SELENOMETHIONINE AS A PROBE OF ITS ENVIRONMENT IN BIOLOGICAL MACROMOLECULES

by

Qingqing Chen

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry and Biochemistry

Summer 2018

SELENOMETHIONINE AS A PROBE OF ITS ENVIRONMENT IN BIOLOGICAL MACROMOLECULES

by

Qingqing Chen

Approved: _____
Brian J. Bahnson, Ph.D.
Chair of the Department of Chemistry and Biochemistry

Approved: _____
George Watson, Ph.D.
Dean of the College of Arts and Sciences

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Approved: _____

Sharon Rozovsky, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Approved: _____

Brian J. Bahnson, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Approved: _____

Neal J Zondlo, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Approved: _____

Shuo Wei, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

It is without a doubt that none of this work would be possible, had I not have received inspiration, guidance, encouragement, and critique from the extraordinary people in my life. My most sincere gratitude to my PhD advisor, Dr. Sharon Rozovsky, your drive has and will continue to inspire me each and everyday. I already miss your hilarious one-liners during group meetings. I also want to thank my PhD committee, Dr. Neal J. Zondlo, Dr. Brian J. Bahnson and Dr. Shuo Wei, and a special thanks to Dr. Bahnson and Dr. Marc Allaire for their collaborations on protein X-ray crystallography.

Thank you to the most amazing mother anyone could ever wish for. You have the strongest will, spirit and determination. Thanks dad for your humors to always keep me lighthearted. I owe my family my eternal gratitude for making me the person I am today. To my colleagues Johnny and Rujin, I miss the late nights in the lab when we labor over long experiments, discussing our research and dreaming about the future. Many thanks to the past, present and affiliated members of the Rozovsky group for their support and inspiration, especially Dr. Jun Liu for insightful scientific discussions.

To my husband Julián, this is my equivalent of that goal in *fútbol*.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NMR	Nuclear Magnetic Resonance
Met	Methionine
SeM	Selenomethionine
GB1	β1 Immunoglobulin Binding Domain of Protein G
DFT	Density Functional Theory
GPx	Glutathione Peroxidase
TrxR	Thioredoxin Reductase
ROS	Reactive Oxygen Species
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
TEV	Tobacco Etch Virus
GB1	Immunoglobulin G-Binding Protein G Domain B1
IPTG	Isopropyl β-D-1-thiogalactopyranoside
PMSF	Phenylmethylsulfonyl Fluoride
IMAC	Immobilized Metal Ion Affinity Chromatography
CV	Column Volume
BME	β-Mercaptoethanol
MWCO	Molecular Weight Cut Off
CP/MAS	Cross-Polarization Magic Angle Spinning
MPD	2-Methyl-2,4-Pentanediol

IPA	Isopropyl Alcohol
XDS	X-ray Detector Software
CCP4	Collaborative Computational Project No. 4
PHENIX	Python-based Hierarchical ENvironment for Integrated Xtallography
COOT	Crystallographic Object-Oriented Toolkit
ADP	Atomic Displacement Parameters
AU	Asymmetric Unit
RMSD	Root Mean Square Distance

ABSTRACT

Selenium can be used as a surrogate atom to sulfur in nuclear magnetic resonance spectroscopy (NMR) in order to understand the biological properties of sulfur, such as that thiols as activating groups in thioester biochemistry and disulfide bonds make up structural features of proteins. The ^{77}Se chemical shielding tensor is highly sensitive to the protein environment and can report on which molecular interactions and degrees of freedom are available for seleno-containing amino acids. Particularly, the substitution of methionine (Met, M) to selenomethionine (SeM) was consistently shown to have minimal impact on structure and dynamics. Therefore, SeM can be used as probe of its local environment in proteins. However, for biological systems, data interpretation has yet to be developed systematically for ^{77}Se NMR to explain which variables most affect the spectra.

To start building a biological ^{77}Se magnetic resonance databank we constructed a library of GB1 variants with a single SeM at dissimilar locations. X-ray crystallography was used to obtain a direct visualization of the local environment surrounding SeM in each variant at atomic resolution. The structures confirmed the conformational flexibility solvent exposed locations at Ile6 and Val29, as well as the local rigidities at Leu5 and Val39. The SeM sidechain at position 34 exists on a continuum of conformations stabilized by aromatic or nonpolar interactions. These structures are used as the basis for the development of DFT calculations to investigate the ^{77}Se NMR parameters.

Chapter 1

SELENIUM IN HUMAN PHYSIOLOGY

1.1 Background

1.1.1 Chemistry of Selenium

Selenium was discovered in 1817 by Swedish chemist Jöns Jacob Berzelius. [1] While working with sulfuric acid he observed a reddish-brown sediment that collected at the bottom of the flask in which the acid was prepared. He noted that this substance possessed similar properties to sulfur and tellurium, and later realized it was the minute presence of this substance that gave tellurium its characteristic odor. The substance was eventually determined to be a new element, which Berzelius named *selenium*, after the Greek moon goddess Selene. During his research Berzelius became physically affected by selenium due to its absorption through his skin and caused him to experience the bad breath typically associated with people who worked with selenium in other capacities.

In earth's crust, selenium is present at four-fold less abundance than sulfur and this difference is reflected also in biological systems. Whereas sulfur is present in humans in gram quantities, selenium in comparison is an order of magnitude less, in

low milligram quantities only. [2] However, since selenium is located directly below sulfur on the periodic table, these two elements are expected to share many physicochemical similarities. For example, there are only minor differences between sulfur and selenium in terms of electronegativity, ionic radius and available oxidation states. Selenium has a larger atomic radius making it more polarizable and can act as both a nucleophile and an electrophile, depending on its oxidation and protonation state. [3]

Some bioavailable forms of selenium include both inorganic selenate and selenite, and organic SeM and Sec. More than 50 Sec-containing proteins have been identified to date and of which the functionally studied ones contain selenocysteine (Sec, U) as the enzymatically active seleno-amino acid. [4,6] Sec has a pKa of 5.3 compared to cysteine (Cys) at 8.3 and is therefore a better nucleophile. The strength of its nucleophilicity has been explored in abiological nad biological systems alike. [5-9] Due to this striking difference in pKa, selenols in selenoproteins typically exist as selenolates under physiological pH, whereas their thiol counterparts are present in the neutral form. Disulfides function as both structural moieties and redox centers in proteins, while diselenide bonds identified so far exist as active functional groups in proteins. [8-9] Selenylsulfide bonds are also present in a number of selenoproteins, the most well studied selenium enzymes are glutathione peroxidase (GPx) and human thioredoxin reductase (TrxR). To date, the redox potential of many selenoproteins possessing this functional moiety have not been determined, largely due to the increased

reactivity of the selenylsulfide bond, and the short lifetime of the reduced selenolate-thiolate species. [2]

1.1.2 Selenium and Health

Selenium was originally recognized as associated with toxicity. First evidence of the health benefits of selenium came from a study by Schwarz and Foltz in 1957 who observed that selenium negated necrotic liver damage in mice. The optimal range of selenium intake for humans is narrow, with both lower and higher intakes associated with disease states. It took another two decades for selenium to be identified as an essential cofactor of GPx, confirming its importance in health. Of the identified mammalian selenoproteins, however, the exact functions of most of them have not been identified. [5-15] The health benefits of selenium include antioxidant properties, modulations of immune systems, detoxification and thyroid hormone metabolism. In addition, due to its suspected antioxidant activity, selenoproteins in brain functions and neurodegenerative diseases have also drawn research interest.

The GPx family of seleno-enzymes are expressed in the neurons and are widely recognized for their roles in protecting against damage by reactive oxygen species (ROS). More specifically, GPx4 is found in the cerebellum, hippocampus and hypothalamus where it protects the cells against lipid peroxide damage. [6,7,10-12]

The TrxR family is another seleno-enzyme family of proteins. TrxRs are abundant in the central nervous system. In the mouse brain, the cytosolic isoform of

TrxR is abundantly expressed in the glial cells while the mitochondrial isoform is found in the cell bodies of neurons. Along with its substrate thioredoxin, these two proteins form the cellular disulfide reductase system. [6,7,10]

Outside of the GPx and TrxR families of seleno-enzymes, another family of selenoproteins are characterized by a conserved Cys-x-x-Sec motif which includes SELENOH, SELENOM, SELENOW, SELENOT and SELENOO. Selenoprotein 15 (Sep15) has an abbreviated Cys-x-Sec motif while SELENOP has a reversed version of Sec-x-x-Cys. Complete physiological roles of this family are not known. It is generally regarded that they act as thiol-disulfide oxidoreductases, but new research in the field have shed light on updated functionalities. [6-7]

Due to the proposed roles of selenoproteins against cellular oxidative challenges, there have been research interest into their roles in neurodegenerative diseases such as Alzheimer's and Parkinson's diseases. Oxidative damage is a central component of Alzheimer's disease. However, to date, research in this field remains inconclusive. [11]

Of the various health benefits that have been attributed to selenium and selenoproteins, its possible ties to cancer has no doubt received the most attention. [7] Over the years, while some clinical trials found decreased incidences of prostate, colon and lung cancers associated with selenium supplementation, others have found no particular positive effects with increased selenium intake. In fact, the largest cancer prevention trial ever taken was terminated early. [12-14] Failure of these studies could partially be attributed to poor understanding of the molecular functions of selenium. In addition, little consideration was given to the type of selenium that was administered in

each case, which may have affected the outcome of each study and possibly even lead to undesired negative effects.

