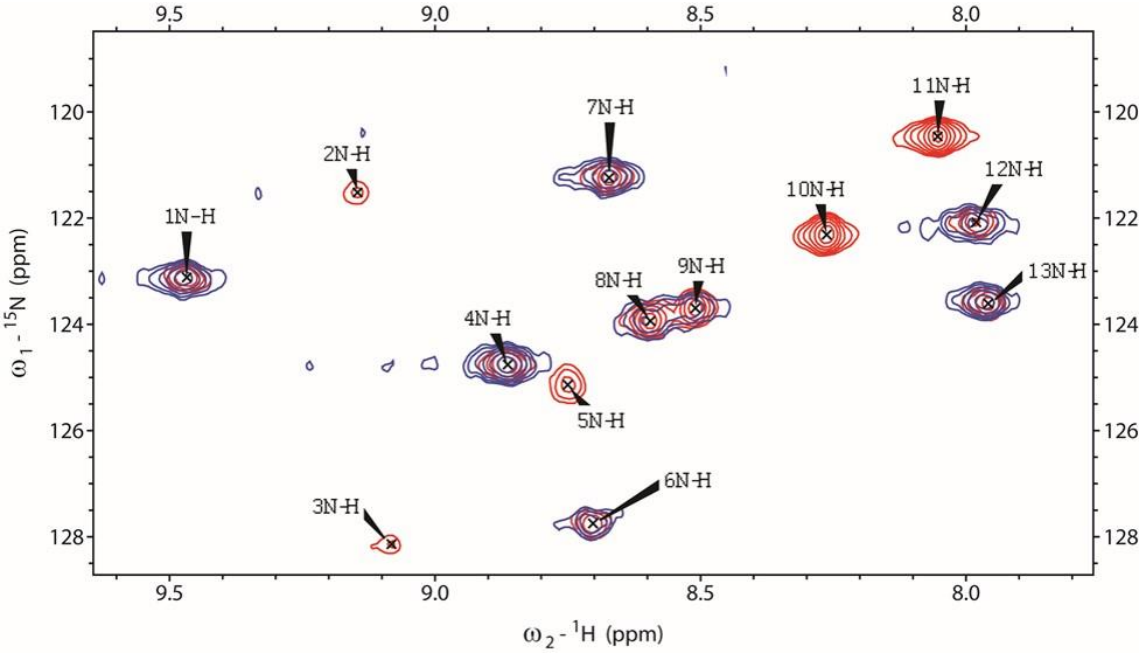
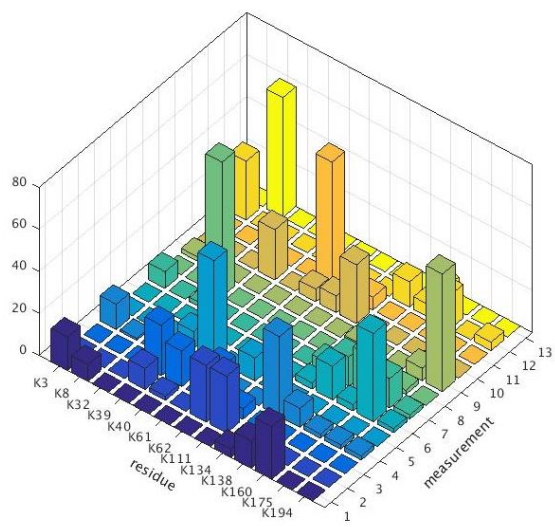
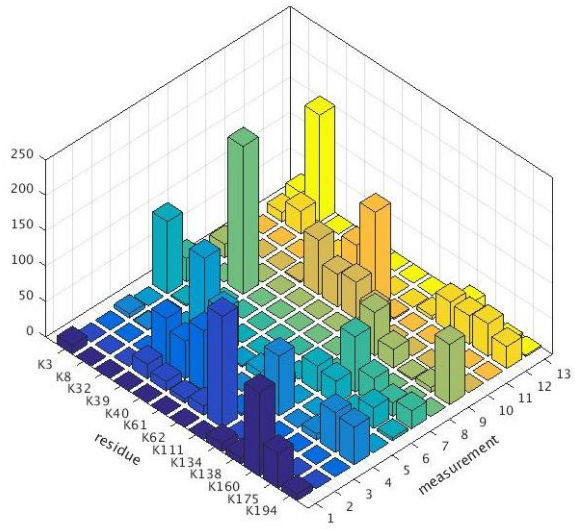


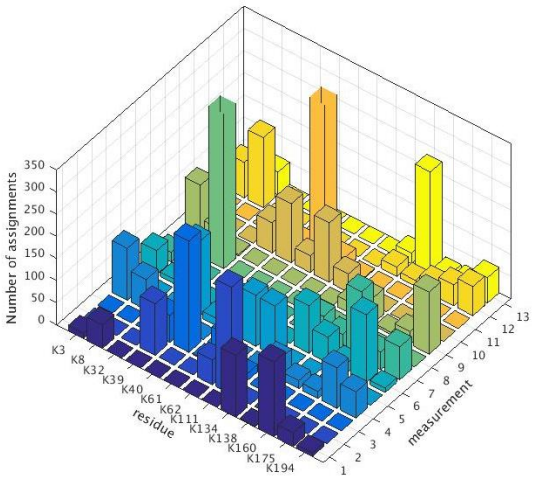
Figure 4.4. Superposition of ^{15}N-^{1}H HSQC spectra of ^{15}N-Lys labeled LAR-Ig1-2, engineered with lanthanide binding peptide loaded with Gd^{3+} (blue).



A



B



C

Figure 4.5. Histograms of statistical results of assignments using (A) 600 ns (B) 800 ns and (C) 1000 ns frames as input structures.

Table 4.4. Assignments summary of ^{15}N-Lys LAR-loop.

Crosspeak	600 ns	800ns	1000 ns	Assignment summary
1	K160	K160	K160,K134	K134(144)
2	K62,K111	K111	K111*	K111(121)

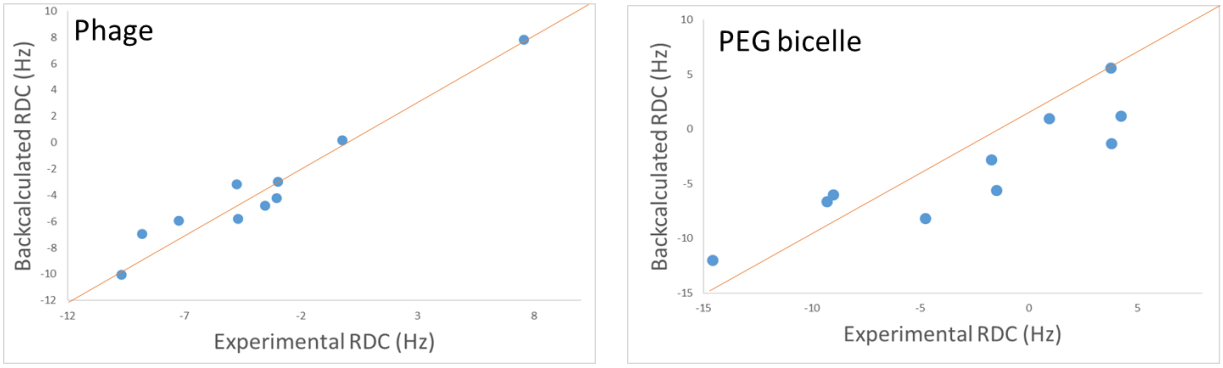
3	K39,K62	K61,K40	K40*	K40(69)
4	K134*	K134	K3,K175	K3(32)
5	K40	K39	K39	K39(68)
6	K175*	K3	K175	K175(185)
7	K160,K175,K138	K138	K160	K160(170)
8	K32	K32*	K32*	K32(61)
9	K194*	K194	K194,K3	K194(204)
10	K111,K39	K40,K62	K62,K40	K62(72)
11	K61*	K62*	K61*	K61(71)
12	K3,K160	K8,K138	K8*	K8(37)
13	K8*	K8*	K138*	K138(148)

*Assignment frequency is higher than half of the total assignments. The residue number used in the table are according to the amber MD naming system. The residue number corresponding to the crystal structure number is listed in parentheses.

4.5.2 Inter-domain geometry of LAR-Ig1-2

The RDC data used in the assignment process is also potentially valuable as a means of assessing the inter-domain geometry of two domain constructs, like our LAR construct. We have implicitly assumed the folded version of inter-domain geometry seen in the crystal structure (see fig. XXX) is preserved in solution, since we needed to combine RDC data for both domains to meet minimum requirements for order parameter determination and back-calculation of RDCs. However, the fact that we find assignments with total scores in the range expected for consistency with all data, suggests that the folded domain geometry is appropriate. Once crosspeaks are assigned, we can also assess the validity of this geometry directly by looking at Q factors for RDC data[36]. These are ratios of root-mean square deviation (RMSD) of measured and calculated RDCs and the root-mean square of the measurements. Figure 4.6 shows a correlation

plot of measure and back-calculated RDCs for our phage and bicelle data. Loop regions are the least ordered in structure, hence the residues located adjacent to these regions are excluded for a RDC calculation. Residues within two of the loop insertion are K69, 71 and 72. K68 is also close (three away), however, its N-H vector does not move significantly during the MD simulation. Therefore, K68 is included in the calculation. 10 pieces of RDC data remain and these were used in the final calculation. Q factors of 0.19 (phage) and 0.43 (PEG) were obtained. A Q factor of 0.4 (based on a larger set of RDC values) has been suggested to reflect a structure comparable to an X-ray structure of approximately 2.5 Å resolution[37]. Hence, we believe these numbers support the contention that the folded LAR structure is preserved in solution.



A B

Figure 4.6. Correlation plot of experimental RDCs and back-calculated RDCs for (A) phage and (B) PEG bicelle media.

Comparison of Q factors back-calculated using the folded LAR structure to those back-calculated using an alternative structure provide another means of assessing the significance of these measurements. The Ig1-2 domains of LAR are structurally similar to the Ig1-2 domains of Robo1, but Robo1 has a more extended structure. So, adjusting LAR inter-domain geometry to superimpose with Robo1 would yield a suitable trial structure. The inter-domain geometry of

Robo1 is quite flexible and it results in different forms of crystal structure (pdb 2v9r and 2v9q, discussed in chapter 2). Our previous study has shown that Robo1 Ig1-2 domains adopt a slightly bent conformation (pdb 2v9r) as opposed to a straight conformation (pdb 2v9q). Therefore, a model of LAR (shown in Figure 4.7) is built based on the Robo1 crystal structure pdb 2v9r. To do this we have broken the linking peptide (which contains no lysines) and aligned each domain of LAR to the respective domains of Robo1. Q factors obtained for phage and bicelle data are now 0.43 and 0.88. The increase confirms that, despite the small number of RDCs, we are sensitive to inter-domain geometry, and the folded LAR structure is the best representation.

