DEVELOPMENT OF METHODOLOGY FOR STRUCTURAL ANALYSIS OF GLYCOPROTEINS USING NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY AND MASS SPECTROMETRY

by

ROBERT V. WILLIAMS

(Under the Direction of I. Jonathan Amster and James H. Prestegard)

ABSTRACT

Glycoproteins are an important class of biomolecule involved in many processes of biomedical interest including cell communication, cancer progression, and pathogen-host interactions. Despite their importance, methods for structural analyses of glycoproteins have lagged behind other classes of proteins. This is largely due to the heterogeneous carbohydrate post-translational modifications typical of glycoproteins. Here, several advancements in methodology for studying glycoproteins with nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) are presented. For NMR, a new pulse sequence for measurement of residual dipolar couplings in ^{13}C-methyl groups using direct carbon observation has been developed. An updated program for resonance assignments of sparsely labeled proteins, Assign_SLP_GUI, has been developed. This program incorporates additional types of data, pseudocontact shifts (PCS) and paramagnetic relaxation enhancements, which can be obtained from paramagnetic tagging experiments. The sparse labeling assignment strategy, along with a structural interpretation of PCSs was then used to study the interdomain orientation of a

glycoprotein construct engineered with a lanthanide binding peptide sequence, hRobo1-Ig1-Ig2-LBP4, both with and without a bound ligand. PCS measurements were used to determine optimal interdomain orientations with and without bound ligand. Lastly, a combination of top-down and bottom-up mass spectrometry was used to measure the glycan occupancy of hCEACAM1-IgV with implications for the function of OST-B, one of the proteins responsible for the initial step of adding an N-glycan to a glycoprotein. Together, these results highlight the complementary nature of NMR and MS for the analysis of glycoproteins.

INDEX WORDS: Glycoproteins, NMR spectroscopy, Mass Spectrometry, RDC, PCS, Isotope Labeling, Native MS, Top-Down MS/MS,

DEVELOPMENT OF METHODOLOGY FOR STRUCTURAL ANALYSIS OF GLYCOPROTEINS USING NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY AND MASS SPECTROMETRY

by

ROBERT V. WILLIAMS

BA, OHIO WESLEYAN UNIVERSITY, 2012

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

2022

DEVELOPMENT OF METHODOLOGY FOR STRUCTURAL ANALYSIS OF GLYCOPROTEINS USING NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY AND MASS SPECTROMETRY

by

ROBERT V. WILLIAMS

Major Professor:	I. Jonathan Amster
	James H. Prestegard
Committee:	Jeffrey Urbauer
	Lance Wells

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

DEDICATION

To Robert H. Lease, my grandfather.

ACKNOWLEDGMENTS

The work described in this dissertation was completed with the help of many people. First and foremost, I would like to thank my co-advisors Dr. Amster and Dr. Prestegard for their advice and support over the last 6 years. I would also like to thank my other committee members Dr. Wells and Dr. Urbauer. It would have very difficult to perform any experiments without the help of my fellow lab members. On the NMR side I'd like to thank Laura Morris, John Glushka, Alex Eletsky, Monique Rogals, Qi Gao, and Joy Zhou. For their assistance in performing mass spectrometry expriments, I would like to thank Yuejie Zhao, Patience Sanderson, Lauren Pepi, Morgan Stickney, Franklin Leach, Brianna Garcia, Tanvir Ahmed, Yiqing Zhang, Jandi Kim, Elijah Roberts, and Johnathan Choi. I would also like to thank Dr. Kelley Moremen, Jeong-Yeh Yang, and Chin Huang for performing protein expression and purification. Finally, I would like to thank my other collaborators Dr. Linda Columbus and Connor McDermott.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Introduction

Glycoproteins are an important class of biomolecule that are typified by the posttranslational modification (PTM) of a polypeptide with a carbohydrate moiety, or glycan. Most secreted proteins are glycosylated.[1] Glycans are also a key component of the extracellular matrix and help drive communication between cells. Pathogens often exploit interactions with surface glycoproteins or matrix components when infecting a cell.[2] Additionally, changes in glycosylation patterns are often observed in cancer.[3] For these reasons understanding glycoprotein structure and the molecular basis of protein-glycan interactions is of great biomedical interest.

Structural studies of glycoproteins are often complicated by the heterogeneity of the carbohydrate modification. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy are two complementary techniques that are well suited for the study of glycoproteins. The results described within this dissertation include advances in NMR and MS methodologies for analyzing glycoproteins.

Background

While glycosylation comes in many flavors, the most widely studied of these protein modifications is N-glycosylation. N-glycans are found attached to asparagine residues by the formation of a C-N linkage between an N-acetylglucosamine (GlcNAc)

residue and the amide side chain of asparagine. A consensus amino acid sequence for N glycans exists: N-X-S/T. For an asparagine residue to be glycosylated it must be followed by a serine or threonine two residues away. The intervening amino acid may be of any type except proline. While this sequence is required for glycosylation, it is not sufficient. Many sequons are not modified or exist as a mixture of modified and unmodified forms. This initial variability is often called macroheterogeneity.[4]

The structure of the carbohydrate portion represents a second level of heterogeneity. All N linked glycans contain a core motif consisting of two N-acetylglucosamine (GlcNAc) and three mannose (Man) residues (Figure 1.1). This starting structure can be further elaborated with a variety of different monosaccharide residues. Three main categories of N-linked glycan have been defined: high mannose, complex, and hybrid (Figure 1.1).

N-glycans are initially attached to a glycoprotein as a branched $Glc_3Man_9GlcNAc_2$ moiety. Subsequent trimming of the terminal glucose residues and several mannose residues leads to the production of high mannose N-glycans. Additional processing of the α3 branch with a GlcNAc residue marks the simplest hybrid N-glycan. Further processing within the Golgi can lead to removal of two remaining mannose residues on the α6 branch and then further elaboration with a GlcNAc and additional monosaccharide residues, creating a complex N-glycan. Within each class there are many different possible glycans. Additionally, many different glycan structures are often observed at a specific glycosylation site on a protein. This second level of variability is termed microheterogeneity.[4]

NMR Spectroscopy

The heterogeneity of glycoproteins poses significant challenges for many structural characterization techniques, which often require highly purified material. Common methods include X-ray crystallography, cryogenic electron microscopy (cryo-EM), and NMR spectroscopy. Each method has its own advantages and disadvantages when applied to glycoproteins. The gold standard of protein structure determination is X-ray crystallography, which claims the largest number of structures deposited in the protein data bank (PDB).[6] Nonetheless, the requirement of crystallizing the analyte can prove problematic for any system. Glycosylation, in particular, can prevent the formation of crystals, in part due to glycan heterogeneity but also a result of the glycan's dynamic nature.[7] A number of structures of glycosylated protein complexes have recently been determined by cryo-EM.[8-9] In most cases, structural information on the carbohydrate is limited by their dynamic motion. Finally, NMR spectroscopy is a solution technique that can probe glycoproteins under native conditions. While glycan heterogeneity can pose obstacles, NMR is well equipped to directly monitor the behavior of the carbohydrates themselves.

One important limitation of NMR spectroscopy is the requirement for isotope-labeled material. Proteins are typically produced using recombinant protein overexpression in *E. coli*. These bacteria can be cultured in media where the sole nitrogen and/or carbon source is enriched with NMR active ^{15}N and/or ^{13}C isotopes, which leads to the production of uniformly isotope labeled protein. These labeling schemes then enable powerful 3D-NMR experiments that correlate signals between backbone hydrogen,

nitrogen, and carbon nuclei to make resonance assignments.[10] These triple resonance experiments form the basis for most biomolecular NMR structure determinations.

While powerful, this approach is not directly applicable to glycoproteins. Bacteria such as *E. coli* do not possess the necessary enzymes to glycosylate a recombinant protein. Several alternative expression systems have been explored for NMR spectroscopy including yeast, insect cells, and mammalian cells.[11-13] Overexpression in mammalian cell culture is the most appealing for human glycoproteins and has been used for NMR studies of glycosylated antibody fragments.[14-16] This comes with the drawback of more complicated isotope labeling procedures. Unlike bacteria, mammalian cells cannot synthesize all amino acids, and some must be included in the growth media. Producing these media with isotope labeled amino acids is quite expensive and limits its application. The costs can be made more reasonable by abandoning the uniform labeling strategy and instead only isotope labeling one or two amino acid residues. With this sparse-labeling approach, resonance assignment from triple resonance datasets is no longer possible and an alternative method is required.

The most straightforward approach to assignment of resonances from sparsely labeled proteins is one-by-one mutagenesis of each residue to a type not selected for labeling. This approach entails simple data analysis but is labor intensive. Instead, Prestegard and coworkers have developed an informatic approach that relies on comparing NMR data with predictions from a known structure.[17-19] Chemical shifts, residual dipolar couplings (RDCs), and nuclear Overhauser effect crosspeaks (NOEs) are readily measured on sparsely labeled glycoproteins. Several tools have been published which use an empirical approach to predict chemical shifts from a protein structure,

including ShiftX2 and ppmOne.[20-21] RDCs can be readily calculated from a set of atomic coordinates. NOEs can be estimated from a static structure but are more accurately predicted from a molecular dynamics trajectory.[22] Once all the necessary data and predictions are obtained, different possible assignments can be compared and scored by the agreement between experimental data and the predicted values. This process was incorporated into a series of MATLAB scripts named ASSIGN_SLP.[18]

Another challenge of the sparse labeling approach is that the reduced number of observed residues can limit the attainable structural information. Paramagnetic lanthanide tags are an appealing avenue to increase the amount of information. These ions introduce a number of structurally dependent effects that are readily measured from 2D NMR spectra of sparsely labeled proteins. These include pseudocontact shifts (PCSs), field-induced RDCs, and paramagnetic relaxation enhancements (PREs). Importantly, these observables are sensitive to changes in protein structure over long ranges of 20 – 40 Å, which greatly expands the utility of these methods for structural biology.[23]

PCSs arise when a molecule contains a paramagnetic ion with an anisotropic magnetic susceptibility tensor, $\Delta\chi$. Nuclei in the vicinity of the paramagnetic site experience a change in their chemical shift. The structural dependence of this phenomenon is well understood and described by the following equation:

Eq. 1.1 $$\Delta\delta^{PCS} = \frac{1}{12\pi r^3}\left[\Delta\chi_{ax}(3\,cos^2\,\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\,sin^2\,\theta\cos 2\phi\right]$$

The parameters $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ describe the magnitude and asymmetry of the magnetic susceptibility tensor in its principal frame, while r is the distance between a nucleus and the paramagnetic site. The paramagnetic lanthanides dysprosium, terbium, thulium, and erbium are commonly used for PCS measurements. PCSs are typically measured by

comparison of peak position with and without the paramagnetic ion, but replacement of the paramagnetic ion with a diamagnetic lanthanide such as Lu^{3+} provides a somewhat better reference state.

RDCs also arise on incorporation of paramagnetic lanthanide ions with anisotropic magnetic susceptibilities. All nuclear magnetic moments experience a strong, distance dependent dipolar interaction (dipolar coupling) with magnetic neighbors that gives rise to the broad lineshapes of solid-state NMR. In solution NMR this interaction is not observed due to the isotropic tumbling of analytes (and the lines are sharp). In the case of proteins binding a paramagnetic lanthanide, a slight preference in orientation is introduced which causes the dipolar interaction to no longer completely average to zero (hence residual). The equation describing the RDC between two nuclei, *i* and *j*, is as follows:

Eq. 1.2 $$D_{ij} = -\frac{hB_0^2\gamma_i\gamma_j}{240\pi^3 k_B T r_{ij}^3}\left[\Delta\chi_{ax}(3\,cos^2\,\theta - 1) + \frac{3}{2}\Delta\chi_{rh}\,sin^2\,\theta\cos 2\phi\right]$$

The form of this equation is similar to that for PCS, but in this case the distance r is between two nuclei. RDCs are most easily measured for directly bonded nuclei such as ^{1}H-^{15}N or ^{1}H-^{13}C spin pairs and manifest themselves as a change in the J-coupling. Determination of the RDC requires measurement under two conditions, typically paramagnetic and diamagnetic forms; however, the magnetic-field dependence can also be used if spectrometers operating at different field strengths are available.

A third effect of paramagnetic tags is that they cause NMR signals to decay more rapidly, which is known as PRE. This phenomenon has a strong inverse distance dependence, proportional to r^{-6} (nucleus to lanthanide distance). One complication of the PRE is nuclei in the vicinity of the paramagnetic site are often broadened beyond

detection. While all paramagnetic lanthanides induce PREs, gadolinium PREs are most structurally informative. This is due to the isotropic magnetic susceptibility tensor of Gd^{3+}, which avoids complications due to cross-correlation effects with other relaxation mechanisms present in other lanthanides.

Initially, the use of paramagnetic lanthanides was limited to metalloproteins, but advances in tagging methodology have broadened their applicability with many other systems. Chemical labeling is commonly used to attach macrocyclic Ln-chelates (e.g. DOTA(Gd^{3+})) to a free thiol via maleimide chemistry.[24] This generally involves introduction of a cysteine residue via mutagenesis, which can greatly complicate protein expression if disulfide bonds are already present. An alternative approach involves a genetically encodable lanthanide binding peptide.[25] This peptide (or tag) was engineered based on a calcium-binding motif using combinatorial peptide synthesis to discover a peptide with nanomolar lanthanide-binding affinity.[26] This optimized amino acid sequence can be inserted into the DNA construct for recombinant protein expression. Some care must be taken when choosing the insertion point in a protein sequence, but one study found that the tag could be inserted into three different loops of Interleukin-1β without disrupting the protein structure.[25] Similar tags have been used to study the interactions of glycosaminoglycan (GAG)-binding proteins.[27-28]

As a part of this thesis, paramagnetic effects have been incorporated into a resonance assignment workflow for sparse-labeled glycoproteins. A novel experiment to measure RDCs in methyl groups was developed and is described in Chapter 3. Additionally, a major rewrite of the AssignSLP software has been undertaken which

incorporates a graphical user interface and adds support for paramagnetic data. This software package, Assign_SLP_GUI, is described in chapter 4.

Mass Spectrometry

Mass spectrometry is widely used to determine the primary structure of proteins and peptides. Several approaches that provide structural information have been developed using bottom-up proteomics as a basis. These include crosslinking MS, hydrogen-deuterium exchange MS, and covalent labeling strategies.[29-31] More recently, methods have emerged that can maintain a protein's folded structure in the gas-phase allowing for more direct study. Termed native MS, this is achieved by electrospray ionization of a protein from aqueous solution conditions that maintain a protein's folded state. Biochemical buffers often contain high concentrations of nonvolatile salts and buffers which lead to intractable mass spectra. Instead, solutions for Native MS incorporate volatile salts, such as ammonium acetate. Native MS has been used to study ligand-binding, protein oligomerization, and protein complexes.[32] A key advantage is the ability to observe these interactions within a heterogeneous sample, such as a glycoprotein.

Mass spectrometry can more directly inform upon the structure of a molecule when combined with ion mobility spectrometry (IM-MS), a gas-phase technique that is sensitive to the size and shape of an ion. Somewhat analogous to electrophoresis in solution, IMS measures the speed with which ions travel through a drift tube under the influence of an electric field. Collisions with an inert buffer gas (typically nitrogen or helium) slow the progress of an ion. The number of collisions (and thus the travel time) an ion experiences is related to its size and shape.[33]

Traveling wave IM-MS has been used to study protein-glycosaminoglycan interactions.[34-36] Experimental details for performing these measurements are provided in Chapter 2. One of these studies found that the glycoprotein, roundabout1 (Robo1), samples two gas-phase conformations. The more compact form was favored in a Robo1-GAG complex. This conformation was significantly different than existing X-ray crystal structures and lead the authors to hypothesize that Robo1 experiences a conformation change upon GAG binding. The NMR methods developed in chapters 3 and 4 have been applied to determining whether this compact form exists in solution and the results are described in chapter 5.

In addition to studying the structure and function of glycoproteins, mass spectrometry is also well suited to the characterization of the glycan PTMs. Top-down mass spectrometry (TDMS) involves MS/MS of a protein without proteolysis. This approach is technically challenging and typically only feasible on FTMS instruments having ultrahigh resolving power. Despite the challenge, TDMS is appealing for the analysis of multiple PTMs. By keeping the protein intact, distinct combinations of PTMs (a.k.a proteoforms) can be observed and the location of their modifications determined through MS/MS. Unlike a bottom-up approach, which first reduces a protein to peptide fragments by proteolysis, information on correlation between two or more modifications is retained.

Collision-based activation methods often produce limited fragmentation of the most labile bonds for large peptides and proteins. In the case of glycoproteins, the glycosidic bonds are often cleaved more efficiently than the peptide backbone, which limits the ability to determine the location of a glycosylation site. Instead, TDMS often

relies on electron-based activation methods for fragmentation, such as electron capture dissociation (ECD) and electron transfer dissociation (ETD). These methods create a radical cation intermediate that leads to cleavage of the protein backbone between N and C_α atoms and the formation of c and z ions. The ECD and ETD processes are stochastic and lead to widespread cleavage across the protein sequence.[37] Importantly, ECD and ETD preserve labile PTMs such as glycans. TDMS has been applied to several glycoproteins, including monocolonal antibodies, RNAseB, and the SARS-COV-2 receptor-binding domain.[38-40]

As a part of this thesis, I have used TDMS to study the glycosylation occupancy of a model glycoprotein construct, the N-terminal immunoglobulin V-like domain of human carcinoembryonic cell adhesion antigen 1 (hCEACAM1-IgV). This protein is variably glycosylated and our results, described in chapter 6, have implications for the mechanism of glycan addition by oligosaccharyltransferase B (OST-B).

Together, the results presented in this dissertation describe recent progress in improving NMR spectroscopy methods for structural studies of glycoproteins. They also highlight the complementary nature of MS measurements, which can more easily monitor ligand-binding and PTM status.

References

1. Apweiler, R.; Hermjakob, H.; Sharon, N., On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1999,** *1473* (1), 4-8.

2. Szymanski C, S. R., Aebi M, Bacterial and Viral Infections. In *Essentials of Glycobiology*, 3rd ed.; Varki A, C. R., Esko JD, et al., Ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2017.
3. Pinho, S. S.; Reis, C. A., Glycosylation in cancer: mechanisms and clinical implications. *Nature Reviews Cancer* **2015,** *15* (9), 540-555.
4. Stanley P, T. N., Aebi M, N-Glycans. In *Essentials of Glycobiology*, 3rd ed.; Varki A, C. R., Esko JD, et al., Ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2017.
5. Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; Varki, A.; Woods, R. J.; The, S. D. G., Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **2019,** *29* (9), 620-624.
6. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000,** *28* (1), 235-242.
7. 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.
8. Ramírez, A. S.; Kowal, J.; Locher, K. P., Cryo-electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. *Science* **2019,** *366* (6471), 1372-1375.

9. Fan, X.; Cao, D.; Kong, L.; Zhang, X., Cryo-EM analysis of the post-fusion structure of the SARS-CoV spike glycoprotein. *Nature Communications* **2020,** *11* (1), 3618.
10. Ikura, M.; Kay, L. E.; Bax, A., A novel approach for sequential assignment of proton, carbon-13, and nitrogen-15 spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **1990,** *29* (19), 4659-4667.
11. Kamiya, Y.; Yamamoto, S.; Chiba, Y.; Jigami, Y.; Kato, K., Overexpression of a homogeneous oligosaccharide with 13C labeling by genetically engineered yeast strain. *Journal of Biomolecular NMR* **2011,** *50* (4), 397-401.
12. Walton, W. J.; Kasprzak, A. J.; Hare, J. T.; Logan, T. M., An economic approach to isotopic enrichment of glycoproteins expressed from Sf9 insect cells. *Journal of Biomolecular NMR* **2006,** *36* (4), 225-233.
13. Barb, A. W.; Meng, L.; Gao, Z.; Johnson, R. W.; Moremen, K. W.; Prestegard, J. H., NMR Characterization of Immunoglobulin G Fc Glycan Motion on Enzymatic Sialylation. *Biochemistry* **2012,** *51* (22), 4618-4626.
14. Kato, K.; Yamaguchi, Y.; Arata, Y., Stable-isotope-assisted NMR approaches to glycoproteins using immunoglobulin G as a model system. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2010,** *56* (4), 346-359.
15. Barb, A. W.; Falconer, D. J.; Subedi, G. P., Chapter Nine - The Preparation and Solution NMR Spectroscopy of Human Glycoproteins Is Accessible and Rewarding. In *Methods in Enzymology*, Wand, A. J., Ed. Academic Press: 2019; Vol. 614, pp 239-261.

16. Yanaka, S.; Yagi, H.; Yogo, R.; Onitsuka, M.; Kato, K., Glutamine-free mammalian expression of recombinant glycoproteins with uniform isotope labeling: an application for NMR analysis of pharmaceutically relevant Fc glycoforms of human immunoglobulin G1. *Journal of Biomolecular NMR* **2022**.
17. 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.
18. Gao, Q.; Chalmers, G. R.; Moremen, K. W.; Prestegard, J. H., NMR assignments of sparsely labeled proteins using a genetic algorithm. *Journal of Biomolecular NMR* **2017,** *67* (4), 283-294.
19. Chalmers, G. R.; Eletsky, A.; Morris, L. C.; Yang, J.-Y.; Tian, F.; Woods, R. J.; Moremen, K. W.; Prestegard, J. H., NMR Resonance Assignment Methodology: Characterizing Large Sparsely Labeled Glycoproteins. *Journal of Molecular Biology* **2019,** *431* (12), 2369-2382.
20. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR* **2011,** *50* (1), 43.
21. Li, D.; Brüschweiler, R., PPM_One: a static protein structure based chemical shift predictor. *Journal of Biomolecular NMR* **2015,** *62* (3), 403-409.
22. Chalmers, G.; Glushka, J. N.; Foley, B. L.; Woods, R. J.; Prestegard, J. H., Direct NOE simulation from long MD trajectories. *Journal of Magnetic Resonance* **2016,** *265*, 1-9.

23. Otting, G., Protein NMR using paramagnetic ions. *Annu Rev Biophys* **2010,** *39*, 387-405.

24. Nitsche, C.; Otting, G., Pseudocontact shifts in biomolecular NMR using paramagnetic metal tags. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2017,** *98–99*, 20-49.

25. 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. *Journal of the American Chemical Society* **2011,** *133* (4), 808-819.

26. Franz, K. J.; Nitz, M.; Imperiali, B., Lanthanide-Binding Tags as Versatile Protein Coexpression Probes. *ChemBioChem* **2003,** *4* (4), 265-271.

27. 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,** *11* (11), 3106-3113.

28. Gao, Q.; Yang, J.-Y.; Moremen, K. W.; Flanagan, J. G.; Prestegard, J. H., Structural Characterization of a Heparan Sulfate Pentamer Interacting with LAR-Ig1-2. *Biochemistry* **2018,** *57* (15), 2189-2199.

29. Leitner, A.; Faini, M.; Stengel, F.; Aebersold, R., Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences* **2016,** *41* (1), 20-32.

30. Trabjerg, E.; Nazari, Z. E.; Rand, K. D., Conformational analysis of complex protein states by hydrogen/deuterium exchange mass spectrometry (HDX-MS): Challenges and emerging solutions. *TrAC Trends in Analytical Chemistry* **2018,** *106*, 125-138.
31. Mendoza, V. L.; Vachet, R. W., Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrometry Reviews* **2009,** *28* (5), 785-815.
32. Tamara, S.; den Boer, M. A.; Heck, A. J. R., High-Resolution Native Mass Spectrometry. *Chemical Reviews* **2021**.
33. Uetrecht, C.; Rose, R. J.; van Duijn, E.; Lorenzen, K.; Heck, A. J., Ion mobility mass spectrometry of proteins and protein assemblies. *Chem Soc Rev* **2010,** *39* (5), 1633-55.
34. Zhao, Y.; Singh, A.; Li, L.; Linhardt, R. J.; Xu, Y.; Liu, J.; Woods, R. J.; Amster, I. J., Investigating changes in the gas-phase conformation of Antithrombin III upon binding of Arixtra using traveling wave ion mobility spectrometry (TWIMS). *Analyst* **2015,** *140* (20), 6980-9.
35. Zhao, Y.; Singh, A.; Xu, Y.; Zong, C.; Zhang, F.; Boons, G. J.; Liu, J.; Linhardt, R. J.; Woods, R. J.; Amster, I. J., Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. *J Am Soc Mass Spectrom* **2017,** *28* (1), 96-109.
36. Zhao, Y.; Yang, J. Y.; Thieker, D. F.; Xu, Y.; Zong, C.; Boons, G.-J.; Liu, J.; Woods, R. J.; Moremen, K. W.; Amster, I. J., A Traveling Wave Ion Mobility

Spectrometry (TWIMS) Study of the Robo1-Heparan Sulfate Interaction. *Journal of The American Society for Mass Spectrometry* **2018**.

37. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **1998,** *120* (13), 3265-3266.
38. Mao, Y.; Valeja, S. G.; Rouse, J. C.; Hendrickson, C. L.; Marshall, A. G., Top-Down Structural Analysis of an Intact Monoclonal Antibody by Electron Capture Dissociation-Fourier Transform Ion Cyclotron Resonance-Mass Spectrometry. *Analytical Chemistry* **2013,** *85* (9), 4239-4246.
39. Bourgoin-Voillard, S.; Leymarie, N.; Costello, C. E., Top-down tandem mass spectrometry on RNase A and B using a Qh/FT-ICR hybrid mass spectrometer. *PROTEOMICS* **2014,** *14* (10), 1174-1184.
40. Roberts, D. S.; Mann, M.; Melby, J. A.; Larson, E. J.; Zhu, Y.; Brasier, A. R.; Jin, S.; Ge, Y., Structural O-Glycoform Heterogeneity of the SARS-CoV-2 Spike Protein Receptor-Binding Domain Revealed by Top-Down Mass Spectrometry. *Journal of the American Chemical Society* **2021,** *143* (31), 12014-12024.

Figures

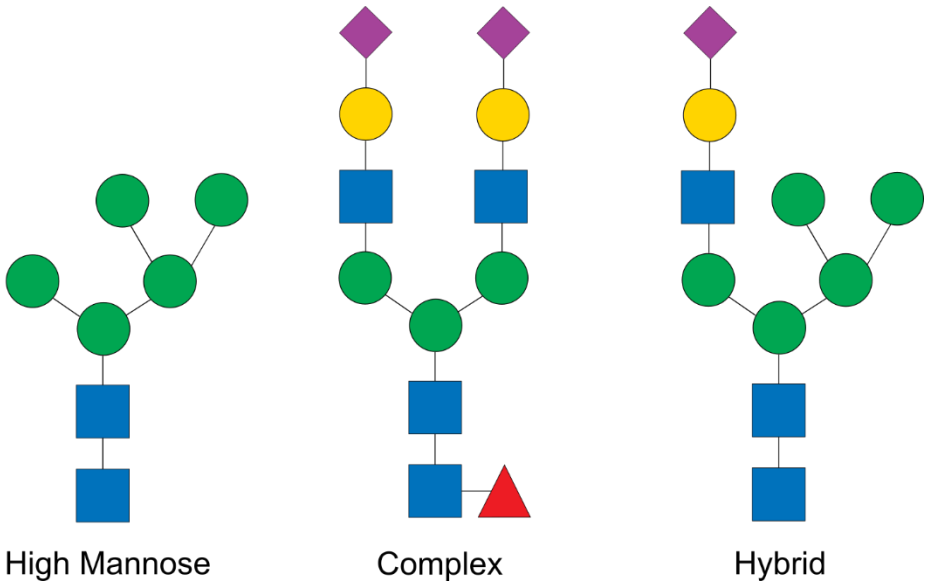

Figure 1.1. Example structures of the main classes of N-glycan. Monosaccharide residues are indicated using the SNFG representation scheme: blue squares represent GlcNAc, green circles represent mannose, yellow circles represent galactose, purple diamonds represent sialic acid, and a red triangle indicates fucose.[5]

CHAPTER 2

ANALYSIS OF PROTEIN-GLYCOSAMINOGLYCAN INTERACTIONS USING TRAVELING WAVE ION MOBILITY MASS SPECTROMETRY[1]

[1] Williams, R. V.; Amster, I. J. (2022) In: Balagurunathan K., Nakato H., Desai U., Saijoh Y. (eds) Glycosaminoglycans. Methods in Molecular Biology, vol 2303. Reprinted here with permission of publisher, 03/01/2022.

Abstract

Traveling wave ion mobility-mass spectrometry (TWIMS) combined with native mass spectrometry (MS) have emerged as a powerful tool for analyzing biomolecules, including complexes of protein and heparan sulfate (HS). This technique allows determination of the stoichiometry of the protein-HS interaction and information on the overall 3D molecular envelope.

Key Words

Traveling Wave Ion Mobility, Native Mass Spectrometry, Heparan Sulfate, Protein-glycosaminoglycan Interactions

Introduction

Traveling wave ion mobility spectrometry (TWIMS) coupled to mass spectrometry (MS) has recently emerged as a convenient tool for studying biomolecules and biomolecular complexes in their native state. A protein, protein-ligand complex, or multimeric protein complex can be introduced into the mass spectrometer in its biologically active folded state using native-spray conditions for electrospray ionization (ESI).[1,2] This approach has great utility for examining glycosaminoglycans (GAGs) and their interactions with target proteins.[3-5] Using native-spray conditions, noncovalent interactions are preserved during the transition from solution into the gas-phase. This allows protein-GAG complexes to be observed, thus providing information on the stoichiometry of the interaction and the relative affinity between different GAG

oligomers can also be assessed. These analyses are aided by the low sample requirements of mass spectrometry compared to other biophysical methods.

While native MS can be quite powerful by itself, the addition of ion mobility provides global structural information through the measurement of a collision cross-section (CCS).[6] The CCS is related to the three-dimensional size and shape of a GAG-Protein complex and can be extracted from the observed arrival time distribution. CCS values are readily calculated from a model structure for comparison.

Ion mobility also allows investigation of how a protein-GAG complex's arrival time distribution changes as a function of applied energy. Such collision induced unfolding (CIU) datasets depend upon the gas-phase stability of the ions of interest.[7-9] Evaluation of CIU heatmaps between different samples allows comparison of their relative stability and can be useful for comparing complexes of a given protein with several GAG ligands.

Materials

Sample Preparation

1. Native MS solution: 20 mM ammonium acetate, pH 6.8. Dissolve 0.15 g ammonium acetate in 90 mL water (*see* **Note 1**). Raise solution to 100 mL with water.
2. Centrifugal filters (*see* **Note 2**).

TWIMS

1. Synapt G2 (Waters).

2. SilicaTip NanoESI emitters (New Objective).

3. Syringe Pump (Harvard Apparatus).

Methods

Sample preparation

1. Using centrifugal filters, perform six rounds of buffer exchange into Native MS solution (*see* **Note 3**). Prepare a final protein solution of 100 µL at 100 µM concentration.
2. Combine protein and GAG to afford a 20 µM concentration of both species (*see* **Note 4**).

TWIMS

1. Infuse sample at a flow rate between 0.5-1.0 µL/min.
2. Set the capillary voltage to initiate electrospray (Typical values between 2 and 3.0 kV).
3. Optimize spectra by decreasing the capillary voltage in 0.1 kV steps (*see* **Note 5**).
4. Optimize gentle conditions by decreasing the sampling cone voltage (*see* **Note 6**).
5. Trap DC can also unfold protein ions, but also has a large impact on ion transmission. (Typical value = 45 V.)
6. Optimize TWIMS parameters by varying the IMS wave velocity and wave height. Typical values are 300 m/s wave velocity and 24 V wave height (*see* **Note 7**).

Collision Induced Unfolding

1. Switch instrument to MS/MS mode and Isolate m/z of interest.

2. Activate ions by increasing the Trap collision energy in identical increments (i.e., 5 V).
3. Record Ion Mobility Mass Spectrum at each voltage level (**Note 8**).
4. Continue until ion intensity is unacceptably low or the ion mobility arrival time distribution no longer changes.
5. A typical CIU series would run from 5 V to 70 V in 5 V increments (15 spectra).

Collision Cross Section Calibration

1. Record ion mobility mass spectrum of a sample of interest.
2. Record spectra of protein calibrants under identical ion mobility instrument settings (*see* **Note 9**).

Data Analysis

1. Third-Party Software
   (a) Vendor software included with commercial mass spectrometers is often not well suited to the analysis of Native MS (and TWIMS) datasets and the development of software has become an active area of research among many groups. The first step in workflows using third-party software is to convert the data from the vendor format to an open-source format.
   (b) The Ruotolo lab has developed the TwimExtract program to convert Waters RAW files into CSV format.[10] The program contains a graphical user interface (GUI) and allows for the extraction of the m/z or drift time dimensions of a TWIMS dataset.

2. Native MS

   Native MS spectra of proteins, GAGs, and protein-GAG complexes contain charge state distributions for the different species, some of which overlap. Manual interpretation is often possible, but software tools designed specifically for Native MS applications are available. Unidec is an open-source package for interpretation of Native MS data, which has many features that support batch processing of many spectra and can aid in high-throughput workflows.[11]

3. CIU

   CIU datasets are usually processed by smoothing functions and displayed as heatmaps. CIUsuite is a collection of python scripts that aid in the postprocessing and plotting of CIU datasets.[9] More recently, CIUsuite2 has been released, which incorporates a GUI and provides additional tools that aid in comparing multiple CIU datasets.[12]

4. CCS

   (a) The TWIMS measured drift time of a particular ion can be converted to a CCS value by first generating a calibration curve. Ruotolo *et al*. have explained this in detail.[6] Briefly, drift times of calibration ions are fit to a power law to generate a calibration curve.

   (b) The drift time is measured at the centroid of the gaussian IM peak. Curve fitting procedures are often beneficial when multiple overlapping peaks are present.

(c) CCS is most informative when compared against a model. Several software programs are available that calculate CCS from an X-ray crystal structure or the output of a molecular dynamics (MD) simulation, including MobCal, PSA, Collidoscope, and IMPACT.[13-16]

Notes

1. Ammonium acetate concentration can be varied between 10 mM to 200 mM. For monomeric protein-GAG complexes, 20 mM is sufficient. For multimeric protein-GAG complexes, 100 mM ammonium acetate would be a good starting solution.
2. The molecular weight cutoff should be approximately half of the molecular weight of the protein.
3. Buffer exchange is a critical step for native MS. Nonvolatile buffers and salts present in commonly used biochemical buffers are not MS friendly and give uninterpretable mass spectra. Centrifuge filtration is often a convenient method of buffer exchange into an ammonium acetate solution. Typically, six rounds of exchange are required to reduce salt concentrations to below 5 μM concentration.
4. A higher molar ratio of GAG to protein may be necessary to obtain strong signals for the protein-GAG complex. Typical conditions range from 1:1 to 4:1 GAG to protein.
5. Typically, optimum conditions are close to the minimum voltage required for ESI. Common settings are in the range of 1.5 – 2.0 kV.

6. Pay close attention to how changes to this voltage affect the ion mobility arrival time distribution. A shift to longer drift times indicates unfolding and the sample cone voltage is too high. There is often a tradeoff between mass spectral quality and maintaining a folded complex. (Typical Value = 30 V.)
7. Optimal IM settings depend on the size of the system being studied and could be quite different from those suggested here. Care should be taken to avoid the IM signal "rolling over." This occurs when ions take longer than a single injection cycle to traverse the IM cell.
8. Common protein calibrants include ubiquitin, cytochrome c, and myoglobin. Literature cross-sections are available from several sources.[6,17]

Acknowledgments

This work was supported by NIH grants P41GM103390 and T32GM107004.

References

1. Hernandez H, Robinson CV (2007) Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. Nat Protoc 2 (3):715-726. doi:10.1038/nprot.2007.73
2. Benesch JLP, Ruotolo BT (2011) Mass spectrometry: come of age for structural and dynamical biology. Current Opinion in Structural Biology 21 (5):641-649. doi:http://dx.doi.org/10.1016/j.sbi.2011.08.002
3. Zhao Y, Singh A, Li L, Linhardt RJ, Xu Y, Liu J, Woods RJ, Amster IJ (2015) Investigating changes in the gas-phase conformation of Antithrombin III upon

binding of Arixtra using traveling wave ion mobility spectrometry (TWIMS). Analyst 140 (20):6980-6989. doi:10.1039/c5an00908a

4. Zhao Y, Singh A, Xu Y, Zong C, Zhang F, Boons GJ, Liu J, Linhardt RJ, Woods RJ, Amster IJ (2017) Gas-Phase Analysis of the Complex of Fibroblast GrowthFactor 1 with Heparan Sulfate: A Traveling Wave Ion Mobility Spectrometry (TWIMS) and Molecular Modeling Study. J Am Soc Mass Spectrom 28 (1):96-109. doi:10.1007/s13361-016-1496-8
5. Zhao Y, Yang JY, Thieker DF, Xu Y, Zong C, Boons G-J, Liu J, Woods RJ, Moremen KW, Amster IJ (2018) A Traveling Wave Ion Mobility Spectrometry (TWIMS) Study of the Robo1-Heparan Sulfate Interaction. Journal of The American Society for Mass Spectrometry. doi:10.1007/s13361-018-1903-4
6. Ruotolo BT, Benesch JL, Sandercock AM, Hyung SJ, Robinson CV (2008) Ion mobility-mass spectrometry analysis of large protein complexes. Nat Protoc 3 (7):1139-1152. doi:10.1038/nprot.2008.78
7. Dixit SM, Polasky DA, Ruotolo BT (2018) Collision induced unfolding of isolated proteins in the gas phase: past, present, and future. Current Opinion in Chemical Biology 42:93-100. doi:https://doi.org/10.1016/j.cbpa.2017.11.010
8. Tian Y, Han L, Buckner AC, Ruotolo BT (2015) Collision Induced Unfolding of Intact Antibodies: Rapid Characterization of Disulfide Bonding Patterns, Glycosylation, and Structures. Anal Chem 87 (22):11509-11515. doi:10.1021/acs.analchem.5b03291
9. Eschweiler JD, Rabuck-Gibbons JN, Tian Y, Ruotolo BT (2015) CIUSuite: A Quantitative Analysis Package for Collision Induced Unfolding Measurements of

Gas-Phase Protein Ions. Anal Chem 87 (22):11516-11522. doi:10.1021/acs.analchem.5b03292

10. Haynes SE, Polasky DA, Dixit SM, Majmudar JD, Neeson K, Ruotolo BT, Martin BR (2017) Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. Analytical Chemistry 89 (11):5669-5672. doi:10.1021/acs.analchem.7b00112
11. Marty MT, Baldwin AJ, Marklund EG, Hochberg GKA, Benesch JLP, Robinson CV (2015) Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. Analytical Chemistry 87 (8):4370-4376. doi:10.1021/acs.analchem.5b00140
12. Polasky DA, Dixit SM, Fantin SM, Ruotolo BT (2019) CIUSuite 2: Next-Generation Software for the Analysis of Gas-Phase Protein Unfolding Data. Analytical Chemistry 91 (4):3147-3155. doi:10.1021/acs.analchem.8b05762
13. Shvartsburg AA, Jarrold MF (1996) An exact hard-spheres scattering model for the mobilities of polyatomic ions. Chemical Physics Letters 261 (1):86-91. doi:https://doi.org/10.1016/0009-2614(96)00941-4
14. Bleiholder C, Contreras S, Bowers MT (2013) A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (IV). Application to polypeptides. International Journal of Mass Spectrometry 354-355:275-280. doi:https://doi.org/10.1016/j.ijms.2013.06.011
15. Ewing SA, Donor MT, Wilson JW, Prell JS (2017) Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method.