Research into the health effects of selenium have largely focused on its role in cancer however it is clear that selenium and selenoproteins participate in a myriad of cellular activities. [6,7,10,12,15-18] A deeper understanding of these roles could help to explain the biological roles of selenium and its relationship to human health, in additional this knowledge can also provide insight for developing protein targets for therapeutics or even new therapeutics themselves.

1.1.3 Selenium and Sulfur

Physicochemical properties of sulfur and selenium are summarized in Table 1. It is immediately clear that the two elements share many commonalities, and therefore not surprising that many sulfur compounds have selenium analogs. A few examples include disulfides for which there exists diselenides, sulfites and selenites, sulfides and selenides, sulfenic and selenic acids and so forth.

Significant differences between the two atoms were reviewed extensively by *Reich and Hondal* in "Why Nature Chose Selenium". [19] These differences arise as a result of the higher polarizability of the heavier selenium atom compared to sulfur. Increased polarizability makes nucleophilic and electrophilic reactions about the selenium atom faster. Bond strengths with selenium are weaker both in general and with hydrogen. The larger atomic size of the selenium atom compared to sulfur, and its

marginally higher electronegativity contribute to the selenol functional group being more basic compared to its thiol counterpart by 3-4 pKa units. For example, the pKa of the free amino acid cysteine is 8.25 but for selenocysteine it is 5.24; and the pKa of hydrogen sulfide is 7.0 while for hydrogen selenide it is 3.74. [19-20]

While the selenol functional groups are more acidic, they are at the same time better nucleophiles by approximately 1 order of magnitude compared to thiolates. [19] Nucleophilicity refers to the propensity of an atom to donate electrons to a foreign nucleus or to supply a pair of electrons when forming a new bond with another atom. The strength of a nucleophile is guided by three major factors: Lewis basicity, polarizability, and access to unshared electrons in a chemical reaction. [21,22] The difference in nucleophilicity between sulfur and selenium becomes more pronounced at physiological pH, as selenols are completely deprotonated to selenolates but thiols remain protonated. Therefore, in the event of a nucleophilic attack, where the deprotonated selenoate or thiolate is the active nucleophile, the selenolate will be the more reactive species of the two, as it possesses both higher nucleophilicity and a larger percentage of it exists in the more nucleophilic anionic form. However, it is worth noting that in the context of the protein, the pKa values of an amino acid can be significantly perturbed as evidenced by the various cysteine-containing enzymes such as papain, caricain and ficin, whose cysteine pKa values are 3.3, 2.9 and 2.5, respectively. [19] The pKa of Sec has also been measured by NMR in the peptide hormone and neurotransmitter vasopressin, which have had two Sec residues artificially incorporated. [23] In the engineered vasopressin, the Sec residues are located four amino acids apart

in the sequence UYFQNUPRG, and form a diselenide bond. The pKa of Sec1 and Sec 6 were found to be 3.3 and 4.3, respectively, both significantly lower than that of the free amino acid. This further corroborates that the local protein environment significantly influences the protonation characteristics of residues. Vasopressin is a small peptide yet its two Sec residues possess remarkably different pKa values. Therefore, it is possible that the pKa of Sec residues are not just dependent on the protein environment, but also the primary amino acid sequences. [23]

1.1.4 Biochemical and Biophysical Characterizations of Selenium and Sulfur

Cys and Met, two canonical amino acids which contain sulfur, and their selenium-containing analogs Sec and SeM, are four amino acids with versatile chemistry and fulfill a wide range of essential biological functions. Together with their chalcogenide congener oxygen, these elements form the functional groups that participate in redox reactions *in vivo*. Redox homeostasis is of paramount importance to cellular health, where an imbalance of ROS species leads to many diseases, neurodegenerative disorders and the aging process. [24-27] Understanding these reactions and diseases processes is of biochemical and pharmacological interest to understand the enzymatic roles of these amino acids.

Functionalized thiol-labeling is an easily accessible biochemical method for determining the redox potential of proteins. Cys residues are sparsely present in proteins and whether they are part of a disulfide or a sulfhydryl, or in any one of sulfur's

oxidation states, these changes can be monitored by biochemical assays. Many selenoproteins contain a selenylsulfide bond (Se-S) formed between a Sec and a Cys residue, in place of the more common disulfide bond (S-S) between two Cys. A selenylsulfide bond typically has a lower reduction potential, making it more resistant to reduction and thus a more stable linkage compared to the disulfide bond. [27-31] To probe the reduced sulfhydryl and selenohydryl groups, there are many reagents that are commercially available with a wide range of selectivity from pH to probe size. Such reagents include maleimide, iodacetate amides or disulfide moieties, and several different reporter choices are also available such as radiolabel, chromophore, fluorophore or affinity tags. [26] In the simplest case, Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis (SDS-PAGE) can be used to separate oxidized proteins from the reduced and label-conjugated version because when conjugated the protein has retarded motion due to increased molecular weight. These assays can be more challenging for selenylsulfide groups due to increased reactivity of selenium, making the reduced form of the protein more difficult to trap. In addition, *in vitro* redox assays are limited to probing only the equilibrium redox potential and cannot provide an accurate reflection of a protein's true behavior intracellularly. Alternative *in vivo* assays are available using fast-conjugation coupled with fast-quenching probes, but these methods are not absolutely quantitative. As a result, important information about turnover rates of a redox active enzyme are lost, as well as dynamic information regarding flux of cellular oxidants and reductants that regulate the redox processes. [26]

While the reactivity of selenium poses challenges in assaying the redox properties of these proteins, both biologically and artificially the properties of selenium has allowed its use as a sulfur substitute to fulfill certain functionalities. [32-33] As previously discussed, of the three biologically observed chalcogenides, selenium is chemically most similar to sulfur. Where the two elements show considerable difference is in their atomic mass number and polarizability. The selenium atom is 10% larger than sulfur (Table 1) which means its electrons are more loosely held compared to the smaller sulfur. Accordingly, the vibrational property of a selenium bond is expected to be considerably different than that with sulfur and this difference was exploited by Meyer *et al.* in studying iron-sulfur proteins using resonance Raman spectroscopy. [33-34] Resonance Raman spectroscopy is a selective and sensitive probe method for chromophores in biological systems. The basic theory of this technique relates the polarizability of an atom to its excitation by an incoming photon. The photon excites the atoms in a molecule through inelastic light scattering, and after interaction with the electron cloud, the sample is left in a different rotational or vibrational state. A key aspect of the Raman scattering technique is that only transitions which lead to a change in the polarization of the bond can be observed. This is because when the incoming photon interacts with the electron cloud of the sample, it causes a momentary distortion to the electron cloud, thus re-shaping the chemical bond temporarily. An induced dipole is created and disappears upon relaxation. This reliance on polarizability explains the popularity of using selenium as a surrogate in studying iron-sulfur clusters. As the rotational and vibrational properties of each selenium isotope is different from the native

sulfur, a systematic series of substitutions were in fact carried out using isotopes of both sulfur and selenium in order to catalog the Fe-S(e) stretching frequencies in a number of iron-sulfur active centers. [33-34]

Two other major techniques which will be discussed in the bulk of this text, that have both derived successes from selenium-sulfur substitution are X-ray crystallography and NMR.

The number of crystal structures solved by X-ray diffraction has increased exponentially since the early 2000s. [35] As the database of known structures expands, the use of molecular replacement (MR) in solving a crystal structure naturally becomes the more popular choice. MR has the advantage of being operationally simple, fast, and highly automated. [36] In fact, MR has been used to solve up to 70% of the structures currently deposited in the PDB. [36-37] The foundation of molecular replacement is to start with a model structure that closely approximates the unknown structure of interest, and calculate the best fit between the data and possible orientations in order to find where the predicted diffraction from the model best matches the observed diffraction of the unknown protein. [37] The phases for the reflections of the model is then lent to obtain the phases for the unknown, from which an initial electron density map cam be calculated.

The derivation of the electron density ρ in a unit cell is empirically straightforward. With unit cell of volume V, the electron density ρ at point (x, y, z) for reflection (h, k, l) is written as: [38]

$$\rho(x,y,z) = \frac{1}{V}\sum\sum|F_{hkl}|\cos[2\pi\,(hx+ky+lz-\alpha_{hkl})]$$

In this expression, $|F_{hkl}|$ represent the structure-factor amplitudes recorded in a diffraction experiment, and α_{hkl} represent the phase information which is lost in the data acquisition process. [39] Therefore, once an X-ray diffraction data set has been recorded, its electron-density map can only be calculated when the phases are obtained either through molecular replacement as previously discussed, or if a *de novo* structure is to be solved, through a process called experimental phasing.