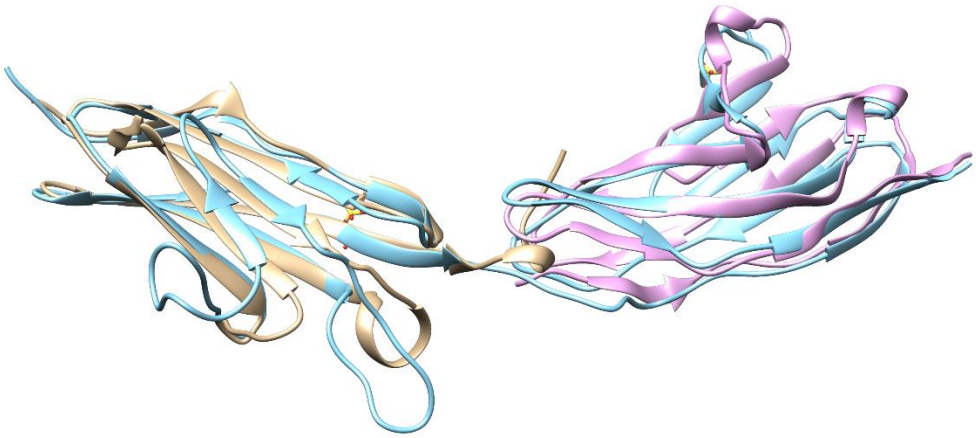
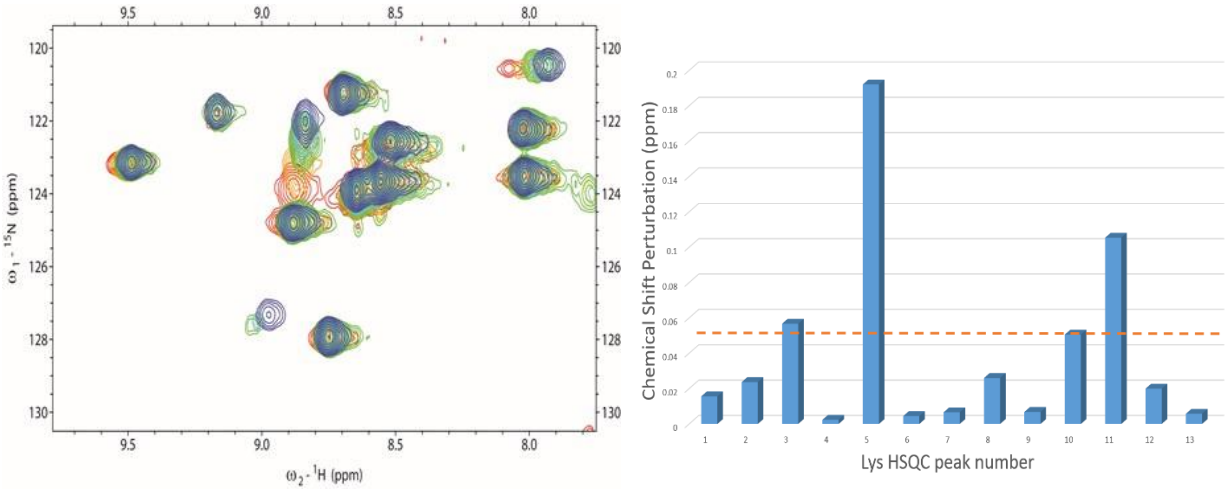


Figure 4.7. LAR-Ig1-2 model based on Robo1 (pdb 2v9r). LAR Ig1 is in beige, Ig2 is in pink and Robo Ig1-2 is in blue.

4.5.3 Protein binding site identified by NMR titration

Direct information on the residue composition of binding sites, as well as disassociation constants (Kd), can be extracted from chemical shift perturbations on addition of ligand to a sample. The overlaid 2D 15N-1H HSQC spectra of LAR with increasing concentrations of fondaparinux are shown in Figure 4.8A. The maximum chemical shift perturbation of each Lys site is plotted in Figure 4.8B.



A B

Figure 4.8. (A) HSQC spectra of 170 μM LAR with increasing concentration of fondaparinux from 0 μM to 550 μM with rainbow colors coded. (B) Total chemical shift perturbation of each Lys residue is plotted against the crosspeak number.

Four residues, belonging to peaks 3, 5, 10 and 11, display substantial chemical shift perturbation (> 0.05 ppm) upon the titration with fondaparinux. Fitting the average of the shifts for these peaks, a disassociation constant (Kd) of 60 ± 24 μM is determined. This is a moderately high affinity, well within the range of dissociation constants found for Robo1 interactions with HS oligomers (see chapter 2). The data and best fit line are shown in Figure 4.9.

The limiting shifts for perturbed crosspeaks can also be used as a very qualitative indicator of residue involved in fondaparinux binding. Crosspeaks 3, 5, 10 and 11 have been assigned to Lys 69, 68, 72 and 71. Previous literature has also reported these four residues to be crucial for binding to heparin by site directed mutagenesis[22]. The crystal structures of LAR with sucrose octasulfate, a mimic compound of heparan sulfate, has also pointed to these four residues[5] . The identification of these residue will be used later as a constraint in generating a model for a fondaparinux-LAR complex.

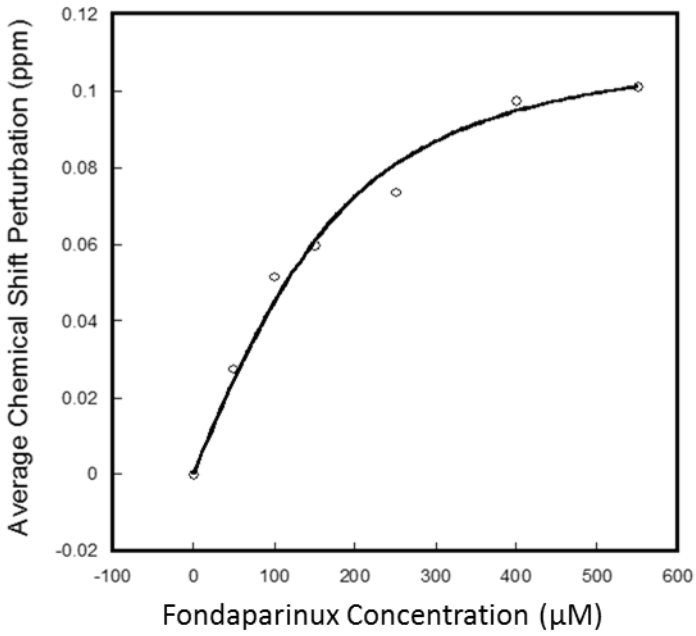


Figure 4.9. Binding affinity of LAR-Ig1-2 for the HS fondaparinux.

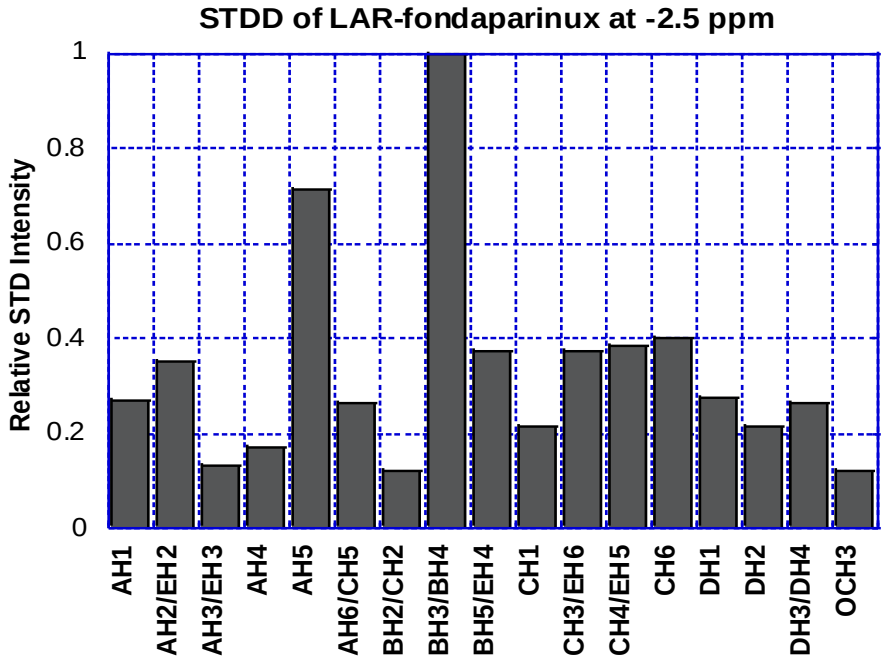
4.5.4 Ligand binding epitopes from STD experiments

Binding epitopes on a ligand can be qualitatively identified by saturation transfer difference (STD) NMR[16, 38-39]. The magnetization saturation of a protein proton can be transferred to a nearby bound-state ligand proton in a $1/r^6$ fashion, where r is the distance between the two protons. The intensity of the resonance associated with this proton then decreases. If the ligand returns back to the free solution state in a time which is short compared to its T_1 relaxation time, its resonance intensities are largely retained and result in diminished intensities of the averaged solution spectrum for the ligand. The larger the intensity loss, the smaller the distance between the ligand proton and the protein proton. Therefore, the binding epitope of ligand can be qualitatively characterized.