Journal of the American Society for Mass Spectrometry 28 (4):587-596. doi:10.1007/s13361-017-1594-2

16. Marklund Erik G, Degiacomi Matteo T, Robinson Carol V, Baldwin Andrew J, Benesch Justin LP (2015) Collision Cross Sections for Structural Proteomics. Structure 23 (4):791-799. doi:https://doi.org/10.1016/j.str.2015.02.010
17. Bush MF, Hall Z, Giles K, Hoyes J, Robinson CV, Ruotolo BT (2010) Collision Cross Sections of Proteins and Their Complexes: A Calibration Framework and Database for Gas-Phase Structural Biology. Analytical Chemistry 82 (22):9557-9565. doi:10.1021/ac1022953

CHAPTER 3

MEASUREMENT OF RESIDUAL DIPOLAR COUPLINGS IN METHYL GROUPS VIA CARBON DETECTION[2]

[2] Williams, R. V.; Yang, J. Y.; Moremen, K. W.; Amster, I. J.; Prestegard, J. H. *Journal of Biomolecular NMR*. **73**, 191-198 (2019). Reprinted here with permission of the publisher, 01/31/2022.

Abstract

Residual dipolar couplings (RDCs) provide both structural and dynamical information useful in the characterization of biological macromolecules. While most data come from the interaction of simple pairs of directly bonded spin-1/2 nuclei (^{1}H-^{15}N, ^{1}H-^{13}C, ^{1}H-^{1}H), it is possible to acquire data from interactions among the multiple spins of ^{13}C-labeled methyl groups ($^{1}H_3$-^{13}C). This is especially important because of the advantages that observation of ^{13}C-labled methyl groups offers in working with very large molecules. Here we consider some of the options for measurement of methyl RDCs in large and often fully protonated proteins and arrive at a pulse sequence that exploits both J-modulation and direct detection of ^{13}C. Its utility is illustrated by application to a fully protonated two domain fragment from the mammalian glycoprotein, Robo1, ^{13}C-methyl-labeled in all valines.

Introduction

Isotope labeling of methyl groups simultaneously in all or a subset of isoleucine, leucine, and valine residues[1-3], or selectively in alanine[4,5] or methionine[6] residues, has emerged as a powerful method for studying large protein systems.[7,8] The improved sensitivity and resolution has been an important driver for this type of isotope labeling. Methyl proton observation is inherently more sensitive because of three equivalent protons contributing to a single signal; in addition, spin relaxation interference (methyl-TROSY effects) narrows lines, especially carbon lines in heteronuclear multiple quantum coherence (HMQC) spectra. This further improves sensitivity. However, when labeling only methyl groups, the reduced number of observable sites in a typical protein has made

the collection of a more diverse set of structural data types important. Among the data that can be collected are residual dipolar couplings (RDCs). RDCs provide information on bond vector orientation relative to the alignment frame of partially oriented molecules.[9] These can in turn be used to refine the global structure of a protein, facilitate resonance assignment, and probe domain orientation and dynamics. Interactions leading to methyl RDCs are potentially a little more complicated than those leading to the more commonly observed ^{1}H-^{15}N RDCs. Rapid methyl rotation leads to projection onto the axis of methyl rotation rather than a C-H bond and the C-H bond is slightly longer than the N-H bond, leading to a reduction by a factor of 0.27. However, this is compensated by the larger ^{13}C moment, resulting in RDCs that are about 70% the size of ^{1}H-^{15}N RDCs. Measurement is also a little more complex because of the increased multiplicity of coupled methyl systems. Here we introduce a method that avoids some of the complexities of dealing with methyl multiplets and yields methyl RDCs with reliable error estimates.

Measurement techniques for methyl RDCs have been developed in the past. Most rely on indirect detection in heteronuclear multiple quantum coherence (HMQC) or heteronuclear single quantum coherence (HSQC) style experiments and measurement of RDCs in a frequency domain.[10-12] Kontaxis and Bax[10] introduced a method based on a constant time (CT)-HSQC pulse sequence with spectral editing in the indirect dimension to produce 4 sub-spectra that could be combined to produce spectra containing a single component of the multiplet. More recent methods have taken advantage of the Methyl-TROSY effect in HMQC type experiments that dramatically improves resolution for large, perdeuterated proteins and protein complexes.[7,8] And, Sprangers and Kay[11]

implemented an HMQC-IPAP experiment that allowed measurement of the coupling constant from the direct ^{1}H frequency dimension. Recently, our lab has utilized an HMQC pulse sequence that introduced J-modulation (actually J+D, where D is the RDC contribution) in the first polarization transfer element in order to extract an RDC by fitting time domain data.[12] While J-Modulation experiments require collection of multiple points in a third time domain, the number of points can be small in number and selected to optimize sensitivity to the J-modulation frequency. Also, programs used to fit the data usually provide a direct estimate of error, something that is important in using RDCs for either resonance assignment or structure refinement.

Indirect detection of a heteronucleus through attached protons, as is done in an HMQC experiment, is not always optimal. These experiments include significant periods of time where proton magnetization is transverse. For large molecules proton transverse relaxation becomes very efficient and significant intensity is lost during these periods. In these cases, the loss may actually outweigh the sensitivity advantages of proton detection. This has been nicely illustrated by the Wagner group where direct detection of ^{15}N in large proteins using an ^{15}N-optimized NMR probe outperforms more conventional experiments using a proton optimized probe.[13,14] This is especially the case with proteins that are not perdeuterated, as transverse proton spin relaxation becomes even more efficient and interference effects leading to TROSY enhancements are outweighed by other relaxation pathways. A similar situation occurs with ^{13}C-methyl detection. Many of the proteins we study are glycosylated proteins that are best expressed in mammalian cell culture where perdeuteration is not an option. Hence, proton relaxation is efficient. Moreover, we frequently use the anisotropic magnetic susceptibility of paramagnetic tags

to induce the partial alignment needed for RDC measurement. This adds additional relaxation mechanisms that further enhance proton relaxation in comparison to ^{13}C relaxation.

Here, we propose an experiment for measuring RDCs that combines the inherent sensitivity of methyl groups with the favorable relaxation properties of carbon observation. We retain a J-modulation element due to the convenience of fitting data in the time domain and eliminate multiplet complexities from this element by making it constant time for evolution of all proton magnetization except that involving carbon. Application is to a ^{13}C-valine-methyl labeled construct containing the two N-terminal Ig-like domains of a fully protonated glycoprotein, Robo1. The N-terminal domain has a lanthanide-binding peptide replacing a short loop between two beta strands.[15] Comparison of paramagnetic (Dy^{3+} or Tb^{3+}) and diamagnetic (Lu^{3+}) complexes, as well as field dependence of the Dy^{3+} complex, is used to extract RDCs. The data are expected to elucidate the inter-domain geometry of this molecule and contribute an understanding of how it may transmit signals that govern axon guidance in developing nerve cells.

Experimental

Isotopic Enrichment, Protein Expression and Purification.

Design, expression, and purification of the two-domain construct of Robo1 containing a lanthanide binding loop has been described elsewhere.[15] Briefly, Robo1 was expressed in lec^- HEK297 cells grown in FreeStyle medium deficient in valine and supplemented with [^{13}C,^{13}C]-methyl-valine. Purified protein was treated with

endoglycosidase F, leaving a single N-acetylglucosamine residue at the one glycosylation site (confirmed by mass spectrometry).

Sample Preparation.

Solutions of the Robo1-Ig1-2-loop construct, ^{13}C labeled in all valine methyls, were prepared in 100% D2O, 25 mM tris buffer, 100 mM KCl, 4 μM DSS, 3 μM NaN3, pH 7.4 (uncorrected) at a final concentration of 300 uM. For the diamagnetic sample, an equimolar amount of lutetium chloride ($LuCl_3$) was added, while the paramagnetic samples were prepared using an equimolar amount of dysprosium chloride ($DyCl_3$) or terbium chloride ($TbCl_3$).

NMR Spectroscopy.

Spectra were recorded on a 600 MHz magnet with a Bruker DCH probe optimized for ^{13}C detection and an Avance III Console or on a 900 MHz magnet with a Bruker TCO probe optimized for $^{15}N/^{13}C$ detection and an Avance III Console. J-Modulated HETCOR data were collected with 2048 x 32 points and spectral widths of 6 and 3 ppm in the direct and indirect dimensions, respectively. All spectra were processed identically using nmrPipe.[16] In the direct dimension, data were zero-filled and apodized with a cosine-bell, while in the indirect dimension the number of points was increased to 64 using linear prediction, zero-filled and apodized with a cosine-squared-bell. Peak picking was performed in Sparky,[17] and integrals were taken using nmrPipe with a footprint fixed in size and position that captured about 80% of modulated crosspeak intensity.

Data Analysis.

J-modulation curves were extracted from the processed data and converted to text format using nmrPipe.[16] Coupling constants were extracted by curve fitting implemented in MATLAB. The curve was assumed to be of the form: $e^{(-R\tau)}\sin(\pi(J+D)2\tau - c)$, where (J+D) is the extracted coupling, τ is the variable modulation delay, R is a decay factor, and c is a phase offset associated with imperfections in the experimental setup; R, (J+D) and c are the fitting parameters.

Results

^{13}C-detected two dimensional ^{1}H-^{13}C spectra of our Robo1 construct (a HETCOR spectrum) are shown in Figure 3.1. Black crosspeaks are from a 300 µM sample with a diamagnetic ion (Lu^{3+}) in its lanthanide-binding loop; this spectrum was collected in approximately 3 hours at 600 MHz for protons on a 300 µL sample. The resolution, particularly in the ^{13}C dimension, is excellent for a fully protonated 25 kDa protein. There are 19 valines in the construct; 36 of the 38 expected peaks are observable. The sensitivity is also excellent and is very comparable to that of a spectrum collected at 600MHz with a proton optimized cryoprobe (manuscript in preparation). The red crosspeaks are from samples with a paramagnetic lanthanide ion in the binding loop (Dy^{3+} in the upper panel and Tb^{3+} in the lower panel). There are notable pseudo contact shifts (PCSs) in both cases. As shifts are approximately the same in both ^{1}H and ^{13}C dimensions, peaks in the presence of diamagnetic and paramagnetic ions can be connected by diagonal lines in most cases. The shifts are slightly different for Dy^{3+} and

Tb^{3+} due to the different magnetic susceptibility anisotropy tensors for the two paramagnetic complexes.[18] Also, the paramagnetically shifted peaks are missing in several cases due to the enhanced relaxation. However, the loss of peaks is significantly less than in corresponding HSQC or HMQC spectra, because of the absence of an additional ^{1}H observation period where paramagnetic contributions to transverse relaxation are larger than for ^{13}C.

There are some subtilties to collection of the spectra in Figure 3.1. Transfer of proton magnetization at the end of the t_1 period results in antiphase components for the four proton-coupled ^{13}C lines. A simulation in the absence of paramagnetic relaxation contributions for the spectrum at this point using MATLAB scripts that calls functions from the SPINACH package[19] is shown in Figure 3.2. The behavior of an isolated $^{13}C^{1}H_3$ methyl group with rapid rotation about its symmetry axis and tumbling with the correlation time of a protein was simulated by giving it an on-axis rotational correlation time of 10 ps and an off-axis correlation time of 10 ns. Note that the inner multiplet components are sharper because of the partial interference of α and β spins in the mixed proton states for these lines. The relative integrals are actually 1:1:1:1 because the experiment begins with proton magnetization and there is three times the magnetization transferred to the outer lines, compensating for the three-fold degeneracy of the inner lines. Immediate decoupling of protons would result in zero signal. Full refocusing of the ^{13}C magnetization before decoupling is not possible but decoupling after a total refocusing delay of 1/3J results in a single line of nearly twice the intensity of the two sharper central lines with no contribution from the outer lines. A slightly shorter delay results in maximum refocusing of intensity in the limit where all lines have equal

widths.[20] Note that RDCs can in principle be measured by acquiring a coupled spectrum immediately after transfer of proton magnetization. However, turning each crosspeak of Figure 3.1 into the four crosspeaks of Figure 3.2 with a mix of positive and negative intensities would be very hard to deconvolute.

There have also been some attempts to measure RDCs directly in the proton dimension.[11] To illustrate the complexities of this approach we have simulated a ^{13}C-coupled proton spectrum using SPINACH functions (see Figure 3.3). As for Figure 3.2, rapid methyl rotation was simulated using anisotropic diffusion. The effects of partial orientation were introduced by defining an order tensor with principal values of 2e-4, 2e-4 and -4e-4 and the z-axis parallel to the methyl rotation axis. Note that, unlike a system with just scalar coupling, which would have two single lines, the presence of proton-proton dipolar couplings produces a ^{1}H proton spectrum with two triplets. Inner lines of the triplets are sharp due to the interference of relaxation contributions from pairs of α and β 1H spins. The outer lines of the triplet are of unequal widths due to interference with relaxation contributions from the ^{13}C spin. Again, measurement of RDCs from the positions of six superimposed crosspeaks for each crosspeak shown in Figure 3.1 would be difficult. Hence, we have chosen to design a pulse sequence that yields a simple time modulation of the intensity of each crosspeak seen in Figure 3.1.

Sequence design.

The sequence we generated is shown in Figure 3.4. It is a modification of a standard HETCOR sequence in which a constant time J-modulation period has been inserted between the end of the ^{1}H t_1 evolution period and transfer to ^{13}C by a pair of ^{1}H

and ^{13}C 90° pulses. Making the J-modulation period, T, constant time removes modulation due to proton-proton couplings. Insertion of a 180° carbon pulse at a time τ before the end of the modulation period allows modulation due to carbon-proton couplings for times that vary from 0 to T as τ is incremented from 0 to T/2. Ideally, the value of T should be long enough to allow at least two zero crossings of the coupling evolution as points near these crossings are most sensitive to the modulation frequency ($^1J_{CH}+^1D_{CH}$ in the presence of an RDC). However, measured intensities, and signal to noise ratios, decay as exp(-R2T) as T is increased, limiting the number of zero crossings that can be accessed. In our application the constant time period, T, was set to 32 ms (4/J for J = 125 Hz) and τ was incremented from 2 to 16 ms in steps of 0.93 ms. The refocusing delays at the end of the sequence allow the antiphase lines of the ^{13}C multiplet to come partially in phase so protons can be decoupled during observation. τ_R was set to 0.8 ms to optimize signal intensity with some sacrifice in resolution over collection of just the inner lines of the ^{13}C multiplet.

A methyl group is an IS_3 spin system and is not rigorously described by the common product operator treatment used to describe experiments on ^{15}N-^1H spin pairs; however, it has been shown elsewhere[7] that the eigen basis contains a spin-1/2 manifold, and for the purposes of describing the expected J-modulation a spin pair product operator treatment is adequate. For this description we will only consider magnetization that is ultimately converted to an observable form. During the indirect evolution period the C-H coupling is refocused and ^1H chemical shift evolution gives $-\cos(\omega_I t_1)\,\hat{I}_y$ at time point (a) in Figure 3.4. At the end of the constant-time INEPT period this magnetization becomes $-\cos(\omega_I t_1)\sin(\pi(J+D)2\tau)\,\hat{I}_z\hat{S}_x$ (Figure 3.4 - point b). The final refocusing

period converts most of the I_zS_x coherence to S_y for observation. One thing that is not included in this treatment is spin relaxation. Proton transverse relaxation effects are largely eliminated by using a constant time. However, ^{13}C T_1 relaxation dephases modulation and results in a slow exponential decay. Also, the 180° ^{13}C pulse that interconverts α and β states interchanges the broad and narrow components of the proton triplets seen in Figure 3.3, potentially resulting in a small rise of intensities at longer τ values. Both of these effects are well modeled in SPINACH simulations which do show a slight decay of the sinusoidal modulation. We model this by including an exponential term exp(-Rτ) in our fitting function. One possible effect not included in the simulation is that of a small three bond ^1H-^{13}C coupling to the second valine methyl group (~4 Hz). In the interest of generality, we have not explicitly included this in our derivation. However, its small effects are likely absorbed in the exponential fitting constant, R and phase correction, c. Hence, after Fourier transform and measurements of crosspeak integrals as a function τ, intensities, I, are fit to equation 3.1.

$$\text{Eq. 3.1.} \qquad \sin(\pi(J+D)2\tau + c)\exp(-R\tau)$$

RDC Measurement.

The above discussion applies to data acquired with any means of alignment. However, our Robo1 construct contains a lanthanide ion binding loop. When paramagnetic ions like Dy^{3+} and Tb^{3+} are present, field induced alignment can be utilized. Valine $^{13}C\gamma$-^1H RDCs were measured on a fully protonated sample of our 25 kDa Robo1 construct using alignment by Dy^{3+} and Tb^{3+} at two magnetic field strengths, 21T for 900 MHz data and 14T for 600 MHz. Fitting of modulation data in the presence of these ions

gave values for (J+D). A third measurement with the diamagnetic ion (Lu^{3+}) provided a measurement of J. Since we do not expect J to vary significantly from site to site or with magnetic field, the average taken from the most precise measurements (those from peaks 19 and 33) was used to evaluate J (125.8 Hz) and extract RDCs from (J+D) measurements. In the diamagnetic sample, 36 peaks were observed (Figure 3.1); Enhanced relaxation in the samples containing paramagnetic ions caused several peaks to become greatly attenuated or disappear entirely. Also, pseudocontact shifts caused additional overlap of peaks, preventing some measurements. Of the original 36 peaks, only 10 allowed extraction of RDCs from the Dy^{3+} sample and only 13 allowed extraction of RDCs from the Tb^{3+} sample. The extracted RDCs are reported in Table 3.1; for the Dy^{3+} complex at 21T they range from approximately -3 to +6 Hz, and for the Tb^{3+} complex they range from approximately 0 to 7 Hz. Figure 3.5 shows the modulation curves of peak 16 in the diamagnetic and paramagnetic Dy^{3+} samples at 21T (900 MHz). This clearly shows the difference in modulation frequency that we attribute to an RDC.

Collection of RDCs at more than one field strength can be useful. There can be field dependent broadening of resonances due to modulation of PCSs by internal motion, making some crosspeaks broader and less intense at higher fields; the known field dependence can also be used to extract an RDC in the absence of a separate measurement of scalar couplings or, when scalar couplings are known, compare results as an additional check on data quality. Both the induced magnetic moments and the interaction energies depend linearly on magnetic field, so RDCs depend on field squared.[1] RDCs are therefore expected to be 2.25 times as large at 21 T as 14 T. In Table 2.1 we have scaled up the 600 MHz RDC values and their estimated errors by this factor and entered this in

columns 4 and 7 so that direct comparisons of data collected at the two field strengths can be made. Three of the 21 possible comparisons lie more than 1 Hz outside the sum of our estimated errors (peak 25 for $Dy^{3+}$and 8 and 16 for Tb^{3+}). All three peaks are in crowded regions of the spectrum and the curve fitting procedure may be negatively impacted by signal overlap. A more sophisticated peak fitting approach that simultaneously fits multiple peaks could improve the quality of extracted intensities in these instances. The estimated errors in many cases seem large compared to the RDCs. However, it is important to point out that in applications such as crosspeak assignment or protein domain orientation, it is the error relative to the range of RDC measurements that counts. The ratio of the average error relative to the range for Dy^{3+} and Tb^{3+} are 0.16 and 0.22 respectively. This precision proves to be useful for many applications.

Discussion

The above data demonstrate J-modulation combined with ^{13}C detection to be a viable method of measuring ^{13}C-methyl RDCs. These RDCs are useful both for assignment of resonances and monitoring of structural changes. For proteins such as Robo1, which is natively glycosylated and best expressed in mammalian cell culture, assignment procedures based on uniform isotopic labeling are often not an option. We have recently introduced a procedure for assignment that uses comparison of easily acquired data, such as NOEs, chemical shifts and RDCs, to predictions based on known domain structures.[2-4] Utilization of the RDCs collected with our ^{13}C-detected J-modulation experiment in assignment of the valine methyl resonances of our Robo1 construct is a high priority moving forward. Once assigned, there are emerging questions

about the inter-domain orientation in our two-domain construct of Robo1. The RDCs we have measured, as well as the PCSs, will be useful in assessing inter-domain geometry.

Our Robo1 construct is not a particularly large protein, but even here ^{13}C-observe methods appear competitive with ^{1}H-observe methods. This is in part a consequence of an inability to perdeuterate natively glycosylated systems. We anticipate that the advantages of ^{13}C-detection methods will only increase as protein targets become larger and the number of glycosylated systems under investigation increases.

The pulse sequence and data fitting protocols presented here are likely to be refined in the future. We have already mentioned the possibility of explicitly including the effects of long-range couplings, such as that from protons of the methyl group observed to the ^{13}C of the second intra-residue valine methyl group, in our fitting function. For methyl labeling, this is a unique problem arising from ^{13}C-enrichment in both the methyls of the valine we used. However, similar problems can arise with leucine, and if uniform labeling of other methyl-containing amino acids is used (uniformly labeled alanine, for example), it will arise there as well. In the latter case, substitution of a selective pulse for the hard 180° pulse in our modulation period may eliminate the problem with long-range couplings.

We have also not collected modulation data in the most time-efficient way. First, we have picked a rather long modulation period for the purpose of showing a more complete modulation curve. In principle, two zero crossings are likely to be adequate; this corresponds to $T = 2/J = 16$ ms, a factor of 2 shorter than that used in our illustration. While the shorter modulation time would sacrifice some precision, we would expect a substantial increase in signal to noise of each point measured. Proton transverse

relaxation is occurring during our modulation period and based on proton line widths from HSQC spectra (not shown), we estimate T_2 to be roughly 11 ms. Assuming mono-exponential decay during the J-modulation element, the signal intensity would be approximately 4 times greater with a 16 ms delay than with 32 ms delay used in our data.

Also, there is no need to collect equally incremented modulation points. Assuming that noise is constant as a function of τ, the contribution a measurement at a particular τ value can make to the accuracy of J (or $J+D$) should be proportional to the variation of intensity with J. This variation is given in equation 3.2, Where the effect on intensity as a function of transverse relaxation (R_2) during the total modulation period, T, is also included.

Eq. 3.2 $$\frac{\partial I}{\partial J} = \pi 2\tau e^{-R_2 T} \cos[\pi J 2\tau]$$

It is clear from equation 3.2 that contributions would have local maxima (or minima) at 2τ = n/J (zero crossings of Figure 3.5, where J = 125 Hz). It is also clear that values at these points would rise linearly with τ, but decrease exponentially as T was made longer to accommodate additional zero crossings. For our estimate of R_2 contributions would drop off rapidly after the first two zero crossings. To account for other factors, including phase offsets and decay as τ is incremented, it would be advisable to also place a few of the observation points near the first maximum and minimum of the curve and clustering the rest near the first two zero-crossings. With these improvements in mind, we expect that applications to protein systems significantly larger than our Robo1 construct should be possible.

Conclusion

We have designed a carbon detected pulse sequence for measurement of methyl ^{1}H-^{13}C RDCs, which uses an intensity modulation scheme, and we have demonstrated this pulse sequence on a fully protonated, two-domain construct from the glycoprotein, Robo1. Our proposed design allows the efficient collection of RDCs and is broadly applicable to any system containing isotopically labeled methyl groups.

References

1. Goto, N.K., Gardner, K.H., Mueller, G.A., Willis, R.C. & Kay, L.E. A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated N-15-, C-13-, H-2-labeled proteins. *Journal of Biomolecular NMR* **13**, 369-374 (1999).
2. Tugarinov, V. & Kay, L.E. An Isotope Labeling Strategy for Methyl TROSY Spectroscopy. *Journal of Biomolecular NMR* **28**, 165-172 (2004).
3. Lichtenecker, R.J. et al. Independent valine and leucine isotope labeling in Escherichia coli protein overexpression systems. *Journal of Biomolecular NMR* **57**, 205-209 (2013).
4. Ayala, I., Sounier, R., Usé, N., Gans, P. & Boisbouvier, J. An efficient protocol for the complete incorporation of methyl-protonated alanine in perdeuterated protein. *Journal of Biomolecular NMR* **43**, 111-119 (2009).
5. Godoy-Ruiz, R., Guo, C. & Tugarinov, V. Alanine Methyl Groups as NMR Probes of Molecular Structure and Dynamics in High-Molecular-Weight Proteins. *Journal of the American Chemical Society* **132**, 18340-18350 (2010).

6. Fischer, M. et al. Synthesis of a 13C-Methyl-Group-Labeled Methionine Precursor as a Useful Tool for Simplifying Protein Structural Analysis by NMR Spectroscopy. *ChemBioChem* **8**, 610-612 (2007).
7. Ollerenshaw, J.E., Tugarinov, V. & Kay, L.E. Methyl TROSY: explanation and experimental verification. *Magnetic Resonance in Chemistry* **41**, 843-852 (2003).
8. Rosenzweig, R. & Kay, L.E. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu Rev Biochem* **83**, 291-315 (2014).
9. Prestegard, J.H., Al-Hashimi, H.M. & Tolman, J.R. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Quarterly Reviews of Biophysics* **33**, 371-424 (2001).
10. Kontaxis, G. & Bax, A. Multiplet component separation for measurement of methyl 13C-1H dipolar couplings in weakly aligned proteins. *Journal of Biomolecular NMR* **20**, 77-82 (2001).
11. Sprangers, R. & Kay, L.E. Probing Supramolecular Structure from Measurement of Methyl $^{1}H–^{13}C$ Residual Dipolar Couplings. *Journal of the American Chemical Society* **129**, 12668-12669 (2007).
12. Pederson, K. et al. NMR characterization of HtpG, the E. coli Hsp90, using sparse labeling with ^{13}C-methyl alanine. *Journal of Biomolecular NMR* **68**, 225-236 (2017).
13. Takeuchi, K., Arthanari, H., Imai, M., Wagner, G. & Shimada, I. Nitrogen-detected TROSY yields comparable sensitivity to proton-detected TROSY for non-deuterated, large proteins under physiological salt conditions. *Journal of Biomolecular NMR* **64**, 143-151 (2016).

14. Takeuchi, K., Arthanari, H., Shimada, I. & Wagner, G. Nitrogen detected TROSY at high field yields high resolution and sensitivity for protein NMR. *Journal of Biomolecular NMR* **63**, 323-331 (2015).

15. Gao, Q. et al. Structural Aspects of Heparan Sulfate Binding to Robo1-Ig1-2. *ACS Chem Biol* **11**, 3106-3113 (2016).

16. Delaglio, F. et al. NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**, 277-293 (1995).

17. Goddard TD, K.D. SPARKY 3. (University of California, San Francisco, 2008).

18. Nitsche, C. & Otting, G. Pseudocontact shifts in biomolecular NMR using paramagnetic metal tags. *Progress in Nuclear Magnetic Resonance Spectroscopy* **98-99**, 20-49 (2017).

19. Hogben, H.J., Krzystyniak, M., Charnock, G.T.P., Hore, P.J. & Kuprov, I. Spinach – A software library for simulation of spin dynamics in large spin systems. *Journal of Magnetic Resonance* **208**, 179-194 (2011).

20. Keeler, J. Understanding NMR Spectroscopy, (Wiley, West Sussex, UK, 2010).

21. Bertini, I., Luchinat, C., Parigi, G. & Pierattelli, R. NMR Spectroscopy of Paramagnetic Metalloproteins. *ChemBioChem* **6**, 1536-1549 (2005).

22. Gao, Q., Chalmers, G.R., Moremen, K.W. & Prestegard, J.H. NMR assignments of sparsely labeled proteins using a genetic algorithm. *Journal of Biomolecular NMR* **67**, 283-294 (2017)

Figures and Tables

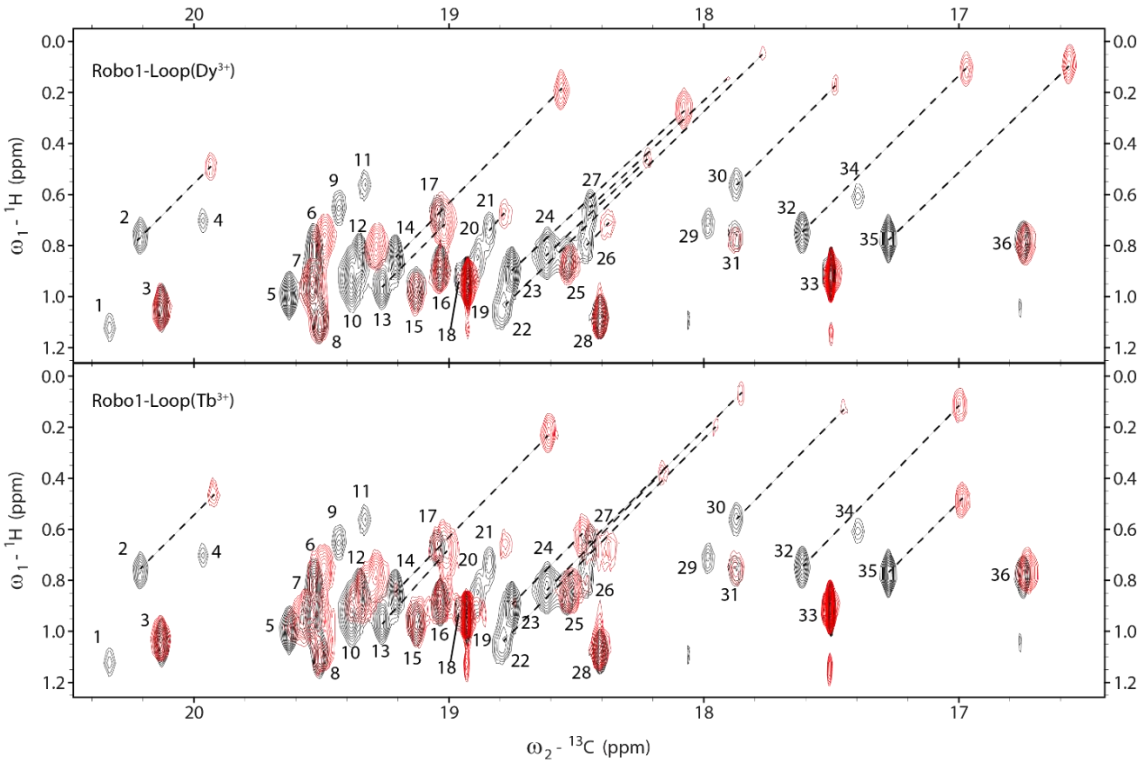

Figure 3.1. HETCOR spectra of [$^{13}CH_3$,$^{13}CH_3$]-Valine-Robo1-Ig1-2-Loop showing overlays of the diamagnetic (black) Lu^{3+} complex with the paramagnetic (red) Dy^{3+} and Tb^{3+} complexes. Spectra were acquired on 300 µM samples in approximately 3 hrs at 600 MHz for protons. Resonance assignments have not been completed and peaks in the diamagnetic spectra are labeled with arbitrary numbers. Several peaks experience large pseudocontact shifts in the paramagnetic spectra and are connected to their diamagnetic counterparts with dashed lines.

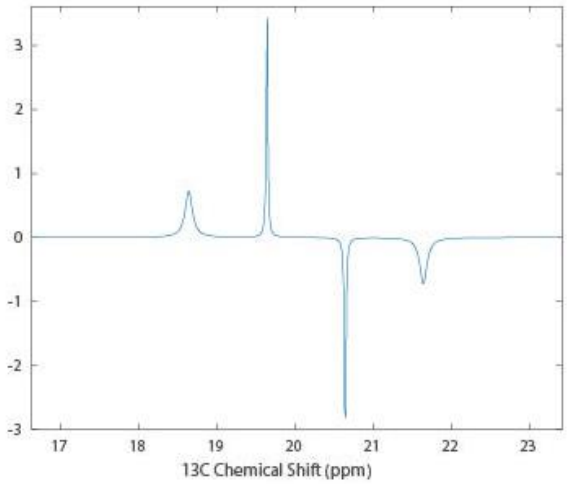

Figure 3.2. SPINACH simulation of ^{13}C magnetization for a proton coupled methyl group in a HETCOR experiment before the final refocusing period.

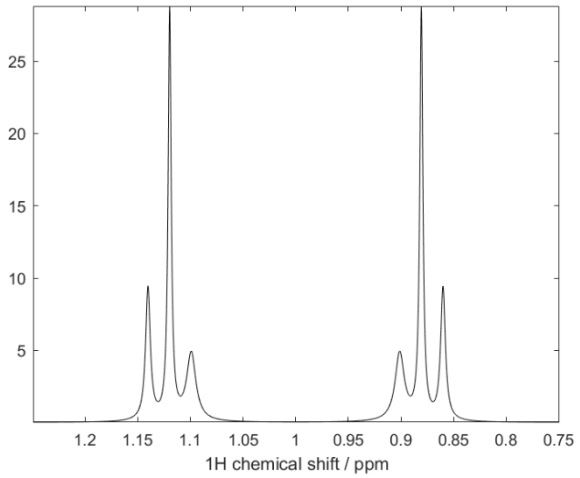

Figure 3.3. SPINACH simulation of a 1D 600 MHz ^{1}H spectrum of an isolated ($^{13}C^{1}H_3$) spin system under conditions of partial alignment. The large splitting is due to $^{1}J_{CH}$. The proton lines are also split into triplets due to a ^{1}H-^{1}H RDC of approximately 12 Hz. Cross-correlation effects between ^{13}C-^{1}H and ^{1}H-^{1}H dipole-dipole relaxation cause the differential linewidths for the inner and outer lines of each triplet.

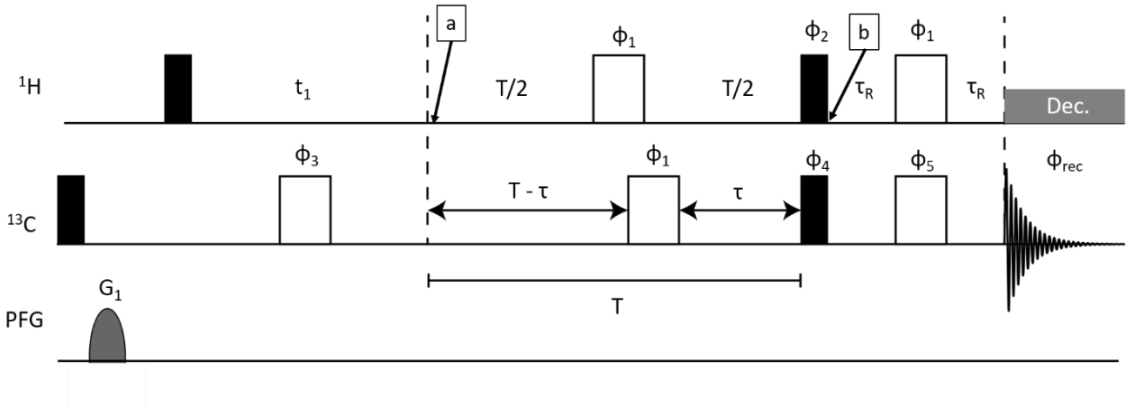

Figure 3.4. J modulated HETCOR pulse sequence diagram. Filled pulses are 90° and unfilled are 180°. Pulse phases are as follows: ϕ_1 = [x, x,-x,-x], ϕ2 = [y,-y], ϕ_3 = [y, y, -y, -y], ϕ4 = [4x,4y,4(-x),4(-y)], ϕ_{rec} = [2(x, -x), 2(y, -y), 2(-x, x), 2(-y, y)]. All other pulses have x-phase. All gradients were 1 ms duration and strength of 52 G/cm.

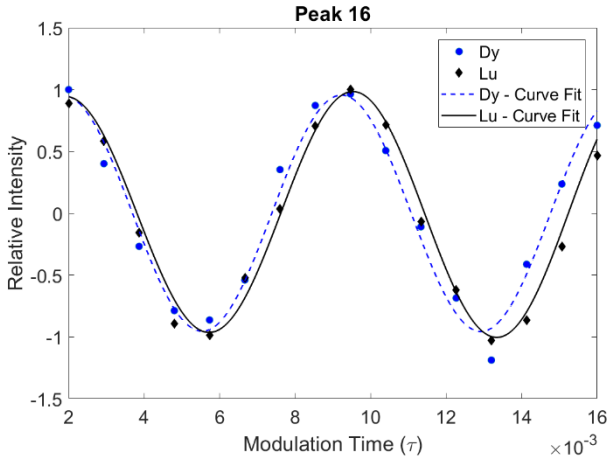

Figure 3.5. Overlay of measured J-modulation curves as a function of τ for peak 16 in paramagnetic Robo1+Dy sample and diamagnetic Robo1+Lu sample recorded using the pulse sequence in Figure 3.4. Both samples were measured at 900 MHz at concentrations of 300μM in approximately 24 hrs.