Multiple approaches have been developed for *de novo* phase estimation: Multiple Isomorphous Replacement (MIR), Single Isomorphous Replacement (SIR), Single-wavelength Anomalous Dispersion (SAD) and Multi-wavelength Anomalous Dispersion (MAD). [39-42] These approaches require the presence of "phasing probes" or heavy atoms in the protein crystal lattice, and the most commonly utilized heavy atoms are gold, mercury, platinum and selenium. [43] when X-rays interact with the electron cloud of an atom, the more electrons there are, the more X-rays will be scattered, and the larger the change in the diffraction pattern compared to the native crystals, which makes obtaining the phase information an easier task.

Methods of incorporation of the probes range from solution diffusion or incorporation into the protein during expression, such is the case when selenomethionine is used. [44] The application of SeM is the most robust and widespread. [38,45-51] The strategy for SeM incorporation is simple, adaptable to a variety of expression systems, and often results in quantitative replacement of Met to

yield a highly homogenous sample suitable for crystallization. Many studies have shown that this modification does not significantly alter the structure of the protein, so the end structure obtained is essentially that of the native protein of interest. [51]

Another structural biology method that benefits from having selenium in place of sulfur is NMR spectroscopy. Cys and Met both participate in a variety of biochemical interactions, ranging from enzymatic catalysis to protein recognition and binding. [52-58] NMR is a popular biophysical technique that can be utilized to probe the mechanistic changes of the sulfur atom during a chemical reaction, because chemical shift measured by NMR is intimately related to the local electronic environment of the nucleus of interest. Unfortunately, the only NMR-active isotope of sulfur, ^{33}S is a quadrupolar nucleus (I = 3/2) whose quadrupole moment results in fast relaxation of the free induction decay (FID) which leads to broad lines that are difficult to analyze. Additionally, its natural abundance is 0.76% which is within the same order of magnitude as ^{13}C and ^{15}N, which combined with its inherent low sensitivity, means that isotopic labelling would always be required to obtain adequate signal to noise ratio in an experiment. [59-61] These challenges can be overcome in small sulfur-containing organic molecules. It has been demonstrated that for highly symmetric molecules such as SO_4^{2-}, the NMR peak is sufficiently sharp, while for asymmetric environments such as a biological macromolecule, the NMR peak is broad and often undetectable. [61-64]

On the contrary, ^{77}Se NMR could potentially be an ideal alternative for examining the nature and interactions of sulfur in biological molecules. ^{77}Se is a spin 1/2 (I=1/2) nucleus with natural abundance of 7.5% and good sensitivity ($6.93x10^{-3}$),

allowing its detection in protein molecules using conventional NMR instrumentation. [65] In fact, ^{77}Se signal can be detected at natural abundance with sample concentration of 1.5 mM which allows straightforward detection using samples containing SeM. [66] As discussed in earlier sections, Cys and Met both have selenium-containing analogs in the form of Sec and SeM, both of which are naturally occurring amino acids and can be used as substitutes to Cys and Met. In addition, unlike NMR with ^{13}C or ^{15}N, in the case of selenium there is unlikely to be significant background contribution from the rest of the protein, as Cys or Met are found less than 4% of the time in protein sequences. [67-68] The selenium chemical shift range is also an order of magnitude wider compared to that of ^{13}C (~ 200 ppm) or ^{15}N (~ 800 ppm) at 3000 ppm, which makes selenium more sensitive to changes in bonding and conformation changes surrounding the nucleus. [69-70] However, this sensitivity brings about a paradoxical challenge as well: the selenium atom has large anisotropies in its chemical shift, which results in efficient spin-spin relaxation and as a consequence has broad peaks and decreased signal to noise ratio. The natural countermeasure for reduced sensitivity is to increase the magnetic field strength. Yet, the efficiency of relaxation through chemical shift anisotropy increases at high field strength which again leads to decreased sensitivity. [69] The feasibility of selenium NMR in biological systems has been demonstrated in multiple systems, using both natural and semi-synthetic systems. The full potential of this technique has not been realized due to challenges in sensitivity, and a lack of available literature aside from the pioneering work taken place some decades ago. [65,66,69-75] Another added layer of complication comes from the fact that while SeM incorporation into proteins

are straightforward, the substitution of Cys with Sec has been a challenge until recently. [65,76]

1.1.5 Biological Interactions of Met and SeM

Research over the past two decades has largely refuted the notion that the amino acid Met is a generic nonpolar residue synonymous to valine (Val), leucine (Leu) and isoleucine (Ile). Lack of branching compared to Leu and Ile bestows flexibility to the sidechain of Met and allows it to adapt to protein partners of different primary sequences but the same secondary structural characteristics. [77-78] In fact, Met engages in specific nonpolar interactions that contribute to stability and molecular recognition events. [80] Pathologies including von Willebrand disease have been traced back to the mutation of key Met residues which results in the loss of protein stability. [81]

Methionine, due to its nonpolar properties, is likely to be found in the interior of proteins, much like the aromatic residues. [77] However, the preference of finding Met residues close to aromatic residues may be due to more than their shared aversion to the polar exterior of a protein solution. The possibility of sulfur interacting preferentially with aromatic residues such as phenylalanine (Phe), tyrosine (Tyr), tryptophan (Trp) and histidine (His) was proposed in 1978. *Morgan et al.* observed high incidence of sulfur-containing amino acids located near other aromatic residues. [81-82] Such interactions are thought to be important in the stabilization of protein secondary and tertiary structures. [83] The energy associated with S-aromatic interactions are 3 to 5

times higher than van der Waals forces alone, therefore are likely to serve functional roles in both non-biological and biological systems. [84-86] Many approaches have been taken to shed light on the chemistry of this type of interactions, including data mining from the Protein Data Bank, model peptides systems, proteins examples, and *ab initio* calculations. [87-96]

The molecular basis specifically for the S-H/π interactions have been attributed to energetically favorable geometries between the aromatic residue HOMO and the S-H residue antibonding LUMO orbitals. [84] When examining the X-ray diffraction data from small molecules deposited in the PDB, *Forbes et al.* concluded that the distance between the proton from a thiol group and the carbon atom of the nearby aromatic residue, specifically the S-H$\cdots C_{aromatic}$ interaction distance is 2.63 Å, below the typical interaction distance for van der Waals interactions at 2.90 Å, suggesting a particularly favorable interaction between a thiol functional group and an aromatic amino acid. [84]

Data mining from the PDB often yields incomplete results because the electron density of a hydrogen atom cannot be observed for about 50% of the structures, particularly for macromolecular systems. [84,97-98] In this case, *Forbes et al.* applied Density Functional Theory (DFT) to investigate the driving force of S-H/π interactions, as well as the energies associated with these systems.

The formation of classical hydrogen bonds is driven by electrostatic and induced dipole/polarization interactions. An X-H/π interaction resembles that of a hydrogen bond, where the negatively charged ring face interacts with the partial positive charge on the hydrogen atom. [84] By comparing two systems, one in which a thiol group

interacts with an aromatic ring and the other system in which the thiol does not, *Forbes et al.* calculated that in the non-interacting system, H_{thiol} carries a partial positive charge of +0.21 and the sulfur a partial negative charge of -0.29, while in the interacting system, the charges on H_{thiol} and sulfur are +0.21 and -0.34, respectively. The partial positive charge on H_{thiol} is also comparable in strength with H on the aromatic ring, indicating the ring face is not strongly electron-withdrawing, therefore the partial negative charge on the ring face would be modest as well. These measurements lead the authors to conclude that a S-H/ π interaction is only modestly electrostatic in nature. Such effect is even further diminished in water - whose H atoms carry larger partial charges than H_{thiol}. [84]

The geometry of sulfur-aromatic interactions has also been the source of debate, again largely due to the absence of electron density around the thiol hydrogen, or in the case of methionine the entire terminal methyl group, thus making the assignment of the bond angle around the sulfur atom ambiguous. [84-85] *Ringer et al.* mapped the optimal geometry of a prototype H_2-S-benzene complex to model S/π interactions, and compared the results to sulfur-aromatic interactions in protein structures. [85] The authors identified that in the small molecule prototype system, the most energetically favorable configuration is to place SH_2 directly over the benzene ring, with hydrogens directed towards the face of the ring. The interaction energy of this configuration is -2.64 kcal/mol. Two other energetic minima were also identified. In one, the hydrogen atoms are directed away from the face of the ring with interaction energy of -1.12 kcal/mol. In the other, the hydrogen atoms are also directed away from the ring but the

SH_2 group is positioned in-plane with the ring face. In this latter case, the interaction energy is -0.74 kcal/mol. [85] This is an updated result compared to *Reid et al.* [80], whose survey of 36 proteins from the PDB determined that the in-plane configuration, where the sulfur atom interact with the edge of the aromatic ring is the most energetically preferable.