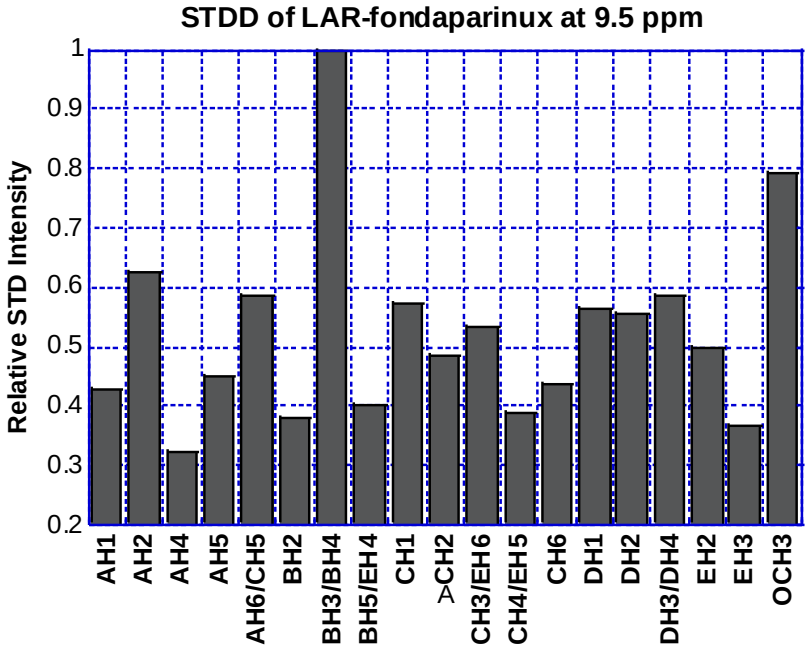
Despite the fact that most protein resonances are suppressed in a normal STD experiment by filtering out broad lines, there are complexities that arise with glycosylated proteins, like LAR. Although our LAR construct has been treated with Endo-H glycosidase to trim off the high-mannose glycans generated from the Lec1 cell line, leaving a single GlcNAc attached to the protein, the resonance signals from GlcNAc are often sharp and can carry through to the STD difference spectrum. Therefore, we conduct a double difference STD experiment and interpret it in a more qualitative manner.

Figure 4.10A and B are the histograms of relative STD intensity of different ligand protons irradiated at -2.5 ppm and 9.5 ppm respectively using a saturation time of 4s. Glycan residues are designated as A to E starting from the non-reducing end as in Figure 4.1B, and protons are numbered starting with the anomeric proton and proceeding sequentially toward the exo-cyclic hydroxymethlene or carboxyl group of each residue. The largest STD intensity losses come from BH3/BH4, and some parts of the A (AH2 and AH5) and C (CH1, CH3 and CH5) residues. These results indicate that these three residues are most proximate to the protein surface, especially the GlcA, with the most significant perturbation from BH3 and BH4 protons, whereas the remainder of the ligand (mainly D and E) suffers less perturbation. The linker residue (-OME) presents a much stronger STD signal, particularly with 9.5 ppm irradiation. This could be associated with a more direct saturation path between aromatic residues and this group in the bound complex; there are phenylalanine and tyrosine groups underlying the putative binding region. However, the longer methyl relaxation times and sharp resonance also makes these groups more susceptible to experimental artifacts, and this methyl group is not characteristic of a native HS fragment. Hence, we will disregard its interactions in our conformational search. In addition to the strongly perturbed resonances, the histogram shows

multiple moderate STD losses throughout the ligand. These effects arise from spin diffusion among sets of proximate protons, especially when binding is tight and release from the protein slow. These effects limit the use of STD information to a more qualitative application. The strongest STD signals (more than 50%) have been implemented as ambiguous restraints for use in the HADDOCK program which will be described later in this chapter.



A



A

B

Figure 4.10. Saturation transfer double difference data on various resonances with a saturation time of 4 s at (A) -2.5 ppm and (B) 9.5 ppm for fondaparinux. Data is normalized to the STD effect for the proton showing the largest saturation in each case.

4.5.5 Ligand bound conformation from trNOE experiments

The transferred nuclear Overhauser effect (trNOE) is a powerful NMR phenomenon to study protein-bound ligand conformations. The rotational correlation time of the ligand in the bound state is longer than that in the free state. Therefore, the efficiency of magnetization transfer is increased in a large complex, and large negative NOEs will dominate over the small positive NOEs found for small ligands in solution. We actually use the change in sign of the

NOE on moving from small to large correlation times to advantage by choosing a temperature at which contributions form free ligands in solution are near zero (30 °C at 600 MHz with a mixing time of 40 ms). The intensity of trNOE signals measured from resonances averaged over populations with high free-state contributions can then be interpreted through a $1/r^6$ dependence on interproton distances of the bound state. We use a distance of 2.5 Å between H2 and H4 of the GlcNAc residue as a reference distance when converting NOE intensity to distance.

Heparan sulfate glycosaminoglycans display a significant level of structural flexibility, primarily due to variation in glycosidic linkage torsion angles and the multiple glycan ring conformations of IdoA. Hence, HS oligomers can adopt conformations in the bound state which are different from the dominant conformation in solution. In Table 4.5, we report trNOE-based distances between pairs of nuclei affected either by glycosidic torsion angles or ring conformations for the bound state of fondaparinux. The distances of same pairs of protons determined from NOEs in the free state are also listed as a comparison.

Table 4.5. Interproton distances of fondaparinux derived in free and bound states from NOE data.

Nuclei pair	Free ligand (Å)	Error (Å)	trNOE (Å)	Error (Å)
AH1-BH4	2.62	0.02	2.86	0.21
CH1-DH4	2.41	0.02	2.65	0.20
CH3-DH1	2.43	0.11	3.05	0.25
DH1-DH2	3.16	0.01	N/A	N/A
DH1-DH3	3.27	0.01	2.72	0.19
DH1-EH4	2.71	0.02	3.14	0.25
DH1-EH61	2.85	0.02	2.91	0.23
DH1-EH62	3.22	0.03	2.49	0.19
Linker H-EH1	2.96	0.02	3.22	0.20

There are some differences between the fondaparinux interproton distances converted from NOEs measured in a free state and the distances converted from trNOE measurements

which are dominated by the bound state. For instance, the measurable transglycosidic distances between CH3 and DH1 are 0.62 Å longer in the bound state (3.05 Å) than that in the free state 2.43 Å. Similarly, a trNOE observed between DH1 and EH4 gives a calculated distance of 3.14 Å, which is also significantly longer than the free state distance (2.71 Å). These two trNOE distances measured from a bound fondaparinux suggest it adopts a preferred conformation which is different from the dominant free state conformation, particularly with respect to the glycosidic linkages between GlcNS3,6S C and IdoA2S D, and IdoA2S D and GlcNS6S E. Besides the transglycosidic distance, another piece of useful information comes from distances between protons on the IdoA2S ring. The distance between IdoA2S D H1 and H2 and D H1 and D H3 are 3.16 Å and 3.27 Å respectively in the free state. These suggest a mixture of a $^{1}C_{4}$ chair and $^{2}S_{0}$ skew-boat conformation. The skew-boat conformation is known to be particularly favorable in solution when a 3-O-sulfate is attached to the preceding GlcNAc residue[40]. A trNOE between IdoA2S D H1 and H3 in the bound state gives a 2.72 Å distance, which is shorter than that of the free state (3.27 Å), indicating a much higher population of the skew-boat conformer and perhaps even some sampling of a true boat conformer. These trNOE data provide another source of structural information that can be applied as restraints in the docking process.

4.5.6 Oligomerization state calculated from rotational correlation times

A previous study showed that heparin was able to induce oligomerization of RPTP σ[41]. In this study, heparin dp8 was the minimum length of a heparin oligosaccharide to promote RPTP σ oligomerization. Under the conditions of our experiment, there is no evidence for oligomer formation. An average rotational correlation time (τ_c) of 8.41 ns was measured from cross-correlation experiments on an LAR-fondaparinux complex suggesting that the protein remains monomeric after interacting with fondaparinux (see Table 4.6). As a general rule of

thumb, the τ_c of a protein in solution in nanoseconds is approximately 0.5 times its molecular weight in kDa. The molecular weight of LAR is 22 KDa giving an estimated τc of 11 ns for monomer and 22 ns for dimeric state. Of course, fondaparinux is only 5 residues long, so we cannot make definitive conclusions about the effect of a longer HS oligomer. LAR is also not identical to RPTP σ, but it is sufficiently similar in sequence and function to look for possible oligomerization modes once we have a structure of the LAR-fondaparinux complex.

Table 4.6. Rotational correlation time τc of Lys residues in LAR in the presence of fondaparinux.

Crosspeak	1	2	3	4	5	6	7	8	9	10	11	12	13
TauC (ns)	8.55	8.15	5.75	7.8	11.65	9.9	9.05	7.25	8.65	7.15	6.95	10.05	8.4
Assignment	K144	K121	K69	K32	K68	K185	K170	K61	K204	K72	K71	K37	K148
Average TauC	8.41 ns												

4.5.7 Complex modeling by computational docking

High Ambiguity Driven biomolecular DOCKing (HADDOCK) is a versatile software package for biomolecule docking that uses a variety of biophysical information[23]. Here we combine all the restraints deduced from different types of NMR measurements, including the chemical shift perturbation data to determine the protein binding pocket, STD-based information on fondaparinux to determine binding epitopes, and trNOE measurements which allow pairwise distance restraints on the bound ligand conformation to give a detailed structure of the LAR-fondaparinux complex. Details concerning the specific docking process are described in the Materials and Methods section. Although we have clear evidence for a $^2S_0$-skew boat conformations in the bound form, two sets of HADDOCK runs were conducted starting with skew-boat and chair conformations separately. The energy minimized input structures for IdoA

were made using the GLYCAM web tool. The overall free energies of binding taken from a molecular mechanics-generalized Born surface area (MM-GBSA) calculation conducted on a 50 ns MD simulation of docked structures were in fact significantly more negative for the skewed-boat structures (-39.34 kcal/mol for the best chair structure versus -43.71kcal/mol for the average of the best three skew-boat structures). Hence, we focus on the skew-boat structures in what follows. The top 5 structures with lowest HADDOCK score, energy as well as smallest rms NOE violations are presented in Figure 4.11. Comparing the top docked structures, we find that most of the structure differences arise from the last two sugars (IdoA2S and the reducing end GlcNS6S, residues D and E). These interact with multiple lysines on the binding loop, taking advantage of the motional freedom of these residues. A single structure representation of the protein residues having close contacts with the ligand is illustrated in Figure 4.12.

4.5.8 Free energy calculation and analysis

Molecular mechanics-generalized Born surface area (MM-GBSA)[34] calculations were conducted as to identify residues making major contributions to a free energy of binding. Even though docking finds the same binding pocket, the top 5 docked structures with IdoA2S in the $^{2}S_{0}$ conformation were not well aligned and showed three types of positioning, therefore 3 structures representing different bound states were chosen for the energy analysis. The energy analysis is shown in Table 4.7. Because the conformational entropy terms are excluded from the calculation and conformations for both protein and ligand are assumed to be the same in the dissociated state as in the associated state, there is typically a large overestimation of binding energies. Therefore, absolute energies are not meaningful, but relative contributions from different interacting groups and types of energy, particularly for the ligand, can be. The free protein part of the overestimation, which is associated with not allowing the protein to adjust

conformation and interact optimally with both solvent and counter-ions in the free state, is fortunately the same for all ligands, and won't affect analysis of the ligand part of the analysis. Solvent interactions for the free ligand may also be better approximated because most surfaces of the ligand are solvent exposed, regardless of conformation. Table 4.8 shows the ligand part of the energy decomposition.