Table 3.1. Field-induced RDCs measured for the Robo1 Dy^{3+} and Tb^{3+} complexes at 900 MHz and 600 MHz. RDCs were measured as the difference in J+D in the paramagnetic complex and an average value of J from the diamagnetic complex.

	Dy^{3+} Complex			**Tb^{3+} Complex**		
Peak Label	**$^{900}D_{CH}$**	**$^{600}D_{CH}$**	**Scaled**	**$^{900}D_{CH}$**	**$^{600}D_{CH}$**	**Scaled**
3	-1.8 ± 1.3	1.4 ± 1.8	3.1 ± 3.6	1.4 ± 1.3	0.2 ± 1.9	0.4 ± 2.8
5	N/A[a]	N/A[a]	N/A[a]	1.8 ± 1.8	0.8 ± 2.3	1.8 ± 3.4
6	0.1 ± 1.6	1.1 ± 1.5	2.5 ± 2.3	-0.3 ± 1.6	3.5 ± 2.3	7.9 ± 3.4
8	2.1 ± 1.8	1.1 ± 1.9	2.5 ± 2.3	5.4 ± 1.3	3.9 ± 0.9	8.8 ± 1.4
10	N/A[b]	N/A[b]	N/A[b]	6.4 ± 1.7	N/A[b]	N/A[b]
12	6.2 ± 3.4	0.8 ± 2.5	1.8 ± 3.8	2.6 ± 1.8	N/A[b]	N/A[b]
15	N/A[a]	N/A[a]	N/A[a]	4.1 ± 2.3	3.5 ± 1.6	7.9 ± 2.4
16	6.4 ± 1.6	3.5 ± 1.4	7.9 ± 2.0	7.1 ± 2.4	6.1 ± 1.6	13.7 ± 2.5
19	0.0 ± 0.1	-0.1 ± 0.2	-0.2 ± 0.3	-0.2 ± 0.2	0.2 ± 0.6	0.4 ± 0.8
25	-2.7 ± 1.8	2.2 ± 1.4	4.9 ± 2.0	2.0 ± 3.2	0.6 ± 2.6	1.3 ± 3.7
28	3.9 ± 1.1	2.5 ± 1.5	5.6 ± 2.3	4.4 ± 0.8	2.9 ± 1.6	6.5 ± 2.3
33	0.0 ± 0.1	0.0 ± 0.3	0.0 ± 0.4	-0.2 ± 0.1	-0.3 ± 0.2	-0.7 ± 0.3
36	2.7 ± 2.3	1.7 ± 2.5	3.8 ± 3.7	1.4 ± 2.1	-1.2 ± 2.4	-2.7 ± 3.5

[a]Peaks were greatly attenuated and reliable measurements could not be made.

[b]Peaks 10 and 12 show significant overlap and could not be fit well in all spectra. Errors are standard deviations propagated from curve fitting estimates.

CHAPTER 4

ASSIGN_SLP_GUI: A SOFTWARE TOOL FOR NMR RESONANCE ASSIGNMENT OF SPARSELY LABELED PROTEINS[3]

[3] Williams, R. V. and Prestegard, J. H. To be submitted to the *Journal of Magnetic Resonance*.

Abstract

Widely used NMR methods for protein analysis rely on recombinant expression in *E. coli* to accomplish the uniform ^{13}C and ^{15}N isotope labeling needed for resonance assignment. However, many proteins of biomedical interest cannot be expressed in their native form in bacterial cultures, such as glycoproteins. An alternative approach based on sparse isotope labeling of one or two amino acids combined with mammalian cell culture enables NMR studies for glycoproteins but requires an alternative approach to resonance assignment. Here we describe an updated version of an assignment tool, ASSIGN_SLP, which incorporates a graphical user interface. This version also incorporates several new features including support for ^{13}C-methyl labeling and inclusion of additional data types (pseudocontact shifts and paramagnetic relaxation enhancements). We demonstrate the utility of the software by assignment of valine and alanine methyl resonances in a glycoprotein construct, hRobo1-Ig1-Ig2-LBP4.

Introduction

NMR remains a useful option for de novo determination of protein structures. In these cases, the process begins with assignment of NMR resonances. Assignment of NMR resonances of proteins is usually carried out by applying triple resonance experiments to samples uniformly labeled with ^{15}N and ^{13}C. For validation and functional studies, however, labeling of all protein residues may not be necessary; sparse labeling with one, or a few, isotopically labeled amino acids may be adequate. As an additional benefit, sparse labeling can improve resolution in applications to larger proteins and reduce costs for production of proteins requiring expression in mammalian

cell cultures due to post-translational and protein folding requirements[1]. On the other hand, sparse labeling is not compatible with assignment strategies based on uniform isotope labeling and triple resonance experiments. A few years ago, we outlined an approach for resonance assignment of sparsely labeled proteins based on measurement of a few NMR parameters, namely chemical shifts, nuclear Overhauser effects (NOEs) and residual dipolar couplings (RDCs), all read from simple extensions of 2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments.[2] We provided a command line software package, ASSIGN_SLP, that integrated C++ code, MATLAB scripts and Bash scripts to make assignments of crosspeaks in the underlying HSQC spectra.[3] It did rely on the existence of a good structural model for the protein, usually an X-ray structure, to make predictions of parameters. A genetic algorithm was then used to make assignments that optimized the fit of experimental to predicted data. Recent advances in computational prediction of structures may well allow replacement of experimental structures with computational structures, greatly expanding the application of the method. Recognizing this possibility we here report a much-improved version that is faster, has a graphical user interface (GUI), incorporates new data types (paramagnetic relaxation enhancements (PREs) and pseudo contact shifts (PCSs)) and has a statistics-based module for estimating confidence levels of assignments. It works for proteins labeled with both ^{15}N and ^{13}C labeled amino acids. More importantly, the ^{13}C can be in methyl groups, allowing application to larger systems. It is entirely MATLAB based and can be provided in a compiled version that only requires a free runtime MATLAB library.

We will demonstrate the program's capabilities on a sparsely labeled sample of a construct containing the two N-terminal Ig-like domains of roundabout 1 (Robo1).

Robo1 is a developmentally important mammalian cell-surface signaling molecule with recognized roles in axon guidance, angiogenesis, and the development of other internal organs.[4] Changes in Robo1 levels are also correlated with tumorigenesis, cancer progression, and metastasis.[5] While there are crystal structures for constructs containing the two N-terminal domains, Robo1 is a glycosylated protein (one N-glycosylation site in the first N-terminal domain), and the X-ray structures lack this glycan. Hence, we present this as an example of the many proteins that may be studied with native posttranslational modifications by expression in mammalian cell culture, supplemented with isotopically labeled amino acids or their precursors. More specifically we have labeled valine methyl groups by supplementation with ^{13}C-methyl valine and alanine methyl groups by supplementation with ^{13}C1-glucose.[6] We have also introduced a lanthanide binding peptide to the second N-terminal domain so that we can illustrate the use of PCSs and PREs.[7] Here, we have used an MD refined derivative of one of the X-ray structures as a basis for resonance assignment and an illustration of sensitivity of our scoring procedure to minor structural differences, but we expect the use of computational models to produce similar results.

Program Description

The program retains the basic genetic algorithm search strategy of its predecessor. It begins with 1000 randomly constructed individuals that have specific pairings of each resonance with each site in the protein that hosts one of the selected amino acid types. Scores (penalties) are calculated for each individual by summing contributions over data type and sequence positions. It then uses an algorithm from the MATLAB library of

functions, ga, to search for the minimum in an objective function. The best scoring individuals are saved for each generation and subjected to mutation and crossover to generate a new set of individuals. Default values for mutation and crossover are set but can be changed by opening an "options" menu. Enhanced speed has been achieved by use of predefined MATLAB functions and eliminating unnecessary file input and output. Proteins with less than 20 crosspeaks to assign will solve on typical laptops in a few minutes.

The objective function contributions are primarily root-mean-square-deviations (RMSDs) between predicted and experimental parameters normalized by user-provided error estimates. An exception is the NOE score; it is the root mean square deviation from 1 (perfect match) of the normalized dot products of vectors representing predicted and experimental measures of intensity versus proton chemical shift. An error estimate is taken as the score for the vector with minimum average NOE intensity. Experimental measures are provided by the user in an Excel workbook. Predictions are initiated by selecting one of the "prepare predictions" buttons in the GUI shown in Figure 4.1. A new window opens, prompting for the PDB file of a trial structure, and a chemical shift file which has been calculated using external (web-based) utilities, SHIFTX2 or PPM_One.[8, 9] The remainder of the predictions are calculated internally and saved as a "prediction" workbook.

The predictions workbook entries for RDCs and PCSs stop at coordinates and order parameters, as the predictions change with the assignments of each genetic individual and must be calculated for each generation of the genetic algorithm. Note that options for predictions include "single frame" and "trajectory". The "single frame"

option is fast, but the slower "trajectory" option accounts for conformational averaging. The latter results in "order parameter" additions to the prediction workbooks for RDCs and PCSs, and NOE predictions that include fast motion effects and relaxation matrix calculations.[10] NOEs for the single frame option are calculated using a simple $1/r^6$ dependence on distance between a proton of interest and other protons within a cutoff distance, typically 4Å.

The new PRE data type is calculated using a more complex function of distance that it based on the assumption that experimental values come from reduction of HSQC peak volumes in the presence of a paramagnetic species. This is given in equation 4.1.

Eq. 4.1. $$\frac{V_{Gd}}{V_{Ref}} = \exp\left(-2.303\left(\frac{r_0}{r}\right)^6\right)$$

The paramagnetic species is identified by entering its PDB symbol in a field on the "Prepare Predictions" screen, r is the distance between this entity and the proton of the crosspeak being assigned and r_0 is a user estimated distance at which crosspeaks are reduced by 90%.

PCSs have the same functional form as RDCs but involve distances between observation sites and lanthanide ions (identified in the paramagnetic species field) and angles relative to axes of susceptibility tensors rather than bond lengths and angles relative to axes of an order tensor. Hence, they are treated in the same way as RDCs during the search. A special case arises when RDCs result from field induced alignment. Then the order tensor can be calculated from a susceptibility tensor, reducing the number of adjustable parameters for the paired RDC-PCS sets from 10 to 5.

The inclusion of ^{13}C-methyl data is important for applications to larger proteins. One doesn't get the full effect of relaxation interference seen in Methyl TROSY

experiments unless perdeuteration is used, however, ^{1}H-^{13}C methyl crosspeaks are inherently sharp due to internal rotation, and they carry the intensity of 3 compared to 1 proton.[11] Implementation of RDCs and NOEs for methyl groups is relatively straightforward. The predicted RDCs use the C-C bond to the methyl group as an RDC vector and scale the calculation for differences in magnetic moments (^{1}H,^{13}C vs ^{13}C,^{13}C), the difference in bond length (C-H vs C-C) and averaging by methyl rotation. The NOEs are approximated using the sum of the contributions calculated for static methyl proton positions in supplied trial structures. Simulation using Spinach software suggests less than a 30% error for single proton – methyl interactions, despite the rapid internal motion and multiple spin states of a methyl proton system. Valine, leucine, and isoleucine residues present a special case in that two methyl groups are present on each residue. These could be treated as independent assignments; however, it is often the case that one knows that a pair belongs to the same residue from NOE or TOCSY data. A constraint sheet is provided in the experiment workbook that penalizes assignment of crosspeaks to different residues when they are known to come from the same residue. A similar sheet provided in the previous version is used to enforce assignment of crosspeaks to a particular amino acid type when these carry different isotopes or come from different sample preparations. Assignment to a particular residue may also be enforced when the assignment is known by other means, such as mutational analysis.

When experiment and prediction workbooks have been prepared, selecting the "AssignSLP: Search" tab on the window shown in Figure 4.1 will open a window that prompts for these workbooks and has a tab that starts the genetic algorithm search. As the algorithm runs a heatmap similar to that shown on the GUI in Figure 4.1 is displayed

and updated for each new generation. Intensities of the colored squares denote the fraction of times the assignment of a particular crosspeak to a particular site is made. The process will stop when no significant improvement in score is achieved or when a maximum number of generations is reached (limits for both can be adjusted under the options tab). The window also provides for display of score contributions from different data types (analysis tab) and the selection of a special running mode called “validation mode”.

The validation mode is an important new addition. In many applications it is not necessary to assign all crosspeaks, just a sufficient number to validate a trial structure or probe protein interaction sites. But one must be confident that these partial assignments are correct. The fraction of individuals carrying a particular assignment (as read from the intensities in the final heat map) might be used to select reliable assignments but the selected fraction is difficult to translate to a true confidence level. One way of deriving an appropriate confidence indicator is to assume the error limits entered for each data type (sum of experimental and predicted error) are good representations of a standard deviation of normally distributed independent variables, sample these variables within a gaussian distribution centered at the experimental value and repeatedly predict assignments using the sampled sets. A heatmap similar to the final output of a single genetic algorithm search is produced showing the percentage of searches that assign a particular site to a particular crosspeak. This process can take some time, so we suggest setting the number of samplings to a practical value that is still much larger than the number of crosspeaks being assigned (~250 for < 25 crosspeaks). One can then estimate a threshold at which one can be 95% confident that assignment percentages greater than

this value provide an indication of confidence. For a sampling of 250 sets, the threshold is about 80%. Lower thresholds require much more sampling. We illustrate the entire procedure in the following application to resonance assignment for the two N-terminal domains of the cell-surface signaling molecule, Robo1.

Materials and Methods

Our two-domain construct is based on the hRobo1 uniprot sequence (Q9Y6N7), residues 61 to 266. To allow us to assign domains separately, without concern about effects of interdomain motion, we also produced a construct containing only the N-terminal domain, residues 61 to 169. For the two domain construct, we introduced a lanthanide ion binding site by replacing a turn between Ig2 anti-parallel β-strands F and G , residues 221 to 224, with a slightly modified version of the original Imperiali Ln-binding peptide (CADTNNDGAYEGDELC, LPB4).[12] Our prior applications using these lanthanide-binding peptides have replaced loops between α-helices[13], and there was some preliminary indication that replacing these shorter turns between β-strands with the original peptide would disrupt protein structure as well as decrease ion binding affinity.[14] The terminal pair of cysteines in LBP4 was introduced so that disulfide bond formation would stabilize both the binding peptide and the β-strands. Peptides with disulfide bonds bridging the ends of the ion binding loop have been made before and demonstrated to maintain affinity.[12] There is also structural data that suggests the introduction of a disulfide bond between cysteines, particularly at $\beta_{A,NHB}$ sites in the strands, would also stabilize the antiparallel strands.[15] Our site was chosen with this information in mind. A

1μs long molecular dynamics (MD) run suggested that the anticipated structural characteristics were maintained.

The constructs were synthesized using codons optimized for mammalian cell expression by GeneArt (Regensburg, Germany) and inserted into a pGEn2 expression vector and purified as previously described.[1] The vector contains codons for an N-terminal hexahistidine-tag and GFP protein separated by a short linker and TEV cleavage site. Cleavage leaves a short scar on the N-terminus of the product (GSGG). HEK293S (GnT1-, MGAT1 knockout) cells (ATCC), which add primarily $Man_5GlcNAc_2$ glycans to N-glycosylation sites, were transfected with this vector. Growth and expression proceeded in 500 mL of a custom version of FreeStyle 293 expression media (which lacked both glucose and amino acids) to which 2.5 g of ^{13}C1-glucose, 70 mg? of ^{13}C-methyl-valine (Cambridge Isotope Labs, Tewksbury, MA) and 40 mg of the other natural abundance amino acids normally in the medium were added. After two rounds of metal affinity chromatography and a round of size exclusion chromatography, the final yield was ~10 mg. Molecular weight was verified by mass spectrometry. For NMR, the protein was exchanged into a buffer composed of 25 mM Tris, 100 mM NaCl, pH 7.4, 0.02% NaN_3, 10 μM DSS, 90/10% H_2O/D_2O at a final protein concentration of 300 μM. For PRE and PCS data near molar equivalents (~0.9) of $GdCl_3$ and $DyCl_3$ respectively were added. Addition of diamagnetic $LuCl_3$ lead to significant protein precipitation and a sample containing no lanthanide was used as a reference instead.

Carbon-detected NMR spectra were acquired on a Bruker AVANCE NEO 900 MHz spectrometer using a triple resonance 5mm TXO cryo probe that is optimized for ^{13}C and ^{15}N observation, while proton-detected spectra were recorded on Bruker

AVANCE 3 800 MHz spectrometer equipped with a 5 mm TCI triple-resonance cryoprobe. The following data sets were acquired using the specified standard pulse sequences. ^{13}C, ^{1}H-HETCOR spectra were recorded with 2048 x 192 points using the standard Bruker hxinepph sequence. Sweep widths of 61 ppm and 3.0 ppm were used for the 13C and 1H dimensions, respectively. INEPT delays were set to 1/(4J) (2.0 ms), while the refocusing delays were set to 1/(12J) (0.667 ms), where J was assumed to be 125 Hz. 96 scans were averaged.

A 3D ^{13}C-edited NOESY-HSQC spectrum was acquired using the Bruker noesyhsqcetgp3d pulse sequence with 2048 x 512 x 128 points and 5% non-uniform sampling. All transfer delays were set to 1/(4J) and a NOE mixing time of 150 ms was used. 64 scans were averaged, resulting in a total experiment time of 73 hours. A 3D 13C-edited TOCSY-HSQC spectrum was acquired using the Bruker dipsihsqcetgp3d pulse sequence. The DIPSI mixing pulse was applied for 60 ms. All other parameters were set identically to the 3D NOESY experiment. RDC data were acquired with a pulse sequence we recently devised for collection of J-modulation data with direct ^{13}C observation.[16] 8 ^{13}C,^{1}H HETCOR planes were collected at modulation delays of 2, 3, 3.5, 4.5, 5, 5.5, 6, and 6.6 ms. Total acquisition time for the pseudo-3D experiment was 22 hours.

Results

Figure 4.2 shows a HETCOR spectrum of Robo1-Ig1-Ig2-LBP4 carrying ^{13}C-enriched methyl groups of valine and alanine. The directly observed nucleus in this case is ^{13}C rather than ^{1}H. While ^{1}H observation is normally more sensitive (about a factor of

2 on this protein, comparing results using the TXO probe to using a TCI cryo probe), the inherent resolution in the carbon dimension is normally higher than that in the proton dimension. There are also advantages when working with samples containing paramagnetic ions because of differences in the time proton coherence is transverse.

We are able to distinguish alanine from valine crosspeaks based on a separate expression in which only valines are labeled. As expected crosspeaks from alanine methyls occupy a proton chemical shift region downfield compared to valine methyls. The alanine crosspeaks are less intense than those for valine, due primarily to scrambling C1 and C6 carbons of glucose during glycolysis of our ^{13}C1-labeled supplement. Valine methyls may also have some additional internal motion that sharpens peaks. In our assignment application we have entered order parameters from our 1μs MD simulation to account for some of this motion. While, not shown in the figure, most sugar residue anomeric carbons present well resolved crosspeaks.

Figure 4.3 superimposes HETCOR spectra of samples containing no diamagnetic ion (black) and a paramagnetic Dy^{3+} ion (blue). Lines denote connections between diamagnetic and paramagnetic crosspeaks from which PCSs can be measured. Unique connections can be made in nearly all cases because of the expected near identical shift in proton and carbon dimensions. These are summarized in Table A.1, along with other data. Values for predicted PCSs are determined for each site in the same manner as values for RDCs, using user entered order parameters and lanthanide and labeled site coordinates as assigned in the current round of genetic algorithm search. The score contribution is the RMSD of predicted and experimental values divided by an estimated error. A major difference arises when an RDC set comes from field induced orientation.

Then the user has the option of either treating PCS data as independent, with its own sheet in experiment and prediction workbooks, or combining the field-induced RDCs and PCS in to a single "RDC+PCS" dataset. To properly scale the two types of data an effective D_{max} (or PCS_{max}) is calculated from equation 4.2. In the application presented, we treated PCS data as independent despite the fact that RDCs came from field induced order.

Eq. 4.2 $$PCS_{max} = \frac{60kT}{12\pi B_0^2} \times 10^{30}$$

The genetic algorithm search was performed for this Robo1 construct using collected chemical shift, NOE, and PCS data. During the search, the current generation of assignments is displayed as a heatmap. Figure 4.4 shows the final state of the heatmap for the normal genetic algorithm run. The header indicates that the search converged in 394 generations. Each row and column have a blue box with intensity proportional to the fraction of individuals in the final generation assigning a particular crosspeak to a particular site. By activating the analysis window of ASSIGN_SLP_GUI one can access a list of the best assignment for each crosspeak.

Figure 4.5 shows a heatmap produced by running in "Validation" mode. The input for this mode is the best assignment list from the final generation of each repeated run (200 runs) in our case. Sites V71, V165, and V246 were assigned by mutagenesis and constrained within AssignSLP and as a result are assigned correctly 100% of the time. The remaining assignments appear to be randomly spread within the provided constraints separating alanine, Ig1 valine, and Ig2 valine methyl groups.

The low confidence levels observed for the remaining peaks is likely a result of overestimation of the errors used for random sampling. Repeated runs of the assignment

search converged on the same answer. The high reproducibility in the absence of random errors suggest that the true confidence is higher than that determined by validation mode. Additional statistical analysis to determine a more realistic Monte Carlo sampling approach is likely warranted. Furthermore, some sources of error are difficult to quantify. All of the predictions depend upon a protein structure, the accuracy of which may vary from site to site.

Discussion

The ability to assign a sparsely labeled protein, even to a partial extent, opens possibilities for structural and functional characterization of an expanded set of proteins, including those of biomedical importance. These include proteins not easily expressed in bacterial hosts, such as glycoproteins and proteins requiring folding environments that facilitate formation of disulfide bonds and integration into membrane structures. We have presented a new version of ASSIGN_SLP that accommodates new data types, has an improved graphical user interface, and assigns a statistics-based confidence level to all suggested assignments. Use has been illustrated with application of a two-domain construct from Robo1, a protein that is glycosylated, has an internal disulfide bond, as well as an additional disulfide bond that is introduced to allow observation of one of the new data types, pseudo contact shifts (PCSs).

The results on Robo1 fall short of complete assignment but mutational assignments will allow use of these in other functional and structural studies. This has already occurred to some extent. Assignments have been used to assess the interdomain geometry of a Robo1-Ig1-2 construct and the binding of a ligand important to Robo1's

function, a medically important mimic of the heparan sulfate that mediates Robo1's function (see chapter 5).

Recent advances in computational prediction of protein structures offer opportunities to advance biological understanding well beyond what can be accomplished with just experimental stuctures[17-19]. However, predictions are not always made with high levels of confidence, particularly for proteins with disordered regions and high levels of post-translational modification (glycoproteins, for example). Hence, experimental validation of predictions remains important. NMR data can provide the needed validation, as well as subsequent characterization of dynamics, ligand binding properties and protein-protein interactions. The single prerequisite for these studies is NMR resonance assignment.

We have not illustrated all features of the presented version, including PREs and the use of multiple or correlated RDC sets. Addition of these data types would certainly result in additional confident assignments and more extended studies of Robo1 function. However, our objective here is to present the software tool and encourage use by a wider scientific community.

Acknowledgments

The author would like to thank Daniel Hall and Huimin Hu for their assistance with statistical analysis relevant to validation mode.

References

1. Moremen, K. W.; Ramiah, A.; Stuart, M.; Steel, J.; Meng, L.; Forouhar, F.; Moniz, H. A.; Gahlay, G.; Gao, Z.; Chapla, D.; Wang, S.; Yang, J.-Y.; Prabhakar, P. K.; Johnson, R.; Rosa, M. d.; Geisler, C.; Nairn, A. V.; Seetharaman, J.; Wu, S.-C.; Tong, L.; Gilbert, H. J.; LaBaer, J.; Jarvis, D. L., Expression system for structural and functional studies of human glycosylation enzymes. *Nature Chemical Biology* **2018,** *14* (2), 156-162.
2. Gao, Q.; Chalmers, G. R.; Moremen, K. W.; Prestegard, J. H., NMR assignments of sparsely labeled proteins using a genetic algorithm. *Journal of Biomolecular NMR* **2017,** *67* (4), 283-294.
3. Chalmers, G. R.; Eletsky, A.; Morris, L. C.; Yang, J.-Y.; Tian, F.; Woods, R. J.; Moremen, K. W.; Prestegard, J. H., NMR Resonance Assignment Methodology: Characterizing Large Sparsely Labeled Glycoproteins. *Journal of Molecular Biology* **2019,** *431* (12), 2369-2382.
4. Blockus, H.; Chédotal, A., Slit-Robo signaling. *Development* **2016,** *143* (17), 3037.
5. Wang, B.; Xiao, Y.; Ding, B.-B.; Zhang, N.; Yuan, X.-b.; Gui, L.; Qian, K.-X.; Duan, S.; Chen, Z.; Rao, Y.; Geng, J.-G., Induction of tumor angiogenesis by Slit-Robo signaling and inhibition of cancer growth by blocking Robo activity. *Cancer Cell* **2003,** *4* (1), 19-29.
6. Rogals, M. J.; Yang, J.-Y.; Williams, R. V.; Moremen, K. W.; Amster, I. J.; Prestegard, J. H., Sparse isotope labeling for nuclear magnetic

resonance (NMR) of glycoproteins using 13C-glucose. *Glycobiology* **2020**.

7. 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. *Journal of the American Chemical Society* **2011,** *133* (4), 808-819.
8. Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S., SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR* **2011,** *50* (1), 43.
9. Li, D.; Brüschweiler, R., PPM_One: a static protein structure based chemical shift predictor. *Journal of Biomolecular NMR* **2015,** *62* (3), 403-409.
10. Chalmers, G.; Glushka, J. N.; Foley, B. L.; Woods, R. J.; Prestegard, J. H., Direct NOE simulation from long MD trajectories. *Journal of Magnetic Resonance* **2016,** *265*, 1-9.
11. Ollerenshaw, J. E.; Tugarinov, V.; Kay, L. E., Methyl TROSY: explanation and experimental verification. *Magnetic Resonance in Chemistry* **2003,** *41* (10), 843-852.
12. Nitz, M.; Franz, K. J.; Maglathlin, R. L.; Imperiali, B., A Powerful Combinatorial Screen to Identify High-Affinity Terbium(III)-Binding Peptides. *ChemBioChem* **2003,** *4* (4), 272-276.
13. 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 Science* **2012,** *21* (10), 1456-1466.

14. M. Rogals, J. Prestegard. *manuscript in preparation*.

15. Hutchinson, E. G.; Sessions, R. B.; Thornton, J. M.; Woolfson, D. N., Determinants of strand register in antiparallel β-sheets of proteins. *Protein Science* **1998,** *7* (11), 2287-2300.

16. Williams, R. V.; Yang, J.-Y.; Moremen, K. W.; Amster, I. J.; Prestegard, J. H., Measurement of residual dipolar couplings in methyl groups via carbon detection. *Journal of Biomolecular NMR* **2019,** *73* (3), 191-198.

17. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.

18. Pereira, J.; Simpkin, A. J.; Hartmann, M. D.; Rigden, D. J.; Keegan, R. M.; Lupas, A. N., High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics* **2021,** *89* (12), 1687-1699.

19. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Dijk, A. A. v.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373* (6557), 871-876.

Figures

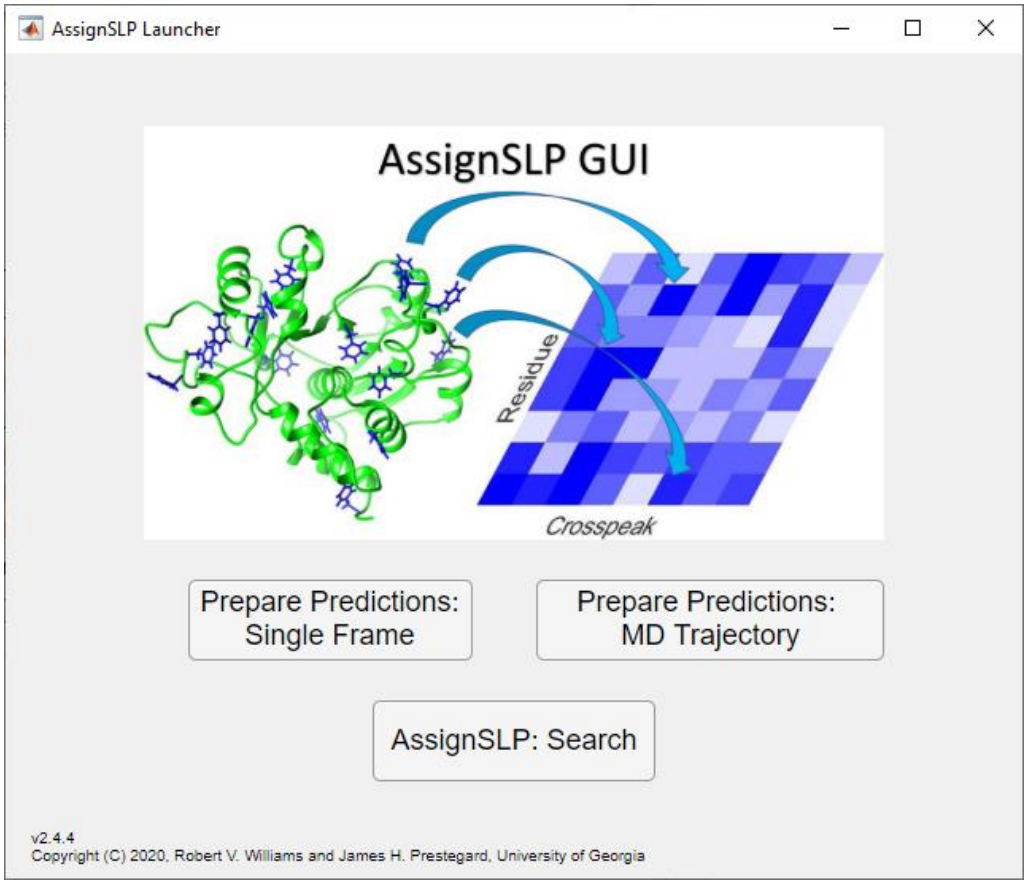

Figure 4.1. Launcher window for Assign_SLP_GUI. The three buttons launch the main features of the program in separate windows.

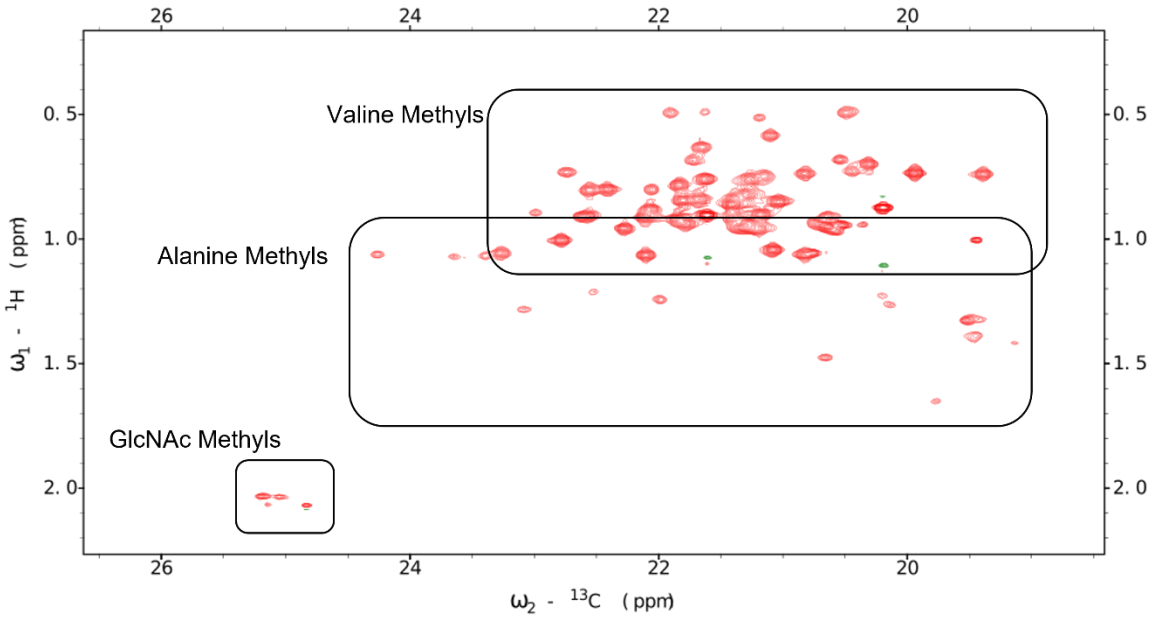

Figure 4.2. $^{13}C,^{1}H$-HETCOR spectrum of Robo1-Ig1-Ig2-LBP4 labeled with $^{13}C_1$-Glucose and ^{13}C-dimethyl-valine. The spectrum is dominated by the strong valine methyl signals which are isotope labeled with high efficiency. Less intense methyl peaks from alanine residues are observed. Additional signals from N-acetylglucosamine methyl groups are observed near 25, 2.0 ppm for ^{13}C and ^{1}H dimensions, respectively.

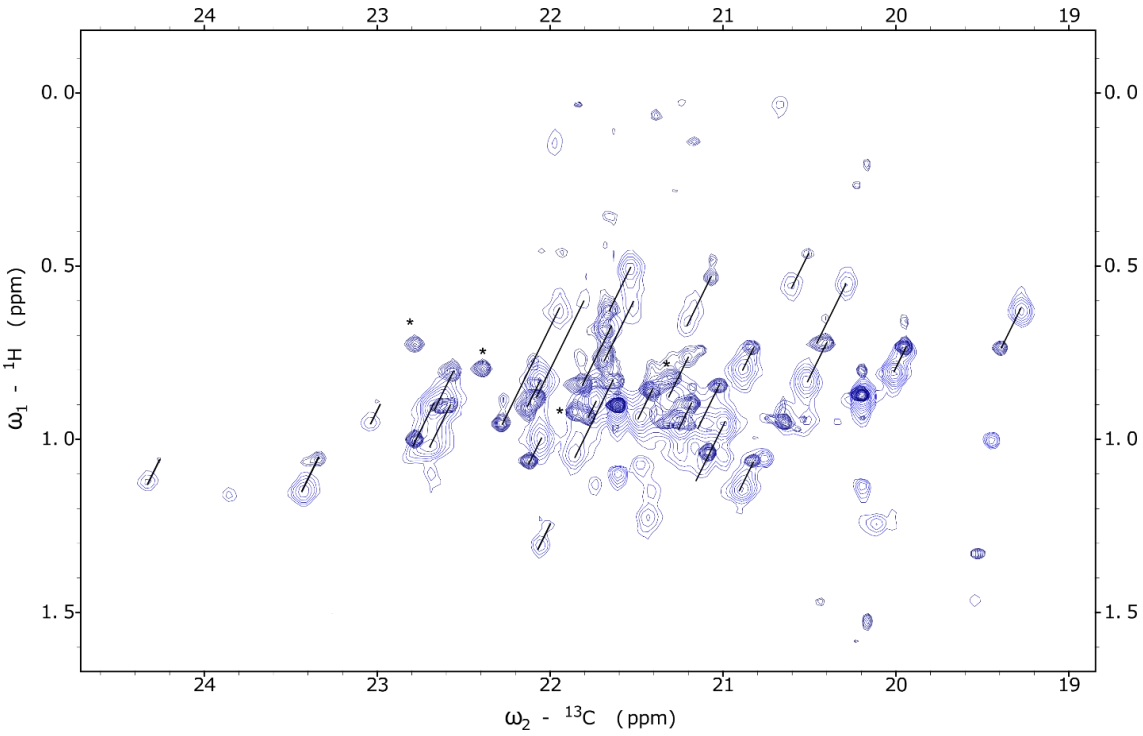

Figure 4.3. Overlay of HETCOR spectra collected on diamagnetic (no lanthanide, dark blue) and paramagnetic (Dy^{3+}, light blue) Robo1-Ig1-Ig2-LBP4 samples.