In biological systems, sulfur-aromatic interactions have been identified in the D2 dopamine receptor, 7TMRs (7-transmembrane receptors), and methionine sulfoxide. [94-96] In the D2 dopamine receptor, a sulfur-arene interaction between Cys and Trp was identified at the conserved ligand binding site. However, the exact geometry of this interaction cannot be determined due to lack of hydrogen resolution in the crystal structure. Computational studies also could not definitively determine if the sulfur atom preferentially interacts with the aromatic ring through its lone pair of electrons or through the H_{thiol}. [94] In 7TMRs, Met/Cys-aromatic and Met-Met interactions together participate in ligand binding during activation and signal transduction. [95] Finally, *Aledo et al.* determined that a S-aromatic motif decreases the reactivity of the sulfur towards oxidants and in turn protects the enzyme during oxidative stress. [96]

Sulfur-aromatic interactions have been shown to be stronger compared to aromatic interactions with other aliphatic amino acids. These interactions are postulated to provide additional driving forces for maintaining protein structure, which was specifically shown to be important for membrane proteins. [99] Sulfur atoms can both act as hydrogen bond donors and acceptors. This property can be harvested in crystal engineering to produce stable crystal structures. [100] Overall, there is much to be

understood about these noncovalent interactions involving the sulfur atom, and research in this field is still ongoing. [101-103]

1.1.6 Open Questions

The two sulfur containing amino acids in nature, Cys and Met, together with their selenium-containing analogs Sec and SeM, are found at enzymatic active sites and have become important targets for therapeutic intervention. Biochemical and biophysical characterizations of sulfur can be challenging depending on the technique of choice, in which case their selenium analogs have been successfully employed to expand the capabilities of certain techniques. The application of SeM substitution is almost ubiquitous but Sec substitution has not yet become generally applicable. This is because a complex molecular machinery is required to decode the UAG codon as Sec instead of a translational stop codon. This limitation makes the incorporation of Sec in proteins a much more challenging task and also severely restricts the amount of proteins that can be obtained using currently available Sec-insertion methods. Furthermore, while NMR is an attractive method for studying protein activity due to its exquisite sensitivity to the electronic structures of the target, selenium NMR is still in its infancy stage and the paucity of available research means data interpretation is difficult.

In the present work, a method for producing milligram quantities of Sec-containing proteins is described using expressed selenoprotein ligation. This method helps to address the limitations on protein availability and paves the way for more Sec-

containing systems to be investigated in the future – using either artificially incorporated Sec or natural human selenoproteins. The second half of this work describes the efforts undertaken to systematically construct a library of SeM-containing model proteins in order to understand the influence of the protein environment on ^{77}Se SeM chemical shifts.

Table 1. Physicochemical properties of sulfur and selenium [20]

	Sulfur	**Selenium**
Atomic Radius (Å)	1.80	1.90
Covalent Radius (Å)	1.04	1.18
Electron Affinity (kJ/mol)	200.41	194.97
Electronegativity (Pauling Scale)	2.58	2.55
Oxidation States	-2, 0, +4, +6	-2, 0, +4, +6
Electron Configuration	[Ne] $3s^2 3p^4$	[Ar] $3d^{10} 4s^2 4p^4$
Free Amino Acid pKa	8.25	5.24

REFERENCES

1. Weeks ME. (1932) The discovery of the elements VI. Tellurium andselenium. J Chem Educ 9 (3): 474. doi: 10.1021/ed009p474

2. Jacob C, Giles GL, Giles NM and Sies H. (2003) Sulfur and selenium: the role of oxidation state in protein structure and function. Angew Chem Int Ed 42 (39): 4742-4758. doi: 10.1002/anie.200300573

3. Mousa R, Nortis Dardashti R and Metanis N. (2014) Selenium and Sec in protein chemistry. Angew Chem Int Ed 56 (50): 15818-15827. doi: 10.1002/anie.201706876

4. Arteel G and Sies H. (2001) Biochemistry of selenium and the glutathione system. Environ Toxicol Pharmacol 10 (4): 153-158. doi: 10.1016/S1382-6689(01)00078-3

5. Birrigen M, Pilawa S and Flohé L. (2002) Trends in selenium biochemistry. Nat Prod Rep 19: 693-718. doi: 10.1039/b205802m

6. Labunskyy V, Hatfield DL and Gladyshev VN. (2014) Selenoproteins: molecular pathways and physiological roles. Physiol Rev 94 (3): 739-777. doi: 10.1152/physrev.00039.2013

7. Barage SH, Deobagkar DD and Baladhye VB. (2018) Characterization of structural and functional role of selenocysteine in selenoprotein H and its impact on DNA binding. Amino Acids 50 (5): 593-607. doi: 10.1007/s00726-018-2543-5

8. Liu J, Zhang Z and Rozovsky S. (2014) Selenoprotein K form an intermolecular diselenide bond with unusually high redox potential. FEBS Lett 588 (18): 3311-3321. doi: 10.1016/j.febslet.2014.07.037

9. Bachrach SM, Demoin DW, Luk M and Miller JV. (2004) Nucleophilic attack at selenium in diselenides and selenosulfides. A computational study. J Phys Chem A 108 (18): 4040-4046. doi: 10.1021/jp037972o

10. Fairweather-Tait SJ, Bao Y, Broadley MR, Collings R, Ford D, Hesketh JE and Hurst R. (2011) Selenium in human health and disease. Antioxid Redox Sign 14 (7): 1337-1383. doi: 10.1089/ars.2010.3275

11. Cardoso BR, Roberts BR, Bush AI and Hare DJ. (2015) Selenium, selenoproteins and neurodegenerative diseases. Metallomics 7: 1213-1228. doi: 10.1039/c5mt00075k

12. Kurokawa S and Berry MJ. (2013) Selenium. Role of the essential metalloid in health. Met Ions Life Sci 13: 499-534. doi: 10.1007/978-94-007-7500-8_16

13. Klein EA, Thompson IM, Tangen CM, Crowley JJ, Lucia MS, Goodman PJ, Minasian L, Ford LG, Parnes HL, Gaziano JM, Karp DD, Lieber MM, Walther PJ, Klotz L, Parsons JK, Chin JL, Darke AK, Lippman SM, Goodman GE, Meyskens FL and Baker LH. (2011) Vitamin E and the Risk of Prostate Cancer: Updated Results of The Selenium and Vitamin E Cancer Prevention Trial (SELECT). JAMA 306 (14): 1549-1556. doi: 10.1001/jama.2011.1437

14. Hatfield DL and Gladyshev VN. (2009) The Outcome of Selenium and Vitamin E Cancer Prevention Trial (SELECT) Reveals the Need for Better Understanding of Selenium Biology. Mol Interv 9 (1): 18-21. doi: 10.1124/mi.9.1.6

15. Norton R and Hoffmann PR. (2012) Selenium and asthma. Mol Aspects Med 33: 98-106.

doi: 10.1016/j.mam.2011.10.003

16. Papp LV, Lu J, Holmgren A and Khanna KK. (2007) From selenium to selenoproteins: synthesis, identity, and their role in human health. Antioxid Redox Sign 9 (7): 775-806. doi: 10.1089/ars.2007.1528

17. Rayman MP. (2000) The importance of selenium to human health. Lancet 356: 233-241. doi: 0.1016/S0140-6736(00)02490-9

18. Roman M, Jitaru P and Barbante C. (2014) Selenium biochemistry and its role for human health. Metallomics 6: 25-54. doi: 10.1039/c3mt00185g

19. Reich HJ and Hondal RJ. (2016) Why nature chose selenium. ACS Chem Biol 11: 821-841. doi: 10.1021/acschembio.6b00031

20. Johansson L, Gafvelin G and Arnér ESJ. (2005) Selenocysteine in proteins—properties and biotechnological use. Biochim Biophys Act 1726: 1-13. doi: 10.1016/j.bbagen.2005.05.010

21. Levander OA. (1992) Selenium and sulfur in antioxidant protective systems: relationships with vitamin E and malaria. Proc Soc Exp Biol Med 200 (2): 255-259. doi: 10.3181/00379727-200-43430

22. Forman HJ, Rotman EI and Fisher AB. (1983) Roles of selenium and sulfur-containing amino acids in protection against oxygen toxicity. Lab Invest 49 (2): 148-153. PMID: 6876743

23. Mobli M, Morgenstern D, King GF, Alewood PF and Muttenthaler M. (2011) Site-specific pKa determination of selenocysteine residues in selenovasopressin by using ^{77}Se NMR spectroscopy. Angew Chem Int Ed 50 (50): 11952-11955. doi: 10.1002/anie.201104169

24. Ayer A, Gourlay CW and Dawes IW. (2014) Cellular redox homeostasis, reactive oxygen species and replicative ageing in *Saccharomyces cerevisiae*. FEMS Yeast Res 14 (1): 60-72. doi: 10.1111/1567-1364.12114

25. Kwon YW, Masutani H, Nakamura H, Ishii Y, Yodoi J. (2003) Redox regulation of cell growth and cell death. Biol Chem 384: 991-996. doi: 10.1515/BC.2003.111

26. Rudyk O and Eaton P. (2014) Biochemical methods for monitoring protein thiol redox states in biologicalsystems. Redox Biol 2: 803-813.
doi: 10.1016/j.redox.2014.06.005

27. Hanold LE, Watkins CP, Ton NT, Liaw P, Beedle AM and Kennedy EI. (2015) Design of a selenylsulfide-bridged EGFR dimerization arm mimic. Bioorg Med Chem 23 (12): 2761-2766. doi: 10.1016/j.bmc.2015.03.040

28. Besse D, Siedler F, Diercks T, Kessler H, Moroder L. (1997) The Redox Potential of Selenocystine in Unconstrained Cyclic Peptides. Angew Chem Int Ed 36: 883–885. doi: 10.1002/anie.199708831