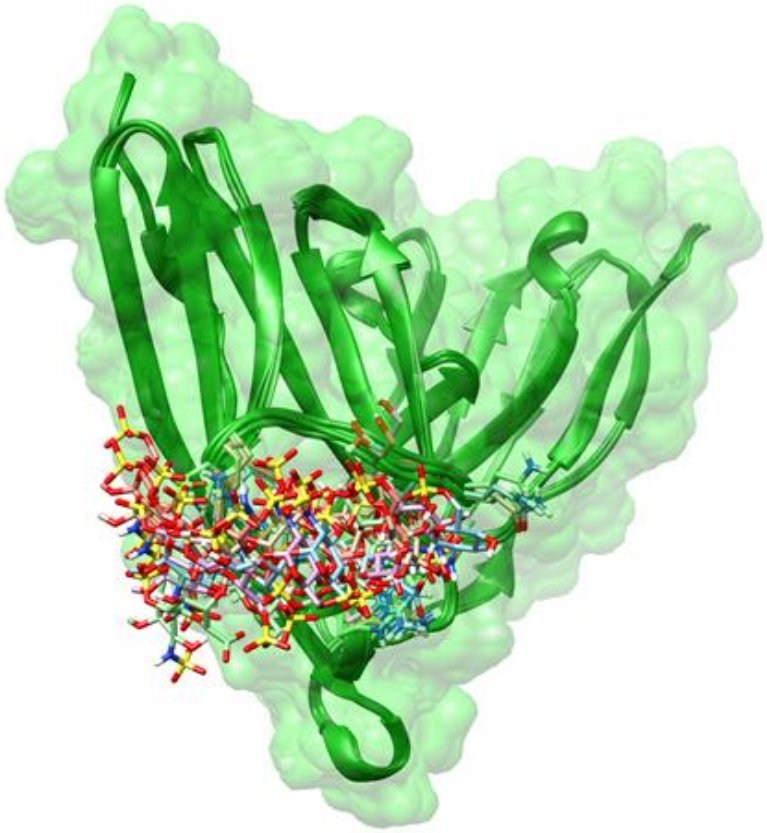


Figure 4.11. Top 5 HADDOCK structures with the highest score and lowest energy for the LAR-fondaparinux complex, starting with IdoA2S in the $^{2}S_{0}$-skew boat conformation.

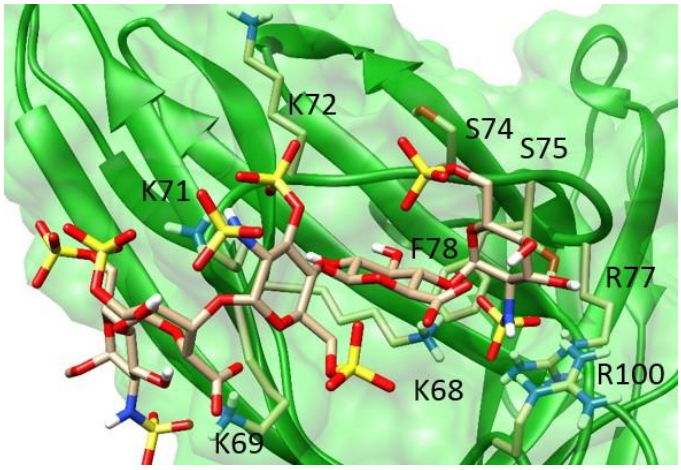


Figure 4.12. Expanded view of the binding pocket for the LAR-fondaparinux HADDOCK structure. While IdoA2S started in a $^{2}S_{0}$ conformation, it is more boat-like in this structure. Nearby interacting residues are shown as stick structures and labeled.

Table 4.7. MM-GBSA energy component analysis of the interactions of the fondaparinux – LAR-Ig1-2 complex.

	Mean value of $^{1}C_{4}$ structure (kcal/mol)	Standard deviation (kcal/mol)	Mean value of $^{2}S_{0}$ structure (kcal/mol)	Standard deviation
ELE	-885.05	92.11	-904.67	87.04
VDW	-27.81	8.05	-29.83	5.99
INT	0	0	0	0
GAS	-912.86	94.87	-934.50	88.23
GBSUR	-4.11	0.82	-4.37	0.47
GB	879.01	88.24	893.26	82.89
GBSOL	874.91	87.75	888.90	82.75
GBELE	-6.04	10.17	-11.41	10.85
GBTOT	-37.96	10.72	-45.61	10.23

ELE, non-bonded electrostatic energy; VDW, non-bonded van der Waals energy; INT, bond, angle, dihedral energies; GAS, ELE+VDW+INT; GBSUR, hydrophobic contributions to solvation free energy for GB calculations; GB, reaction field energy calculated by GB; GBSOL=GBSUR+GB; GBELE=GB+ELE; GBTOTAL=GBSOL+GAS. All energies are in kcal/mol.

Table 4.8. Per-residue free energy decomposition of fondaparinox by MM-GBSA. All energies are in kcal/mol.

		van der Waals	Electrostatic	Polar Solvation	Non-polar Solv	TOTAL
Structure1	GlcNS6S E	-0.829706	-9.24199	11.914534	-0.138422246	1.7044
	Ido2S D	-0.779146	-52.633566	51.377852	-0.138602693	-2.173
	GlcNS3,6S C	-1.748396	-60.81628	61.25015	-0.084021235	-1.399
	GlcA B	-3.733636	-58.353608	60.063738	-0.534431174	-2.558
	GlcNS6S A	-5.626716	-85.480442	81.523724	-1.220501218	-10.8
	6-O-sulfate A	-1.246506	-38.146192	41.600684	-0.446028149	1.762
	6-O-sulfate C	-0.541652	-47.614176	47.671596	-0.322706866	-0.807
	3-O-sulfate C	-0.218276	-58.772708	58.359308	-0.249511738	-0.881
	2-O-sulfate D	-0.105096	-31.274244	31.751916	-0.092826749	0.2797
	6-O-sulfate E	-0.094084	-9.010922	9.661438	-0.020360866	0.5361
Structure 2	GlcNS6S E	-1.13236	-25.294204	27.950062	-0.176792198	1.3467
	Ido2S D	-0.687954	-95.27199	87.688796	-0.19624968	-8.467
	GlcNS3,6S C	-1.828848	-85.77829	87.909882	-0.153272736	0.1495
	GlcA B	-2.085772	-54.348434	57.066436	-0.286577122	0.3457
	GlcNS6S A	-3.717194	-81.411562	78.264068	-0.973291421	-7.838
	6-O-sulfate A	-0.440018	-17.4071	19.116184	-0.118821456	1.1502
	6-O-sulfate C	-0.203734	-31.224506	31.790806	-0.040662864	0.3219
	3-O-sulfate C	-0.098286	-87.50416	82.923378	-0.392379998	-5.071
	2-O-sulfate D	-0.571136	-47.084156	49.026078	-0.18705744	1.1837
	6-O-sulfate E	-0.113428	-16.217698	16.929708	-0.017257925	0.5813
Structure 3	GlcNS6S E	-1.324558	-10.40416	12.539278	-0.178529674	0.632
	Ido2S D	-0.795402	-76.98578	73.92209	-0.255227861	-4.114
	GlcNS3,6S C	-2.482934	-88.210108	88.921224	-0.19269635	-1.965
	GlcA B	-2.06481	-48.01781	49.793296	-0.380465568	-0.67

	GlcNS6S A	-3.587294	-52.581182	55.575192	-0.768571992	-1.362
	6-O-sulfate A	-0.459228	-35.713552	37.626984	-0.286873301	1.1673
	6-O-sulfate C	-0.427192	-32.967002	34.623002	-0.058167734	1.1706
	3-O-sulfate C	-0.44977	-85.213384	81.721704	-0.456844622	-4.398
	2-O-sulfate D	-0.125008	-47.048892	47.692304	-0.173710642	0.3447
	6-O-sulfate E	-0.306512	-11.823726	13.198436	-0.096766934	0.9714
Average	GlcNS6S E	-1.095541	-14.980118	17.467958	-0.164581373	1.2277
	Ido2S D	-0.7541673	-74.963779	70.996246	-0.196693411	-4.918
	GlcNS3,6S C	-2.0200593	-78.268226	79.36041867	-0.143330107	-1.071
	GlcA B	-2.6280726	-53.573284	55.64115667	-0.400491288	-0.961
	GlcNS6S A	-4.3104013	-73.157729	71.78766133	-0.987454877	-6.668
	6-O-sulfate A	-0.7152506	-30.422281	32.781284	-0.283907635	1.3598
	6-O-sulfate C	-0.3908593	-37.268561	38.028468	-0.140512488	0.2285
	3-O-sulfate C	-0.255444	-77.163417	74.33479667	-0.366245453	-3.45
	2-O-sulfate D	-0.26708	-41.802431	42.82343267	-0.151198277	0.6027
	6-O-sulfate E	-0.1713413	-12.350782	13.263194	-0.044795242	0.6963

4.6 Discussion

There is no crystal structure of LAR complexed with heparan sulfate available to date; only one with sucrose octasulfate as an HS mimic (pdb 2YD8)[5]. Sucrose octasulfate is very different from a true HS oligomer; it has just two sugar residues, neither one occurring in HS. Fondaparinux is more representative of HS for at the least highly sulfated segments, sharing a repeating disaccharide of GlcNAc and IdoA or GlcA. The highly sulfated characteristic shared with sucrose octasulfate likely is responsible for the similarity in binding sites. In the sucrose octasulfate structure, three residues are within 1 Å van der Waals contacts: K68, K69 and R77. In our study, the docked structure of fondaparinux has some part within 1 Å of van der Walls

contact with residues K68, K69, K71, K72, S74, S75, R77, F78 and R100. The three residues identified in the sucrose octasulfate complex structure are also identified here. Six of the residues in contact with fondaparinux are positively charged. Fondaparinux is a highly negatively charged ligand. There are 8 different sulfation groups and two carboxylate groups distributed on both sides of the ligand, giving many opportunities for electrostatic interactions with positively charged protein residues. Examining the protein structure we find that all of the binding site residues, except F78 and R100, are located on loops. Significant motional flexibility of these loops could also facilitate numerous contributions to electrostatic interactions. The predicted average free energies of binding given by MM-GBSA calculations for the three $^{2}S_{0}$ conformations were, in fact dominated by the electrostatic component of the free energy of binding (GBELE). The flexibility of the loops combined with the flexibility of fondaparinux may contribute to our finding several conformations lying close in calculated free energies of binding (all within +/- 3.8 kcal).