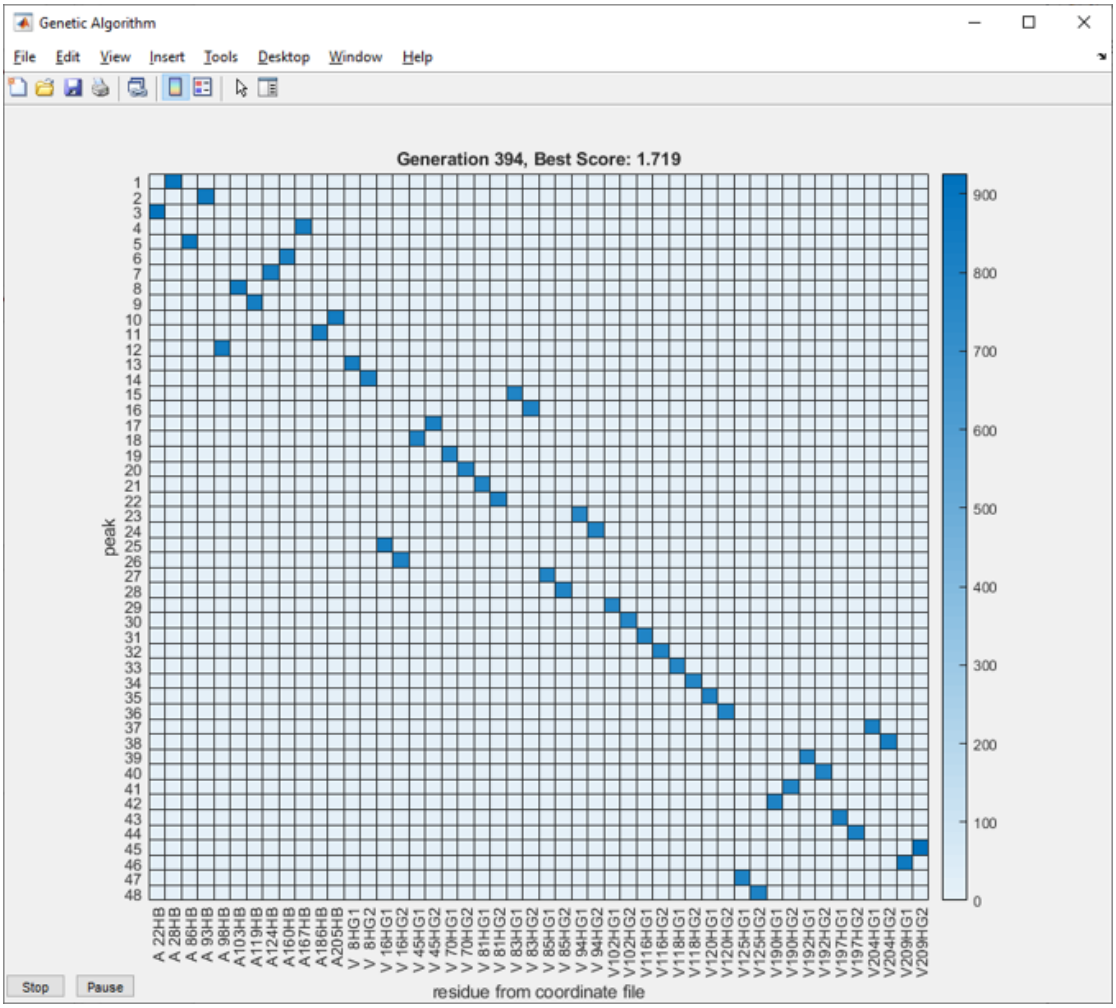

Figure 4.4. Assignment heatmap. A heatmap indicating the final population of 1000 possible assignments at the end the genetic algorithm search is shown in panel A. One blue box is observed for each row and column, indicating the search has converged on a solution. Panel B shows a different heatmap displaying the results of validation mode.

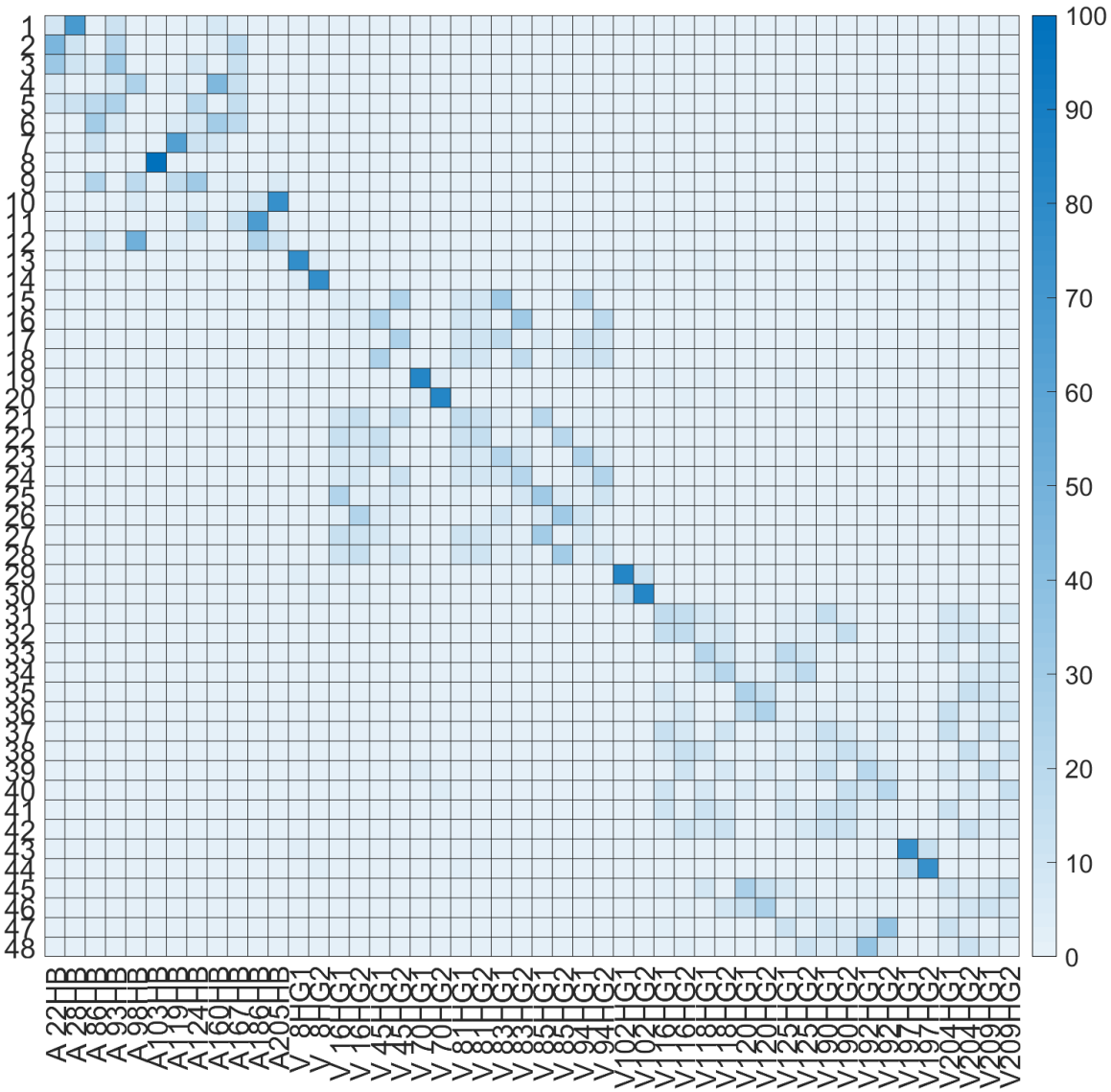

Figure 4.5. Validation heatmap showing the results of 200 trial searches, each using data with random errors. The bins show the distribution of assignments from each run.

CHAPTER 5

INVESTIGATION OF DOMAIN ORIENTATION OF hROBO1-Ig1-2 VIA SPARSE ISOTOPE LABELING: EFFECT OF HEPARAN SULFATE[4]

[4] Williams, R. V.; Huang, C.; Moremen, K. W.; Amster, I. J.; Prestegard, J.H. To be submitted to *Scientific Reports*.

Abstract

Human roundabout 1 (hRobo1) is an extracellular receptor glycoprotein that plays important roles in angiogenesis, organ development, and tumor progression. Interaction between hRobo1 and heparan sulfate (HS) have been shown to be essential for its biological activity. To better understand the effect of HS binding we engineered a lanthanide-binding peptide sequence into the Ig2 domain of hRobo1. Native mass spectrometry was used to verify that loop introduction did not inhibit HS binding or conformational changes previously suggested by gas phase ion mobility measurements. NMR experiments measuring long-range pseudocontact shifts (PCSs) were then performed on ^{13}C-methyl labeled hRobo1-Ig1-2-LBP4 in HS-bound and unbound forms. The magnitude of most PCSs for methyl groups in the Ig1 domain increase in the bound state confirming a change in the distribution of interdomain geometries. A grid search over Ig1 orientations produced two similar conformers for the bound and unbound form, both of which differ from existing X-ray crystal structures.

Introduction

Human roundabout 1 (hRobo1) is a highly glycosylated extracellular receptor protein, first identified in Drosophila as important for immune system development.[1] Robo1 was later discovered to participate in angiogenesis and organ development.[2, 3] hRobo1 achieves these biological functions through interaction with another secreted glycoprotein, Slit2, and heparan sulfate (HS).[4] A ternary signaling complex is formed between the first two immunoglobulin-like (Ig1 and Ig2) domains of hRobo1, the second leucine-rich-repeat domain (D2) of Slit2, and HS.

Much is known about the molecular basis for these interactions. The binding interface between hRobo1 and Slit2 was determined from a cocrystal structure of hRobo1-Ig1 and Slit2-D2.[5] Some information on the HS-binding site was revealed by an X-ray crystal structure of Drosophila Robo1-Ig1-2 in complex with a heparin-derived oligosaccharide of incompletely defined composition.[1] However, Slit2 was not in the crystal and the fragment bound between two molecules in the crystal structure leaving details of the binding site in question. Hydroxy radical footprinting mass spectrometry further characterized the HS binding site,[6] and solution nuclear magnetic resonance (NMR) spectroscopy provided a more detailed model of a complex between hRobo1-Ig1-2 and a synthetic HS tetrasaccharide.[7] More recently, studies of larger hRobo1 constructs have shed light on the mechanisms that propagate a signal from binding inside the cell. A combination of x-ray crystallography for hRobo1-Ig1-4 and small angle X-ray scattering (SAXS) of the entire hRobo1 ectodomain determined that full-length hRobo1 forms a tetrameric assembly which undergoes a large-scale conformation change upon binding Slit2.[8]

These structural studies have helped flesh out the biophysical roles of the two proteins involved in the ternary signaling complex, but the role of HS binding remains unclear. One intriguing result was reported from a study employing ion mobility mass spectrometry (IM-MS).[9] hRobo1-Ig1-2 was observed to exist in two conformations, one significantly more compact than the available crystal structures, and one more consistent with elongated forms consistent with the X-ray data. Analysis of a hRobo1-Ig1-2 complexed with an HS hexasaccharide showed a shift towards the more compact conformation. These results lead the authors to conclude that HS may cause a

conformation change in hRobo1-Ig1-2. An important caveat of these results is that IM-MS is a gas phase technique, and the observed conformations may not be reflective of the solution state of the protein.

To further investigate the possibility of an HS-induced conformation change of hRobo1-Ig1-2, we performed solution NMR spectroscopy. Our strategy involved engineering a lanthanide-binding peptide sequence into the Ig2 domain of hRobo1. Similar sequences have been engineered into loop regions of several proteins.[10-12] This construct allowed measurement of pseudocontact shifts (PCSs) for methyl groups in both domains. PCSs are long-range angle and distance ($1/r^3$) dependent changes in resonance positions (chemical shifts) that can report on interdomain geometry. Measurements were performed with and out a Arixtra (fondaparinux) a highly sulfated, synthetic pentasaccharide. Comparison of the two sets of PCS measurements allowed us to determine that the protein does not experience a conformation change; however, the data suggests that Robo1-Ig1-2 becomes more rigid when binding Arixtra.

Materials and Methods

Protein Expression and Purification

The hRobo1-Ig1-2-LBP4 construct was synthesized using codons optimized for mammalian cell expression by GeneArt (Regensburg, Germany). The lanthanide-binding loop was derived from the original Imperiali construct by replacement of amino acids at both ends with cysteines, deletion of an isoleucine group and replacement of a tryptophan with an alanine.[1] This resulted in the amino acid sequence CADTNNDGAYEGDELC, which was inserted between residues R221 and K223 while residues G222 and G223

were deleted (numbering from Uniprot entry Q9Y6N7). The construct was inserted into a pGEn2 expression vector. The vector contains codons for an N-terminal hexahistidine-tag and GFP protein separated by a short linker and TEV cleavage site. Both protein expression and purification proceeded as previously described.[13] Cleavage with TEV leaves a short scar on the N-terminus of the product (GSGG). A preliminary expression was performed in HEK293F cells, while a second isotope-labeled batch was expressed in HEK293S (GnT1$^-$, MGAT1 knockout) cells (ATCC), which add primarily $Man_5GlcNAc_2$ glycans to N-glycosylation sites. Growth and expression of this batch proceeded in 500 mL of a custom version of FreeStyle 293 expression media (which lacked both glucose and amino acids) to which 2.5 g of ^{13}C1-glucose, 75 mg of ^{13}C-methyl-valine (Cambridge Isotope Labs, Tewksbury, MA) and other natural abundance amino acids normally in the medium were added. After two rounds of metal affinity chromatography and a round of size exclusion chromatography, the final yield was ~10 mg.

Protein concentration was determined by UV/Vis absorbance measurement at 280 nm. An extinction coefficient of 20315 M^{-1} cm^{-1} was predicted using the ProtParam tool on the Expasy webserver and used in the calculations.[14]

Native Mass Spectrometry

Purified hRobo1-Ig1-2-LBP4 from HEK293F cells was buffer exchanged into 10 mM ammonium acetate, pH 6.8 using Amicon microconcentrators (10 kDa molecular weight cutoff) for Native MS and IM-MS measurements. Samples were immediately frozen and stored at -20°C prior to analysis.

Assessment of lanthanide binding was performed using a 12 T Bruker Solarix FT-ICR-MS instrument. The instrument was calibrated for high-m/z using 1 mg/mL sodium perfluoroheptanoate (Sigma-Aldrich) in 50:50 (v/v) % acetonitrile/water. Samples were prepared with 10 µM protein and 0, 10, 20, and 30 µM of $LuCl_3$. Samples were infused by syringe pump at a rate of 2.0 µL/min and ionized via electrospray (ESI) at a voltage drop of 4500 V. Ions were desolvated with a skimmer 1 voltage of 100 V and accumulated in the collision cell for 1.0 s prior to injection into the ParaCell, where ions were trapped after a 1.4 ms time of flight delay. Ions were excited for broadband detection by a chirp waveform at 40% excitation power. Spectra were collected from m/z 500 to 5000 with 1M points and a transient duration of 1.4 s. Each spectrum was the sum of 100 scans. Spectra were analyzed using Bruker DataAnalysis software.

Ion mobility mass spectrometry data were collected using a Waters Synapt G2-S instrument. The protein concentration was 10 µM. Robo1-Arixtra complex spectra were collected on a solution also prepared with 10 µM Arixtra (fondaparinux sodium salt, Sigma-Aldrich). Solutions were directly infused at flow rates of 0.5 - 1.0 µL/min. NanoESI was achieved using fused sillica emitter tips (I.D. 15 um, new objective) and a Waters Zspray source operating at capillary voltages of 1.8 - 2.5 kV. To ensure gentle ionization conditions the source block temperature was set to 30°C, the sampling cone was kept at 30 V, and the extraction cone was set to 1.0 V. The traveling wave in the IMS cell had a velocity of 300 m/s and height of 21 V. IM-MS data was exported to a csv file using TWIMExtract.[15]

NMR Spectroscopy

NMR spectroscopy was performed on ^{13}C1-Glucose, ^{13}C-dimethyl-valine labeled hRobo1-Ig1-2-LBP4. For NMR, the protein was exchanged into a buffer composed of 25 mM Tris, 100 mM NaCl, pH 7.4, 0.02% NaN_3, 10 µM DSS, 90/10% H_2O/D_2O at a final protein concentration of 300 µM. For PCS data a near molar equivalent (~0.9) of $DyCl_3$ was added. Addition of diamagnetic $LuCl_3$ lead to prohibitive levels of protein precipitation and a sample containing no lanthanide was used as a reference instead. For Arixtra binding experiments, Arixtra was titrated to a final concentration of 600 µM.

NMR spectra were acquired on a Bruker AVANCE NEO 900 MHz spectrometer using a triple resonance TXO cryo probe that is optimized for ^{13}C and ^{15}N observation. ^{13}C, ^{1}H-HETCOR spectra were recorded with 2048 x 192 points using the standard Bruker hxinepph sequence. Sweep widths of 61 ppm and 3.0 ppm were used for the ^{13}C and ^{1}H dimensions, respectively. INEPT delays were set to 1/(4J) (2.0 ms), while the refocusing delays were set to 1/(12J) (0.667 ms), where J was assumed to be 125 Hz. 96 scans were averaged. All spectra were processed using nmrPipe.[16] Peak picking was performed with NMRFAM-Sparky.[17]

Model Building and Molecular Dynamics simulations

A model of hRobo1-Ig1-2-LBP4 was built starting from an available X-ray crystal structure (PDB 2V9R) using UCSF chimera.[5, 18] Initial coordinates for the LBP were taken from the X-ray crystal structure of interleukin-1β with a lanthanide-binding insertion sequence (PDB 3LTQ, residues 53A through 53P) and modified to match our modified construct.[10] Phi and psi angles near the fusion were adjusted to smoothly extend

the beta sheet. Hydrogen atoms were added with the reduce tool.[19] Aspartic acid residue 97 was changed to asparagine and a $Man_5GlcNAc_2$ glycan was added using tleap.[20]

The system for a conventional molecular dynamics (cMD) simulation was prepared and run using Amber2018 with the ff99SB forcefield for amino acids and Glycam-06j forcefield for carbohydrates. The hRobo1-Ig1-2-LBP4 model was solvated with a truncated octahedron of TIP5P water and neutralized by the addition of sodium cations.[21] The resulting system was energy minimized with the Sander module using 25,000 steps of steepest descent minimization followed by 46,996 steps of conjugate gradient minimization. The system was slowly heated to 300 K over 1 ns and then the density was equilibrated for 1 ns in the NPT ensemble. The system was further equilibrated by 50 ns of NVT simulation. The production cMD run consisted of 1 μs of NVT simulation.

A gaussian accelerated MD (GaMD) simulation was performed starting from the same equilibrated system with randomized initial velocities.[22] Acceleration was performed with a dual-boost scheme applied to both the total potential and the dihedral potential. Energy statistics were calculated from an initial 14 ns NVT simulation and both σ_{0D} and σ_{0P} were set to 6.0. Likely conformers were selected by clustering the frames via their C_α RMSD in MATLAB using functions from MDToolBox .[23] Clusters containing more than 300 frames were chosen for energy reweighting, which was also performed in MATLAB using the approach described by Miao et al.[22]

PCS Analysis

PCS data were fit to PDB structures using custom MATLAB code. The magnetic susceptibility tensor was determined from the following linear system of equations:

$$\text{Eq. 5.1} \qquad M\chi = P$$

Where M is a n x 5 matrix with row corresponding to a single atomic nucleus, χ is a 5 x 1 matrix containing five independent elements of the magnetic susceptibility tensor, and P is a n x 1 matrix containing the measured PCS values. The elements are constructed from the cartesian coordinates of the nucleus as follows:

$$\text{Eq. 5.2} \qquad M = \begin{bmatrix} \frac{1}{2}(2z^2 - x^2 - y^2) & \frac{1}{2}(x^2 - y^2) & 2xy & 2xz & 2yz \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

With at least 5 PCS values, the above system can be solved for χ. This approach is quite similar to that used by REDCAT (albeit for RDC calculations) and Paramagpy.[24-25] Agreement between a given structure and the PCS measurements was assessed by calculating a Q-factor, which is given by equation 5.3.

$$\text{Eq. 5.3} \qquad Q = \frac{\sum(PCS_{obs} - PCS_{calc})^2}{\sum(PCS_{obs})^2}$$

Ig1 orientational grid search

A grid search over possible Ig1 orientations was conducted using a MATLAB script. Different orientations of the Ig1 domain were generated using Euler angle rotations (ZXZ convention) about the alpha carbon within residue Ile 104. For each Euler angle 50 values were used, creating a 3-dimensional grid of 125,000 possible orientations. Each orientation was scored by calculating a Q-factor (equation 5.3) for the

Ig1 PCS data. The $\Delta\chi$ tensor used to back-calculate PCS values was determined from Ig2 data.

Results

We set out to investigate the domain orientation of hRobo1-Ig1-2 using a construct containing an inserted lanthanide-binding amino acid sequence. A construct incorporating a lanthanide-binding peptide (LPB) into a loop in the Ig1 domain was previously used to study ligand binding of a model HS tetrasaccharide.[7] This construct proved unsuitable for interrogating the domain orientation as few pseudocontact shifts (PCSs) were observed from Ig2-domain residues. This was likely due to a poor choice of position, but we also suspected a high degree of loop motion and relatively low binding affinity. Hence, we designed a new construct better suited to our purpose.

An ideal construct would contain the LBP in a position that places ^{13}C-labeled sites in both domains within the sphere of PCS influence ~30 Å. A previous study showed that lanthanide sequences could replace β-turn motifs in interleukin-1β with minimal perturbation of the native structure. Inspection of a Robo1-Ig1-2 X-ray crystal structure found 4 candidate locations for inserting the LBP (Figure 5.1). Both turns in the Ig1 domain face away from Ig2, which likely place the lanthanide too far to produce significant PCSs in the Ig2 domain. The turn between Ig2 sheets F and G faces towards the other domain but is very close to the interface. Insertion at this position would likely perturb the domain orientation and was not considered further. Lastly, the short turn between Ig2 sheets D and E also faces the Ig1 domain but is far enough away to accommodate insertion of the LBP. Thus, we selected this location to insert the LBP.

To stabilize the binding loop and minimize possible loop motion, we introduced cysteines at both ends of the inserted sequence, which we now refer to as LBP4. Formation of a disulfide bond between ends would stabilize the structure. Stabilization of the LBP in this way had been suggested previously and even a version omitting the original isoleucine had been tested and found to retain lanthanide binding affinity.[26] However, it was not clear that compatibility with turns at the ends of antiparallel β-strands had been tested. Fortunately, there is data suggesting that $\beta_{A,NHB}$ sites (alternate positions in the β-strands) were compatible with disulfide bond formation.[27] Hence, we selected the pair of sites (222 , 223) for insertion. Addition of cysteine residues to a protein construct can complicate protein expression and purification. With E. coli expression products refolding procedures are often necessary to ensure proper disulfide linkages; however, in our case the mammalian cell culture properly forms the disulfide bonds without further processing.

As an initial test of this proposed Robo1-Ig1-2-LBP4 construct, an atomic model was built and a long 1 μs molecular dynamics (MD) simulation was performed. A bound dysprosium (Dy^{3+}) ion was included in the model and remained in the binding site for the entire trajectory. The simulation also showed minimal disruption of the Ig2 fold as the beta sheet between strands D and E remained intact.

Robo1-Ig1-2-LBP4 was initially expressed in HEK293F cells as a test batch. Native mass spectrometry (MS) on a 10 μM sample was used to verify lanthanide binding activity. The spectra showed a heterogeneous mixture of complex-type glycoforms due to the single N-glycosylation site present in the Ig1 domain. Addition of lanthanide further increased the complexity of the spectra, but several glycoforms were

sufficiently well-resolved to observe lanthanide binding. Figure 5.2A shows the results of titrating our Robo1 construct with lutetium chloride for a single glycoform. After addition of 2 molar equivalents of lanthanide, the protein+Lu^{3+} complex is observed. With 3 molar equivalents of lanthanide roughly equal proportions of bound and unbound species are observed. Based on these results measured at 10 µM we hypothesized that this level of lanthanide affinity would be sufficient at the elevated concentrations used for NMR (~300 µM). Furthermore, these data likely underestimate the affinity due to competition with formation of lutetium-acetate complexes from the ammonium acetate native MS solution.

Native IM-MS was also used to verify that our lanthanide-binding construct maintains heparan sulfate binding ability. To this end, we tested the construct's ability to bind Arixtra (fondaparinux), a pentasaccharide drug reproducing a sulfation pattern found in heparin. In Figure 5.2B a comparison of native mass spectra with (bottom) and without (top) an equimolar addition of Arixtra is shown. Several new species are observed upon addition of Arixtra, which are consistent with Robo1-Ig1-2-LBP4 + Arixtra complexes for the various glycoforms. In the top spectrum, the two most abundant species have masses of 26345 ± 2 and 26963 ± 1 Da which agrees with Robo1-Ig1-2-LPB4 $GlcNAc_4Hex_3Fuc_1$ and $GlcNAc_4Hex_5SiaFuc$ glycoforms (glycan composition only). In the bottom, new peaks appear with a mass increase of 1509 ± 2 Da, which agrees with fully deprotonated Arixtra ($C_{31}H_{43}N_3O_{49}S_8$). These results confirm that the lanthanide-binding construct retains HS binding activity.

Additionally, ion mobility data were collected. Travel of ions through a drift tube field with low pressure N_2 gas is impeded in a manner dependent on cross-sectional area,

and hence the conformation of ions.[2] Example arrival time distributions are shown in Figure 5.2C. For both bound and unbound species, two features were observed which can be interpreted as a compact and extended conformation of the ion. For the unbound state, the extended conformation was the more abundant feature, while for the bound form the distribution shifted towards faster drift times and a more compact conformation. This shift is consistent with prior observations on an unmodified Robo1-Ig1-2 construct and further validates use of this lanthanide-binding construct going forward.[9]

Having confirmed both lanthanide and HS-binding activity of our Robo1 construct, we expressed an isotope labeled sample for NMR spectroscopy. The expression media was enriched with ^{13}C1-glucose and ^{13}C-dimethyl-valine, which leads to ^{13}C incorporation in all alanine and valine methyl groups. NMR spectra were recorded on samples prepared with and without paramagnetic Dy^{3+} and also with and without Arixtra (4 samples in total) allowing measurement of PCS data in bound and unbound states. Resonance assignments were determined using an updated version of the Assign_SLP software. Briefly, resonances were assigned by comparison of measured chemical shifts, NOEs, and TOCSY correlations with predicted values to find the most optimal assignment.

PCS values were measured from ^{13}C,^{1}H-HETCOR spectra (Figure 5.3). Direct observation of carbon nuclei proved useful for hRobo1-Ig1-Ig2 by maximizing the resolution in the ^{13}C dimension which accounted for most of the dispersion in the observed methyl signals. While peak overlap increases in the paramagnetic spectra, the majority of PCSs could be measured unambiguously by exploiting the 1:1 shift in each dimension. Several paramagnetic peaks were observed with no obvious diamagnetic

counterpart and were not included. Comparison of PCS values measured with and without Arixtra show several changes. Valine methyl groups in the Ig2 domain, where the LBP is located, showed little change in their PCS value, while many methyl groups in the Ig1 domain experienced a larger PCS in the Arixtra-bound state (Figure 5.4). V144-Me2 showed the largest change in PCS from 0.148 to 0.218 ppm. The number of changes in PCS values suggests some global change in the domain orientation.

The measured PCS data was next used to find a model of the domain orientation that best fits the data. The initial model was the representative frame of the most probable cluster from the GaMD simulation. A grid search of possible conformers was generated by rotating the Ig1 domain, while leaving the Ig2 static. The pivot point for the rotations was chosen as the alpha carbon of isoleucine 104 (model numbering system) which is centered in the linker between domains. Each conformer was scored against the PCS data by solving for the magnetic susceptibility tensor using the Ig2-domain measurements and back-calculating Ig1-domain PCS values. The agreement between observed and back-calculated Ig1 PCS was measured using a Q factor, which is defined by equation 5.3. Lower values of Q indicate better agreement between a structure and the PCS data.

The results of the orientational grid search using the PCS data measured on the unbound form of Robo1 are shown in Figure 5.5A. The initial structure had a Q factor of 0.63 for Ig1 data, indicating poor agreement with the PCS data. After optimization this improved to a Q factor of 0.32. Comparison of the top 20 best fitting Ig1 orientations with the starting structure shows an Ig1 orientation that is significantly different from the starting structure (Figure 5.5C). These 20 models all fit the data quite well with small variations in Q and fall along a plane of orientations. While it is tempting to conclude that

the spread of structures reflects underlying molecular motion, this is not necessarily true. More likely, the PCS data and the arrangement of valine methyls groups does not create a strong constraint on the orientation along the shown plane of orientations.

This process was repeated using the PCS data measured on the bound-state (Figure 5.5B). The starting structure also showed poor agreement to the Ig1 PCS data, with a Q value of 0.72. The grid search found a slightly different Ig1 orientation to be most consistent with the PCS data, and a lower Q score of 0.25. The 20 lowest scoring conformers are shown in Figure 5.5D, which adopt a similar orientation as that observed with the unbound data; however, in this case the bundle of Ig1 orientations is more tightly clustered. In this case, the lower Q factor indicates that the PCS data better fits a single rigid structure. The reduced spread of the 20 best models is a reflection of the deeper "well" of Q values observed in the contour plot (5B).

Discussion

We have used PCS data to determine the average orientation of Robo1-Ig1-2 domains both with and without the presence of an HS ligand. Interestingly, the best fitting structures show a significant deviation from previously determined X-ray crystal structures. Crystal packing forces may be responsible for the observed differences in domain orientation.

These results show that the observed changes in PCS upon binding Arixtra result in a small change in the domain orientation of the Ig1 and Ig2 domains, highlighting the sensitivity of PCS data to small structural changes. It is important to note this analysis was performed using a single structure and does not include the effects of averaging over

multiple conformations. Our MD simulations indicate that hRobo1-Ig2 exhibits some flexibility and likely exists in multiple states in solution. As such, these results inform upon changes in the average domain orientation over the ensemble of conformers. Nonetheless, the improvement in Q score for the Arixtra-bound PCS dataset suggests that the flexibility has decreased. HS is known to bind near the hinge region connecting the two domains and it is reasonable that the presence of a ligand would limit the extent of interdomain motion.

Additionally, these results are not consistent with the ion mobility findings. This discrepancy is likely the result of gas-phase compaction of hRobo1-Ig1-Ig2. Experimental collision-cross sections (CCS) of globular proteins are often found to be smaller (~10%) than predictions from X-ray or NMR structures.[28] This small degree of compaction is often attributed to "self-solvation" and rearrangement of surface side-chains. More extreme deviations in CCS values have been observed for non-globular proteins and proteins containing flexible linkers (e.g. IgG antibodies).[29] While hRobo1-Ig1-2 does not have a disordered linker between domains, there may be sufficient flexibility to collapse into a more spherical structure upon transfer to the gas phase.

In light of the relatively small change in interdomain orientation upon binding Arixtra, it seems unlikely that an HS-induced conformation change plays an important role in Robo-Slit signaling pathways. Instead, one function of HS-binding may be to facilitate diffusion of Robo1 and Slit2 within the extracellular matrix. Diffusion along a one-dimensional HS "rail" would be much more efficient then in three dimensions and aid the two proteins finding one another. HS may also strengthen the interaction by bridging both proteins.

This report also demonstrates the utility of the sparse labeling approach. When combined with paramagnetic lanthanide tag, numerous long-range structural restraints can be measured. Similar strategies could be adopted to study the structure and function of other post-translationally modified proteins produced in native eukaryotic cell culture. Our structural models started from previously determined X-ray crystal structures, but this may not always be necessary. Recent advancements in the accuracy of structure prediction algorithms suggest that predicted models could instead serve as a starting point for structural investigations.

References

1. 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. *J Biol Chem* **2008,** *283* (23), 16226-34.
2. 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.
3. Andrews, W.; Liapi, A.; Plachez, C.; Camurri, L.; Zhang, J.; 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), 2243-52.
4. Blockus, H.; Chédotal, A., Slit-Robo signaling. *Development* **2016,** *143* (17), 3037.

5. 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* **2007,** *104* (38), 14923-14928.

6. Li, Z.; Moniz, H.; Wang, S.; Ramiah, A.; Zhang, F.; Moremen, K. W.; Linhardt, R. J.; Sharp, J. S., High structural resolution hydroxyl radical protein footprinting reveals an extended Robo1-heparin binding interface. *J Biol Chem* **2015,** *290* (17), 10729-40.

7. 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,** *11* (11), 3106-3113.

8. Aleksandrova, N.; Gutsche, I.; Kandiah, E.; Avilov, S. V.; Petoukhov, M. V.; Seiradake, E.; McCarthy, A. A., Robo1 Forms a Compact Dimer-of-Dimers Assembly. *Structure* **2018,** *26* (2), 320-328.e4.

9. Zhao, Y.; Yang, J. Y.; Thieker, D. F.; Xu, Y.; Zong, C.; Boons, G.-J.; Liu, J.; Woods, R. J.; Moremen, K. W.; Amster, I. J., A Traveling Wave Ion Mobility Spectrometry (TWIMS) Study of the Robo1-Heparan Sulfate Interaction. *Journal of The American Society for Mass Spectrometry* **2018**.

10. 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. *Journal of the American Chemical Society* **2011,** *133* (4), 808-819.

11. 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 Science* **2012,** *21* (10), 1456-1466.

12. Gao, Q.; Yang, J.-Y.; Moremen, K. W.; Flanagan, J. G.; Prestegard, J. H., Structural Characterization of a Heparan Sulfate Pentamer Interacting with LAR-Ig1-2. *Biochemistry* **2018,** *57* (15), 2189-2199.

13. Moremen, K. W.; Ramiah, A.; Stuart, M.; Steel, J.; Meng, L.; Forouhar, F.; Moniz, H. A.; Gahlay, G.; Gao, Z.; Chapla, D.; Wang, S.; Yang, J.-Y.; Prabhakar, P. K.; Johnson, R.; Rosa, M. d.; Geisler, C.; Nairn, A. V.; Seetharaman, J.; Wu, S.-C.; Tong, L.; Gilbert, H. J.; LaBaer, J.; Jarvis, D. L., Expression system for structural and functional studies of human glycosylation enzymes. *Nature Chemical Biology* **2018,** *14* (2), 156-162.

14. Duvaud, S.; Gabella, C.; Lisacek, F.; Stockinger, H.; Ioannidis, V.; Durinx, C., Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Research* **2021,** *49* (W1), W216-W227.

15. Haynes, S. E.; Polasky, D. A.; Dixit, S. M.; Majmudar, J. D.; Neeson, K.; Ruotolo, B. T.; Martin, B. R., Variable-Velocity Traveling-Wave Ion Mobility Separation Enhancing Peak Capacity for Data-Independent Acquisition Proteomics. *Analytical Chemistry* **2017,** *89* (11), 5669-5672.

16. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A., NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **1995,** *6* (3), 277-293.

17. Lee, W.; Tonelli, M.; Markley, J. L., NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **2014,** *31* (8), 1325-1327.

18. 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. *Journal of Computational Chemistry* **2004,** *25* (13), 1605-1612.

19. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation11Edited by J. Thornton. *Journal of Molecular Biology* **1999,** *285* (4), 1735-1747.

20. D.A. Case, I. Y. B.-S., S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman *AMBER 2018*, University of California, San Francisco: 2018.

21. Mahoney, M. W.; Jorgensen, W. L., A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics* **2000,** *112* (20), 8910-8922.

22. Miao, Y.; Feher, V. A.; McCammon, J. A., Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation* **2015,** *11* (8), 3584-3595.

23. Matsunaga, Y.; Sugita, Y., Refining Markov state models for conformational dynamics using ensemble-averaged data and time-series trajectories. *The Journal of Chemical Physics* **2018,** *148* (24), 241731.

24. Valafar, H.; Prestegard, J. H., REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **2004,** *167* (2), 228-41.

25. Orton, H. W.; Huber, T.; Otting, G., Paramagpy: software for fitting magnetic susceptibility tensors using paramagnetic effects measured in NMR spectra. *Magn. Reson.* **2020,** *1* (1), 1-12.

26. Nitz, M.; Franz, K. J.; Maglathlin, R. L.; Imperiali, B., A Powerful Combinatorial Screen to Identify High-Affinity Terbium(III)-Binding Peptides. *ChemBioChem* **2003,** *4* (4), 272-276.

27. Hutchinson, E. G.; Sessions, R. B.; Thornton, J. M.; Woolfson, D. N., Determinants of strand register in antiparallel β-sheets of proteins. *Protein Science* **1998,** *7* (11), 2287-2300.

28. Rolland, A. D.; Prell, J. S., Computational insights into compaction of gas-phase protein and protein complex ions in native ion mobility-mass spectrometry. *TrAC Trends in Analytical Chemistry* **2019,** *116*, 282-291.

29. Hansen, K.; Lau, A. M.; Giles, K.; McDonnell, J. M.; Struwe, W. B.; Sutton, B. J.; Politis, A., A Mass-Spectrometry-Based Modelling Workflow for Accurate Prediction of IgG Antibody Conformations in the Gas Phase. *Angewandte Chemie International Edition* **2018,** *57* (52), 17194-17199.