29. Wessjohann LA, Schneider A, Abbas M, Brandt W. (2007) Selenium in chemistry and biochemistry in comparison to sulfur. Biol Chem. 388:997–1006. doi: 10.1515/BC.2007.138

30. Koide T, Itoh H, Otaka A, Yasui H, Kuroda M, Esaki N, Soda K, Fuji N. (1993) Synthetic Study on Selenocystine-Containing Peptides. Chem Pharm Bull 41:502–506. doi: 10.1248/cpb.41.502

31. Muttenhaler M. (2008) Selenopeptide chemistry. J Pepti Sci 14:1223–1239. doi: 10.1002/psc.1075

32. Jacob C, Giles GI, Giles NM and Sies H. (2003) Sulfur and selenium: the role of oxidation state in protein structure and function. Angew Chem Int Ed 42: 4742-4758. doi: 10.1002/anie.200300573

33. Meyer J, Moulis JM, Gaillard J and Lutz M. (1992) Replacement Of Sulfur By Selenium In Iron-Sulfur Proteins. Adv Inorg Chem 38: 73-115. doi: 10.1016/S0898-8838(08)60062-8

34. Meyer J, Moulis JM and Lutz M. (1986) Resonance Raman spectroscopy of $[2Fe{-}2X]^{2+}$ (X = S, Se) clusters in ferredoxins. Biochim Biophys Acta Protein Struct Mol Enzymol 873 (1): 108-118. doi: 10.1016/0167-4838(86)90196-2

35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE. (2000) The Protein Data Bank. Nucleic Acids Research, 28: 235-242. URL: www.rcsb.org

36. Evans P and McCoy A. (2008) An introduction to molecular replacement. Acta Crystallogr D Biol Crystallogr 64 (Pt 1): 1-10. doi: 10.1107/S0907444907051554

37. Mao B, Guan R and Montelione GT. (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 19 (6): 757-766. doi: 10.1016/j.str.2011.04.005

38. Boggon TJ and Shapiro L. (2000) Screening for phasing atoms in protein crystallography. Structure 8 (7): R143-R149. doi: 10.1016/S0969-2126(00)00168-4

39. Dickerson RE, Kendrew JC and Strandberg BE. (1961) The crystal structure of myoglobin: Phase determination to a resolution of 2 Å by the method of isomorphous replacement. Acta Cryst 14: 1188-1195. doi: 10.1107/S0365110X61003442

40. Wang BC. (1985) Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol 115: 90-112. doi: 10.1016/0076-6879(85)15009-3

41. Hendrickson WA and Teeter MM. (1980) Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. Nature 290: 107-113. doi: 10.1038/290107a0

42. Hendrickson WA, Horton JR, Murthy HMK, Pahler A and Smith JL. (1989) Multiwavelength Anomalous Diffraction as a Direct Phasing Vehicle in Macromolecular Crystallography. Basic Life Sci 51: 317-324. doi: 10.1007/978-1-4684-8041-2_28

43. Pike AC, Garman EF, Krojer T, von Delft F and Carpenter EP. (2016) An overview of heavy-atom derivatization of protein crystals. Acta Crystallogr D Struct Biol 72 (Pt3): 303-318. doi: 10.1107/S2059798316000401

44. Hendrickson WA, Horton JR and LeMaster DM. (1990) Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): A vehicle for direct determination of three dimensional structure. EMBO J 9 (5): 1665-1672. doi:

45. Sheng J and Huang Z. (2008) Selenium derivatization of nucleic acids for phase and structure determination in nucleic acid X-ray crystallography. Int J Mol Sci 9 (3): 258-271. PMCID: PMC2635675

46. Jiang J, Sheng J, Carrasco N and Huang Z. (2006) Selenium derivatization of nucleic acids for crystallography.Nucleic Acids Res 35 (2): 477-485. doi: 10.1093/nar/gkl1070

47. Du Q, Carrasco N, Teplova M, Wilds CJ, Egli M and Huang Z. (2002) Internal Derivatization of Oligonucleotides with Selenium for X-ray Crystallography Using MAD. J Am Chem Soc 124 (1): 24-25. doi: 10.1021/ja0171097

48. Wilds CJ, Pattanayek R, Pan C, Wawrzak Z and Egli M. (2002) Selenium-Assisted Nucleic Acid Crystallography: Use of Phosphoroselenoates for MAD Phasing of a DNA Structure. J Am Chem Soc 124 (50): 14910-14916. doi: 10.1021/ja021058b

49. Zhang W, Szostak JW and Huang Z. (2016) Nucleic acid crystallization and X-ray crystallography facilitated by single selenium atom. Front Chem Sci Eng 10 (2): 196-202. doi: 10.1007/s11705-016-1565-3

50. Luo Z. (2016) Selenourea: a convenient phasing vehicle for macromolecular X-ray crystal structures. Sci Rep 6: 37123. doi: 10.1038/srep37123

51. Barton WA, Tzvetkova-Robev D, Erdjument-Bromage H, Tempst P and Nikolov DB. (2006) Highly efficient Sem labeling of recombinant proteins produced in mammalian cells. Protein Sci 15 (8): 2008-2013. doi: 10.1110/ps.062244206

52. Klomsiri C, Karplus PA and Poole LB. (2011) Cys-based redox switches in enzymes. Antioxid Redox Signal 14 (6): 1065-1077. doi: 10.1089/ars.2010.3376

53. Nagahara N. (2011) Catalytic Site Cyss of Thiol Enzyme: Sulfurtransferases. J Amino Acids 2011: 709404. doi:10.4061/2011/709404

54. Giles NM, Watts AB, Giles GI, Fry FH, Littlechild JA, Jacob C. (2003) Metal and redox modulation of Cys protein function. Chem Biol 10 (8): 677-693. doi: 10.1016/S1074-5521(03)00174-1

55. Windle CL, Simmons KJ, Ault JR, Trinh CH, Nelson A, Pearson AR and Berry A. (2017) Extending enzyme molecular recognition with an expanded amino acid alphabet. Proc Natl Acad Sci U S A 114 (10): 2610-2615. doi: 10.1073/pnas.1616816114

56. Bernsteine HD, Poritz MA, Strub K, Hoben PJ, Brenner S and Walter P. (1989) Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature 340: 482-486. doi: 10.1038/340482a0

57. O'Neil KT, Erickson-Viitanen S, DeGrado WF. (1989) Photolabeling of calmodulin with basic, amphiphilic alpha helical peptides containing p-benzoylphenylalanine. J Biol Chem 264 (24): 14571-14578.

58. Gellman SH. (1991) On the role of Met residues in the sequence-independent recognition of nonpolar protein surfaces. Biochemistryy 30 (27): 6633-6636. doi: 10.1021/bi00241a001

59. Barbarella G, Bongini A, Chatgilialoglu C, Rossini S and Tugnoli V. (1986) Sulfur-33 NMR of cyclic sulfides, sulfoxides and sulfones. J Org Chem 52 (17): 3857-3860. doi: 10.1021/jo00226a025

60. Hoyle J, Grossert JS, Hooper DL and Sotheeswaran S. (1986) Sulphur-33 nuclear magnetic resonance spectroscopy of sulphones. Can J Chem 64: 1581-1685. doi: 10.1139/v86-261

61. Hobo F, Takahashi M, Saito Y, Sato N, Takao T, Koshiba S and Maeda H. (2010) ^{33}S nuclear magnetic resonance spectroscopy of biological samples obtained with a laboratory model 33S cryogenic probe. Rev Sci Instrum 81 (5): 0504302. doi: 10.1063/1.3424853.

62. Barbarella G. (1993) Sulfur-33 NMR. Prog Nucl Magn Reson Spectrosc 25 (4): 317-343. doi: 10.1016/0079-6565(93)80001-A

63. Berger S, Braun S and Kaliowsky HO. (1997) NMR Spectroscopy of the Nonmetallic Elements. Wiley, Chichester.

64. Musio R and Sciacovelli O. (2001) Detection of Taurine in Biological Tissues by ^{33}S NMR Spectroscopy. J Magn Res 153 (2): 259-261. doi: 10.1006/jmre.2001.2437

65. Schaefer SA, Dong M, Rubenstein RP, Wilkie WA, Bahnson BJ, Thorpe C and Rozovsky S. (2013) ^{77}Se Enrichment of Proteins Expands the Biological NMR Toolbox. J Mol Biol 425: 222-231. doi: 10.1016/j.jmb.2012.11.011

66. Zhang M and Vogel HHJ. (1994) Two-dimensional NMR studies of selenomethionyl calmodulin. J Mon Biol 239 (4): 545-554. doi: 10.1006/jmbi.1994.1393

67. Go YM, Chandler JD and Jones DP. (2016) The cysteine proteome. Free Radic Biol Med 84: 227-245. doi: 10.1016/j.freeradbiomed.2015.03.022

68. Dyer KF. (2010) The Quiet Revolution: A New Synthesis of Biological Knowledge. J Biol Edu 5 (1): 15-24. doi: 10.1080/00219266.1971.9653663

69. Gettins P and Wardlaw SA. (1991) NMR relaxation properties of ^{77}Se-labeled proteins. J Biol Chem 266 (6): 3422-3426. PMID: 1995605