It is useful for us to compare the results we got from our Robo1-HS tetramer study to the present research on LAR interaction with fondaparinux. A binding constant of 45 μM was obtained from the interaction of Robo1 and an HS tetramer which only contains two N-sulfates, two 6-O-sulfates and two carboxylates. Given that fondaparinux is a much more highly sulfated GAG and more positive protein residues are involved, a binding constant of 60 μM may suggest that fondaparinux is not an ideal ligand for LAR interactions. Some insight into which groups may be more or less important could arise from a per-residue decomposition of MM-GBSA free energies and an examination of the ligand portion of this decomposition. We treated the 2-O-sulfate 3-O-sulfate and 6-O-sulfates as separate residues in this decomposition so their contributions can be explicitly examined (see Table 4.8).

Note that, binding energy contributions from the sugar rings also come primarily from electrostatic interactions. The high electrostatic energy can, however, be compensated for by high desolvation energies and even result in a positive total energy in some cases. A similar phenomenon was observed in our study of Robo1 interacting with an HS tetramer. To be more specific, among the 5 decomposed sulfates in fondaparinux (three 6-O-sulfates, one 3-O-sulfate and one 2-O-sulfate), only the 3-O-sulfate on the middle GlcNS3,6S shows an average favorable total energy -3.45kcal/mol). All of the other sulfates, although having very large electrostatic contributions, have this more than offset by a higher desolvation penalty. Among the residual sugar rings, which contain either a negatively charged carboxylate, in the case of GlcA and IdoA residues, or N-sulfates, in the case of the three GlcNAc residues, only the non-reducing end GlcNAc consistently shows a favorable energy of interaction, while that of the reducing end is always unfavorable. This could of course result from a distortion caused by the unusual concentration of charge around the central GlcNAc; 3-O-sulfation is actually rare in native HS[42] . However, examining the current structure, one finds that Lys 68 is well positioned to interact with a 2-O-sulfate if the GlcA residue near the non-reducing end were replaced with IdoA2S. Extending the oligomer at the non-reducing with an additional IdoA2S may also add favorable contacts with Arg 97. These predictions of changes that may increase affinity are worthy of examination.

Previous studies have shown that heparan sulfate may induce LAR ectodomain oligomerization in solution. In our case, however, an average correlation time of 8.4 ns (see data in Table 4.6) suggests that the protein still remains monomeric in solution when it interacts with fondaparinux. This doesn't mean that longer oligomers couldn't induce aggregation. The results from previous research on RPTP σ binding with depolymerized heparin showed that dp8 was the

minimum length to assemble RPTP σ oligomers. In our model, if we orient the two domains with the C-terminus pointing down toward a hypothetical membrane and the N-terminus pointing up, the binding site is near the top of the two domain construct. LAR molecules could assemble along a charge-rich portion of an extended HS chain that is well outside the membrane surface. Also, with just a slight tilt one can have an HS chain insert between a pair of LAR molecules with its non-reducing end contacting R100, R77, R97 and pointing toward the membrane. In this geometry the HS chain could facilitate formation of either a 2:1 or a 2:2 LAR-HS complex with the HS chains contacting lysine and arginine residues on both LAR molecules. One does note that several of the HS sulfates of the monomer in Figure 4.12 do point out toward the solution and could complex with a second LAR molecule in a dimer structure. Among the residues that could contact sulfates normally exposed to solution are Lys 72 and Arg 63. A model for the 2:2 complex is shown in Figure 4.13. Examination of possibilities for multimer formation would require work with a set of longer, well-defined HS oligomers.

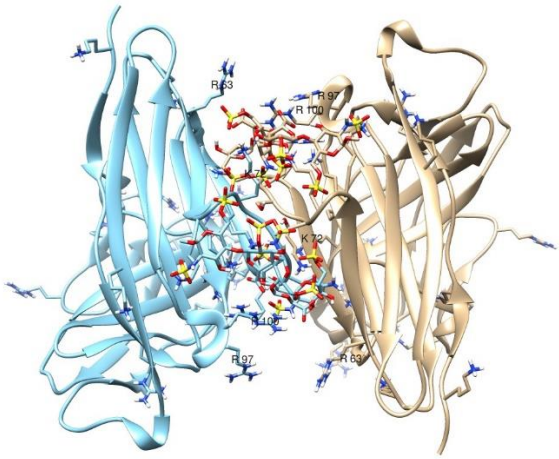


Figure 4.13. A model for LAR-fondaparinux in a 2:2 binding mode. The view is looking down toward the membrane surface from above the LAR molecules.

4.7 Conclusions

Biomolecular NMR measurements combined with high ambiguous docking has provided a detailed model for the interaction of a well-defined heparan sulfate pentasaccharide (fondaparinux) with the first two Ig domains of human LAR. Compared with a crystal structure where LAR is co-crystalized with sucrose octasulfate, the structure produces a more physically plausible set of molecular interactions, and is able to give some direction to future studies of LAR, its signaling mechanism, and its interaction with signaling partners, such as TrkC. The methodology developed and documented here, particularly that for the assignment of spectra from sparsely labeled proteins, opens application to a large number of glycosylated that are best expressed in mammalian cells. Many of the other type IIa RPTPs fall in this class.

4.8 References

1. Mohebiany, A. N.; Nikolaienko, R. M.; Bouyain, S.; Harroch, S., Receptor-type tyrosine phosphatase ligands: looking for the needle in the haystack. *Febs J* **2013,** *280* (2), 388-400.

2. Coles, C. H.; Jones, E. Y.; Aricescu, A. R., Extracellular regulation of type IIa receptor protein tyrosine phosphatases: mechanistic insights from structural analyses. *Semin Cell Dev Biol* **2015,** *37*, 98-107.

3. Fisher, D.; Xing, B.; Dill, J.; Li, H.; Hoang, H. H.; Zhao, Z. Z.; Yang, X. L.; Bachoo, R.; Cannon, S.; Longo, F. M.; Sheng, M.; Silver, J.; Li, S. X., Leukocyte Common Antigen-Related Phosphatase Is a Functional Receptor for Chondroitin Sulfate Proteoglycan Axon Growth Inhibitors. *J Neurosci* **2011,** *31* (40), 14051-14066.

4. Coles, C. H.; Mitakidis, N.; Zhang, P.; Elegheert, J.; Lu, W. X.; Stoker, A. W.; Nakagawa, T.; Craig, A. M.; Jones, E. Y.; Aricescu, A. R., Structural basis for extracellular cis and trans RPTP sigma signal competition in synaptogenesis. *Nat Commun* **2014,** *5*.

5. Coles, C. H.; Shen, Y.; Tenney, A. P.; Siebold, C.; Sutton, G. C.; Lu, W.; Gallagher, J. T.; Jones, E. Y.; Flanagan, J. G.; Aricescu, A. R., Proteoglycan-specific molecular switch for RPTPsigma clustering and neuronal extension. *Science* **2011,** *332* (6028), 484-8.

6. Dunah, A. W.; Hueske, E.; Wyszynski, M.; Hoogenraad, C. C.; Jaworski, J.; Pak, D. T.; Simonetta, A.; Liu, G.; Sheng, M., LAR receptor protein tyrosine phosphatases in the development and maintenance of excitatory synapses. *Nat Neurosci* **2005,** *8* (4), 458-467.

7. Takahashi, H.; Craig, A. M., Protein tyrosine phosphatases PTPδ, PTPσ, and LAR: presynaptic hubs for synapse organization. *Trends in neurosciences* **2013,** *36* (9), 522-534.

8. Tonks, N. K., Protein tyrosine phosphatases: from genes, to function, to disease. *Nat Rev Mol Cell Bio* **2006,** *7* (11), 833-846.

9. Mustelin, T.; Vang, T.; Bottini, N., Protein tyrosine phosphatases and the immune response. *Nature Reviews Immunology* **2005,** *5* (1), 43-57.

10. Takahashi, H.; Arstikaitis, P.; Prasad, T.; Bartlett, T. E.; Wang, Y. T.; Murphy, T. H.; Craig, A. M., Postsynaptic TrkC and Presynaptic PTP sigma Function as a Bidirectional Excitatory Synaptic Organizing Complex. *Neuron* **2011,** *69* (2), 287-303.

11. Capila, I.; Linhardt, R. J., Heparin–protein interactions. *Angewandte Chemie International Edition* **2002,** *41* (3), 390-412.

12. Raman, R.; Sasisekharan, V.; Sasisekharan, R., Structural insights into biological roles of protein-glycosaminoglycan interactions. *Chemistry & biology* **2005,** *12* (3), 267-277.

13. Brown, J. M.; Xia, J.; Zhuang, B. Q.; Cho, K. S.; Rogers, C. J.; Gama, C. I.; Rawat, M.; Tully, S. E.; Uetani, N.; Mason, D. E.; Tremblay, M. L.; Peters, E. C.; Habuchi, O.; Chen, D. F.; Hsieh-Wilson, L. C., A sulfated carbohydrate epitope inhibits axon regeneration after injury. *P Natl Acad Sci USA* **2012,** *109* (13), 4768-4773.

14. Arnold, J. N.; Wormald, M. R.; Sim, R. B.; Rudd, P. M.; Dwek, R. A., The impact of glycosylation on the biological function and structure of human immunoglobulins. *Annu. Rev. Immunol.* **2007,** *25*, 21-50.

15. Prestegard, J. H.; Agard, D. A.; Moremen, K. W.; Lavery, L. A.; Morris, L. C.; Pederson, K., Sparse labeling of proteins: Structural characterization from long range constraints. *J Magn Reson* **2014,** *241*, 32-40.

16. Gao, Q.; Chen, C. Y.; Zong, C.; Wang, S.; Ramiah, A.; Prabhakar, P.; Morris, L. C.; Boons, G. J.; Moremen, K. W.; Prestegard, J. H., Structural Aspects of Heparan Sulfate Binding to Robo1-Ig1-2. *ACS Chem Biol* **2016,** (DOI: 10.1021/acschembio.6b00692).

17. Prestegard, J. H.; Sahu, S. C.; Nkari, W. K.; Morris, L. C.; Live, D.; Gruta, C., Chemical shift prediction for denatured proteins. *J Biomol Nmr* **2013,** *55* (2), 201-9.

18. Feng, L.; Lee, H. S.; Prestegard, J. H., NMR resonance assignments for sparsely 15N labeled proteins. *J Biomol Nmr* **2007,** *38* (3), 213-9.