Figures

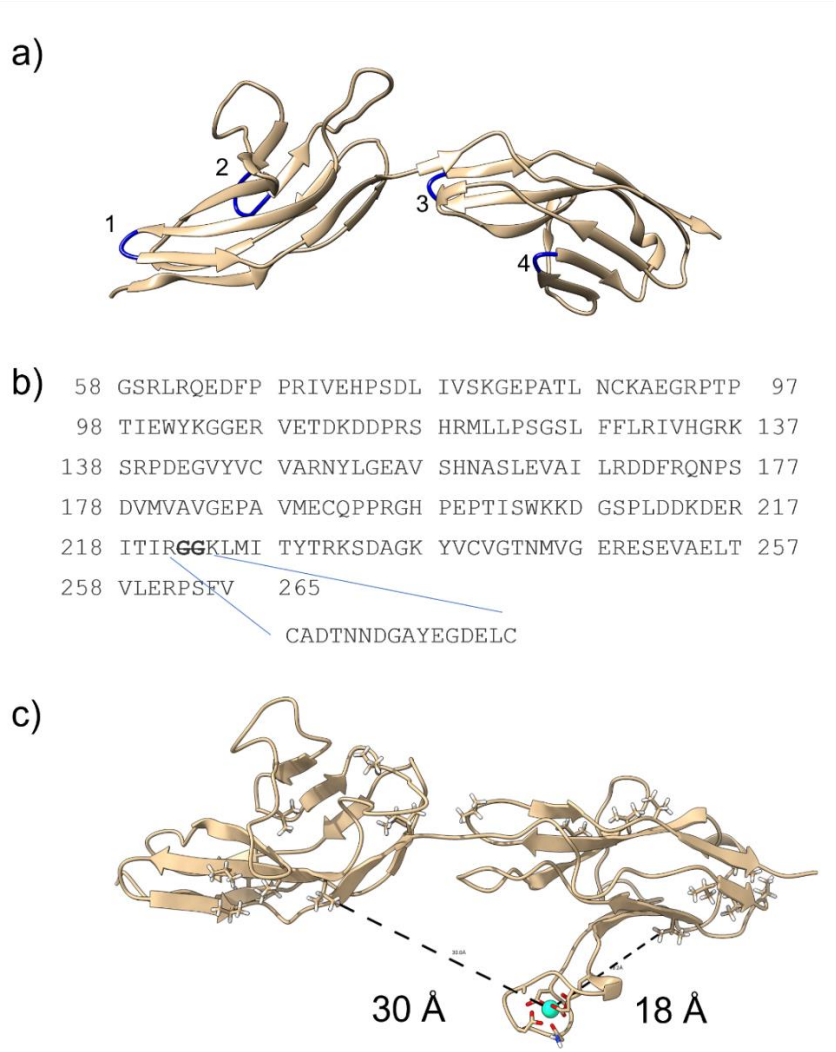

Figure 5.1. Design of Robo1-Ig1-Ig2-Loop4. A) Structure of native Robo1-Ig1-2 (PDB 2V9R). The four sites considered for LBP insertion are colored blue and labeled 1 through 4. B) The modelled sequence of Robo1 is shown with the inserted lanthanide-binding sequence. Amino acid numbers from Uniprot entry Q9YN67 are used. C) Modelled structure of Robo1-Ig1-2-Loop4 taken from a cMD simulation. The dashed lines show distances between the modelled Dy^{3+} ion and two valine methyl groups, one in each domain, which suggests the loop is well positioned to inform on the domain orientation

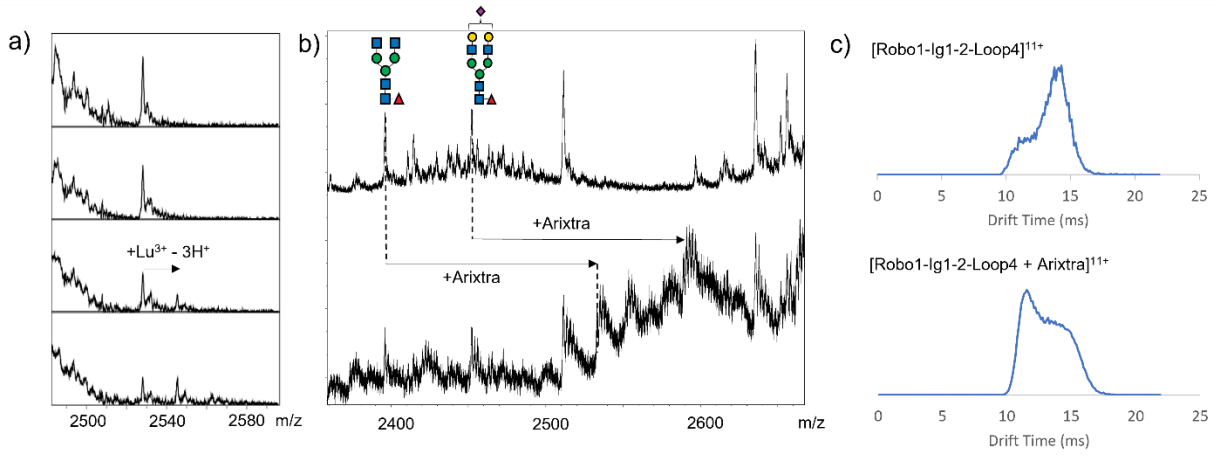

Figure 5.2. Native MS of Robo1-Ig1-2-Loop4. A) Titration with $LuCl_3$. Panels show the same region of the native mass spectrum with increasing levels of lanthanide (0, 10, 20, and 30 µM from top to bottom). A peak consistent with the lanthanide-bound complex is observed at 20 µM $LuCl_3$ and higher. B) Comparison of native mass spectra without addition of Arixtra (top) and with an equimolar addition (bottom). Several new peaks appear in the presence of Arixtra. Protein-Arixtra complexes are indicated for two glycoforms. A plausible glycan consistent with the molecular weight is shown for each species. C) Comparison of representative ion mobility arrival time distributions of a single glycoform alone and in complex with Arixtra. In both cases two peaks are observed corresponding to compact and extended conformations of the molecule. The distribution shifts toward the compact form in the protein-Arixtra complex.

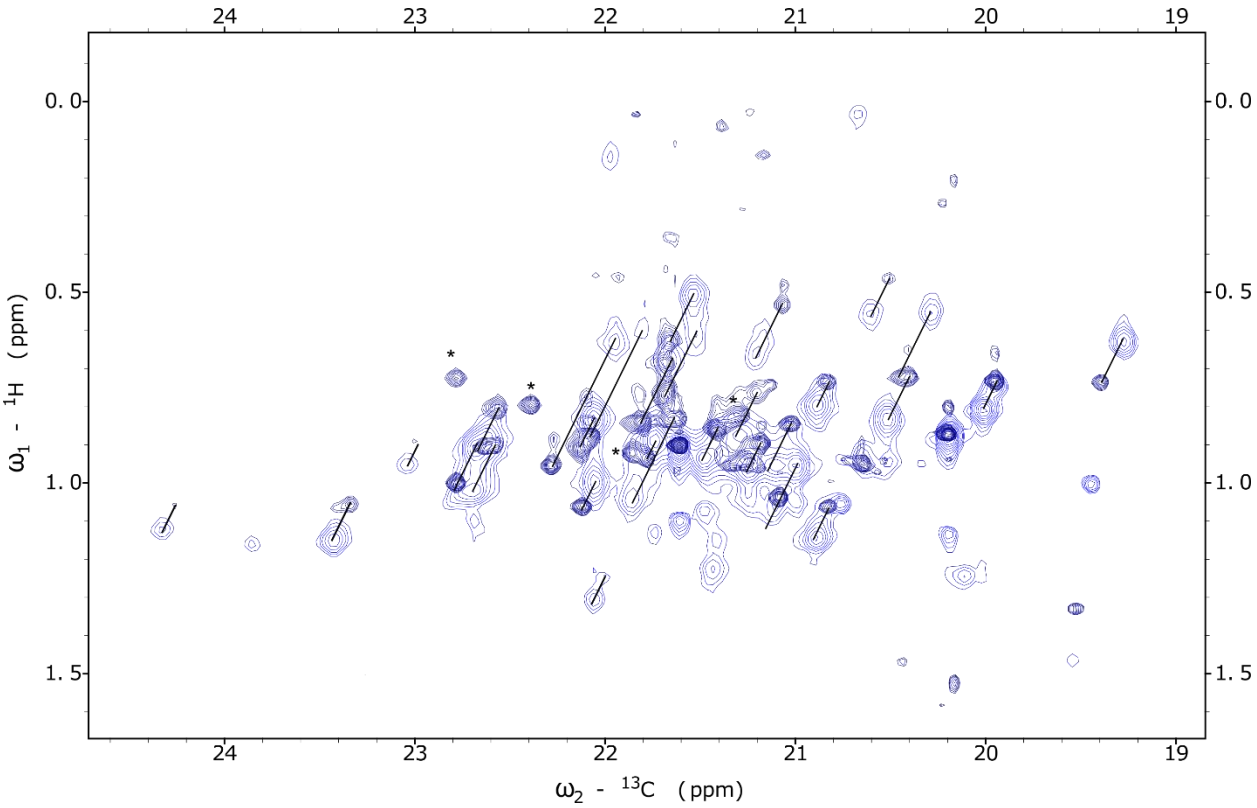

Figure 5.3. HETCOR spectrum of hRobo1-Ig1-Ig2-loop4 + Arixtra with (dark blue) and without (light blue) Dy^{3+}. The black lines connect diamagnetic and paramagnetic peaks for the same resonance. Several signals disappear in the Dy^{3+} spectrum and are indicated by asterisks.

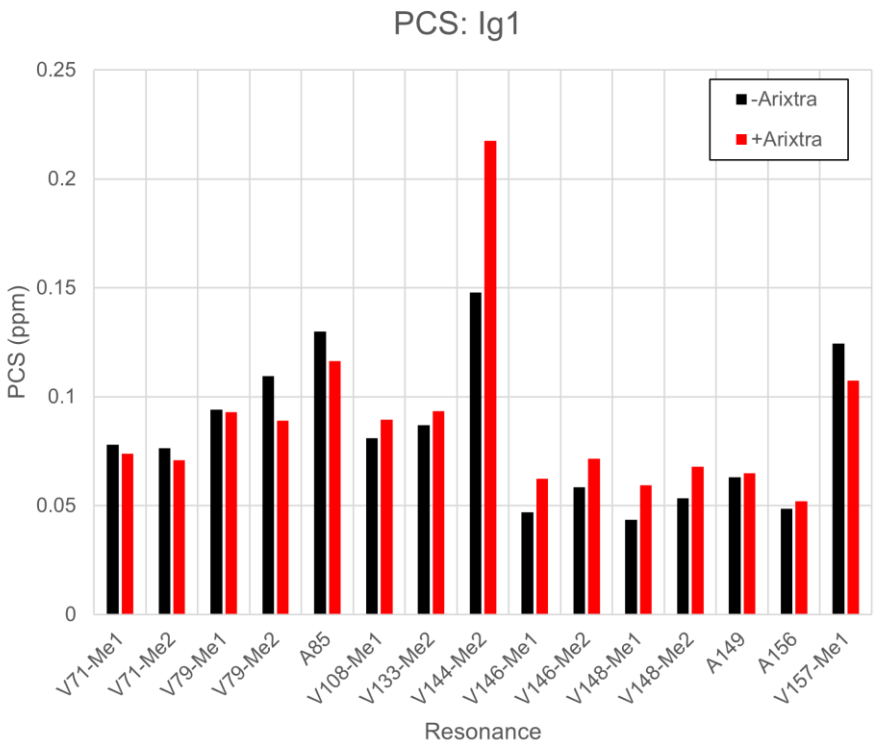

Figure 5.4. Comparison of Ig1 PCS measurements with (red) and without (black) Arixtra. Only those residues observed in both sets of data are shown.

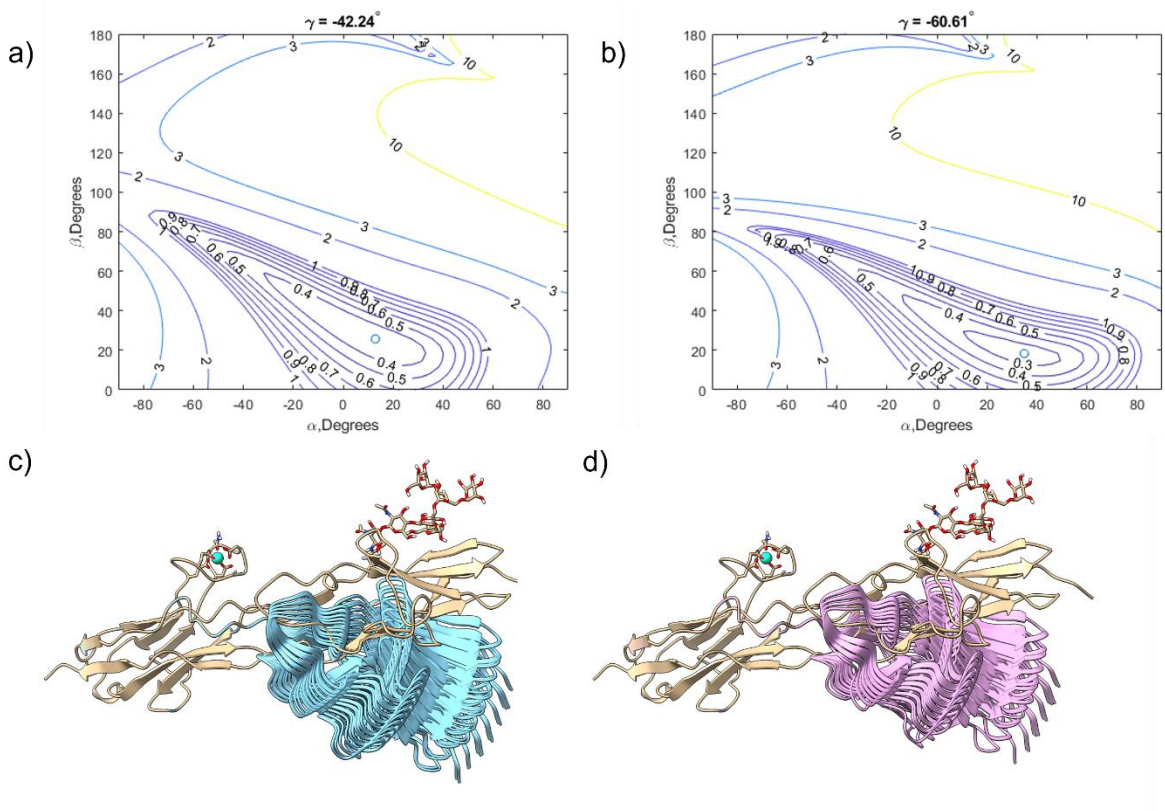

Figure 5.5. Results of Ig1 orientation grid search. a) A contour plot showing the value of Q calculated using PCS data for unbound Robo1-Ig1-2-LBP4 as a function of Euler angles α and β while γ is held constant. The best model is indicated as an open circle. b) A second contour plot showing Q calculated using PCS data for Arixtra-bound Robo1-Ig1-2-LBP4. c) Overlay of the initial Robo1-Ig1-2-LBP4 conformation (tan) with the 20 lowest scoring models (cyan) found with unbound PCS data. D) Overlay of the initial Robo1-Ig1-2-LBP4 conformation (tan) with the 20 lowest scoring models (pink) found with Arixtra-bound PCS data.

CHAPTER 6

SITE-TO-SITE CROSSTALK IN OST-B GLYCOSYLATION OF hCEACAM1-IgV[5]

[5]Williams, R. V.; Huang, C.; McDermott, C.; Ahmed, T.; Columbus, L.; Moremen, K. W.; Prestegard, J. H.; Amster, I. J. Submitted to the *Proceedings of the National Academy of Sciences*, 02/18/2022.

Abstract

N-glycosylation is a common post-translational modification of secreted proteins in eukaryotes. This modification targets asparagine residues within the consensus sequence, N-X-S/T. While this sequence is required for glycosylation, the initial transfer of a high mannose glycan by oligosaccharyl transferases A or B (OST-A or OST-B) can lead to incomplete occupancy at a given site. Factors that determine the extent of transfer are not well understood and understanding them may provide insight into the function of these important enzymes. Here, we use mass spectrometry to simultaneously measure relative occupancies for three N-glycosylation sites on the N-terminal IgV domain of the recombinant glycoprotein, hCEACAM1. We demonstrate that addition is primarily by the OST-B enzyme and propose a kinetic model of OST-B N-glycosylation. Fitting the kinetic model to the MS data yields distinct rates for glycan addition at each site and suggests a largely stochastic initial order of glycan addition. The model also suggests that glycosylation at one site influences the efficiency of subsequent modifications at the other sites and glycosylation at the central or N-terminal site leads to dead-end products that will not lead to full glycosylation of all three sites. Only one path of progressive glycosylation, one initiated by glycosylation at the C-terminal site, can lead to full occupancy for all three sites. Thus, the hCEACAM1 domain provides an effective model system to study site-specific recognition of glycosylation sequons by OST-B and

suggests that the order and efficiency of post-translational glycosylation is influenced by steric crosstalk between adjoining acceptor sites.

Significance Statement

N-Glycosylation is a common posttranslational modification of extracellular proteins. It plays a role in protein folding in the endoplasmic reticulum and later participates in important protein-carbohydrate interactions. Viruses, in particular, are often highly glycosylated. Knowledge of the determinants of modification at a particular site is important for better understanding these processes. One challenge in obtaining this information is separating the effect of the two glycosyltransferases, OST-A and OST-B. Our study focuses on OST-B and reveals how multiple, closely spaced glycosylation sites can influence one another in the context of OST-B glycosylation.

Introduction

Glycosylation is a common post-translational modification of eukaryotic proteins in which a carbohydrate is attached to the sidechains of specific residues. In mammals, most secreted and membrane proteins are O or N-glycosylated with complex structures.[1] In addition, single sugars attached to cytosolic proteins play important regulatory roles in the cytoplasm and nucleus.[2] N-glycosylation is perhaps best understood; it occurs on asparagine residues found within a consensus motif of N-X-S/T and plays important biological roles in both protein folding and cell signaling processes.[3] However, not all consensus sequons are glycosylated and occupancy at some sites can be incomplete.

Omission can have biological consequences; here we present data that can uncover factors leading to these omissions.[4]

Some specificity in N-glycosylation may arise in the initial transfer of a high mannose glycan ($Glc_3Man_9ClcNAc_2$) from a dolichol-phosphate donor to the consensus asparagine. This step is catalyzed by the oligosaccharyltransferase (OST) complex which comes in two forms (OST-A and OST-B) (Figure 6.1A). OST-A associates with the translocon and glycosylates nascent polypeptides in a co-translational manner, while OST-B acts post-translationally.[5] Additionally, OST-B forms transient disulfide bonds with protein substrates via an oxidoreductase subunit.[6] The differential activity of these two complexes has been investigated at the proteome level in combination with CRISPR/Cas9 mutants deficient for either catalytic subunit (STT3A or STT3B). This study found that many N-glycosylation sites modified by OST-A are fully occupied and proposed that OST-B glycosylates primarily a subset of sequons that were skipped by OST-A.[7] Most of these skipped sites are within 65 residues of the C-terminus of a protein, a stretch of residues that may be required to retain attachment to the ribosome while OST-A acts.

Since OST-B acts post-translationally, it is not constrained to an ordered addition to substrate glycosylation sites from N-terminal to C-terminal, as is the case for co-translational addition by OST-A. In addition, steric constraints may contribute to substrate access since folding of protein substrates may be initiated or completed prior to OST-B action (see Figure 6.1A). Here, we use mass spectrometry (MS) to examine the

glycosylation of a small protein, the N-terminal IgV domain of human Carcinoembryonic Antigen Cell Adhesion Molecule 1, hCEACAM1.

hCEACAM1 is a highly glycosylated extracellular protein receptor, which has been implicated in gastrointestinal autoimmune disorders and host-pathogen interactions.[8-9] The N-terminal immunoglobulin-V-like domain (IgV) contains three N-glycosylation sites. While our initial study of hCEACAM1 was motivated by its biological function, the hCEACAM1-IgV construct proves a convenient model system for understanding OST catalyzed glycosylation. First, the small size of the N-terminal domain (12 kDa) makes it an accessible target for top-down MS/MS analysis of the sites of glycosylation. Second, despite the small size of the protein, this construct contains three consensus sequons all within 35 residues of the C-terminus, sites N104, N111, N115 of the Uniprot sequence P13688, suggesting that all glycosylation may be dependent on OST-B. Lastly, hCEACAM1-IgV does not possess any cysteine residues and therefore will not involve the oxido-reductase activity of OST-B; in their absence glycosylation levels are lower allowing the level of glycosylation to report more directly on accessibility of the OST-B catalytic site.

Top-down mass spectrometry (TDMS) is an advanced mass spectrometry method in which an intact protein is analyzed directly, as opposed to analysis of proteolytic fragments.[10] The top-down approach is especially beneficial when a protein has multiple modifications on different regions of the protein. More widely used bottom-up approaches rely on proteolysis which can, in principle, separate PTMs onto different peptides. Tandem mass spectrometry is more easily accomplished for peptides than intact proteins, but information can be lost by enzymatic digestion, specifically, the correlation

between sites of modification for sub-stoichiometric PTMs. TDMS allows for the localization of multiple glycosylation sites simultaneously and informs upon the relative abundance of distinct glycan occupancy states for the intact protein. It does, however, require extensive and controlled fragmentation that focuses on backbone cleavage and does not dissociate the PTMs of interest. Electron-based fragmentation techniques, such as electron capture dissociation (ECD) or electron transfer dissociation (ETD), are employed here to improve sequence coverage and to preserve labile PTMs.[11-13] Using these MS approaches, relative occupancies for three N-glycosylation sites on the N-terminal IgV domain of the recombinant glycoprotein, hCEACAM1 were simultaneously measured.

The resulting information was used to develop a kinetic model of OST-B catalyzed glycosylation (Figure 6.1B). Surprisingly, glycosylation at the central or more N-terminal site led to dead-end products that could not be fully glycosylated at all three sites. Only a single path of ordered glycosylation of the three sites, starting with the C-terminal site, followed by the central site and finally the N-terminal site, led to a fully glycosylated recombinant product. Thus, the order of site-specific modification by OST-B is largely stochastic, and subsequent glycosylations appear to be sterically impacted by the position of the initial glycosylation event. Only a select path of progressive glycan site occupancy could lead to full occupancy in contrast to the efficient N- to C-terminal

glycan modification that is characteristic of co-translational OST-A catalyzed glycosylation.

Results

Glycosylation occupancy

After the initial transfer of a high-mannose glycan to CEACAM1-IgV sites, subsequent trimming in the endoplasmic reticulum (ER) and further glycan additions in the Golgi cisternae would normally lead to a complex mixture of glycoforms that could complicate MS analysis. To avoid this, recombinant hCEACAM1-IgV was expressed in a MGAT1 null HEK293 cell line (HEK293S GnT1$^-$) that only produces $Man_5GlcNAc_2$ glycans.[14] These glycans are susceptible to cleavage between the two N acetylglucosamine (GlcNAc) residues by endoglycosidase-F1 (EndoF1) leaving a single GlcNAc "scar" at any glycosylated site.[15] The simplified mass spectrum of intact hCEACAM1-WT, after EndoF1 treatment, showed a mixture of three species with deconvoluted monoisotopic masses of 12417.22, 12620.29, and 12823.37 Da, which correspond to the protein with 1, 2, or 3 GlcNAc modifications. These species were observed with relative proportions of 29% , 46%, and 26%, respectively (Figure 6.2A). The species lacking glycans was not observed, and presumably is present at a level below noise.

The glycosylation positions of species with one and two GlcNAcs are indeterminate in MS spectra of the intact protein. To determine these positions, top-down ECD MS/MS was applied. An example of the ECD MS/MS fragmentation coverage is shown in Figure 6.2D for the monoglycosylated components. Fragments were observed across a large

portion of the protein backbone. More importantly, fragmentation between the three glycosylation sites is observed allowing determination of site-specific glycan occupancy.

By combining the relative abundances from the intact mass spectrum with the fragmentation patterns from the MS/MS spectra, the abundances of each potential "occupancy state" were determined for wild type, wild type in the presence of an OST-B inhibitor and three glycosylation site mutants (Table 6.1). These states are denoted by a 3-bit binary code in which the bits correspond to sites N104, N111, and N115, and "1" indicates glycosylation and "0" indicates the absence of glycosylation. We first focus on wild type data under normal OST-A and OST-B activity. The results show that the locations of glycans are not random. Instead, there are clear patterns for modification preference. In the case of the species with 1 GlcNAc, the modification was predominantly located at the central N111 sequon, [010]. While in the case of the 2 GlcNAc species, a mixture of [110] and [101] in similar amounts was observed and [011] was observed at a lower level.

To test the reproducibility of these results, hCEACAM1-IgV-WT was expressed in duplicate by two different laboratories. The standard deviations in glycoform distributions are listed in Table 6.1. While there is some variation, there is consistency in that species [010], [110], [101] and [111] were highly populated and [000], [100], [001] and [011] were absent or sparingly populated.

Glycosite Mutants

To further explore the patterns in N-glycosylation, we produced three mutant versions of hCEACAM1-IgV where each glycosylated asparagine residue was individually

mutated to a glutamine, thus eliminating the possibility of glycosylation at the respective site. The top-down MS analysis was then repeated on each of these samples. The intact mass spectra of each mutant show significant changes compared to the wild type construct (Figure 6.2B). In each case, species were observed corresponding to the protein with 1 asparagine to glutamine mutation (+14 Da) and 1 or 2 GlcNAc modifications. The most dramatic change is observed when site 2 (N111) is mutated. Near complete glycosylation was observed with the fully occupied 2 GlcNAc form comprising 98% of the protein signal. Fragmentation patterns of the 1 GlcNAc form of each mutant give insight into pair-wise correlations between glycosylation at the remaining sites. For the N104Q mutant the single GlcNAc was primarily observed at N111. A similar trend was seen for the N115Q mutant where the glycan modification was only observed at N111.

Support for OST-B exclusive glycosylation

Using these data, we sought to develop a simple model describing the glycosylation events; however, the details of this model depend on whether hCEACAM1-IgV is glycosylated co-translationally by OST-A or post-translationally by OST-B. Based on prior literature we expected OST-A to be unable to glycosylate this construct due to the close proximity of the N-glycosylation sequons to the C terminus. However, to test this expectation, we prepared a two-domain hCEACAM1 construct (hCEACAM1-IgV-IgC) that appended an additional 95 amino acid IgC domain C-terminal to the IgV domain. The added domain carries an additional 6 N-glycosylation sites, making 9 sites in total. Intact mass measurement of this sample showed a mixture of species consistent with 5 to 9 GlcNAc residues (Supplemental Figure B.3A). ECD fragmentation spectra revealed that

the first three glycosylation sites, corresponding to those on the IgV domain, were now highly occupied. No fragment ions were observed without GlcNAc modification at these sites (Supplemental Figure B.3B). This observation is consistent with the fact that these three sites are now more than 65 residues away from the C terminus and that glycosylation by both OST-A and OST-B is possible. Also, the fourth glycosylation site is sufficiently far from the C-terminus for OST-A glycosylation and was also observed to be completely occupied by top-down ECD MS/MS. The remaining 5 glycosylation sites are within 65 residues of the C-terminus and, as expected, were variably modified.

One potential caveat regarding the evidence from the 2-domain construct is that the second domain contains a disulfide bond. This raises the possibility that the increased glycosylation occupancy of sites 1, 2, and 3 may not be solely due to engaging OST-A but also may have enhanced OST-B activity, due to the oxidoreductase subunit of OST-B. To further test the mechanism of N-glycosylation, we expressed the hCEACAM1-IgV construct while treating the HEK293 expression culture with an OST-B inhibitor that has been shown to selectively repress glycosylation by OST-B.[16] Intact mass spectra show that the inhibitor-treated protein sample is significantly less glycosylated than untreated samples (Figure 6.2C). While suppression of glycosylation is not complete, this may reflect some site specificity in suppression as opposed to residual activity of OST-A.[16] The inhibitor is not believed to be an active site inhibitor, which would uniformly suppress glycosylation, but is believed to act allosterically, a manner that could be site specific.

Moreover, it was previously observed that administration of the inhibitor to an STT3A knockout did not completely prevent glycosylation of the reporter protein.[16]

Tandem mass spectrometry was also performed on the inhibitor-treated samples resulting in glycoform distributions shown in Table 6.1. The results show that the most abundant form in the inhibitor treated sample contains a single glycan at the central glycosylation site (N111). Interestingly, small amounts of [100] and [001], previously unobserved, were also detected.

Kinetic Modelling

Direct interpretation of the relative amounts of the hCEACAM1-IgV glycoforms is difficult. Several minor species are observed in small amounts and, thus, may be transient intermediates that are quickly glycosylated. Another explanation is that less abundant species are produced slowly for reasons due to the mechanism or structure of the enzyme. To gain further insight into the mechanism of N-glycosylation of hCEACAM1-IgV we developed a kinetic model of OST-B catalyzed addition.

We started modeling with a hypothesis that post-translational addition of N-glycans would be stochastic with 12 possible reactions, each with a unique rate constant. In addition, to mimic the cell expression process, the model employed a single rate describing the protein entering the ER via translation and a single common rate for the secretion of each species. Finally, since the MS analysis gives only ratios of the various glycosylated species, only the relative magnitudes of the constants are meaningful.

We have increased the amount of modeling data by examining glycosylation of mutants that eliminate glycosylation at each of the three consensus sites (N to Q mutations).

If we assume that mutation acts only by elimination of the respective glycosylation site, while kinetic constants for glycosylation of the remaining sites are the same as for wild type, we can combine the MS data for the wild type and mutant hCEACAM1 samples. Thus, we have 19 data points that can be used to solve the resulting system of differential equations to obtain values for the 14 rate constants. The obtained values for each rate constant are shown in Figure 6.3 in which the determined rates have replaced the kinetic constants in Figure 6.1B, and our binary codes have replaced the nodes in Figure 6.1B. Note that the rates of substrate entering and sum of products leaving are exactly the same. Any ER-associated degradation (ERAD) processes are not explicitly included in our model but may lead to artificially low values for some rate constants. Regardless of the limitations of our model, these results suggest distinct pathways for site-specific occupancy of hCEACAM1-IgV. Glycosylation of site 1 and site 2 proceeds efficiently, but further addition at site 3 is unfavorable. Initial glycosylation of site 2 is also favorable, but when this occurs, this leads to a dead end where further addition at either of the remaining sites is very slow. In fact, the only efficient pathway to full glycosylation proceeds via site 3, then site 2, and finally site 1.

These results can be tested by the glycoform distribution measured on the hCEACAM1 IgV sample expressed during OST-B inhibitor treatment (Table 6.1). If we assume that inhibition slows the rate of all glycosylation steps equally then one would expect an accumulation of intermediate states. Our model predicts that [100] and [001] are two such intermediates. The data shows that these species are present in larger amounts than without inhibitor treatment; however, they are still minor glycoforms compared to the [010] species. This discrepancy may reflect a more complex mechanism of inhibition that

impacts some glycosylation sites more severely than others. Alternatively, some specificity in ERAD may play a significant role in the low abundance of [000], [100] and [001] glycoforms.

Discussion

A single dominant path to full glycosylation implies there are restraints on the process and mechanism. Analysis of the structure may provide insight. The structure of the human OST-B complex has been solved by cryo-EM with electron density consistent for an average peptide acceptor sequon bound in the catalytic site.[17] To our knowledge there is no evidence that OST-B acts in a processive fashion on an unfolded protein. Also, un-glycosylated versions of hCEACAM1-IgV readily fold and several crystal structures of bacterially expressed protein have been determined.[18-19] Hence, it seems reasonable to examine the possibility that OST-B acts on a form of hCEACAM1-IgV that can theoretically range between a fully folded to an un-folded state. A folded model based on an X-ray crystal structure (PDB 4QXW) of hCEACAM1-IgV was aligned with the bound peptide at each of the three glycosylation sites. In each case, significant clashes between the folded hCEACAM1-IgV and the OST-B complex were observed (not shown), suggesting that the protein must be at least partially unfolded during glycosylation.

is intriguing that OST-A can more efficiently glycosylate the three sequons of hCEACAM1 IgV than OST-B if a C-terminal extension beyond the IgV domain is provided. The key difference is that OST-A mediated co-translational glycosylation enforces rapid glycosylation as the peptide emerges from the SEC61 translocon prior to

protein folding. Hence, partial folding may be responsible for the slower differential glycosylation at certain sites.

Next, the possible effect of prior glycosylation at adjacent sites was explored. An unfolded polypeptide containing the three N-glycosylation sites was built and aligned so that site 1 (N104) was inside the OST-B catalytic site and then a $Glc_3Man_9GlcNAc_2$ glycan was modelled on site 2 (N111) (Figure 6.4). The 14-residue oligosaccharide is quite large, and it is clear that steric interactions could easily inhibit entry of additional hCEACAM1-IgV sequons.

Unfortunately, without more structural data, it is not possible to build more specific models. In particular, we have ignored the possible motion of OST-B domains. Crystal structures of the bacterial OST PglB (homologous to STT3B) in complex with model substrates have shown that a large loop (EL5) undergoes significant rearrangement between disordered and ordered states as substrates bind and products are released.[20-22] Additional indications of internal motion are present in the cryo-EM structure OST-B17. Several loops in STT3B did not produce strong electron density, and even the entire catalytic domain of the oxidoreductase MagT1 was not observed.

The assumptions of our kinetic model can clearly be debated. We assume that substrates, such as the glycosylated dolichol pyrophosphate and un-glycosylated protein reach steady state levels early in the process and remain at those levels for the majority of expression time. It is unclear if this assumption would be true for protein overexpression during our typical 6 day culture period. Also, glycosylation could alter the folding trajectory of hCEACAM1 and some forms may be preferentially secreted. Additionally, there are quality control processes in the ER that target poorly folded and possibly under-

glycosylated proteins through an ERAD pathway for destruction by cytosolic proteosomes.[23] The loss of un-folded species through the ERAD pathway may be significant as previous expression of unglycosylated hCEACAM1-IgV in E. coli formed inclusion bodies.[24]

Some of these flaws could be corrected by the addition of more processes (and unknown parameters) to the kinetic model. However, the number of independent variables in even this simple model (14) exceeds the amount of MS data on the initial hCEACAM1-IgV construct (7 recognizing that measurements on the 8 species are actually ratios). We were able to compensate by adding data from mutants missing glycosylations sites, but this required additional assumptions regarding a lack of perturbation to remaining sites. Nevertheless, some level of site-to-site interaction is clear and this interaction likely has structural implications.

The interplay of glycans on different sites of glycoproteins could well have implications for other biological functions. Proteins other than CEACAMs are very heavily glycosylated and glycans may well interact in the course of stabilizing certain conformations or protecting sites from interactions with other proteins as in the case of glycan shields of certain virus surfaces from antigenic development.[25] Site specific information on occupancy and what glycans occupy each site could be very important to understanding glycan function in these instances. The top-down MS strategy used in our studies could easily supply these data.

Here we have measured the glycan occupancy of hCEACAM1-IgV using a combination of top-down and bottom-up mass spectrometry. Based on the combination of increased occupancy in a two-domain construct, significantly reduced glycosylation

after treatment with an OST-B specific inhibitor, and literature precedence for inefficient OST-A activity toward sequons proximal to the C-terminus, we conclude that N-glycosylation of this protein is catalyzed by OST B. Using these data, we have proposed a kinetic model of OST-B-driven glycosylation that predicts the only efficient pathway to full glycosylation proceeds from C to N-terminus. The N-glycosylation process is a complex interplay between the activity of the OST complexes, protein folding, and subsequent trafficking through the ER and/or Golgi apparatus. Our results suggest that folding trajectories are glycoform specific and would be ripe for future inquiry. Similar studies on other glycoproteins may shed further light on these complex processes.

Materials and Methods

Protein expression and purification

hCEACAM1 expression constructs were synthesized using human codon optimization (ThermoFisher) and were inserted into a pGEn2 expression vector.[26] The resulting expression products contain an N-terminal signal sequence followed by a His-tag, GFP domain, TEV protease recognition site, and an SGG linker followed by the hCEACAM1-IgV domain (UniProt P13688, residues 34-141). Three glycosylation site mutants (N104Q, N111Q and N115Q) were generated using a Q5 site-directed mutagenesis kit (New England Biolabs) and a two domain hCEACAM1-IgV-IgC (residues 34-236) construct was also prepared in the same vector backbone. All constructs were expressed by transient transfection in HEK293S (GnT1$^-$) cells (ATCC) leading to secretion of the recombinant product into conditioned medium.[14, 26] The HEK293S (GnT1$^-$) recombinant host is defective in the glycan processing enzyme,

MGAT1, and leads to secretion of glycosylated products harboring $Man_5GlcNAc_2$ glycan structures that can be subsequently cleaved by endoglycosidase F1 (EndoF1) to result in a single GlcNAc-Asn linkage at the respective glycosylation site. hCEACAM1 expression products were harvested six days post-transfection and the conditioned media was subjected to centrifugation prior to protein purification. For OST-B inhibition studies, inhibitor C19 (Enamine, cat#: Z26531254) was added to the cell culture medium to a final concentration of 25 μM with 0.1% (v/v) of DMSO. For protein purification, the crude medium was loaded onto Ni-NTA Superflow (Qiagen, Germantown, MD) column, washed with 25 mM HEPES, 300 mM NaCl, 20 mM imidazole, pH 7.0, and eluted with 25 mM HEPES, 300 mM NaCl, 300 mM imidazole, pH 7.0. Purified protein was concentrated and treated with recombinant TEV protease and EndoF1 (both His-tagged) simultaneously for 24 h at 4°C to remove the GFP fusion tag and N-glycans. The digestion products were then passed through a Ni-NTA Superflow column to remove the cleaved GFP fusion tag, TEV protease, and EndoF1 to result in a purified untagged protein in the flow-through fraction. Protein was further purified over a Superdex-75 (GE Healthcare Life Sciences, Chicago, IL) column and concentrated to 1-3 mg/ml. Each purification step was checked by SDS-PAGE. Human CEACAM1 variants were expressed and purified in two different laboratories independently in University of Georgia and University of Virginia following the same protocol to act as biological replicates.

Mass spectrometry

Following protein expression and purification, protein samples were buffer exchanged into 10 mM ammonium acetate and diluted to a concentration of 1 µM with 50/49/1 (v/v/v)% methanol-water-formic acid solution for MS analysis. All mass spectra were acquired using a 12 T Bruker SolariX FT-ICR-MS. Sample solutions were directly infused via a syringe pump at a flow rate of 2.0 µL/min and ionized via electrospray with a capillary voltage of 4500 V and end plate offset of -500 V. Ions were accumulated in the collision cell for 0.1 s for MS spectra and 1.0 s for MS/MS prior to transfer to the analyzer cell. A 1.0 ms time of flight delay was set before trapping ions. For MS spectra, 50 scans were co-added while for MS/MS 300 scans were acquired. ECD MS/MS spectra were achieved with an electron bias of 1.0 V, lens voltage of 10.0 V, and electron pulse length of 15-20 ms. The FT-ICR time domain transients contained 2M data points with a length of 0.8389 s, which corresponds to a resolving power of approximately 77,000 at m/z 1000.

For bottom-up confirmation of selected data sets, protein digestion was performed using sequencing grade trypsin (Promega). A 1 mg/mL hCEACAM1 solution was mixed with trypsin at a 50:1 protein to enzyme ratio (w/w) and incubated at 37°C overnight. The resulting digest solution was diluted 100-fold with 50/50/0.1 (v/v/v)% methanol, water, formic acid solution and infused directly into the mass spectrometer. Identical instrument settings were used as above with the exception that for ECD MS/MS experiments the glycopeptide precursor ions were irradiated with electrons for 100 ms and 200 scans were averaged.