70. McFarlane HCE and McFarlane W. (1983) Sulfur, Selenium, and Tellurium. In: Mason J. (eds) Multinuclear NMR. Springer, Boston, MA. doi: 10.1007/978-1-4613-1783-8_15

71. Luthra NP, Costello RC, Odom JD and Dunlap RB. (1982) Demonstrating the feasbility of observing nuclear magnetic resonance signals of 77Se covalently attached to proteins. J Biol Chem 257 (3): 1142-1144. PMID: 6276374

72. Mullen GP, Dunlap RB, Odom JD. (1986) Synthesis and characterization of a selenium-containing substrate of alpha-chymotrypsin. Selenium-77 nuclear magnetic resonance observation of an acyl-alpha-chymotrypsin intermediate. Biochemistry 25 (19): 5625-5632. PMID: 3778877

73. Boles JO, Tolleson WH, Schmidt JC, Dunlap RB and Odom JD. (1992) Selenomethionyl dihydrofolate reductase from Escherichia coli. Comparative biochemistry and ^{77}Se nuclear magnetic resonance spectroscopy. J Biol Chem 267 (31): 22217-22223. PMID: 1429574

74. Gajda J, Pacholczyk J, Bujacz A, Bartoszak-Adamska E, Bujacz G, Ciesielski W and Potrzebowski MJ. (1999) Structure and dynamics of L-selenomethionine in the solid state. J Phys Chem B 110 (51): 25692-25701. doi: 10.1021/jp063332k

75. Struppe J, Zhang Y, Rozovsky S. (2015) (77)Se chemical shift tensor of L-selenocystine: experimental NMR measurements and quantum chemical investigations of structural effects. J Phys Chem B 119 (9): 3643-3650. doi: 10.1021/jp510857s

76. Liu J, Chen Q and Rozovsky S. (2017) Utilizing Selenocysteine for Expressed Protein Ligation and Bioconjugations. J Am Chem Soc 139 (9): 3430-3437. doi: 10.1021/jacs.6b10991

77. Gellman SH. (1991) On the role of methionine residues in the sequence-independent recognition of nonpolar protein surfaces. Biochemistry 30 (27): 6633-6636. PMID: 2065050

78. Bernstein HD, Poritz MA, Strub K, Hoben PJ, Brenner S, Walter P. (1989) Model for signal sequence recognition from amino-acid sequence of 54K subunit of signal recognition particle. Nature 340 (6233): 482-486. doi: 10.1038/340482a0

80. Reid KSC, Lindley PF and Thornton JM. (1985) Sulfur-aromatic interactions in proteins. FEBS 190 (2): 209-213. doi: 10.1016/0014-5793(85)81285-0

81. Morgan RS, Tatsch CE, Gushard RH, McAdon JM and Warme PK. (1978) Chains of alternating sulfur and π-bonded atoms in eight small proteins. Chem Biol Drug Des 11 (3): 209-217. doi: 10.1111/j.1399-3011.1978.tb02841.x

82. Morgan RS and McAdon JM. (1980) Predictor for sulfur-aromatic interactions in globular proteins. Chem Biol Drug Des 15 (2): 177-180. doi: 10.1111/j.1399-3011.1980.tb02566.x

83. Meyer EA, Castellano RK and Diederich F. (2003) Interactions with aromatic rings in chemical and biological recognition. Angew Chem Int Ed Engl 42 (11):1210-50. doi: 10.1002/anie.200390319

84. Forbes CR, Sinha SK, Ganguly HK, Bai S, Yap GP, Patel S and Zondlo NJ. (2017) Insights into Thiol-Aromatic Interactions: A Stereoelectronic Basis for S-H/π Interactions. J Am Chem Soc 139 (5): 1842-1855. doi: 10.1021/jacs.6b08415

85. Ringer AL, Senenko A and Sherrill CD. (2007) Models of S/π interactions in protein structures: Comparison of the H2S–benzene complex with PDB data. Protein Sci 16 (10): 2216-2223. doi: 10.1110/ps.073002307

86. Biswal HS and Wategaonkar S. (2009) Nature of the N-H...S hydrogen bond. J Phys Chem A 113 (46): 12763-12773. doi: 10.1021/jp907658w

87. Lebl M, Sugg EE and Hruby VJ. (1987) Proton NMR spectroscopic evidence for sulfur-aromatic interactions in peptides. Int J Pept Protein Res 29 (1): 40-45. PMID: 3570653

88. Waters ML. (2004) Aromatic interactions in peptides: impact on structure and function. Biopolymers 76 (5): 435-445. doi: 10.1002/bip.20144

89. Brosnan JT1, Brosnan ME. (2006) The sulfur-containing amino acids: an overview. J Nutr 136 (6 Suppl): 1636S-1640S. doi: 10.1093/jn/136.6.1636S.

90. Imai YN, Inoue Y, Yamamoto Y. (2007) Propensities of polar and aromatic amino acids in noncanonical interactions: nonbonded contacts analysis of protein-ligand complexes in crystal structures. J Med Chem 50 (6): 1189-1196. doi: 10.1021/jm061038a

91. Johnson RM1, Hecht K, Deber CM. (2007) Aromatic and cation-pi interactions enhance helix-helix association in a membrane environment. Biochemistry 46 (32): 9208-9214. doi: 10.1021/bi7008773

92. Valley CC, Cembran A, Perlmutter JD, Lewis AK, Labello NP, Gao J and Sachs JN. (2012) The methionine-aromatic motif plays a unique role in stabilizing protein structure. J Biol Chem 287 (42): 34979-34991. doi: 10.1074/jbc.M112.374504

93. Iwaoka M and Isozumi N. (2012) Hypervalent nonbonded interactions of a divalent sulfur atom: implications in protein architecture and the functions. Molecules 17 (6): 7266-7283.
doi: 10.3390/molecules17067266

94. Daeffler KN, Lester HA and Dougherty DA. (2012) Functionally important aromatic-aromatic and sulfur-π interactions in the D2 dopamine receptor. J Am Chem Soc 134 (36): 14890-14896. doi: 10.1021/ja304560x

95. Cordomí A, Gómez-Tamayo JC, Gigoux V and Fourmy D. (2013) Sulfur-containing amino acids in 7TMRs: molecular gears for pharmacology and function. Trends Pharmacol Sci 34 (6): 320-331. doi: 10.1016/j.tips.2013.03.008

96. Aledo JC, Cantón FR and Veredas FJ. (2015) Sulphur Atoms from Methionines Interacting with Aromatic Residues Are Less Prone to Oxidation. Sci Rep 5: 16955. doi: 10.1038/srep16955

97. Zhou AQ, O'Hern CS and Regan L. (2014) Predicting the side-chain dihedral angle distributions of nonpolar, aromatic, and polar amino acids using hard sphere models. Proteins 82 (10): 2574-2584. doi: 10.1002/prot.24621

98. Virrueta A, O'Hern CS and Regan L. (2016) Understanding the physical basis for the side-chain conformational preferences of methionine. Proteins 84 (7): 900-911.
doi: 10.1002/prot.25026

99. Gómez-Tamayo JC, Cordomí A, Olivella M, Mayol E, Fourmy D, Pardo L. (2016) Analysis of the interactions of sulfur-containing amino acids in membrane proteins. Protein Sci 25 (8): 1517-1524. doi: 10.1002/pro.2955

100. Hützler WM and Bolte M. (2018) Sulfur as hydrogen-bond acceptor in cocrystals of 2-thio-modified thymine. Acta Crystallogr C Struct Chem 74(Pt 1): 21-30. doi: 10.1107/S2053229617017181

101. Motherwell WB, Moreno RB, Pavlakos I, Arendorf JRT, Arif T, Tizzard GJ, Coles SJ and Aliev AE. (2018) Noncovalent Interactions of π Systems with Sulfur: The Atomic Chameleon of Molecular Recognition. Angew Chem Int Ed Engl 57(5):1193-1198. doi: 10.1002/anie.201708485

102. Zetterberg FR, Peterson K, Johnsson RE, Brimert T, Håkansson M, Logan DT, Leffler H and Nilsson UJ. (2018) Monosaccharide Derivatives with Low-Nanomolar Lectin Affinity and High Selectivity Based on Combined Fluorine-Amide, Phenyl-Arginine, Sulfur-π, and Halogen Bond Interactions. ChemMedChem 13 (2): 133-137. doi: 10.1002/cmdc.201700744

103. Shulga DA, Titov OI, Pisarev SA and Palyulin VA. (2018) Multipole models of sulphur for accurate anisotropic electrostatic interactions within force fields. SAR QSAR Environ Res 29 (1): 21-42. doi: 10.1080/1062936X.2017.1406403