19. Nkari, W. K.; Prestegard, J. H., NMR Resonance Assignments of Sparsely Labeled Proteins: Amide Proton Exchange Correlations in Native and Denatured States. *J Am Chem Soc* **2009,** *131* (14), 5344-5349.

20. Qi Gao, G. R. C., Kelley W. Moremen, and James H. Prestegard, NMR Assignments of Sparsely Labeled Proteins Using a Genetic Algorithm *submitted*.

21. Barthelmes, K.; Reynolds, A. M.; Peisach, E.; Jonker, H. R. A.; DeNunzio, N. J.; Allen, K. N.; Imperiali, B.; Schwalbe, H., Engineering Encodable Lanthanide-Binding Tags into Loop Regions of Proteins. *J Am Chem Soc* **2011,** *133* (4), 808-819.

22. Aricescu, A. R.; McKinnell, I. W.; Halfter, W.; Stoker, A. W., Heparan sulfate proteoglycans are ligands for receptor protein tyrosine phosphatase sigma. *Mol Cell Biol* **2002,** *22* (6), 1881-1892.

23. Dominguez, C.; Boelens, R.; Bonvin, A. M. J. J., HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **2003,** *125* (7), 1731-1737.

24. Subedi, G. P.; Johnson, R. W.; Moniz, H. A.; Moremen, K. W.; Barb, A., High Yield Expression of Recombinant Human Proteins with the Transient Transfection of HEK293 Cells in Suspension. *JoVE (Journal of Visualized Experiments)* **2015,** (106), e53568-e53568.

25. Tjandra, N.; Grzesiek, S.; Bax, A., Magnetic field dependence of nitrogen-proton J splittings in N-15-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J Am Chem Soc* **1996,** *118* (26), 6264-6272.

26. Liu, Y. Z.; Prestegard, J. H., Direct measurement of dipole-dipole/CSA cross-correlated relaxation by a constant-time experiment. *J Magn Reson* **2008,** *193* (1), 23-31.

27. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., Nmrpipe - a Multidimensional Spectral Processing System Based on Unix Pipes. *J Biomol Nmr* **1995,** *6* (3), 277-293.

28. Goddard, T.; Kneller, D., SPARKY 3. *University of California, San Francisco* **2004,** *15*.

29. Woods, R., glycam Web. *Complex Carbohydrate Research Center. Athens, GA: University of Georgia* **2005**.

30. Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham Iii, T.; Darden, T.; Duke, R.; Gohlke, H., Amber 14. **2014**.

31. Case, D.; VB JTB, B. R.; Cai, Q.; Cerutti, D.; Cheatham III, T.; Darden, T.; Duke, R.; Gohlke, H.; Goetz, A.; Gusarov, S., The FF14SB force field. *AMBER* **2014,** *14*, 29-31.

32. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **2004,** *25* (13), 1605-1612.

33. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; Gonzalez-Outeirino, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: A generalizable Biomolecular force field. Carbohydrates. *J Comput Chem* **2008,** *29* (4), 622-655.

34. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts Chem Res* **2000,** *33* (12), 889-897.

35. Li, D. W.; Bruschweiler, R., PPM_One: a static protein structure based chemical shift predictor. *J Biomol Nmr* **2015,** *62* (3), 403-409.

36. Lipsitz, R. S.; Tjandra, N., Residual dipolar couplings in NMR structure analysis. *Annu Rev Bioph Biom* **2004,** *33*, 387-413.

37. Bax, A., Weak alignment offers new NMR opportunities to study protein structure and dynamics. *Protein Sci* **2003,** *12* (1), 1-16.

38. Bhunia, A.; Bhattacharjya, S.; Chatterjee, S., Applications of saturation transfer difference NMR in biological systems. *Drug discovery today* **2012,** *17* (9-10), 505-13.

39. Pederson, K.; Mitchell, D. A.; Prestegard, J. H., Structural Characterization of the DC-SIGN-Lewis(X) Complex. *Biochemistry-Us* **2014,** *53* (35), 5700-5709.

40. Hsieh, P. H.; Thieker, D. F.; Guerrini, M.; Woods, R. J.; Liu, J., Uncovering the Relationship between Sulphation Patterns and Conformation of Iduronic Acid in Heparan Sulphate. *Sci Rep-Uk* **2016,** *6*.

41. Coles, C. H.; Shen, Y. J.; Tenney, A. P.; Siebold, C.; Sutton, G. C.; Lu, W. X.; Gallagher, J. T.; Jones, E. Y.; Flanagan, J. G.; Aricescu, A. R., Proteoglycan-Specific Molecular Switch for RPTP sigma Clustering and Neuronal Extension. *Science* **2011,** *332* (6028), 484-488.

42. Thacker, B.; Lawrence, R.; Liu, J.; Esko, J. D., Heparan Sulfate 3-O-sulfation: A Rare Modification in Search of a Function. *Glycobiology* **2012,** *22* (11), 1648-1648.

CHAPTER 5

IMPROVING LANTHANIDE BINDING TAGS

5.1 Acknowledgement

The following work involves collaboration with Dr. Kelly Moremen's laboratory. Thanks go to Pradeep Prabhakar for expressing the Robo1-loop F protein. We also want to thank Dr. Daniel Häussinger from University of Basel and Dr. Gottfried Otting from Australian National University for sharing disulfide forming lanthanide binding chelates for our experimental explorations.

5.2 Abstract

Paramagnetic effects produced by lanthanide ions can provide rich structural information. They are particularly useful in protein structural investigations by NMR spectroscopy. In order to optimize paramagnetic effects in a target protein, a metal ion must be chelated in a fixed position without disrupting the protein. A lanthanide binding loop has been successfully inserted to two protein targets, and achieved reasonable paramagnetic effects in the previous work. However, there are still limitations regarding the loop stability and metal binding affinity. In this chapter, we discuss the efforts in improving the performance of lanthanide binding peptides by screening new constructs using MD simulation. In addition, an alternate approach of introducing the metal ion by using chelates carrying a sulfhydryl group has also been explored.

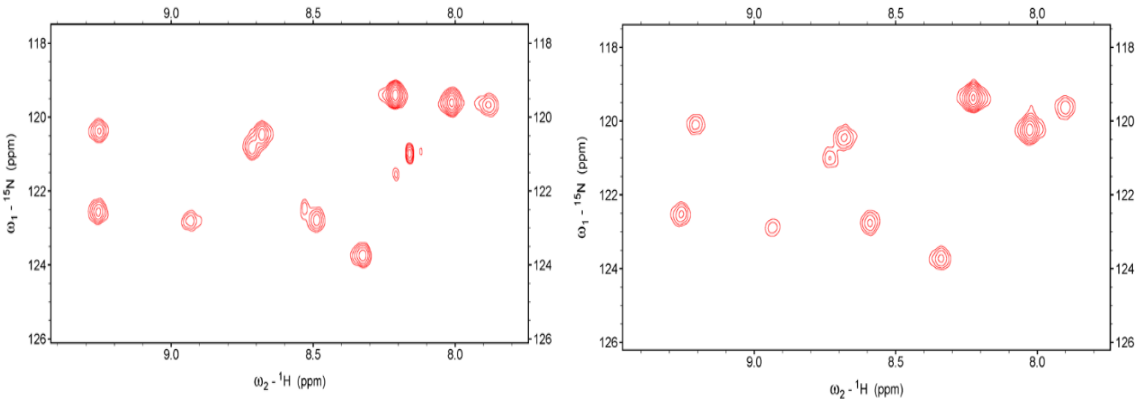
5.3. Methods and Results

Pseudo contact shifts (PCS) are very useful in that they perturb chemical shifts of resonances of both ligands and proteins in a distance and angle dependent manner. A detailed

description of these effects is given in chapter 2 where we have used them to advantage in making resonance assignments and docking a heparan sulfate ligand to Robo1. In order to introduce PCSs to a protein system, a paramagnetic metal site with a high magnetic susceptibility anisotropy ($\Delta\chi$) and short electron spin lifetime must be engineered into the protein. High $\Delta\chi$s also provide field-induced RDCs. Inducing RDCs in this way avoids the usual use of liquid crystal media to induce alignment which can sometimes have problematic interactions with proteins or ligands. It also has the advantage of inducing orientation from the point of one domain, reducing the number of unknowns needed to characterize a dynamic system. To accomplish metal site introduction for work described in chapter 2 we replaced a short peptide loop in the native structure with a lanthanide binding peptide. Using lanthanides with longer electron spin relaxation times (Gd^{3+}) paramagnetic relaxation effects (PREs) purely distance dependent effects arise. This was used to advantage in chapter 4. However, inserting these peptide loops can be challenging. They can interfere with ligand binding, destabilize protein structure or lose lanthanide binding affinity. Therefore, it is important to consider better ways to design loop insertions and completely different means of binding lanthanides.

Using a metal ion-binding chelate is an alternate approach that has been used to introduce a lanthanide into the protein. Chelates carrying a sulfhydryl group, for example, IDA-SH, NTA-SH [1] or DOTA-M8-SH [2], can be attached via a disulfide bond. A single native or mutated cysteine site is needed in order to apply this method. Quite often mammalian glycosylated proteins have at least one internal pair of disulfide bonds that is required to stabilize the fold and tertiary structure of the protein. Adding an additional cysteine and making this specifically reactive in chelate addition is often problematic. Nevertheless, we explored this option and results are described in this chapter.

Experimentally, two mutants (S203C and S162C) were produced separately for the mammalian glycosylated protein Robo1. The ^{1}H-^{15}N HSQC spectra of ^{15}N lysine labeled Robo1 mutants were nearly identical to the native protein, except for perturbation of a peak at 8.0/119.5 ppm that is assigned to K205 near and on the same side of the β-strand as S203. This indicates that the presence of the extra cysteine does not disrupt the protein structure except by very local contact (Figure 5.1).