Tables of all intact mass distributions and ECD MS/MS fragment ion assignments are included as supplemental information (see Appendix B).

Kinetic modelling

In-house MATLAB scripts were used to integrate the system of kinetic equations and to find optimal rate constants. The system of linear kinetic equations was solved numerically using the built-in "ode45" function. An optimal set of rate constants was found using a genetic algorithm search. Potential sets of rate constants were scored by predicting the glycoform distribution at long integration time and calculating the RMSD with the measured values for hCEACAM1-WT, N104Q, N111Q, N115Q. The scripts are included as supplemental information (see Appendix B). Data from four separate models were included in this search. The rate equations are as follows:

hCEACAM1 WT

Eq. 6.1 $$\frac{d}{dt}[000] = k_{in} - (k_1 + k_2 + k_3 + k_{out})[000]$$

Eq. 6.2 $$\frac{d}{dt}[100] = k_1[000] - (k_{1\to2} + k_{1\to3} + k_{out})[100]$$

Eq. 6.3 $$\frac{d}{dt}[010] = k_2[000] - (k_{2\to1} + k_{2\to3} + k_{out})[010]$$

Eq. 6.4 $$\frac{d}{dt}[001] = k_3[000] - (k_{3\to1} + k_{3\to2} + k_{out})[001]$$

Eq. 6.5 $$\frac{d}{dt}[110] = k_{1\to2}[100] + k_{2\to1}[010] - (k_{12\to3} + k_{out})[110]$$

Eq. 6.6 $$\frac{d}{dt}[101] = k_{1\to3}[100] + k_{3\to1}[001] - (k_{13\to2} + k_{out})[101]$$

Eq. 6.7 $$\frac{d}{dt}[011] = k_{2\to3}[010] + k_{3\to2}[001] - (k_{23\to1} + k_{out})[011]$$

Eq. 6.8 $$\frac{d}{dt}[110] = k_{1\to2}[100] + k_{2\to1}[010] - (k_{12\to3} + k_{out})[110]$$

Eq. 6.9 $$\frac{d}{dt}[111] = k_{12\to3}[110] + k_{23\to1}[011] + k_{13\to2} - k_{out}[111]$$

hCEACAM1 N104Q

Eq. 6.10 $$\frac{d}{dt}[Q00] = k_{in} - (k_2 + k_3 + k_{out})[Q00]$$

Eq. 6.11 $$\frac{d}{dt}[Q10] = k_2[Q00] - (k_{2\to3} + k_{out})[Q10]$$

Eq. 6.12 $$\frac{d}{dt}[Q01] = k_3[Q00] - (k_{3\to2} + k_{out})[Q01]$$

Eq. 6.13 $$\frac{d}{dt}[Q11] = k_{2\to3}[Q10] + k_{3\to2}[Q01] - k_{out}[Q11]$$

hCEACAM1 N111Q

Eq. 6.14 $$\frac{d}{dt}[0Q0] = k_{in} - (k_1 + k_3 + k_{out})[0Q0]$$

Eq. 6.15 $$\frac{d}{dt}[1Q0] = k_1[0Q0] - (k_{1\to3} + k_{out})[1Q0]$$

Eq. 6.16 $$\frac{d}{dt}[0Q1] = k_3[0Q0] - (k_{3\to1} + k_{out})[0Q1]$$

Eq. 6.17 $$\frac{d}{dt}[1Q1] = k_{1\to3}[1Q0] + k_{3\to1}[0Q1] - k_{out}[1Q1]$$

hCEACAM1 N115Q

Eq. 6.18 $$\frac{d}{dt}[00Q] = k_{in} - (k_1 + k_2 + k_{out})[00Q]$$

Eq. 6.19 $$\frac{d}{dt}[10Q] = k_1[00Q] - (k_{1\to2} + k_{out})[10Q]$$

Eq. 6.20 $$\frac{d}{dt}[01Q] = k_2[00Q] - (k_{2\to1} + k_{out})[01Q]$$

Eq. 6.21 $$\frac{d}{dt}[11Q] = k_{1\to2}[10Q] + k_{2\to1}[01Q] - k_{out}[11Q]$$

Acknowledgments

This work was supported by NIH grants R01-GM033225 and R35 GM131829.

References

1. Apweiler, R.; Hermjakob, H.; Sharon, N., On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1999,** *1473* (1), 4-8.
2. Hart, G. W.; Housley, M. P.; Slawson, C., Cycling of O-linked β-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **2007,** *446* (7139), 1017-1022.
3. Stanley P, T. N., Aebi M, N-Glycans. In *Essentials of Glycobiology*, 3rd ed.; Varki A, C. R., Esko JD, et al., Ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2017.
4. Contessa, J. N.; Bhojani, M. S.; Freeze, H. H.; Rehemtulla, A.; Lawrence, T. S., Inhibition of N-Linked Glycosylation Disrupts Receptor Tyrosine Kinase Signaling in Tumor Cells. *Cancer Research* **2008,** *68* (10), 3803.
5. Ruiz-Canada, C.; Kelleher, D. J.; Gilmore, R., Cotranslational and Posttranslational N-Glycosylation of Polypeptides by Distinct Mammalian OST Isoforms. *Cell* **2009,** *136* (2), 272-283.
6. Mohorko, E. et al., Structural Basis of Substrate Specificity of Human Oligosaccharyl Transferase Subunit N33/Tusc3 and Its Role in Regulating Protein N-Glycosylation. *Structure* **2014,** *22* (4), 590-601.
7. Cherepanova, N. A.; Venev, S. V.; Leszyk, J. D.; Shaffer, S. A.; Gilmore, R., Quantitative glycoproteomics reveals new classes of STT3A- and STT3B-dependent N-glycosylation sites. *The Journal of Cell Biology* **2019,** *218* (8), 2782-2796.

8. Gray-Owen, S. D.; Blumberg, R. S., CEACAM1: contact-dependent control of immunity. *Nat Rev Immunol* **2006,** *6* (6), 433-46.
9. Kuespert, K.; Pils, S.; Hauck, C. R., CEACAMs: their role in physiology and pathophysiology. *Current Opinion in Cell Biology* **2006,** *18* (5), 565-571.
10. Siuti, N.; Kelleher, N. L., Decoding protein modifications using top-down mass spectrometry. *Nat Methods* **2007,** *4* (10), 817-21.
11. Zubarev, R. A., Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology* **2004,** *15* (1), 12-16.
12. Zubarev, R. A. et al., Electron Capture Dissociation for Structural Characterization of Multiply Charged Protein Cations. *Analytical Chemistry* **2000,** *72* (3), 563-573.
13. Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W., Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* **1998,** *120* (13), 3265-3266.
14. Reeves, P. J.; Callewaert, N.; Contreras, R.; Khorana, H. G., Structure and function in rhodopsin: High-level expression of rhodopsin with restricted and homogeneous N-glycosylation by a tetracycline-inducible *N*-acetylglucosaminyltransferase I-negative HEK293S stable mammalian cell line. *Proceedings of the National Academy of Sciences* **2002,** *99* (21), 13419.
15. Meng, L. et al., Enzymatic Basis for *N*-Glycan Sialylation: STRUCTURE OF RAT α2,6-SIALYLTRANSFERASE (ST6GAL1) REVEALS CONSERVED AND UNIQUE FEATURES FOR GLYCAN SIALYLATION *Journal of Biological Chemistry* **2013,** *288* (48), 34680-34698.

16. Rinis, N. et al., Editing N-Glycan Site Occupancy with Small-Molecule Oligosaccharyltransferase Inhibitors. *Cell Chemical Biology* **2018,** *25* (10), 1231-1241.
17. Ramírez, A. S.; Kowal, J.; Locher, K. P., Cryo-electron microscopy structures of human oligosaccharyltransferase complexes OST-A and OST-B. *Science* **2019,** *366* (6471), 1372-1375.
18. Huang, Y.-H. et al., CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. *Nature* **2015,** *517* (7534), 386-390.
19. Fedarovich, A.; Tomberg, J.; Nicholas, R. A.; Davies, C., Structure of the N-terminal domain of human CEACAM1: binding target of the opacity proteins during invasion of Neisseria meningitidis and N. gonorrhoeae. *Acta Crystallogr D Biol Crystallogr* **2006,** *62* (Pt 9), 971-9.
20. Lizak, C.; Gerber, S.; Numao, S.; Aebi, M.; Locher, K. P., X-ray structure of a bacterial oligosaccharyltransferase. *Nature* **2011,** *474* (7351), 350-355.
21. Napiórkowska, M.; Boilevin, J.; Darbre, T.; Reymond, J.-L.; Locher, K. P., Structure of bacterial oligosaccharyltransferase PglB bound to a reactive LLO and an inhibitory peptide. *Scientific Reports* **2018,** *8* (1), 16297.
22. Napiórkowska, M. et al., Molecular basis of lipid-linked oligosaccharide recognition and processing by bacterial oligosaccharyltransferase. *Nature Structural & Molecular Biology* **2017,** *24* (12), 1100-1106.
23. Qi, L.; Tsai, B.; Arvan, P., New Insights into the Physiological Role of Endoplasmic Reticulum-Associated Degradation. *Trends in Cell Biology* **2017,** *27* (6), 430-440.

24. Zhuo, Y.; Yang, J. Y.; Moremen, K. W.; Prestegard, J. H., Glycosylation Alters Dimerization Properties of a Cell-surface Signaling Protein, Carcinoembryonic Antigen-related Cell Adhesion Molecule 1 (CEACAM1). *J Biol Chem* **2016,** *291* (38), 20085-95.
25. Zhao, P. et al., Virus-Receptor Interactions of Glycosylated SARS-CoV-2 Spike and Human ACE2 Receptor. *Cell Host & Microbe* **2020,** *28* (4), 586-601.
26. Moremen, K. W. et al., Expression system for structural and functional studies of human glycosylation enzymes. *Nature Chemical Biology* **2018,** *14* (2), 156-162.

Figures and Tables

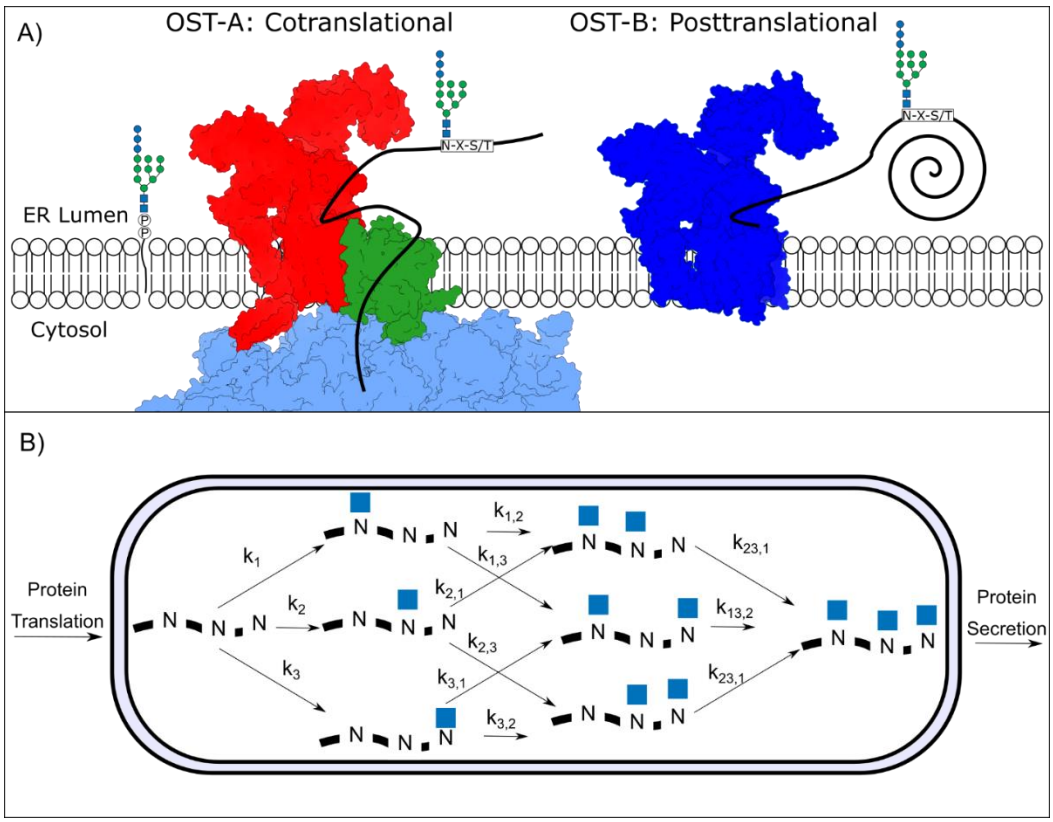

Figure 6.1. A) Diagram of OST-A and OST-B function. On the left, a nascent polypeptide is shown emerging from the ribosome (light blue, PDB 6FTI), passing into the endoplasmic reticulum (ER) lumen via the sec61 channel (green) and being scanned by OST-A (red, PDB 6S7O). On the right, OST-B (blue, PDB 6S7T) is shown interacting with a partially folded peptide substrate. A dolichol-pyrophosphate linked oligosaccharide donor is shown in the membrane to the left of OST-A. B) Model of glycosylation pathways for hCEACAM1-IgV via OST-B. After protein translation glycans can be added in any order to each of the three N-glycan sequons (indicated by an 'N' in the linear polypeptide representation) and each addition is given its own rate constant. Glycans are shown with the SNFG symbol for GlcNAc for simplicity.

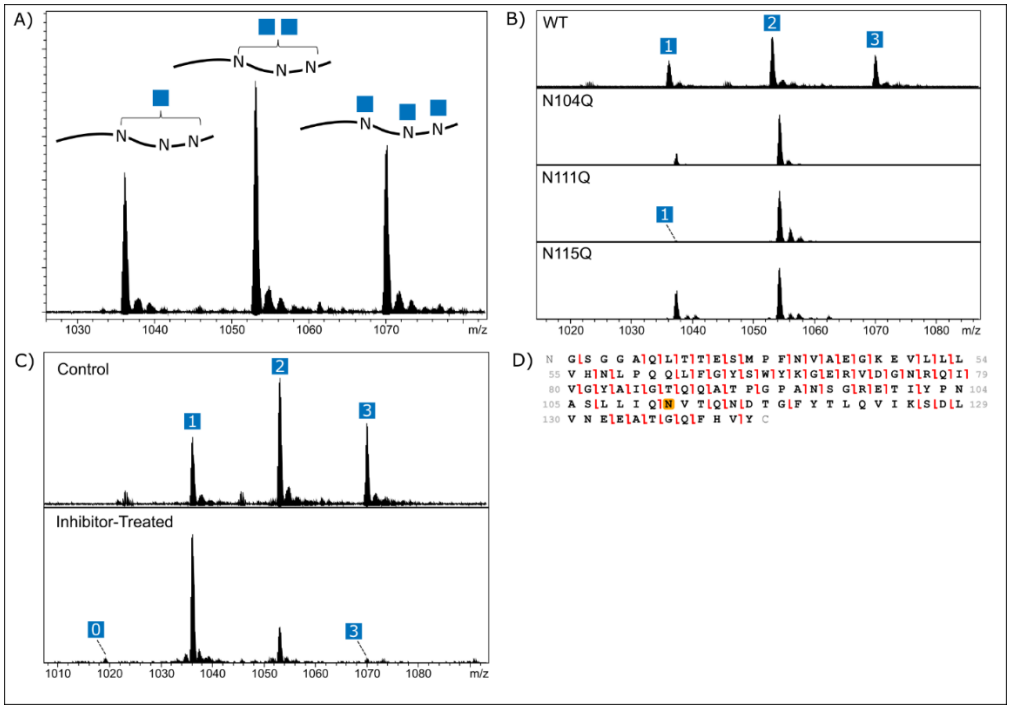

Figure 6.2. A) Mass spectrum of hCEACAM1-IgV-WT showing a region containing the 12+ charge distribution. Three peaks are observed corresponding to the protein with 1, 2, or 3 GlcNAc residues (blue squares). In the case of 1 or 2 GlcNAc residues the location of the modification is ambiguous and indicated with brackets. B) Comparison of mass spectra for hCEACAM1 glycosylation site mutants. The number within the blue square indicates the number of GlcNAc residues. C) Comparison of mass spectra for hCEACAM1-WT samples expressed in cells with and without the presence of OST-B inhibitor. D) Example fragmentation map from top-down ECD MS/MS spectra of monoglycosylated hCEACAM1-IgV. Lines drawn between residues indicate an assigned fragment ion. The notch at the top of a line indicates a c-type fragment ion while a notch at the bottom indicates an assigned z ion. Residue N111 is highlighted, indicating the presence of a GlcNAc PTM.

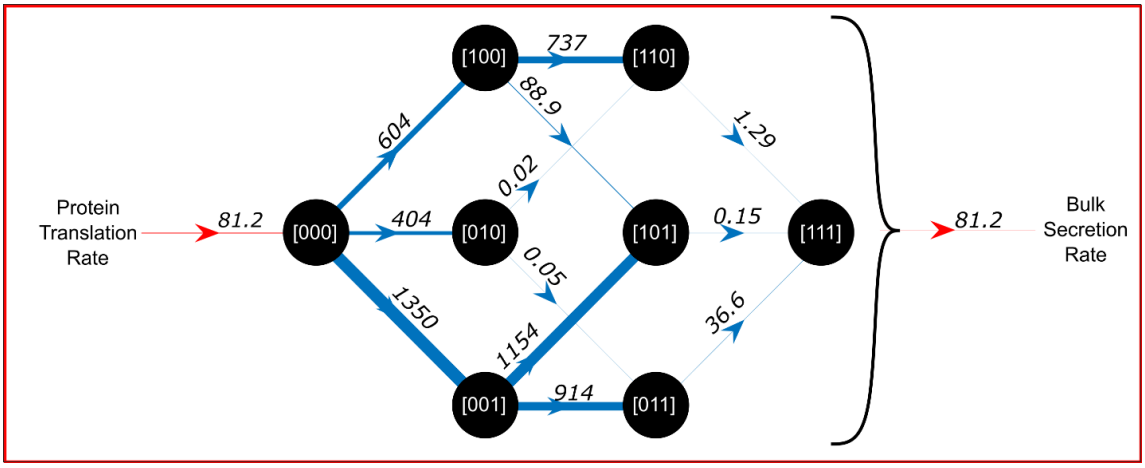

Figure 6.3. Flowchart describing the kinetic model of OST-B glycosylation. Glycosylation state is shown as a 3-bit binary representation for simplicity. The order of the three bits (e.g. [010]) corresponds to sites N104, N111, and N115, respectively, where a '0' represents an unglycosylated residue and a '1' represents the presence of a GlcNAc at the respective site. Each arrow indicates a separate reaction. In the case of the processes removing protein from the ER only one arrow is drawn for clarity, but the model includes a separate process for each species to exit the ER. The numbers above each arrow indicate the value of the corresponding optimized rate constant from the kinetic model. The thickness of each line has been scaled in proportion to the value of the rate constant. The orientation of the graph mirrors the diagram in panel A of Figure 6.1B.

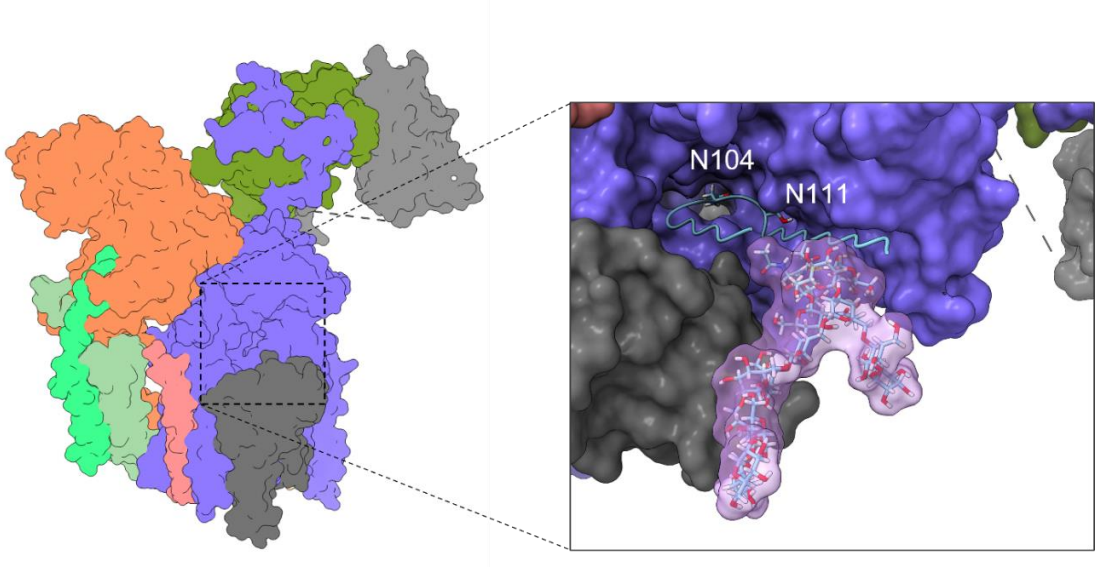

Figure 6.4. Model of unfolded hCEACAM1-IgV peptide (blue ribbon) docked into the catalytic site of STT3B (blue surface). Site 1 (N104) is docked in the catalytic site and a Glc¬3Man9GlcNAc2 N-glycan (transparent surface) is attached at site 2 (N111). The structure of OST-B was taken from PDB 67ST.

Table 6.1. Glycoform distributions of hCEACAM1-IgV variants

	Relative Proportion (%)							
Sample	**[000]**	**[100]**	**[010]**	**[001]**	**[110]**	**[101]**	**[011]**	**[111]**
WT[a]	0[b] ± 0	2 ± 2	27 ± 12	0 [b] ± 0	23 ± 5	19 ± 9	6 ± 6	25 ± 10
Inhibitor-Treated	1	6	70	3	8	7	3	1
N104Q	0 [b]	n/a[c]	19	1	n/a[c]	n/a[c]	81	n/a[c]
N111Q	0 [b]	2	n/a[c]	0 [b]	n/a[c]	98	n/a[c]	n/a[c]
N115Q	0[b]	0[b]	39	n/a[c]	61	n/a[c]	n/a[c]	n/a[c]

[a]*WT sample is reported as the average ± standard deviation of all biological replicates (n = 4).*

[b]*Not observed*

[c]*Impossible due to NΔQ mutation*

CHAPTER 7

CONCLUSION

This dissertation has outlined several advancements in methodology for structural studies of glycoproteins using both NMR spectroscopy and mass spectrometry. The combination of the two techniques has proved useful for the analysis of glycoproteins, as demonstrated by application to Robo1 in chapters 4 and 5. These same approaches should be useful for other glycoproteins. Sparse labeling of methyl groups combined with direct carbon observation should allow investigation of even larger proteins.

The examples described here relied on existing structural data from X-ray crystallography, but recent advancements in the accuracy of computational structure prediction suggest that this need not be the case in future studies. Both NMR spectroscopy and native MS can be used to validate predicted structures. The investigation of hRobo1-Ig1-Ig2 domain orientation serves as an example of this. For proteins with multiple domains programs such as AlphaFold can predict each domain with high accuracy, but the overall organization may be poorly defined. A similar strategy to the one used here with paramagnetic NMR and/or ion mobility mass spectrometry could be used to refine the global geometry of such proteins or protein complexes.

Additionally, we have used a combination of top-down and bottom-up mass spectrometry to characterize the occupancy of glycosylation sites in hCEACAM1-IgV. This protein construct contained three sites, but many glycoproteins contain more sites.

The top-down approach could likely be extended to larger systems. Fragmentation of this construct was less extensive than is typically achieved for proteins of similar molecular weight (~12 kDa), which may be due to its low isoelectric point and relatively low charge states observed in ESI mass spectra. Larger, more basic proteins would likely produce higher charge states and lead to better fragmentation coverage. The data collected on the two-domain construct show improved fragmentation coverage, which supports this conclusion. Nonetheless, it is clear this approach cannot be scaled to analyze the occupancy of an arbitrary number of glycosylation sites.

Overall, the results presented here represent a small but significant step forward in the analysis of glycoproteins. Future applications could further leverage the combination of NMR and MS by incorporating MS measurements into the AssignSLP methodology. As structural biology moves towards ever larger and more complex targets, no single technique will provide sufficient information alone. The work presented here serves as a framework for future studies.

APPENDIX A

SUPPLEMENTAL INFORMATION FOR CHAPTER 4

Mutagenesis Analysis

We first expressed a ^{13}C-dimethyl valine labeled Robo1-Ig1-2 construct without any lanthanide-binding sequence. A HETCOR spectrum shows the expected 38 methyl signals. Despite the close spacing of the observed signals, direct observation of ^{13}C magnetization provides excellent resolution. Two methyl peaks have narrower lines than the others and most likely belong to V265 which is the C-terminal residue of our construct. A reference spectrum is shown in figure A.1.

To aid the Assign_SLP assignment process we expressed a set of valine to isoleucine mutants as well: V71I, V133I, V144I, V165I, V188I, V241I, and V246I. HETCOR spectra of these mutant samples are provided in figures A.2-A.9. In most cases, two methyl resonances vanish from the spectrum allowing straightforward assignment. However, in the case of V144I two methyl peaks vanish and two either move to overlapping positions or are broadened to a point beyond detection. There are other peaks that move on mutation of V144 to isoleucine supporting the latter suggestion.

Figures

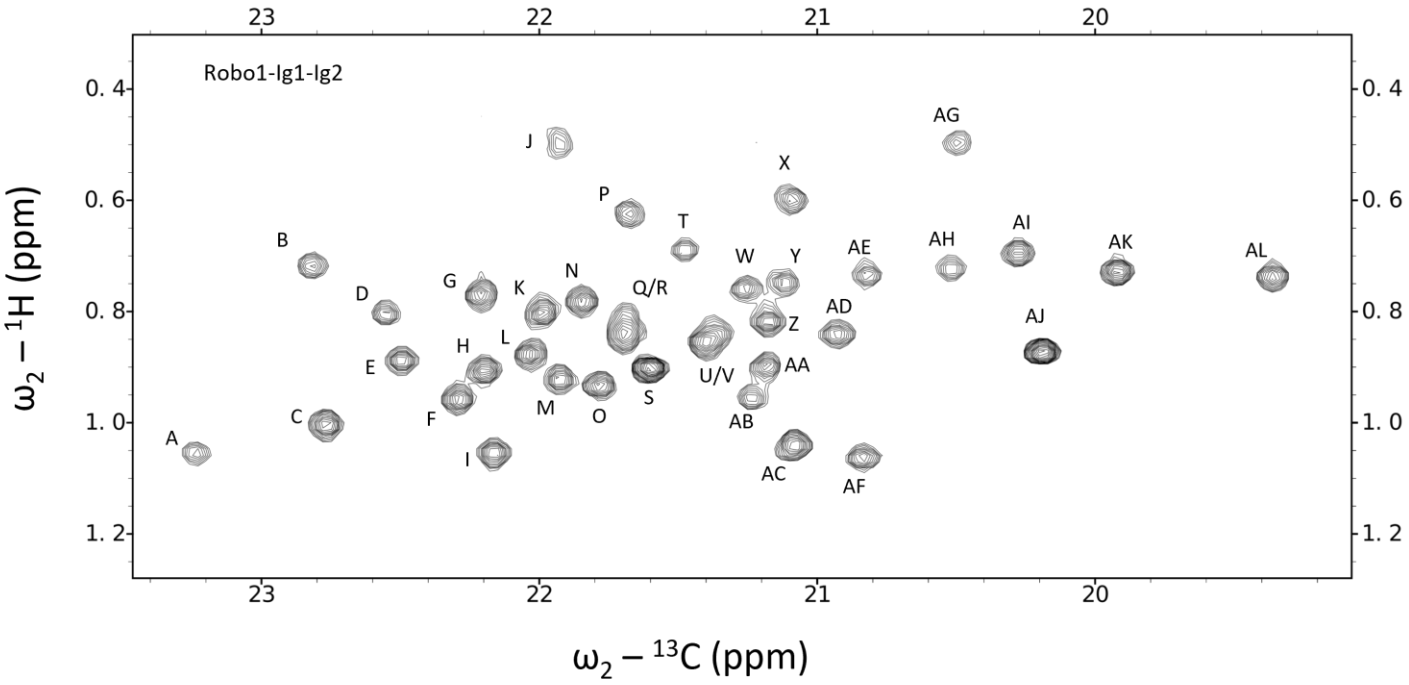

Figure A.1. HETCOR spectrum of WT Robo1-Ig1-Ig2 showing valine methyl signals. Peaks are given arbitrary labels from left to right.

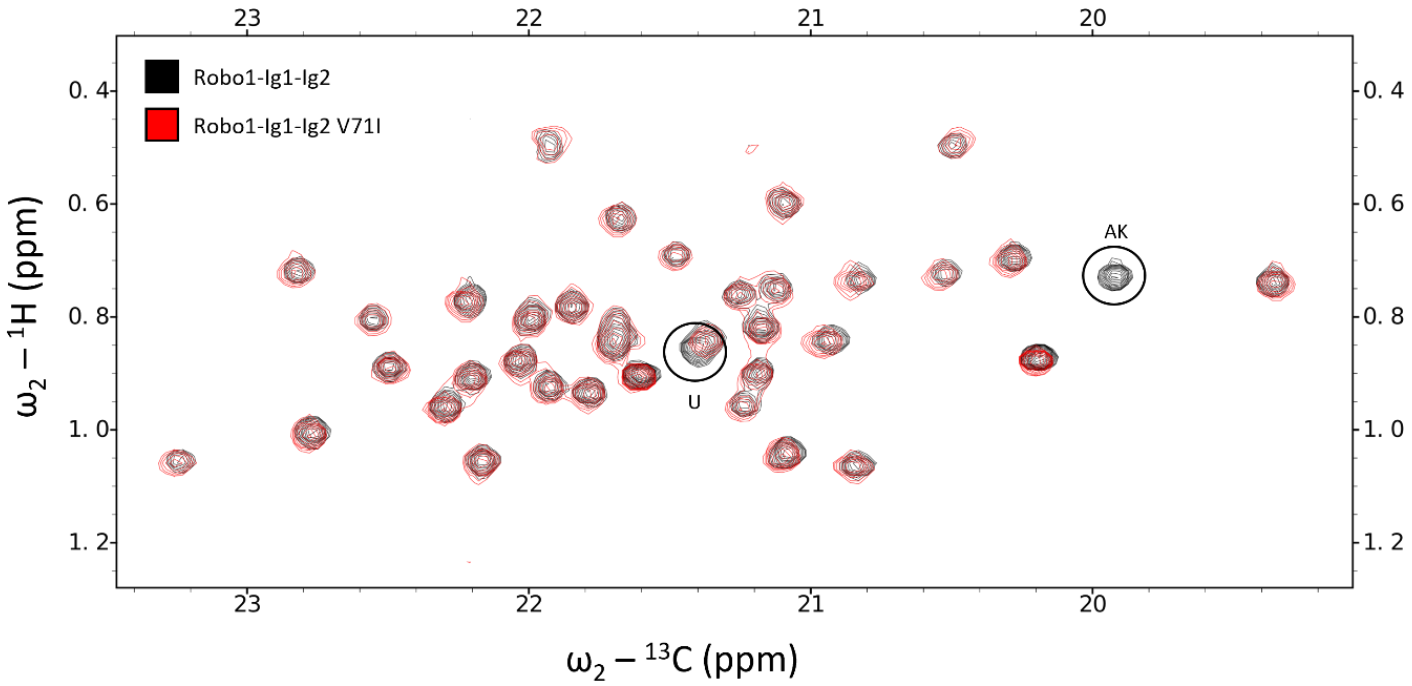

Figure A.2. Overlay of HETCOR spectrum showing V71 assignment. Robo1-Ig1-2 V71I is shown in blue and WT Robo1-Ig1-2 is shown in black. Methyl signals U and AK have vanished in the mutant spectrum (circled).

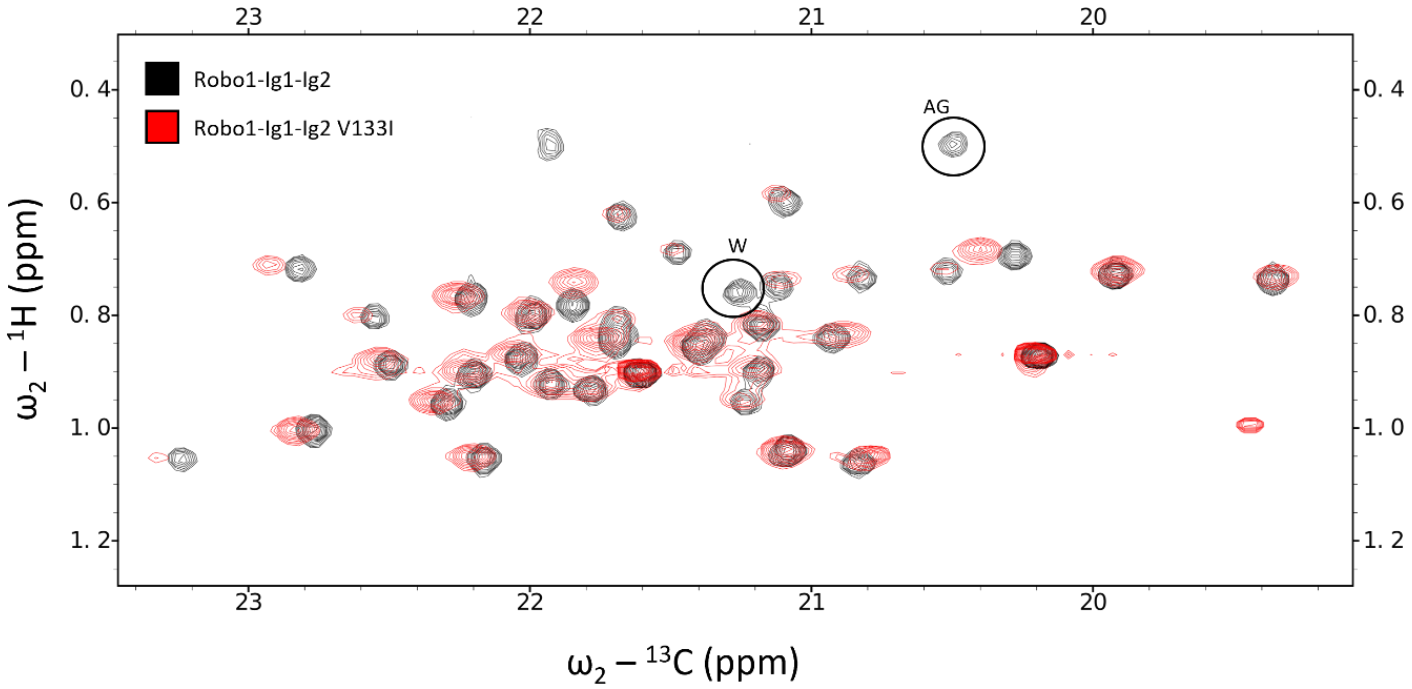

Figure A.3 – Overlay of WT Robo1-Ig1-2 HETCOR spectrum (black) and V133I mutant HSQC spectrum. HSQC Axes have been transposed to aid comparison with HETCOR spectra. Peaks W and AG have vanished in the mutant spectrum (indicated by circles). Note that peak J is observed but not visible in the HSQC at the plotted contour level.

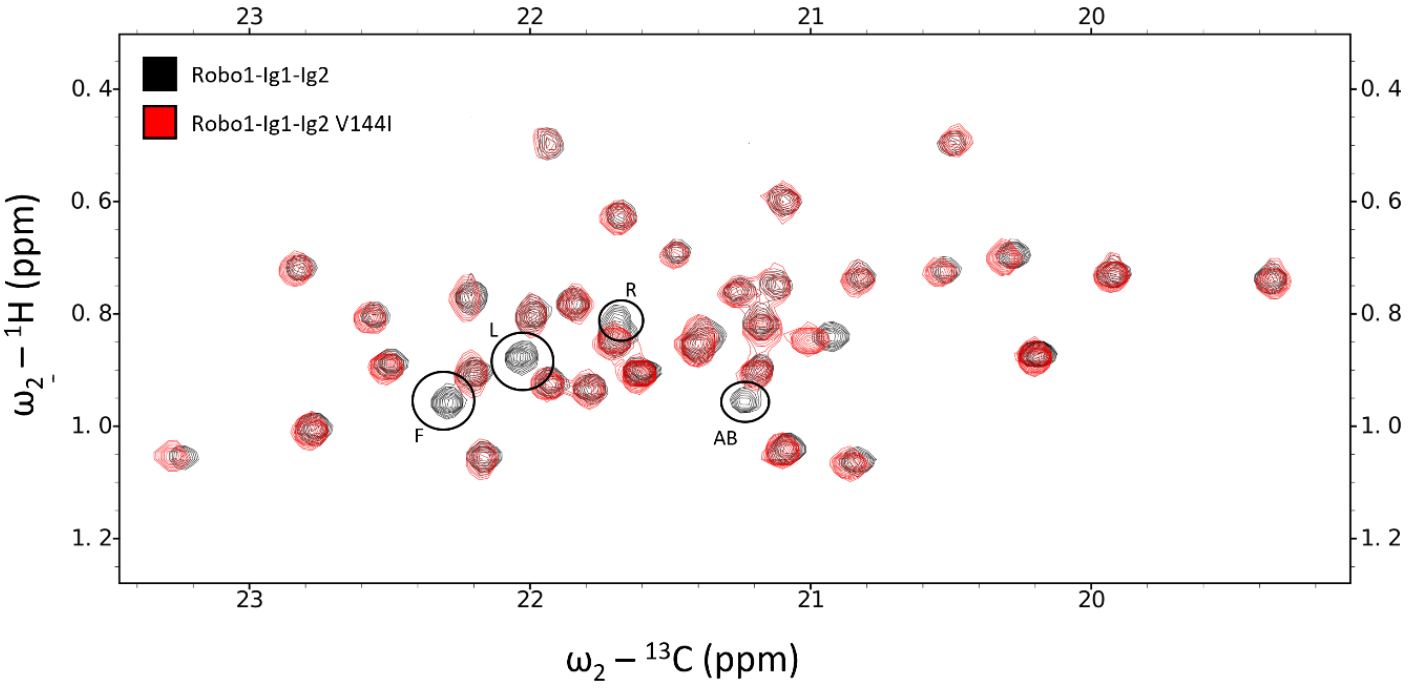

Figure A.4. Overlay of HETCOR spectrum for WT Robo1-Ig1-2 (black) and V144I mutant (blue). Peaks F, L, Q and AB have vanished in the mutant spectrum (circled).