Chapter 2

SELENOCYSTEINE MEDIATED EXPRESSED PROTEIN LIGATION OF SELENOPROTEIN M

2.1 Background

A large percentage of the selenoproteome is composed of selenoproteins whose catalytic Sec resides in a thioredoxin (Trx) fold. The fundamental Trx fold is composed of a four stranded β-sheets, packed on one side flanked by two α-helices, although variations to this basic unit are common. [1] This fold encompasses several superfamilies of enzymes with functions from detoxification, to the formation and reduction of disulfide bonds in proteins and hydroperoxides. [2] In the majority of enzymes possessing the Trx fold, the redox-active site is typically positioned right before an α helix, capping it (Figure 1). In contrast, in a subset of selenoproteins, the redox active motif Cys-X-X-Sec (where X stands for any amino acid and Sec stands for Sec) is positioned immediately after the first β-sheet. [3] A conformational switch following the reduction of the intramolecular selenylsulfide bond was proposed to

trigger a signaling event or binding of protein partners. [4] In humans, SELENOH, SELENOM, SELENOT, SELENOV, SELENOW, and SELENOF exhibit this common Trx fold and placement of the selenoredox motif. [5,6] This chapter details the preparation of a representative member of this group, SELENOM, in a manner that can be generalized to other selenoproteins with a minimal Trx fold. [7]

Like many selenoproteins, SELENOM (UniProt Q8WWX9) is tied to the management of cellular oxidative stress. [8-14] SELENOM resides in the endoplasmic reticulum [15] and its overexpression governs the expression of antioxidant enzymes like glutathione peroxidase and superoxide dismutase. [10] SELENOM knock-out mice showed increases in weight gain and changes in metabolism. [16] It was proposed to regulate energy metabolism through participation in leptin signaling. [17] The NMR-derived structure of *Mus musculus* SELENOM U48C mutant was reported (Figure 1a) but the redox properties and precise function of the wild type SELENOM with its CGGU motif remain undetermined. [4]

This chapter presents a versatile chemical approach based on native chemical ligation (NCL) to prepare the wild-type selenium-containing SELENOM. [18] In Sec-driven NCL, a Sec residue mediates the formation of the peptide bond from two respective protein fragments to form the target protein. [19-22] As shown in Figure 2, an N-terminal Sec from one fragment initiates the nucleophilic attack of the C-terminal thioester from the complementary fragment and subsequently undergoes a Se→N acyl shift to generate the native amide bond (Figure 2). The ligation rate of Sec-mediated NCL is faster and more pH tolerant than Cys-mediated ligation, and the resulting

ligation efficiency is often higher than that of Cys-mediated reactions. [19,23] Sec-mediated NCL is an attractive method for preparing selenoproteins because it is possible to achieve high yield, and in addition it is compatible with introducing posttranslational modifications or chemical tags into the protein. However, since the preparation of selenium-containing protein fragments has traditionally relied on chemical synthesis, this method was restricted by limitations on the size of the Sec-containing peptide fragment and its solubility. [24-27,21] The production of SELENOM discussed here is distinguished from previous approaches in that all fragments are prepared by heterologous expression in *E. coli* and rely only on standard protein expression and purification. [7,28] This reduces cost and simplifies preparation by alleviating the need to synthesize the seleno-containing part of the protein.

To generate the Sec-containing fragment of SELENOM (SELENOM^{CT} (residues 48-145)), a cytosolic maltose binding protein (MBP) is coupled to its N-terminus for enhanced expression and solubility, generating MBP-SELENOM^{CT}. A Tobacco Etch Virus (TEV) protease cleavage site is introduced between the MBP and SELENOM^{CT} so that SELENOM^{CT} can be efficiently released since TEV protease has good tolerance at its P1' position in the recognition site ENLYFQ/P1' (where P1' stands for any canonical amino acid except P). [29] The Sec residue is incorporated into the fragment by mutating the Sec to Cys and subsequently expressing the protein in *E. coli* grown on defined growth medium supplemented with L-selenocystine. [30] The cysteinyl-tRNA is misloaded with Sec, which becomes incorporated instead of Cys. The thioester-containing fragment of SELENOM is produced by intein technology. The N-

terminal fragment of SELENOM (residues 25-47) is fused to the *Saccharomyces cerevisiae* Vacuolar Membrane ATPase (Sce VMA1; 454 residues; 56 kDa) intein. [31] SELENOM is then prepared by ligation of the $SELENOM^{CT}$ to $SELENOM^{NT}$ under native conditions (Figure 2). Following ligation, SELENOM is purified to homogeneity and characterized by SDS-PAGE, mass spectrometry, and circular dichroism spectroscopy.

2.2 Materials and Methods

2.2.1 Gene, Plasmids and Molecular Cloning

The expression vector pMAL-c5X, designed to produce fusion proteins with MBP, is from New England Biolabs (NEB). To aid purification, a general cloning vector was prepared by introducing a hexahistidine tag to the N-terminus of MBP. [32] In addition, a TEV protease site was inserted at the C-terminus of MBP to facilitate efficient release of the target protein or protein fragment. The VMA intein was introduced following MBP and the TEV cleavage site allowing cloning of the protein of interest fused to MBP, the VMA intein or both. This pMAL-c5X-VMA cloning vector is made available through Addgene (plasmid #86590). [7]

The gene of *Homo sapiens* SELENOM (accession number: NP_536355.1) optimized for *E. coli* expression can be obtained from Addgene (plasmid #86579). The

gene was inserted into a pMAL vector to allow expression of SELENOM U48C fused to hexahistidine-tagged MBP creating the plasmid pMAL-c5X-SELENOM^{U48C}. A TEV protease cleavage site exists between MBP and SELENOM U48C so that SELENOM U48C can be efficiently released from MBP. [7]

Molecular cloning of the SELENOM N and C termini were done using the SELENOM U48C as the starting point. The MBP-SELENOM^{NT} gene, which includes MBP, TEV cleavage site and SELENOM residues 25-47, was cloned into vector pMAL-c5X-VMA using restriction sites NdeI and SapI to create the vector pMAL-c5X-*SELENOM^{NT}* for the expression of SELENOM^{NT} fragment (Figure 2). A TEV protease cleavage site assists the efficient release of SELENOM^{NT} fragment from the fusion construct.

The expression vector for SELENOM^{CT} is prepared by deleting residues 25-47 from pMAL-c5X-*SELENOM^{U48C}* to generate pMAL-c5X-*SELENOM^{CT}* (SELENOM residues 48-145, beginning with U48 mutated to C48). Restriction enzymes SapI and NdeI are from NEB as well as the Q5™ site-directed mutagenesis kit which was used in all mutagenesis of this work.

2.2.2 Protein Expression and Purification

Expression and purification of SELENOM U48C For protein expression of SELENOM U48C mutant, the plasmids were transformed into *E. coli* strain

BL21(DE3). Cells were grown in low salt LB (10 g tryptone, 5 g yeast extract, 5 g NaCl), supplemented with 0.2% glucose at 37 °C, with good aeration and relevant antibiotic selection (100 μg/mL ampicillin). When the optical density (OD) at 600 nm reached 0.5, the temperature was lowered to 18 °C, and the cells were allowed to shake at the lower temperature for an additional 1 h. Protein expression was induced with 0.5 mM isopropyl 1-thio-β-D-galactopyranoside (IPTG). The cells were harvested pH 7.5) and 200 mM NaCl, pH 7.5 (amylose buffer), supplemented with 0.5 mM benzamidine, 1 mM phenylmethanesulfonyl fluoride (PMSF), and 2 mM ethylenediaminetetraacetic acid (EDTA). Cells were lysed using a high-pressure homogenizer (Emulsi-Flex-C5, Avestin) on ice, and all subsequent procedures were conducted at 4 °C. Cell debris was removed by centrifugation at 20000 g for 1 h. The supernatant was loaded onto an amylose column, and the column was washed with the amylose buffer. The SELENOM U48C fusion protein was eluted using amylose buffer containing 20 mM maltose. Cleavage of the fusion partner cMBP was conducted by incubating a 1:20 molar ratio of TEV protease to MBP-SELENOM U48C at 4 °C overnight. Following cleavage, the protein was dialyzed to remove EDTA and loaded onto a 5 mL histidine affinity column (Histrap FF, GE Healthcare) to remove MBP and TEV protease. The flowthrough was pooled, concentrated and loaded onto a size exclusion column (HiPrep 16/60 Sephacryl S-100 HR column, GE Healthcare), as a polishing step to further purify SELENOM U48C to homogeneity. Protein purity, as determined with 16% Tris-glycine sodium dodecyl sulfate−polyacrylamide gel electrophoresis (SDS−PAGE) gel, was greater than 95%.