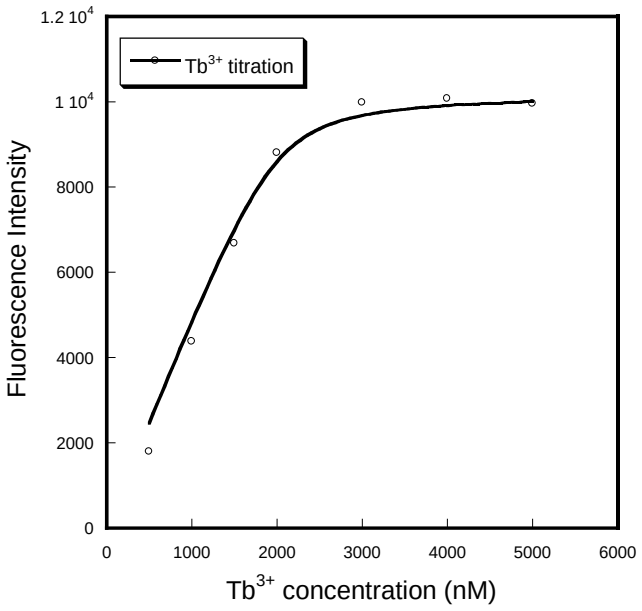
(A) (B)

Figure 5.1. ^{15}N-^{1}H HSQC spectra of (A) ^{15}N-Lys labeled Robo1 Ig1-2 S162C and (B) ^{15}N-Lys labeled Robo1 Ig1-2 S203C.

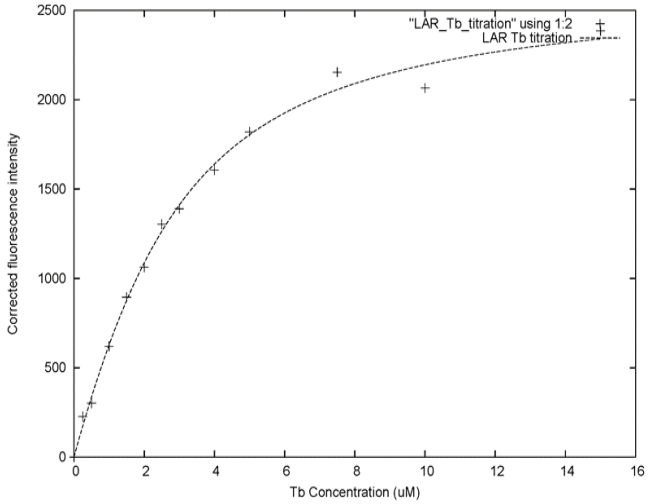
Disulfide bonds in proteins are formed by oxidation of the thiol groups of cysteine residues. The introduced cysteine can also be oxidized by the dissolved O_2 in the solution, which can reach a concentration as high as 0.4 mM when exposed to room temperature air. As a consequence disulfide-linked dimers can form. Therefore, the biggest challenge in successfully employing this ligation method is to reduce the mutant cysteine without breaking the native intra-molecular di-sulfide bonds in Robo1. Different treatments were explored to search for suitable conditions: All the buffers used during the expression process were degassed by

injecting argon into the buffer for a period of time to expel oxygen and an equivalent amount of reducing agent (TCEP used here) was added before attempting ligation of the chelate and filtered out by centrifugation. The level of free SH groups were monitored using Ellmans's reagent during these treatments to assure the existence of one free sulfhydryl. Despite these efforts, trials with ligation of the two mutants were not successful. Neither produced a ligated product. One Cys mutation (S162C) was likely too close to the glycosylation site and the other Cys mutation (S203C) is somewhat buried in a depression on the proteins surface. Either could have made the cysteine less reactive. What's more, enzymatic digestion and peptide analysis by mass spectrometry suggested that either the mutated cysteine had been partially oxidized or the disulfide bonds had been broken during the reduction process. These results reveal potential problems and uncertainties in using disulfide forming chemistry to introduce paramagnetic metal chelates into proteins that contain native disulfide bonds.

Another way to improve paramagnetic metal ion binding properties is to stay with lanthanide-binding peptides, but engineer shorter polypeptide-based lanthanide binding tags (LBTs), into the protein[3]. The polypeptide approach is advantageous because it can be encoded into a protein expression vector and expressed via standard molecular biology methods. The original LBTs were designed by the Imperiali laboratory[4-5]and they comprise 17- 23 amino acid including 6-8 carboxyl groups and a tryptophan. The tryptophan is used to induce luminescence of the bound Tb^{3+} through energy transfer and the disassociation constant can be obtained by luminescence monitoring during titration. An LBT with a sequence of SYIDTNNDGAYEGDELSG was originally engineered into glycosylated mammalian proteins, Robo1 and LAR. Using luminescent data (irradiating at 277 nm and observing at 542 nm a dissociation constant of 62 nM and 1.7 μM were determined respectively (see figure 5.2).



A



B

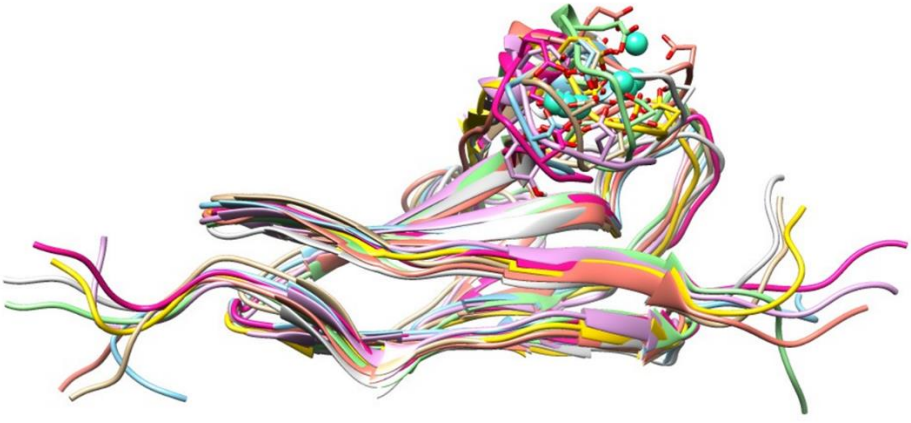
Figure 5.2. Tb^{3+} binding affinity of (A) Robo1 with the original LBT construct and (B) LAR with the original LBT construct.

There were sound reasons to introduce the LBT as a replacement for a native loop. Attachment of the peptide to either the N- or C-terminus of the protein usually introduces motional degrees of freedom between protein and tag that reduce PCSs and make both PREs and PCSs more difficult to interpret. Insertion into a loop usually reduces motional freedom, but this has its own challenges. It must be done without causing structural changes or unfolding of the protein. It also must be done without distorting the ion coordinating geometry and reducing ion affinity. This is not easy to accomplish since even 1.5 kcal of distortion can change affinities by an order of magnitude. To optimize loop insertion in Robo1 different insertion positions were tried, finally settling on replacing the short loop β-strands C and D of Ig domain 1. Also several loop modifications have been made on the original polypeptide-based LBT (SYIDTNNDGAYEGDELSG) in an effort to shorten the loop without disrupting the protein structure, but still reduce peptide motion and preserve ion binding affinity. These designs are shown in Table 5.1. To avoid some of the effort in expressing and characterizing multiple constructs, structural screening was done by running long (1 us) MD trajectories and examining the LBT peptide position every 100 ns of the trajectories. Superposition of these snapshots for some of the constructs is shown in Figure 5.3. The parameters chosen for the MD simulations are the same as those described in chapter 4 for the work on LAR. Some of the variations in peptide design included removing the end serines and a glycine. The glycine originally located on the protein was replaced by a valine to improve the coordination, and the tryptophan that extended into solution was moved to replace a tyrosine that seemed to make good contact with the protein surface.

Table 5.1. Designed LBT peptide constructs for Robo1 Ig1-2 domains.

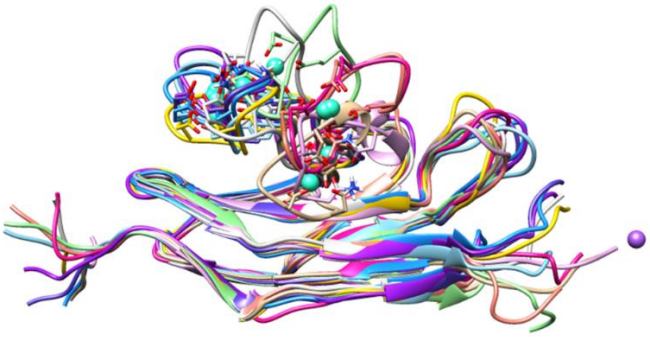
Design	Protein sequence	LBT peptide
Original sequence	NCKAEGRPTPTIEWYKG	SYIDTNNDGAYEGDELSG
Construct a	NCKAEGRPTPTIEWYKG	YIDTNNDGWYEGDEL
Construct b	NCKAEGRPTPTIEWYKG	SYIDTNNDGWYEGDELS
Construct c	NCKAEGRPTPTIEWYKG	YIDTNNDGWYEGDELS
Construct d	NCKAEGRPTPTIEWYKV	YIDTNNDGWYEGDELS
Construct e	NCKAEGRPTPTIEWYKV	YIDTNNDGWYEGDESS
Construct f	NCKAEGRPTPTIEWYKV	WIDTNNDGSYEGDELS

(Original construct)

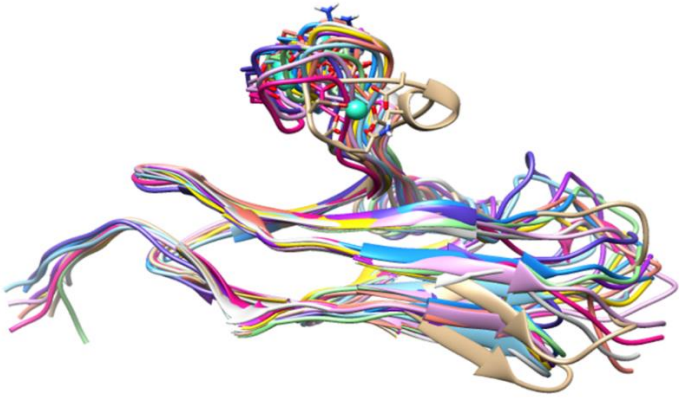


(Construct c)

(Construct d)



(Construct e)



(Construct f)

Figure 5.3. Superposition of different snapshots for some of the constructs. The MD simulations lasted for 1 μs. The constructs are designated: original, c, d, e and f. The sequences are given in Table 5.1.

Based on the MD simulation, the motion of the loop was reduced when construct a, c, d and f were used. There were interactions with the GAG binding loop located at the back of the LBT loop for construct c. The G44V mutation shifted the loop position to be away from the GAG binding loop in d and the coordination was improved, but in construct d Y45 was exposed in solution and the alpha helix on the loop was corrupted on construct e. In construct f, the metal ion was restricted along the entire simulation without any structural corruption or interaction. Hence, we inserted this LBT loop (construct f) to Robo1 Ig1-2 domain construct between strands C and D of the Ig1 domain to experimentally determine the metal binding affinity and resulting paramagnetic effects.