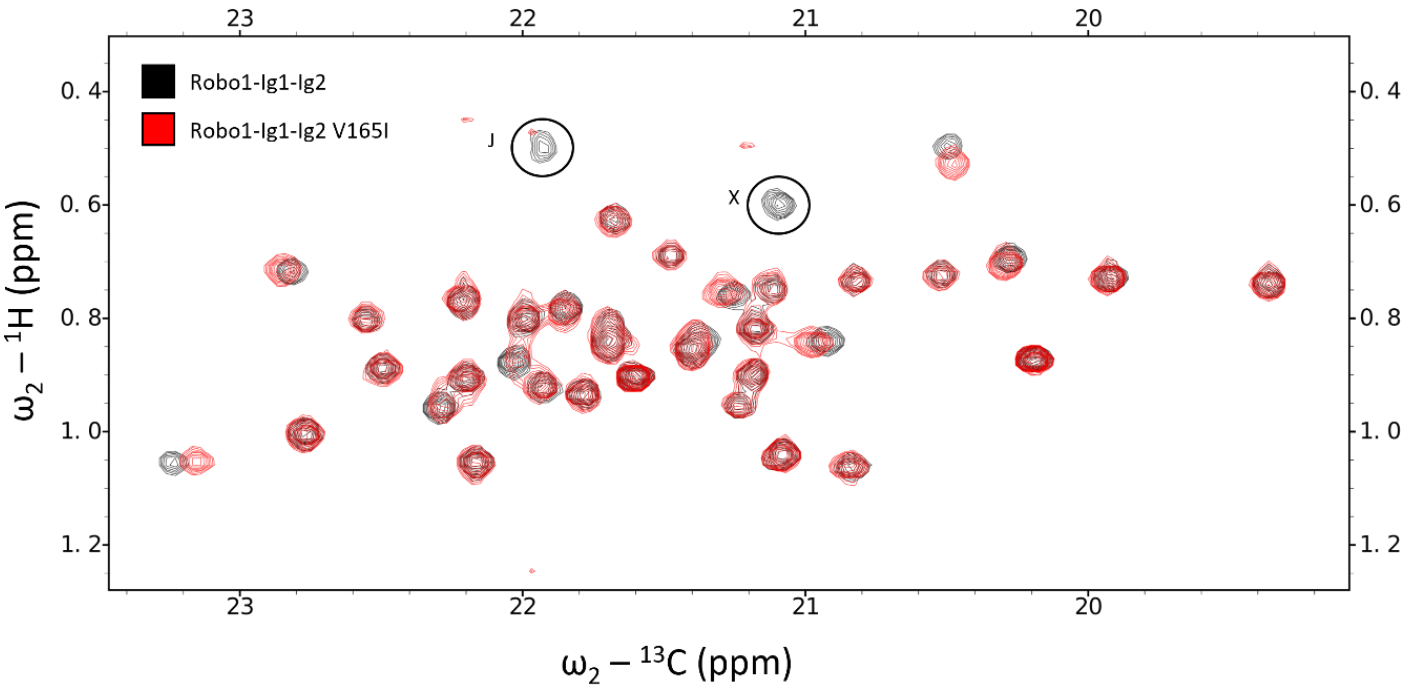

Figure A.5. Overlay of HETCOR spectra for WT Robo1-Ig1-2 (black) and V165I mutant (blue). Peaks J and X have vanished (open circles) and are assigned to V165.

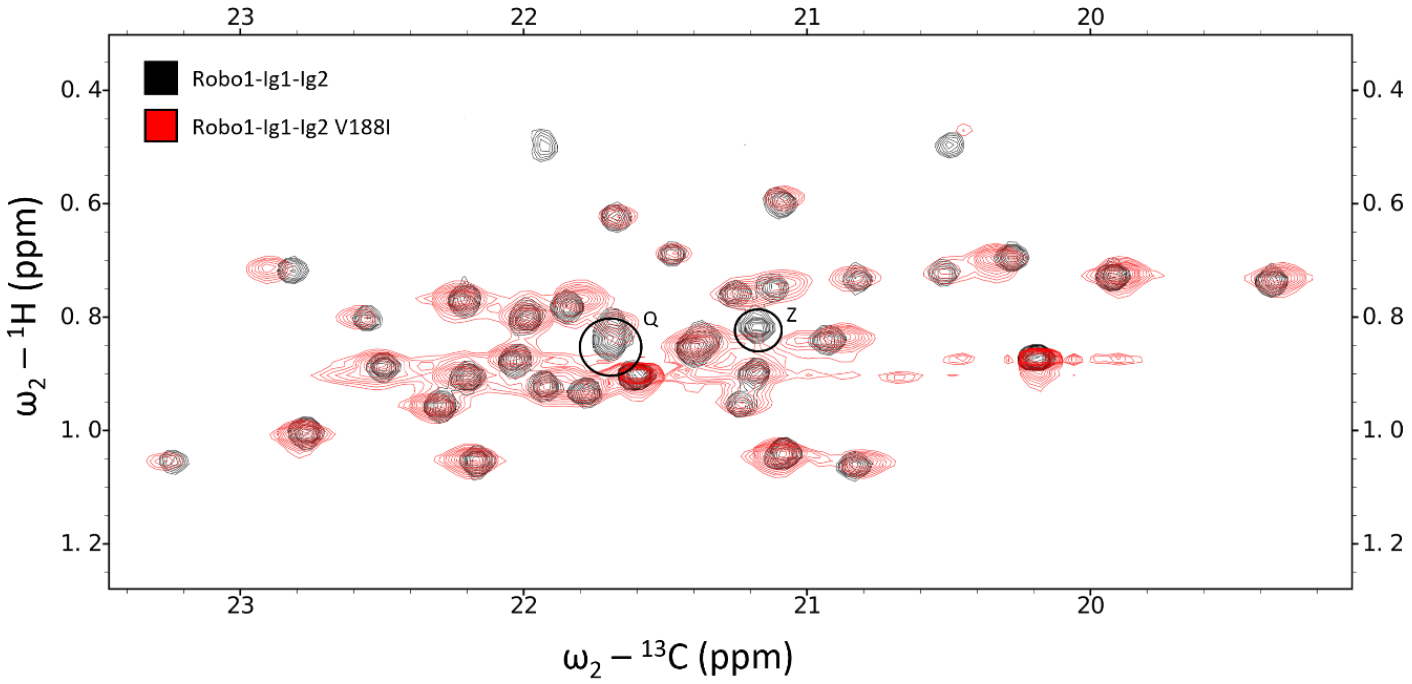

Figure A.6. Overlay of WT Robo1-Ig1-2 HETCOR spectrum (black) and V188I mutant HSQC spectrum (blue). HSQC Axes have been transposed to aid comparison with HETCOR spectra. Peaks Q and Z have vanished (indicated by circles). Note that peak J is observed but not visible in the HSQC at the plotted contour level.

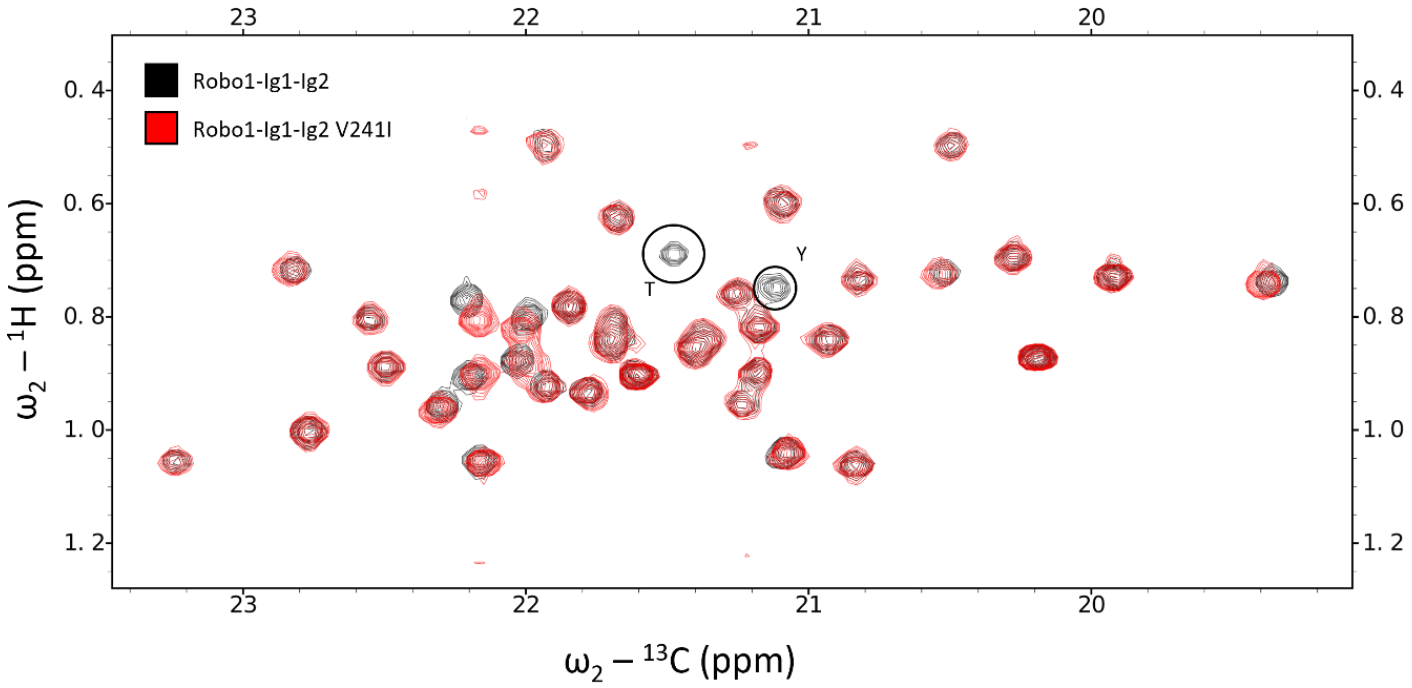

Figure A.7. Overlay of HETCOR spectra from WT Robo1-Ig1-2 (black) and the V241I mutant (blue). Peaks T and Y have vanished.

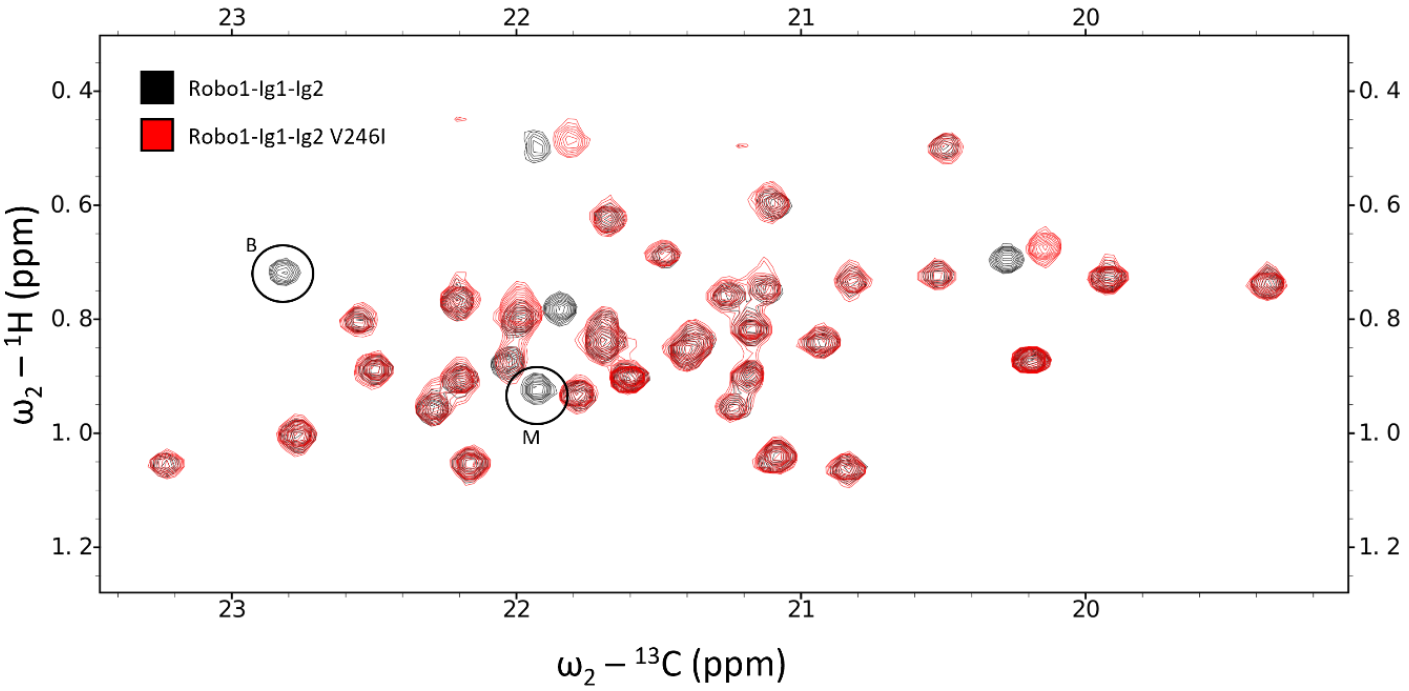

Figure A.8. Overlay of HETCOR spectra for WT Robo1-Ig1-2 (black) and the V246I mutant (blue). Peaks B and M have vanished.

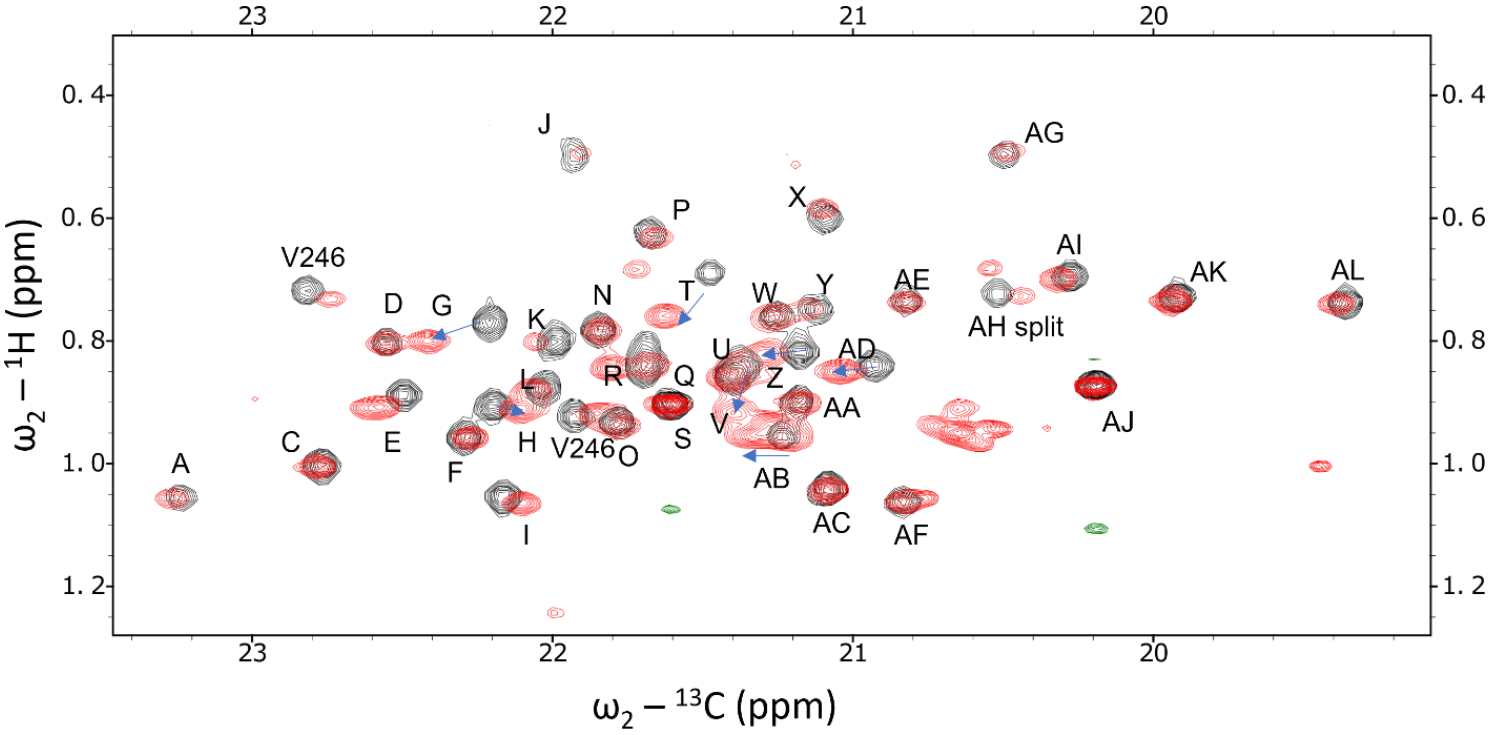

Figure A.9. Overlay of HETCOR spectra for WT Robo1-Ig1-2 (black) and Robo1-Ig1-Ig2-LBP4 (red).

Tables

Table A.1. Experimental chemical shift and PCS data used as part of AssignSLP input. Missing data is indicated as 999.

Peak	δ_{13C} (ppm)	δ_{1H} (ppm)	PCS (ppm)
A1	24.262	1.061	0.063
A2	23.648	1.07	0.13
A3	23.392	1.066	999
A4	23.083	1.282	0.2225
A5	22.991	0.895	0.0485
A6	22.529	1.214	999
A7	21.992	1.243	999
A8	20.656	1.476	999
A9	20.141	1.264	-0.01275
A10	19.768	1.65	-0.20875
A11	19.515	1.326	999
A12	19.45	1.388	999
U	21.406	0.86	0.078
AK	19.938	0.736	0.0765
E	22.576	0.907	0.094
AD	21.039	0.85	0.1095
A	23.265	1.057	0.081
V	21.398	0.917	999
W	21.261	0.762	999
AG	20.492	0.493	0.087
R	21.678	0.841	0.139
N	21.837	0.784	0.148
AA	20.825	1.063	0.047
AF	21.174	0.901	0.0585
D	22.556	0.805	0.0435
AE	20.823	0.737	0.0535
AI	21.306	0.952	0.1245
AB	20.315	0.7	999
X	21.1	0.585	999
J	21.907	0.494	999
P	21.659	0.631	-0.15375
AH	20.44	0.727	-0.1235
O	21.783	0.936	-0.11525
AL	19.39	0.74	-0.053
AC	21.079	1.043	-0.0875

C	22.782	1.005	-0.0985
Q	21.291	0.822	-0.155
Z	21.62	0.759	999
G	22.414	0.8	999
K	22.058	0.801	999
T	21.802	0.845	-0.155
Y	21.153	0.748	999
B	21.852	0.918	-0.3195
M	22.741	0.732	999
H	22.103	1.064	-0.071
I	22.113	0.912	-0.0775
F	22.275	0.958	-0.2945
L	22.067	0.881	-0.2735

Table A.2. Experimental NOE peak list used as part of AssignSLP input. Missing data is indicated as 999.

Assignment	w1	w2	w3	S/N
A1	0.635	24.235	1.036	14
	0.39	24.206	1.039	12
	1.345	24.229	1.042	14
	1.719	24.24	1.036	9
	1.037	24.239	1.04	42
A2	999	999	999	999
A3	999	999	999	999
A4	0.833	23.06	1.261	13
	2.015	23.038	1.268	8
	2.332	23.044	1.264	8
	2.995	23.053	1.261	7
	3.741	23.051	1.259	6
	4.885	23.031	1.259	15
	9.609	23.048	1.259	7
	1.257	23.052	1.26	44
A5HB	999	999	999	999
A6	1.451	22.482	1.189	10
	1.938	22.534	1.2	8
	4.738	22.55	1.21	7
	1.188	22.517	1.192	25
A7	1.539	21.975	1.223	12
	0.372	21.974	1.223	11
	1.896	21.993	1.221	8
	4.458	21.986	1.223	8
	4.737	21.978	1.221	17
	5.39	21.983	1.222	7
	8.39	21.985	1.222	10
	8.859	21.969	1.22	6
	1.224	21.97	1.221	61
A8	1.632	20.644	1.454	18
	1.951	20.638	1.452	8
	2.679	20.666	1.451	8
	3.193	20.629	1.452	8
	4.448	20.638	1.452	34
	4.69	20.631	1.454	11
	8.411	20.636	1.453	12
	1.456	20.634	1.453	328

A9	1.508	20.108	1.245	11
	2.238	20.115	1.243	12
	1.952	20.121	1.239	7
	4.635	20.124	1.243	12
	4.838	20.122	1.243	15
	7.89	20.137	1.238	7
	8.443	20.114	1.242	12
	1.241	20.118	1.243	102
A10	7.561	19.74	1.631	11
	4.222	19.73	1.627	11
	4.462	19.757	1.622	5
	1.818	19.743	1.631	7
	0.701	19.776	1.632	9
	0.514	19.729	1.638	10
	1.629	19.741	1.631	59
A11	2.788	19.462	1.302	11
	4.31	19.467	1.303	49
	4.666	19.469	1.303	14
	8.028	19.448	1.299	11
	1.3	19.461	1.303	1236
A12	4.356	19.473	1.364	18
	1.358	19.442	1.364	643
U	0.091	21.381	0.844	38
	1.071	21.357	0.841	38
	1.407	21.371	0.832	27
	1.609	21.377	0.839	76
	1.965	21.379	0.84	116
	2.228	21.378	0.84	76
	3.134	21.377	0.836	23
	4.284	21.394	0.841	83
	5.331	21.37	0.84	20
	8.726	21.374	0.833	19
	0.829	21.374	0.841	1431
AK	0.878	19.92	0.714	147
	1.067	19.925	0.715	31
	1.367	19.926	0.713	44
	1.64	19.924	0.714	126
	1.961	19.923	0.714	103
	2.231	19.918	0.714	79
	2.966	19.924	0.715	21
	3.125	19.958	0.712	12

	3.994	19.925	0.713	31
	4.271	19.925	0.714	92
	4.537	19.922	0.713	25
	4.806	19.914	0.715	28
	5.287	19.918	0.714	36
	6.243	19.947	0.719	8
	7.01	19.935	0.71	7
	7.567	19.927	0.714	40
	7.456	19.958	0.714	11
	8.336	19.912	0.714	17
	8.732	19.925	0.714	56
	9.231	19.927	0.717	16
	0.093	19.931	0.714	31
	0.723	19.926	0.714	1124
E	1.956	22.568	0.886	62
	2.538	22.584	0.888	21
	2.725	22.534	0.889	27
	4.367	22.584	0.888	45
	4.719	22.589	0.885	15
	5.347	22.561	0.886	46
	8.866	22.624	0.889	17
	0.872	22.57	0.888	658
AD	1.96	21.013	0.827	55
	2.744	21.023	0.828	25
	0.992	21.013	0.828	76
	2.525	21.005	0.835	30
	3.172	21.076	0.83	11
	4.353	21.014	0.828	19
	4.711	21.006	0.826	19
	5.04	20.961	0.824	12
	5.353	20.999	0.826	22
	5.738	21.024	0.829	13
	6.673	21.005	0.828	34
	8.367	21.003	0.827	13
	8.906	21.026	0.825	24
	9.293	21.067	0.831	20
	0.835	21.012	0.828	774
A	0.827	23.247	1.037	46
	1.421	23.241	1.032	13
	1.902	23.245	1.04	24
	2.82	23.24	1.037	15

	3.84	23.214	1.036	18
	5.065	23.257	1.04	12
	1.029	23.261	1.036	91
V	999	999	999	999
W	0.484	21.235	0.74	59
	1.354	21.207	0.742	18
	1.623	21.202	0.736	13
	1.852	21.201	0.737	46
	3.665	21.221	0.739	21
	3.95	21.231	0.742	23
	4.466	21.261	0.737	14
	4.671	21.217	0.738	54
	9.612	21.214	0.741	8
	0.733	21.21	0.739	413
AG	0.738	20.483	0.473	48
	1.888	20.479	0.474	17
	0.656	20.446	0.477	16
	3.033	20.474	0.473	9
	3.949	20.48	0.477	15
	4.67	20.458	0.473	25
	4.504	20.478	0.474	8
	0.468	20.462	0.472	159
R	1.513	21.809	0.763	46
	2.021	21.813	0.762	45
	0.182	21.824	0.763	42
	4.597	21.83	0.763	40
	0.757	21.81	0.762	308
N	9.269	21.652	0.82	10
	2.767	21.666	0.821	14
	5.337	21.691	0.818	12
	4.817	21.689	0.818	24
	8.918	21.666	0.82	10
	3.746	21.671	0.82	15
	3.899	21.671	0.819	11
AA	0.679	21.094	0.88	107
	1.9	21.161	0.878	98
	4.748	21.134	0.876	48
	10.914	21.134	0.869	11
	0.848	21.119	0.878	726
AF	0.847	20.782	1.042	205
	1.912	20.814	1.042	73

	4.695	20.806	1.039	32
	8.4	20.808	1.042	18
	8.952	20.789	1.037	12
	9.348	20.802	1.039	23
	1.028	20.811	1.039	383
D	4.644	22.519	0.779	16
	6.675	22.537	0.784	46
	4.745	22.523	0.782	23
	6.675	22.537	0.784	46
	0.774	22.544	0.782	235
AE	1.085	20.778	0.716	14
	1.557	20.795	0.717	16
	1.903	20.802	0.716	62
	3.109	20.797	0.716	9
	4.67	20.807	0.717	18
	5.512	20.807	0.718	8
	6.675	20.793	0.717	20
	8.933	20.802	0.717	17
	8.678	20.849	0.717	7
	0.734	20.795	0.716	203
AI	0.916	21.272	0.924	4749
	2.066	21.287	0.929	82
AB	0.176	20.308	0.677	49
	0.01	20.307	0.678	19
	1.497	20.307	0.677	27
	2.036	20.309	0.677	33
	3.818	20.315	0.675	9
	4.59	20.31	0.678	42
	5.067	20.305	0.675	16
	8.358	20.297	0.681	16
	8.519	20.306	0.677	18
	0.684	20.309	0.678	305
J	0.172	21.81	0.474	10
	1.354	21.873	0.472	14
	2.14	21.856	0.476	9
	2.497	21.874	0.471	7
	4.578	21.879	0.471	10
	4.741	21.849	0.476	6
	8.48	21.874	0.474	9
	0.505	21.861	0.473	54
X	1.38	21.068	0.563	19

	1.76	21.084	0.563	20
	2.136	21.076	0.566	10
	2.894	21.078	0.565	11
	3.065	21.09	0.563	9
	4.561	21.084	0.563	19
	8.484	21.097	0.562	20
	0.552	21.084	0.563	147
P	1.023	21.623	0.608	47
	1.223	21.619	0.607	18
	1.629	21.635	0.611	19
	1.876	21.629	0.612	31
	3.046	21.654	0.614	8
	3.395	21.638	0.617	7
	3.859	21.63	0.611	10
	4.283	21.657	0.609	8
	4.842	21.626	0.609	20
	8.189	21.62	0.618	7
	9.177	21.645	0.609	11
	0.619	21.633	0.611	100
AH	1.143	20.446	0.703	22
	1.624	20.373	0.708	20
	1.894	20.362	0.708	21
	4.283	20.363	0.709	8
	4.833	20.357	0.705	14
	7.54	20.355	0.7	9
	0.707	20.446	0.704	125
O	2.246	21.766	0.915	72
	0.736	21.766	0.915	124
	4.669	21.768	0.914	53
	8.394	21.765	0.921	40
AL	0.529	19.363	0.718	36
	0.956	19.364	0.718	88
	1.157	19.374	0.719	50
	1.889	19.378	0.722	13
	2.243	19.383	0.718	31
	3.948	19.37	0.716	15
	4.376	19.338	0.717	13
	4.662	19.4	0.717	28
	4.844	19.379	0.718	16
	5.445	19.365	0.719	10
	8.443	19.362	0.715	29

	0.716	19.388	0.718	180
AC	0.617	21.076	1.021	42
	1.283	21.036	1.021	20
	1.624	21.055	1.021	33
	1.86	21.065	1.021	65
	2.265	21.063	1.022	16
	3.065	21.044	1.018	17
	3.398	21.057	1.022	36
	4.305	21.061	1.021	28
	3.959	21.093	1.025	15
	4.716	21.064	1.02	27
	7.881	21.055	1.021	18
	8.405	21.022	1.023	9
	8.81	21.053	1.021	14
	1.014	21.06	1.021	1022
C	0.623	22.763	0.984	70
	1.226	22.755	0.985	23
	1.63	22.763	0.984	35
	1.863	22.764	0.984	73
	2.24	22.774	0.982	17
	3.044	22.755	0.981	23
	3.384	22.772	0.985	29
	3.848	22.759	0.984	13
	4.3	22.765	0.982	32
	4.726	22.752	0.981	30
	7.878	22.766	0.982	28
	8.794	22.751	0.981	12
	9.19	22.763	0.985	12
	0.986	22.763	0.983	508
Q	0.793	21.269	0.8	494
	1.385	21.274	0.8	24
	5.076	21.271	0.801	29
	4.676	21.272	0.801	35
	5.425	21.227	0.8	11
	7.838	21.249	0.802	13
	8.738	21.276	0.804	18
	8.932	21.31	0.802	14
	1.6	21.272	0.799	46
	1.847	21.266	0.8	65
	2.228	21.231	0.798	24
	2.895	21.243	0.798	15

Z	0.568	21.592	0.736	36
	1.016	21.59	0.737	31
	1.276	21.602	0.737	54
	1.467	21.61	0.739	26
	1.748	21.595	0.736	43
	1.93	21.594	0.738	35
	2.626	21.591	0.738	18
	2.827	21.583	0.737	7
	4.072	21.58	0.735	20
	4.247	21.592	0.737	23
	4.606	21.576	0.734	11
	4.878	21.584	0.736	19
	5.558	21.567	0.737	13
	7.361	21.588	0.733	9
	8.326	21.594	0.737	10
	8.616	21.587	0.735	21
	9.04	21.584	0.735	14
	9.271	21.607	0.734	10
	1.277	21.69	0.662	18
	0.735	21.596	0.737	595
G	1.24	22.387	0.778	30
	1.909	22.464	0.781	59
	2.245	22.428	0.779	15
	2.576	22.39	0.778	10
	3.145	22.444	0.779	8
	3.519	22.441	0.778	5
	4.071	22.407	0.779	21
	4.272	22.4	0.778	22
	4.886	22.417	0.78	19
	5.544	22.466	0.778	7
	8.911	22.491	0.784	21
	1.457	22.386	0.778	18
	0.777	22.447	0.78	456
K	1.088	21.952	0.775	18
	1.513	21.959	0.78	20
	0.777	21.958	0.778	208
T	1.174	21.769	0.82	28
	1.402	21.787	0.822	26
	1.575	21.771	0.822	50
	1.853	21.763	0.822	92
	2.277	21.77	0.82	46

	2.91	21.814	0.817	9
	4.649	21.772	0.82	35
	5.072	21.768	0.822	55
	7.834	21.781	0.822	26
	8.789	21.773	0.822	22
	0.812	21.768	0.821	775
Y	3.948	21.112	0.727	8
	4.949	21.132	0.724	15
	2.005	21.112	0.727	28
	9.002	21.155	0.723	11
	1.188	21.129	0.729	10
B	0.898	22.745	0.71	49
	1.507	22.714	0.71	12
	1.979	22.677	0.708	8
	2.192	22.727	0.71	22
	3.845	22.702	0.712	24
	3.845	22.702	0.712	24
	4.542	22.682	0.715	14
	7.907	22.703	0.711	11
	8.187	22.688	0.712	7
	7.51	22.67	0.706	8
	0.718	22.719	0.712	91
M	1.201	22.699	0.712	31
	1.562	21.861	0.902	33
	2.208	21.861	0.898	28
	3.849	21.825	0.897	42
	0.892	21.897	0.898	617
H	1.047	22.08	0.888	155
	1.216	22.066	0.887	33
	1.517	21.977	0.891	31
	1.845	22.083	0.889	49
	4.366	22.068	0.888	52
	4.665	22.043	0.89	22
	8.459	22.127	0.89	32
	0.886	22.057	0.884	786
I	999	999	999	999
F	0.756	22.26	0.938	111
	1.248	22.263	0.937	34
	1.608	22.263	0.938	39
	2.108	22.2	0.931	24
	4.081	22.26	0.937	34

	4.52	22.338	0.919	25
	1.234	22.031	0.857	29
	4.782	22.075	0.868	32
	1.583	22.064	0.86	57
	1.898	22.048	0.862	82
	5.061	22.056	0.871	60
	8.321	22.012	0.87	35
	1.42	22.242	0.938	9
	1.907	22.25	0.936	92
	0.919	22.249	0.936	703
L	0.859	22.056	0.864	342

APPENDIX B

SUPPLEMENTAL INFORMATION FOR CHAPTER 6

MATLAB Code. MATLAB scripts used for fitting kinetic model to MS data.

```
% continuous flow kinetics

% optimize rate constants with genetic algorithm

p=parcluster('local');

p.NumWorkers=25;

ppool=parpool(p,24);


% input glycoform MS data

WT = [0,0,29,0,15,31,0,26];

N104Q = [0,19,1,81];

N111Q = [0,2,0,98];

N115Q = [0,0,39,61];


data = [WT, N104Q, N111Q, N115Q];


f = @(x) gafun(x,data);

nvars = 14;

% bound rate constants b/t [0,inf]

lb = zeros(nvars,1);

ub = Inf*ones(nvars,1);

options =

optimoptions('ga','PlotFcn',@gaplotbestf,'UseParallel',true,'InitialPop

ulationMatrix',k);


fileID = fopen('rate_constants.txt','w');

formatSpec =

'%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f,%.2f\

n';
```

```
fprintf(fileID,formatSpec,k);


rng('default') % for reproducibility
[x, fval, exitflag, output, population, scores] =
ga(f,nvars,[],[],[],[],lb,[],[],options);
while exitflag == 0
    options = optimoptions(options,'InitialPopulation',population);
    [x,fval,exitflag,output,population,scores] =
ga(f,nvars,[],[],[],[],lb,[],[],options);
    fprintf(fileID,formatSpec,x);
end
fclose(fileID);
p = gcp;
delete(p)


%% declarations
function score = gafun(k,data)
glycans = glycoNetwork(k);
score = sqrt(mean((data - glycans).^2));
end


function glycans = glycoNetwork(k)
tspan = [0,15];
y0 = zeros(16,1);


% k_in = 1; % protein translation
% k_out = 1; % protein exits ER and secreted
% k = [k_in;k_out;ones(12,1)];
```

```
f = @(t,y) odefun_WT(t,y,k);
[t,y] = ode45(f,tspan,y0);

% get secreted concentrations and convert to percent of total
cols = 9:16;
total = sum(y(end,cols),2);
p = y(end,cols)./total*100;

f = @(t,y) odefun_N70Q(t,y,k);
y0 = zeros(8,1);
[t,y] = ode45(f,tspan,y0);
cols = 5:8;
total = sum(y(end,cols),2);
p = [p, y(end,cols)./total*100];

f = @(t,y) odefun_N77Q(t,y,k);
[t,y] = ode45(f,tspan,y0);
cols = 5:8;
total = sum(y(end,cols),2);
p = [p, y(end,cols)./total*100];

f = @(t,y) odefun_N81Q(t,y,k);
[t,y] = ode45(f,tspan,y0);
cols = 5:8;
total = sum(y(end,cols),2);
p = [p, y(end,cols)./total*100];
```

```
glycans = p;

end


function dydt = odefun_WT(t,y,k)

% k = [k_in, k_out, k1, k2, k12, k21]

k_in = k(1);

k_out = k(2);

k1 = k(3);

k2 = k(4);

k3 = k(5);


k12 = k(6);

k21 = k(7);


k23 = k(8);

k32 = k(9);


k13 = k(10);

k31 = k(11);


k123 = k(12);

k231 = k(13);

k132 = k(14);


n = 8;

dydt = zeros(n,1);

dydt(1) = k_in - (k1 + k2 + k3 + k_out)*y(1); % 000 ER

dydt(2) = k1*y(1) - k12*y(2) -k13*y(2) - k_out*y(2);  % 100 ER
```

```
dydt(3) = k2*y(1) - k23*y(3) -k21*y(3) - k_out*y(3);  % 010 ER
dydt(4) = k3*y(1) - k32*y(4) -k31*y(4) - k_out*y(4);  % 001 ER
dydt(5) = k12*y(2)+ k21*y(3) - k123*y(5) -k_out*y(5); % 110 ER
dydt(6) = k13*y(2)+ k31*y(4) - k132*y(6) -k_out*y(6); % 101 ER
dydt(7) = k23*y(3) + k32*y(4)- k231*y(7) -k_out*y(7); % 011 ER
dydt(8) = k123*y(5)+ k231*y(7) + k132*y(6) -k_out*y(8); % 111 ER
dydt(9) = k_out*y(1); % 000 secreted
dydt(10) = k_out*y(2); % 100 secreted
dydt(11) = k_out*y(3); % 010 secreted
dydt(12) = k_out*y(4); % 001 secreted
dydt(13) = k_out*y(5); % 110 secreted
dydt(14) = k_out*y(6); % 101 secreted
dydt(15) = k_out*y(7); % 011 secreted
dydt(16) = k_out*y(8); % 111 secreted
end

function dydt = odefun_N70Q(t,y,k)
% k = [k_in, k_out, k1, k2, k12, k21]
k_in = k(1);
k_out = k(2);
k1 = k(3);
k2 = k(4);
k3 = k(5);

k12 = k(6);
k21 = k(7);

k23 = k(8);
```

```
k32 = k(9);


n = 8;
dydt = zeros(n,1);
dydt(1) = k_in - (k2 + k3 + k_out)*y(1); % Q00 ER
dydt(2) = k2*y(1) - k23*y(2) - k_out*y(2); % Q10 ER
dydt(3) = k3*y(1) - k32*y(3) - k_out*y(3); % Q01 ER
dydt(4) = k23*y(2) + k32*y(3)- k_out*y(4); % Q11 ER
dydt(5) = k_out*y(1); % Q00 secreted
dydt(6) = k_out*y(2); % Q10 secreted
dydt(7) = k_out*y(3); % Q01 secreted
dydt(8) = k_out*y(4); % Q11 secreted


end


function dydt = odefun_N77Q(t,y,k)
% k = [k_in, k_out, k1, k2, k12, k21]
k_in = k(1);
k_out = k(2);
k1 = k(3);
k2 = k(4);
k3 = k(5);


k12 = k(6);
k21 = k(7);


k23 = k(8);
k32 = k(9);
```

```
k13 = k(10);
k31 = k(11);

n = 8;
dydt = zeros(n,1);
dydt(1) = k_in - (k1 + k3 + k_out)*y(1); % 0Q0 ER
dydt(2) = k1*y(1) - k13*y(2) - k_out*y(2); % 1Q0 ER
dydt(3) = k3*y(1) - k31*y(3) - k_out*y(3); % 0Q1 ER
dydt(4) = k13*y(2) + k31*y(3)- k_out*y(4); % 1Q1 ER
dydt(5) = k_out*y(1); % 0Q0 secreted
dydt(6) = k_out*y(2); % 1Q0 secreted
dydt(7) = k_out*y(3); % 0Q1 secreted
dydt(8) = k_out*y(4); % 1Q1 secreted

end

function dydt = odefun_N81Q(t,y,k)
% k = [k_in, k_out, k1, k2, k12, k21]
k_in = k(1);
k_out = k(2);

k1 = k(3);
k2 = k(4);
k3 = k(5);

k12 = k(6);
k21 = k(7);
```

```
n = 8;

dydt = zeros(n,1);

dydt(1) = k_in - (k1 + k2 + k_out)*y(1); % 00Q ER

dydt(2) = k1*y(1) - k12*y(2) - k_out*y(2); % 10Q ER

dydt(3) = k2*y(1) - k21*y(3) - k_out*y(3); % 01Q ER

dydt(4) = k12*y(2)+ k21*y(3) - k_out*y(4); % 11Q ER

dydt(5) = k_out*y(1);% 00Q secreted

dydt(6) = k_out*y(2);% 10Q secreted

dydt(7) = k_out*y(3);% 01Q secreted

dydt(8) = k_out*y(4);% 11Q secreted


end
```

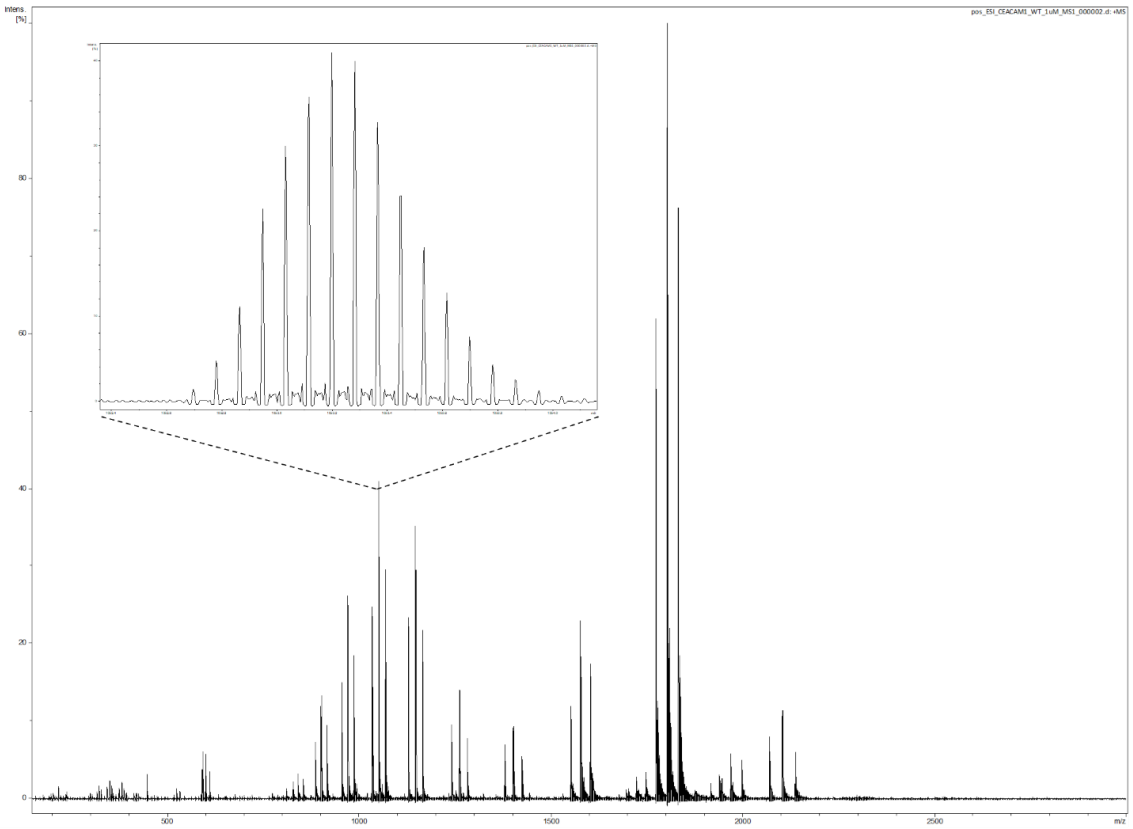

Figure B.1. FT-ICR Mass spectrum of intact hCEACAM1-IgV, batch 1. A close-up on isotope distribution for the diglycosylated species (12+ charge state) is shown in the inset. The resolving power is sufficient to baseline resolve the isotope peaks.

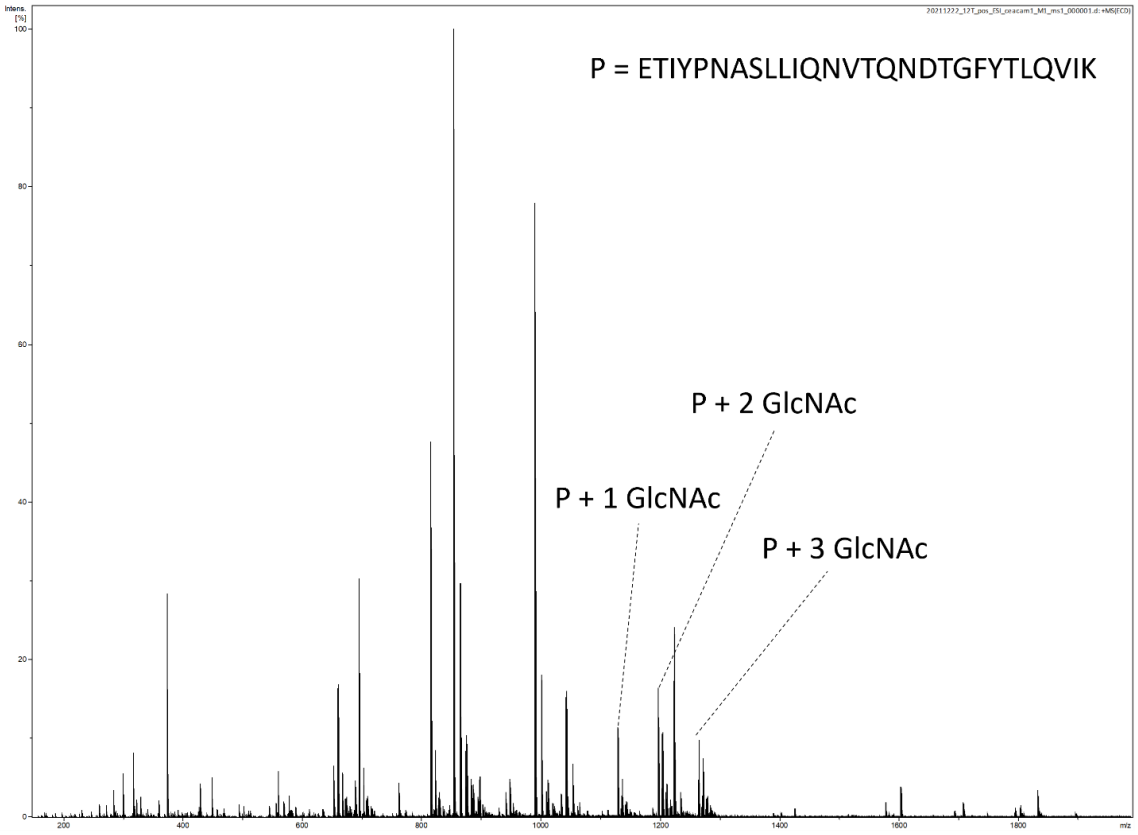

Figure B.2. Mass spectrum of a tryptic digest of hCEACAM1-IgV. The glycopeptide is indicated by its glycan composition.

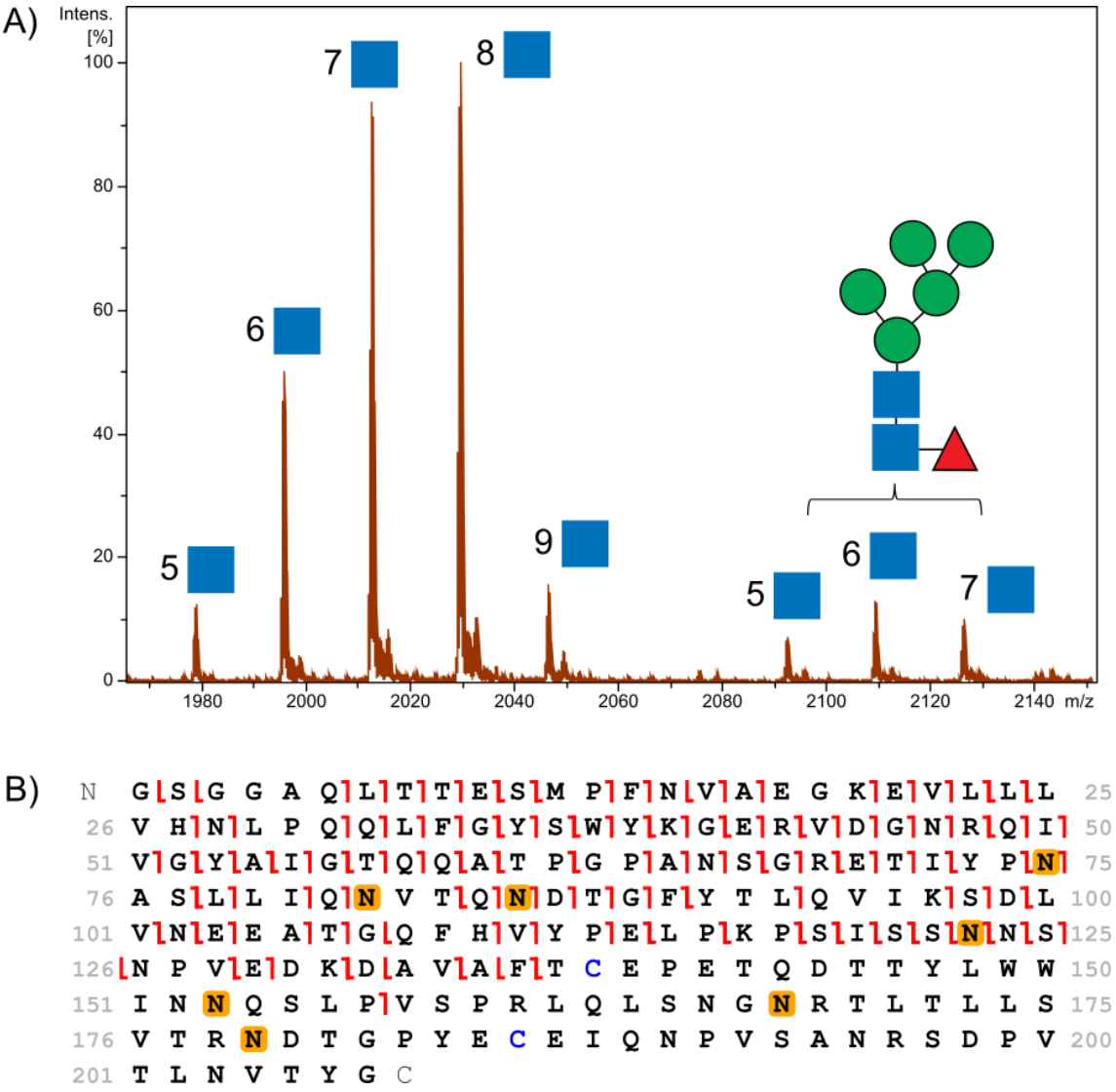

Figure B.3. A) Mass spectra 2 domain hCEACAM1-IgV-IgC construct, focusing on the 12+ charge state. A mixture of glycoforms is observed due to variable glycan occupancy and incomplete EndoF cleavage. B) Example fragmentation map showing ECD MS/MS results for the 7 GlcNAc species.

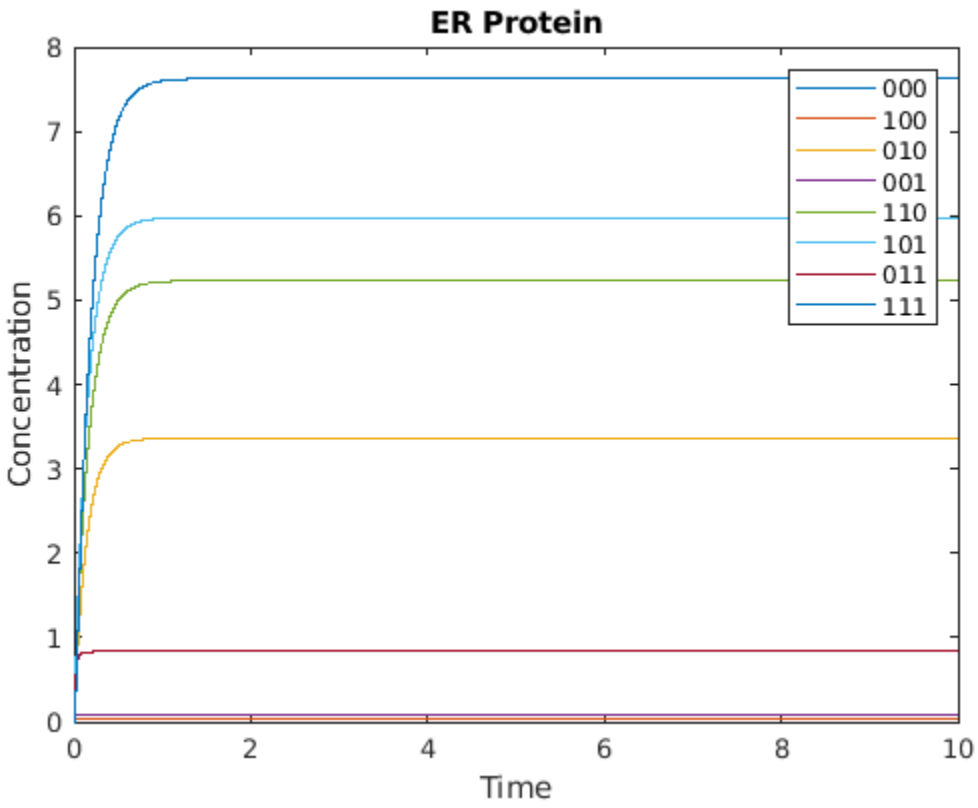

Figure B.4. Time course of CEACAM1 glycoform concentrations within the ER from a solution to the kinetic model. Concentrations quickly converge to a steady-state level. Arbitrary units are used for both axes.

Table B.1. Intact mass distribution for intact hCEACAM1-IgV, batch 1.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	12417.22	62.89	27
2 GlcNAc	12620.29	100.00	43
3 GlcNAc	12823.37	70.44	30

Table B.2. Diagnostic fragment ions from top-down ECD MS/MS of WT, 1 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7392	2	9890000	1 GlcNAc	z	27	2	0.0
1024.8285	3	14500000	1 GlcNAc	z	27	3	-0.1
1067.5150	3	2210000	1 GlcNAc	z	28	3	0.3
1395.6973	3	2170000	1 GlcNAc	z	35	3	-0.4
1088.5608	8	3150000	1 GlcNAc	c	81	8	0.3
1558.6276	6	1720000	1 GlcNAc	c	85	6	0.0
1169.2226	8	11000000	1 GlcNAc	c	85	8	0.1
1336.1115	7	5270000	1 GlcNAc	c	85	7	0.8

Table B.3. Diagnostic fragment ions from top-down ECD MS/MS of WT, 2 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7388	2	3.29E+07	0 GlcNAc	z	27	2	-0.2
1024.8279	3	1.60E+07	0 GlcNAc	z	27	3	-0.6
1638.2785	2	3.76E+07	1 GlcNAc	z	27	2	-0.2
1092.5215	3	6.01E+06	1 GlcNAc	z	27	3	-0.2
1702.3089	2	4.81E+06	1 GlcNAc	z	28	2	0.4
1859.3875	2	2.69E+07	1 GlcNAc	z	31	2	-0.1
1239.9272	3	1.58E+07	1 GlcNAc	z	31	3	-0.3
1960.9274	2	1.81E+06	2 GlcNAc	z	31	2	0.0
2036.5012	2	1.58E+06	1 GlcNAc	z	34	2	0.1
2093.5456	2	2.91E+06	1 GlcNAc	z+1	35	2	-0.6
1113.9452	8	2.98E+07	1 GlcNAc	c	81	8	-0.1
990.2854	9	1.03E+07	1 GlcNAc	c	81	9	-0.1
1153.2147	8	4.47E+06	1 GlcNAc	c	84	8	-0.5
1336.1106	7	4.89E+06	1 GlcNAc	c	85	7	0.1
1169.2223	8	1.85E+07	1 GlcNAc	c	85	8	-0.2
1039.4206	9	1.69E+07	1 GlcNAc	c	85	9	-0.3
1365.1224	7	6.20E+06	2 GlcNAc	c	85	7	0.4
1194.6074	8	1.02E+07	2 GlcNAc	c	85	8	-0.1
1061.9853	9	4.31E+06	2 GlcNAc	c	85	9	0.1

Table B.4. Intact mass distribution for intact hCEACAM1-IgV N104Q mutant.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	12431.43	24.08	19
2 GlcNAc	12634.54	100	81

Table B.5. Diagnostic fragment ions from top-down ECD MS/MS of N104Q 1 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7395	2	3.75E+07	0 GlcNAc	z	27	2	0.2
1024.8287	3	4.12E+07	0 GlcNAc	z	27	3	0.1
1067.5153	3	3.88E+06	1 GlcNAc	z	28	3	0.5
1638.2791	2	3.67E+06	1 GlcNAc	z	27	2	0.1
1092.5223	3	1.44E+06	1 GlcNAc	z	27	3	0.6
1145.4635	8	1.32E+06	0 GlcNAc	c-1	85	8	-0.1
1319.8188	7	1.56E+06	1 GlcNAc	c	84	7	0.2
1170.974	8	2.63E+07	1 GlcNAc	c	85	8	-0.5
1338.1126	7	1.30E+07	1 GlcNAc	c	85	7	-0.1

Table B.6. Intact mass distribution for intact hCEACAM1-IgV N111Q mutant.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	12431.2	1.94	1.9
2 GlcNAc	12634.28	100	98.1

Table B.7. Diagnostic fragment ions from top-down ECD MS/MS of N111Q, 1 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7388	2	5.53E+07	0 GlcNAc	z	27	2	-0.2
1024.8284	3	3.43E+06	0 GlcNAc	z	27	3	-0.2
1638.2788	2	2.39E+07	1 GlcNAc	z	27	2	0.0
1092.5224	3	3.10E+06	1 GlcNAc	z	27	3	0.6
1294.9824	3	5.52E+06	0 GlcNAc	z	34	3	0.4
1942.4733	2	2.96E+06	0 GlcNAc	z+1	34	2	0.1
1998.5126	2	2.66E+06	0 GlcNAc	z	35	2	0.7
2100.0511	2	1.72E+06	1 GlcNAC	z	35	2	0.1
1400.3690	3	1.84E+06	1 GlcNAC	z	35	3	-0.5
1193.7449	7	2.19E+06	1 GlcNAc	c	76	7	0.2
1206.1771	7	2.24E+06	1 GlcNAc	c	77	7	-0.6
1560.9639	6	7.00E+06	1 GlcNAc	c	85	6	0.2
1338.1122	7	1.97E+07	1 GlcNAc	c	85	7	-0.4
1170.9739	8	2.27E+07	1 GlcNAc	c	85	8	-0.5

Table B.8. Intact mass distribution for intact hCEACAM1-IgV N115Q mutant.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	12431.18	63.07	39
2 GlcNAc	12634.27	100	61

Table B.9. Diagnostic fragment ions from top-down ECD MS/MS of N115Q 1 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
2043.5079	2	4.28E+06	1 GlcNAc	z	34	2	-0.5
1362.6753	3	2.07E+06	1 GlcNAc	z	34	3	0.2
1400.3687	3	6.41E+06	1 GlcNAc	z	35	3	-0.7
2100.0509	2	6.26E+06	1 GlcNAc	z	35	2	0.0
1193.3228	7	3.32E+06	0 GlcNAc	c	78	7	1.2

Table B.10. Intact mass distribution for intact hCEACAM1-IgV WT, inhibitor-treated.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	12214.16	3	2
1 GlcNAc	12417.23	100	75
2 GlcNAc	12620.32	28	21
3 GlcNAc	12823.40	2	2

Table B.11. Diagnostic fragment ions from top-down ECD MS/MS of WT, Inhibitor-treated, 1 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1024.8286	3	1.62E+08	0 GlcNAc	z	27	3	0.0
1536.7396	2	1.92E+08	0 GlcNAc	z	27	2	0.3
1067.5147	3	2.04E+07	0 GlcNAc	z	28	3	0.0
1600.7693	2	1.59E+07	0 GlcNAc	z	28	2	0.5
1651.2921	2	5.52E+06	0 GlcNAc	z	29	2	-0.1
1923.9193	2	5.15E+06	1 GlcNAc	z+1	32	2	-0.8
1395.6974	3	1.69E+07	1 GlcNAc	z	35	3	-0.3
1088.5602	8	2.63E+07	0 GlcNAc	c	81	8	-0.2
1153.2154	8	8.58E+06	1 GlcNAc	c	84	8	0.1
1558.6269	6	1.54E+07	1 GlcNAc	c	85	6	-0.5
1039.4206	9	1.54E+07	1 GlcNAc	c	85	9	-0.2
1169.2226	8	9.32E+07	1 GlcNAc	c	85	8	0.0
1336.1109	7	5.16E+07	1 GlcNAc	c	85	7	0.3
1870.1528	5	2.37E+06	1 GlcNAc	c	85	5	0.6

Table B.12. Diagnostic fragment ions from top down ECD MS/MS of WT, Inhibitor-treated, 2 GlcNAc species.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7393	2	5.47E+07	0 GlcNAc	z	27	2	0.1
1024.8284	3	2.78E+07	0 GlcNAc	z	27	3	-0.2
1600.7682	2	7.14E+06	0 GlcNAc	z	28	2	-0.2
1067.5148	3	3.48E+06	0 GlcNAc	z	28	3	0.1
1638.2789	2	5.37E+07	1 GlcNAc	z	27	2	0.0
1092.5216	3	1.08E+07	1 GlcNAc	z	27	3	0.0
1651.2921	2	4.79E+06	0 GlcNAc	z	29	2	-0.1
1702.3083	2	9.52E+06	1 GlcNAc	z	28	2	0.1
1859.3875	2	4.66E+07	1 GlcNAc	z	31	2	-0.1
1239.9274	3	2.05E+07	1 GlcNAc	z	31	3	-0.2
1047.0249	4	7.94E+06	1 GlcNAc	z	35	4	-0.2
1395.6993	3	1.19E+07	1 GlcNAc	z	35	3	1.1
1113.9452	8	2.31E+07	1 GlcNAc	c	81	8	-0.2
990.2853	9	6.90E+06	1 GlcNAc	c	81	9	-0.3
1169.2224	8	1.46E+07	1 GlcNAc	c	85	8	-0.1
1039.4208	9	1.10E+07	1 GlcNAc	c	85	9	-0.1
1336.1109	7	6.00E+06	1 GlcNAc	c	85	7	0.3
1194.6070	8	7.57E+06	2 GlcNAc	c	85	8	-0.4

Table B.13. Glycopeptide distribution for hCEACAM1-IgV WT, batch 1.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	3386.75842	10.74	28
2 GlcNAc	3589.83886	17.15	45
3 GlcNAc	3792.92190	9.84	26

Table B.14. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 1, 1 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1536.7393	2	5.47E+07	0 GlcNAc	z	27	2	0.1
1024.8284	3	2.78E+07	0 GlcNAc	z	27	3	-0.2
1600.7682	2	7.14E+06	0 GlcNAc	z	28	2	-0.2
1067.5148	3	3.48E+06	0 GlcNAc	z	28	3	0.1
1638.2789	2	5.37E+07	1 GlcNAc	z	27	2	0.0
1092.5216	3	1.08E+07	1 GlcNAc	z	27	3	0.0
1651.2921	2	4.79E+06	0 GlcNAc	z	29	2	-0.1
1702.3083	2	9.52E+06	1 GlcNAc	z	28	2	0.1
1859.3875	2	4.66E+07	1 GlcNAc	z	31	2	-0.1
1239.9274	3	2.05E+07	1 GlcNAc	z	31	3	-0.2
1047.0249	4	7.94E+06	1 GlcNAc	z	35	4	-0.2
1395.6993	3	1.19E+07	1 GlcNAc	z	35	3	1.1
1113.9452	8	2.31E+07	1 GlcNAc	c	81	8	-0.2
990.2853	9	6.90E+06	1 GlcNAc	c	81	9	-0.3
1169.2224	8	1.46E+07	1 GlcNAc	c	85	8	-0.1
1039.4208	9	1.10E+07	1 GlcNAc	c	85	9	-0.1
1336.1109	7	6.00E+06	1 GlcNAc	c	85	7	0.3
1194.607	8	7.57E+06	2 GlcNAc	c	85	8	-0.4

Table B.15. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 1, 2 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1511.7731	1	1.18E+06	0 GlcNAC	z+1	13	1	-0.62
1563.8279	1	2.60E+06	1 GlcNAc	c	12	1	0.93
2006.045	1	3.72E+06	1 GlcNAc	c	16	1	0.47
2383.238	1	1.76E+06	1 GlcNAc	z+1	19	1	-0.47
2583.3539	1	1.33E+06	1 GlcNAc	z+1	21	1	-0.53

Table B.16. Glycopeptide distribution for hCEACAM1-IgV WT, batch 2.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	3386.7971	10.12	28
2 GlcNAc	3589.8764	19.09	53
3 GlcNAc	3792.9743	6.88	19

Table B.17. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 2, 1 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
893.4362	1	3.87E+06	0 GlcNAc	c	8	1	-0.1
1006.5208	1	3.15E+06	0 GlcNAc	c	9	1	0.4
1383.7149	1	2.49E+07	0 GlcNAc	z+1	12	1	-0.4
1611.8129	1	2.20E+06	0 GlcNAc	z	14	1	-0.6
1710.883	1	4.74E+06	0 GlcNAc	z	15	1	0.4
2005.0371	1	2.55E+07	0 GlcNAc	c-1	16	1	0.4
2028.0057	1	9.16E+06	1 GlcNAc	z	16	1	0.5
2156.0623	1	4.57E+06	1 GlcNAc	z	17	1	-0.4
2270.1542	1	1.88E+06	1 GlcNAc	z+1	18	1	-0.4
2293.2435	1	1.58E+06	1 GlcNAc	z+1	20	1	-0.2

Table B.18. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 2, 2 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1383.7164	1	9.46E+06	0 GlcNac	z+1	12	1	0.7
1563.8267	1	2.75E+06	1 GlcNAc	c	12	1	0.1
2006.0446	1	6.71E+06	1 GlcNAc	c	16	1	0.3
2029.0107	1	3.97E+06	1 GlcNAc	z+1	16	1	-0.9
2081.0667	1	1.32E+06	2 GlcNAc	c	15	1	0.9
2156.0618	1	1.10E+06	1 GlcNAc	z	17	1	-0.7
2209.1251	1	7.08E+06	2 GlcNAc	c	16	1	0.8
2496.3246	1	1.36E+07	1 GlcNAc	z+1	20	1	0.6
2654.3938	1	1.69E+07	1 GlcNAc	z+1	22	1	0.6
2699.4032	1	9.18E+06	2 GlcNAc	z+1	20	1	0.3
2786.432	1	1.82E+06	2 GlcNAc	z+1	21	1	-0.9

Table B.19. Glycopeptide distribution for hCEACAM1-IgV WT, batch 3.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	3386.7654	5.14	8
2 GlcNAc	3589.8397	32.96	51
3 GlcNAc	3792.9276	26.04	41

Table B.20. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 3, 1 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
893.4367	1	1.07E+07	0 GlcNAc	c	8	1	0.4
1006.5201	1	9.77E+06	0 GlcNAc	c	9	1	-0.3
1232.6888	1	5.24E+06	0 GlcNAc	c	11	1	0.2
1360.7463	1	1.78E+07	0 GlcNAc	c	12	1	-0.6
1383.715	1	7.61E+07	0 GlcNAc	z+1	12	1	-0.3
1511.7731	1	1.80E+07	0 GlcNAc	z+1	13	1	-0.6
1711.8886	1	1.46E+07	0 GlcNAc	z+1	15	1	-0.9
1802.9651	1	1.13E+06	0 GlcNAc	c	16	1	0.2
2005.0369	1	2.40E+07	1 GlcNAc	c-1	16	1	0.3
2156.0621	1	1.51E+07	1 GlcNAc	z	17	1	-0.5

Table B.21. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 3, 2 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1383.7056	1	6.12E+06	0 GlcNAc	z+1	12	1	-0.3
2005.0193	1	949529	1 GlcNAc	c-1	16	1	0.4
2028.9932	1	2.20E+06	1 GlcNAc	z+1	16	1	-0.5
2209.1007	1	5.30E+06	2 GlcNAc	c	16	1	-0.4
2383.2144	1	3.81E+06	1 GlcNAc	z+1	19	1	0.2
2496.2951	1	9.90E+06	1 GlcNAc	z+1	20	1	-0.4
2583.3282	1	4.98E+06	1 GlcNAc	z+1	21	1	0.9
2654.3624	1	1.41E+07	1 GlcNAc	z+1	22	1	0.2
2699.3691	1	2.42E+06	2 GlcNAc	z+1	20	1	-0.7

Table B.22. Glycopeptide distribution for hCEACAM1-IgV WT, batch 4.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	N/A	N/A	0
1 GlcNAc	3386.7827	20.41	47
2 GlcNAc	3589.8567	19.9	46
3 GlcNAc	3792.9467	2.93	7

Table B.23. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 4, 1 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
806.4009	1	2.46E+06	0 GlcNAc	c	7	1	0.8
893.4318	1	1.07E+07	0 GlcNAc	c	8	1	0.0
1006.5144	1	9.77E+06	0 GlcNAc	c	9	1	-0.5
1232.6813	1	5.24E+06	0 GlcNAc	c	11	1	0.3
1360.7378	1	1.78E+07	0 GlcNAc	c	12	1	-0.4
1383.7063	1	7.61E+07	0 GlcNAc	z+1	12	1	0.0
1511.7632	1	1.80E+07	0 GlcNAc	z+1	13	1	-0.2
1711.8768	1	1.46E+07	0 GlcNAc	z+1	15	1	-0.2
1776.9227	1	9.40E+06	1 GlcNAc	c	14	1	-0.7
1877.9697	1	1.95E+07	1 GlcNAc	c	15	1	-0.3
2006.0264	1	8.35E+07	1 GlcNAc	c	16	1	-0.3
2028.9945	1	3.32E+07	1 GlcNAc	z+1	16	1	-0.3
2156.0456	1	1.51E+07	1 GlcNAc	z	17	1	0.8
2270.1343	1	4.79E+06	1 GlcNAc	z+1	18	1	0.2
2293.2217	1	1.28E+06	0 GlcNAc	z+1	20	1	-0.3
2383.2159	1	5.62E+07	1 GlcNAc	z+1	19	1	-0.1
2496.2991	1	9.52E+07	1 GlcNAc	z+1	20	1	0.4
2583.3292	1	2.41E+07	1 GlcNAc	z+1	21	1	0.2
2654.3652	1	2.88E+07	1 GlcNAc	z+1	22	1	0.4

Table B.24. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, Batch 4, 2 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1383.7152	1	1.03E+07	0 GlcNAc	z+1	12	1	-0.2
1913.9631	1	1.40E+06	1 GlcNAc	z	15	1	0.7
2006.0446	1	8.42E+06	1 GlcNAc	c	16	1	0.3
2029.0112	1	5.23E+06	1 GlcNAc	z+1	16	1	-0.6
2157.0698	1	2.70E+06	1 GlcNAc	z+1	17	1	-0.6
2209.123	1	1.36E+07	2 GlcNAc	c	16	1	-0.2
2231.0833	1	1.40E+06	2 GlcNAc	z	16	1	-0.3
2496.3226	1	1.79E+07	1 GlcNAc	z+1	20	1	-0.2
2585.3086	1	1.58E+06	2 GlcNAc	z	19	1	-0.8
2654.3931	1	2.13E+07	1 GlcNAc	z+1	22	1	0.3
2699.403	1	1.46E+07	2 GlcNAc	z+1	20	1	0.2
2785.4269	1	2.27E+06	2 GlcNAc	z	21	1	0.1

Table B.25. Glycopeptide distribution for hCEACAM1-IgV WT, inhibitor-treated.

Species	Observed Monoisotopic Mass	Rel. Intensity (%)	Proportion of total signal (%)
0 GlcNAc	3183.6992	0.52	1
1 GlcNAc	3386.7769	39	80
2 GlcNAc	3589.8632	9	19
3 GlcNAc	3792.963	1	1

Table B.26. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, inhibitor-treated, 1 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
735.367	1	1.60E+06	0 GlcNAc	c	6	1	-0.1
806.4037	1	3.54E+06	0 GlcNAc	c	7	1	0.2
893.4352	1	1.33E+07	0 GlcNAc	c	8	1	0.7
1006.5177	1	1.14E+07	0 GlcNAc	c	9	1	0.4
1231.6751	1	3.53E+06	0 GlcNAc	c-1	11	1	0.9
1360.7373	1	1.88E+07	0 GlcNAc	c	12	1	-0.4
1383.7058	1	7.81E+07	0 GlcNAc	z+1	12	1	0.1
1511.7608	1	1.86E+07	0 GlcNAc	z+1	13	1	-0.3
1711.8716	1	1.45E+07	0 GlcNAc	z+1	15	1	-0.5
1776.9171	1	9.40E+06	1 GlcNAc	c	14	1	-0.8
1802.9439	1	4.95E+06	0 GlcNAc	c	16	1	-0.5
1825.9111	1	5.60E+06	0 GlcNAc	z+1	16	1	-0.3
1877.9617	1	2.02E+07	1 GlcNAc	c	15	1	-0.7
2006.0152	1	8.81E+07	1 GlcNAc	c	16	1	-0.4
2028.9833	1	3.05E+07	1 GlcNAc	z+1	16	1	-0.1
2157.0373	1	1.43E+07	1 GlcNAc	z+1	17	1	-0.8
2179.1203	1	3.13E+06	0 GlcNAc	z	19	1	0.5
2293.2062	1	1.10E+07	0 GlcNAc	z+1	20	1	0.7
2379.2273	1	1.87E+06	0 GlcNAc	z	21	1	0.6
2383.1966	1	5.68E+07	1 GlcNAc	z+1	19	1	-0.4
2496.2755	1	1.02E+08	1 GlcNAc	z+1	20	1	-0.2
2583.3043	1	2.49E+07	1 GlcNAc	z+1	21	1	-0.2
2654.3368	1	3.19E+07	1 GlcNAc	z+1	22	1	-0.5

Table B.27. Diagnostic fragment ions from bottom-up ECD MS/MS of WT, inhibitor-treated, 2 GlcNAc glycopeptide.

m/z	z	Intensity	PTM	Ion Type	Index	Charge	Error
1383.7052	1	9.85E+06	0 GlcNAc	z+1	12	1	-0.3
1585.7715	1	3.47E+06	1 GlcNAc	z	12	1	-0.5
1877.963	1	1.61E+06	1 GlcNAc	c	15	1	0.3
2027.9765	1	950634	1 GlcNAc	z	16	1	-0.1
2081.0333	1	1.29E+06	2 GlcNAc	c	15	1	-0.5
2157.0362	1	1.66E+06	1 GlcNAc	z+1	17	1	-0.8

Table B.28. Predicted glycoform distribution from best fitting kinetic model.

Glycoform	**Relative Proportion (%)**
[000]	0.2
[100]	0.0
[010]	15.1
[001]	0.0
[110]	17.9
[101]	33.4
[011]	3.7
[111]	29.7
[Q00]	0.2
[Q10]	23.0
[Q01]	0.5
[Q11]	76.3
[0Q0]	0.4
[1Q0]	0.1
[0Q1	0.1
[1Q1]	99.5
[00Q]	0.2
[10Q]	0.2
[01Q]	40.4
[11Q]	59.2