Luminescence data show the site to have a binding affinity of 3.3 μM (Figure 5.4), indicating a decreased affinity comparing the original loop affinity of 60 nM. The superimposed spectra of lysine labeled Robo1-Ig1-2-loopF with Dy^{3+} (a paramagnetic lanthanide) and Lu^{3+} (diamagnetic lanthanide) are shown in Figure 5.5. The unique diagonal shifts in peak positions are used to pair the resonances in each spectrum. Only moderate PCSs (< 0.1 ppm) were measured, possibly because the lanthanide ion has been moved further away from most labeled lysine sites, but also there may be residual motion not seen in the timescale of the MD simulation. These results suggest that MD may be of value in eliminating some constructs that clearly won't work, but precision and length of simulations are not sufficient to exclude loop motion and prevent modest alteration of ion binding constants. Constructs still need to be examined experimentally to determine the exact properties of the LBT loop and the paramagnetic effects. It may be possible that energy decomposition analysis would guide iterations in construct design that could improve stability an ion affinity further.

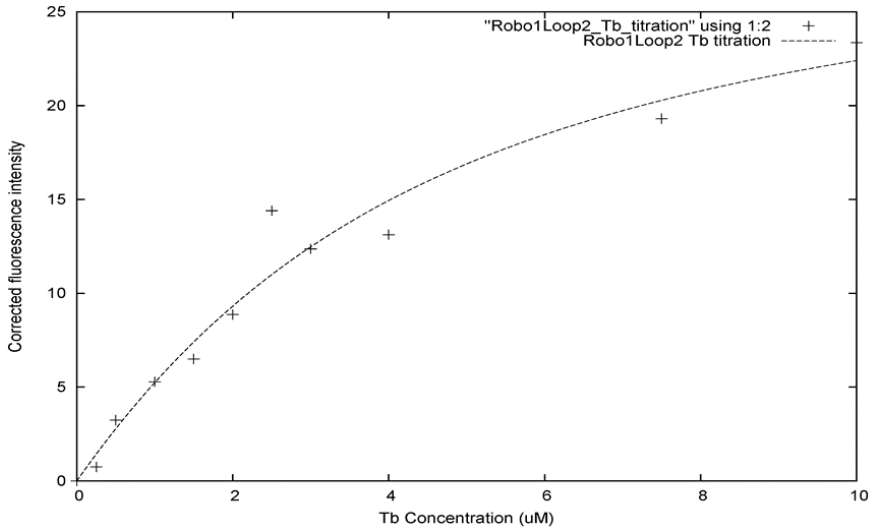


Figure 5.4. Tb^{3+} binding affinity of Robo1 LBT construct F.

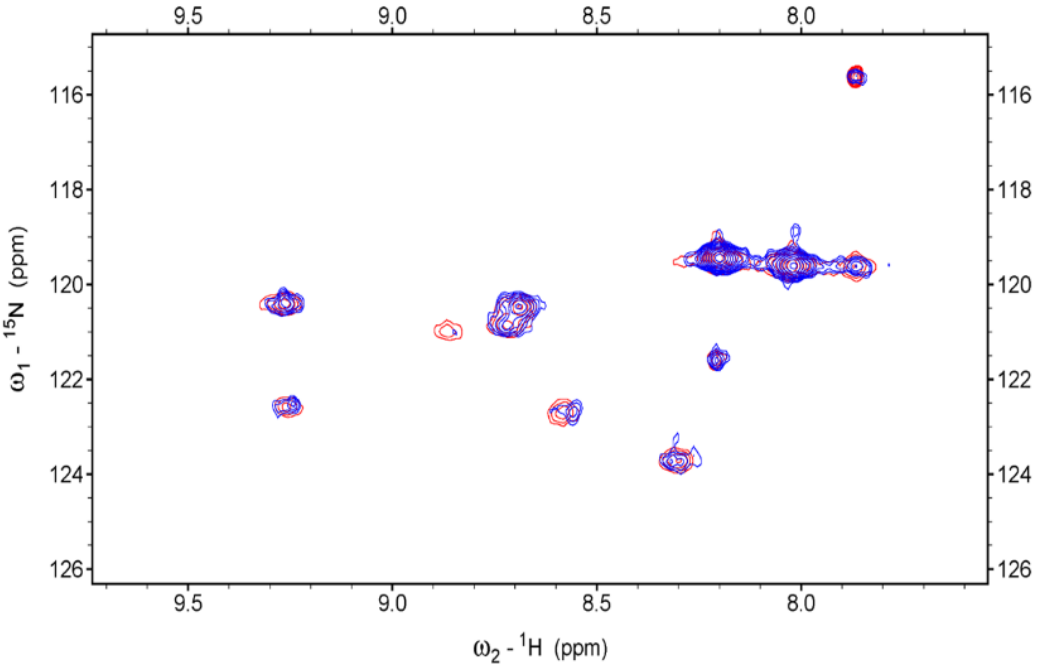


Figure 5.5. Superposition of ^{15}N-^{1}H HSQC spectra of ^{15}N-Lys labeled Robo1-Ig1-2 engineered with lanthanide binding peptide construct F loaded with Lu^{3+} (red) or Dy^{3+} (blue).

5.4 References

1. Yagi, H.; Maleckis, A.; Otting, G., A systematic study of labelling an alpha-helix in a protein with a lanthanide using IDA-SH or NTA-SH tags. *Journal of biomolecular NMR* **2013,** *55* (2), 157-66.

2. Haussinger, D.; Huang, J. R.; Grzesiek, S., DOTA-M8: An extremely rigid, high-affinity lanthanide chelating tag for PCS NMR spectroscopy. *J Am Chem Soc* **2009,** *131* (41), 14761-7.

3. Barthelmes, K.; Reynolds, A. M.; Peisach, E.; Jonker, H. R. A.; DeNunzio, N. J.; Allen, K. N.; Imperiali, B.; Schwalbe, H., Engineering Encodable Lanthanide-Binding Tags into Loop Regions of Proteins. *J Am Chem Soc* **2011,** *133* (4), 808-819.

4. Nitz, M.; Sherawat, M.; Franz, K. J.; Peisach, E.; Allen, K. N.; Imperiali, B., Structural origin of the high affinity of a chemically evolved lanthanide-binding peptide. *Angew Chem Int Edit* **2004,** *43* (28), 3682-3685.

5. Martin, L. L.; Sculimbrene, B. R.; Nitz, M.; Imperiali, B., Rapid combinatorial screening of peptide libraries for the selection of lanthanide-binding tags (LBTs). *Qsar Comb Sci* **2005,** *24* (10), 1149-1157.

CHAPTER 6

CONCLUDING REMARKS

Despite the prevalence of protein glycosylation in mammalian systems and the importance of glycan-protein interactions in signaling events, the structural characterization of glycosylated systems remains rare. This is primarily due to challenges in preparing glycoproteins and well defined glycan ligands. This thesis has focused on developing new solution-based NMR methodology to illuminate the structure and function of glycosylated mammalian proteins and their interactions with glycosaminoglycans. The application of these methods to terminal domains from two membrane anchored signaling proteins has begun to lay a solid foundation for understanding biomolecular functions, including cell signaling processes that begin with glycan-protein interactions at the cell surface.

The first phase of this research was the development of structural characterization approaches for glycosylated mammalian proteins and their application to Robo1. Robo1 is a cell surface signaling molecule important in axon guidance during mammalian development. Its interaction with heparan sulfate (HS) chains and members of the Slit protein family is essential to its activity, making characterization of these interactions by structural methods such as NMR highly desirable. From these studies, we obtained a detailed model of Robo1-Ig1-2 interacting with a heparan sulfate tetramer and an explanation for how heparan sulfate may facilitate the interaction between Robo1 and its signaling partner, Slit2. In the process, sparse labeling with NMR active isotopes was tested, along with a number of structurally sensitive NMR expetiments that can be applied to sparsely labeled systems. The methods used set a precedent which can

facilitate studies of a large number of other glycosylated protein complexes found on the surfaces of mammalian cells.

The second phase of research involved the development of a software package that facilitates NMR resonance assignment of sparsely labeled proteins. Sparse isotopic labeling of proteins for NMR studies offers advantages in spectral resolution and economical expression of glycoproteins in mammalian cells. However, a major limitation is that the one-bond connectivity between isotopically labeled sites is lost; making resonance assignment by traditional triple resonance approaches are not applicable. A particular milestone in our research is the development of a complete NMR resonance assignment strategy that does not rely on triple resonance experiments. We designed the "ASSIGNments for Sparsely Labeled Proteins (ASSIGN_SLP)" program, (built on a MATLAB platform), to use a genetic algorithm to search for an optimal paring of HSQC crosspeaks with labeled sites in proteins having known domain structures. Its objective function is based on differences between readily accessible experimental data and predictions from known domain structures. This assignment method is applicable to a number of large and glycosylated proteins that benefit from sparse isotope labeling.

The final stage of this research involved the application of the resonance assignment strategy and structurally sensitive NMR experiments to a second glycoprotein-GAG interaction system. Leukocyte common antigen-related (LAR) protein is a type IIa receptor protein tyrosine phosphatase (RPTP) that is important for signal transduction at the axon surfaces. Glycosaminoglycans are known to modulate LAR signaling. We successfully constructed a detailed model that demonstrates the interaction of a heparan sulfate pentasaccharide (fondaparinux) with LAR-Ig1-2 by combining various biomolecular NMR measurements with a docking procedure that uses highly ambiguous constraints. The assignments of crosspeaks from

the HSQC spectrum of sparsely labeled LAR were achieved using ASSIGN_SLP. This modeled structure offers important guidance for future studies of LAR with respect to its signaling mechanism and its interaction with signaling partners. This second successful application confirms the value of our approach for characterizing glycoproteins that are best expressed in mammalian cells.

Future structural studies of protein-GAG interactions would require longer and well-defined glycosaminoglycan chains to fully characterize binding sites, receptor oligomerization and interactions with other proteins. Expressing proteins with different glycan forms also would be important for demonstrating the full impact of glycosylation upon oligomerization and ligand interaction. The current work used protein with heterogeneous glycosylation in the Robo1 study and homogeneous glycosylation but with just a single GlcNAc in the LAR study. The number of different NMR measurements and the precision of those measurements can also be improved. Chapter 5 clearly outlined steps that could be taken to facilitate the measurement of paramagnetic constraints. These have not been added to ASSIGN_SLP yet but their addition should be straight forward. Labeling proteins with a larger number of amino acid types will also improve the accuracy of future models. With these additions, we believe that the approach described here will evolve into a robust procedure for the characterization of glycosylated proteins and protein-glycan complexes.