Expression and purification of MBP-SELENOM^{NT} thioester For protein expression of the MBP-SELENOM^{NT}-VMA, the plasmid was transformed into *E. coli* strain BL21(DE3). Cells were grown in low salt LB, supplemented with 0.2% glucose at 37 °C, with good aeration and the relevant antibiotic selection (100 μg/mL ampicillin). When the optical density (OD) at 600 nm reached 0.5, the temperature was lowered to 18 °C, and the cells were allowed to shake at the lower temperature for an additional half hour. Protein expression was induced with 0.5 mM IPTG. The cells were harvested after 18−20 h. Cell paste (13 g/L) was resuspended in amylose buffer supplemented with 0.5 mM benzamidine, 1 mM PMSF, and 2 mM EDTA. Cells were lysed using a high-pressure homogenizer on ice, and all subsequent procedures were conducted at 4 °C. Cell debris was removed by centrifugation at 20000 g for 1 h. The supernatant was loaded on an amylose column, and the column was washed first with 15 column volumes (CV) of the amylose buffer followed by 2 CV with 25 mM 2-(N-morpholino)ethanesulfonic acid (MES), 200 mM NaCl, 2 mM EDTA, pH 6.5 (wash buffer). The MBP-SELENOM^{NT}-VMA fusion protein was eluted using wash buffer containing 20 mM maltose. The MBP-SELENOM^{NT} thioester was generated either by on column or in solution cleavage by 2-mercaptoethanesulfonic acid (MESNA). For on column cleavage, the elution from the amylose column was loaded onto a 40 mL chitin column, then incubated with 2 CV 25 mM MES, 200 mM NaCl, 2 mM EDTA, 75 mM MESNA, pH 6.5 (thioester cleavage buffer) after the column was drained completely. For off column cleavage, 75 mM MESNA was added to the above MBP-SELENOM^{NT}

amylose elution, and left to react at room temperature (RT) for 12 h. The cleavage mixture was then loaded onto a 40 mL chitin column, and MBP-SELENOM^{NT} was collected in the flow through when the column was washed with 2 CV of wash buffer 2. Thioester cleavage buffer as well as wash buffer should have pH at or below 6.5 as higher pH accelerates thioester hydrolysis. The thioester cleavage process was repeated twice to obtain maximum yield. The formation of MBP-SELENOM^{NT} thioester was confirmed by mass spectrometry (Figure 3).

Expression and purification of MBP-SELENOM^{CT} For protein expression of MBP-SELENOM^{CT}, the expression vector was freshly transformed into *E. coli* strain BL21(DE3). The procedure for Sec incorporation was modified from previous protocols. (5,6) In brief, 1 mL of an overnight culture in LB broth was spun at 2500 g for 5 min, then resuspended in modified MDAG media from *Studier et al.* as listed in Table 2 step 1. [7] Cells were grown at 37 °C with good aeration. When the OD at 600 nm reached 1.5 (about 10-14 h), the growth medium was supplemented as detailed in Table 2 step 2, and continuously shaken for an additional 10 min. The temperature was then lowered to 18 °C, and the cells were allowed to recover for 10 min. Protein expression was induced with 0.5 mM IPTG, and the cells were grown for 20-24 h at 18 °C. Cells were harvested by centrifugation and subsequent purification by amylose affinity chromatography was the same as described for SELENOM U48C. When adding L-Sec at Step 2, our experience with various suppliers of L-Sec suggest that the compound is best purchased through Sigma-Aldrich or Acros Organics. Batches acquired through other vendors may contain contaminants that impede cell growth and

reduce Sec incorporation. Purification of MBP-SELENOM^{CT} is carried out under non-reducing conditions, and because the Sec-containing MBP-SELENOM^{CT} forms an intermolecular diselenide bond faster than the formation of disulfide bond by the Cys-containing variant, extended wash can help to enrich the fraction of Sec-containing MBP-SELENOM^{CT} due to tighter binding of the dimeric diselenide-containing form to affinity columns over that of the monomeric Cys-containing form. This is often a helpful strategy especially for protein targets unable to achieve high Sec incorporation. During handling of the Sec-containing terminus, whenever possible avoid the use of TCEP and tris(hydroxypropyl)phosphine (THP) which can promote selenium elimination from Sec. The yield of MBP-SELENOM^{CT} using this method ranged from 60 to 80 mg per L. The ratio of Sec incorporation into MBP-SELENOM^{CT} was evaluated by mass spectrometry after its reduction by 100 molar excess DTT at 37 °C for 20 min. MBP-SELENOM^{CT} was cleaved by TEV protease at a 1:5 molar ratio of TEV protease to MBP-SELENOM^{CT} at 25 °C for 16 h in amylose elution buffer supplemented with 10 mM DTT. The cleavage was monitored by 16% Tris-Glycine SDS-PAGE (Figure 4).

2.2.3 Expressed Protein Ligation of SELENOM

The purified MBP-SELENOM^{NT} thioester in 25 mM MES, 200 mM NaCl, 2 mM EDTA, pH 6.5, and the MBP-SELENOM^{CT} with TEV protease in 50 mM sodium phosphate, 200 mM NaCl, 2 mM EDTA, pH 7.5, were mixed together in the molar ratio of SELENOM^{CT} to SELENOM^{NT} thioester of 1:5. The ligation reaction was initiated

by adding to this mixture 100 mM MESNA, 50 mM MPAA (4-Mercaptophenylacetic acid), and 5 mM DTT. The pH of the ligation mixture should be about 7 without additional adjustments. Reaction progress was monitored by the disappearance of $SELENOM^{CT}$ and the appearance of the full length SELENOM at 14.1 kDa using 16% Tris-Glycine SDS-PAGE or Tricine-SDS-PAGE. Ligation typically reaches 90% completion within 1 day at 25 °C and pH 7.0 (Figure 5). However, the optimal pH as seen in Figure 6 was at 6.5.

Upon completion of the reaction, the MES-phosphate buffer mixture of the reaction was exchanged into the IMAC binding buffer of 50 mM sodium phosphate, 200 mM NaCl, pH 7.5, either by desalting column such as HiPrep 26/10 or by dialysis. This buffer exchange was necessary in order to remove EDTA and DTT which are incompatible with IMAC resin at their current concentrations of 2 mM and 10 mM, respectively. The desalted protein mixture was loaded onto another IMAC column and washed with 5 CV IMAC binding buffer. The tag-free full length SELENOM could be collected from the column flow through.

Fractions containing SELENOM was pooled and concentrated to 5 mg/ml using an Amicon Ultra-15 centrifugal filter unit with MWCO of 3 kDa. The concentrated protein was loaded onto an appropriate size exclusion column suitable for small proteins (<20 kDa) such as Superdex™10/300 GL or Sephacryl™ S-100, pre-equilibrated with gel filtration buffer such as 50 mM sodium phosphate, 200 mM NaCl, 2 mM EDTA, pH 7.5. Fractions containing SELENOM was collected and visualized on 16% Tris-Glycine SDS-PAGE. Select fractions with purity > 90% were pooled and SELENOM

protein concentration was determined by UV-Vis spectroscopy with an extinction coefficient of 18450 M^{-1} cm^{-1}.

2.2.4 Removal of Co-purified Cellular Thiols

Co-purified cellular thiols may present an obstacle to downstream applications. An estimated 0.1 – 1 mM of cellular free thiols is usually present in the purified SELENOM protein sample after size exclusion chromatography. It is recommended that this free thiol population be removed to avoid downstream interference with protein activity. Purified SELENOM was reduced with 10 mM DTT for 2 h at 25 °C or overnight at 4 °C. The protein was then extensively dialyzed against 50 mM sodium phosphate, 200 mM NaCl, 2 mM EDTA, pH 7.5, until the free thiol count of the protein is undetectable using 5,5'-dithiobis-(2-nitrobenzoic acid) (DTNB) assay. [35]

2.2.5 Intact-Mass Spectrometry Characterization

Generally speaking, incorporation of selenium is best carried out by Inductively Coupled Plasma (ICP) - atomic emission spectrometry or ICP - mass spectrometry. However, in simple cases such as proteins with a single amino acid substitution (such as Cys vs. Sec), intact mass spectrometry provides a good estimation. During acquisition, avoid oversaturating the detector which causes an inaccurate reflection in the respective ion counts of the Cys and Sec-containing species, which in turn leads to

errors in deconvolution and misrepresented peak intensities of identified masses. Sec incorporation in SELENOM typically exceeds 90% (Figure 7b).

2.2.6 Tandem mass sequencing

The position of the Sec residue is confirmed by sequencing the Sec-containing peptide using a Q Exactive Orbitrap mass spectrometer interfaced with an Ultimate 3000 LC system (or equivalent set up). In brief, 6 µg of SELENOM in 60 µL of trypsin digest buffer was reduced by the addition of 5 mM DTT. The protein was incubated for 5 min at 95 °C to achieve full reduction and denaturation. When sample returned to ambient temperature, 12 mM IAM was added immediately and incubated for an additional 20 min to fully alkylate exposed Sec and Cys residues, as well as the remaining DTT.

Total of 0.3 µg of the trypsin enzyme was added into the above mixture and digested at 37 °C for 16 h. A ZipTip™ with 15 µm, 200 Å pore size was employed to desalt the trypsin digested protein sample using the vendor procedure. 1 µg of trypsin digested SELENOM was loaded onto a reverse-phase LC column such as C18 suitable for small peptide analysis. The LC performed an acetonitrile gradient supplemented with 0.1% formic acid over a run time of at least 45 minutes to ensure complete elution of all digested fragments. Mass spectrometry data was acquired using a stepped normalized collision energy of 28, 30, 35 eV. A representative sequencing of SELENOM Sec-containing peptide is shown in Figure 7c.

2.2.7 Circular dichroism spectroscopy of SELENOM

SELENOM was exchanged into the CD buffer and interfering chloride ions are removed during this buffer exchange. The far-UV spectra of the sample was recorded using a 1 mm path-length cell from 190 nm to 250 nm at 20 °C. Three accumulation scans were taken for baseline, and an average of eight accumulation scans were taken for SELENOM. Mean residual ellipticity was plotted against wavelength for each sample. A representative CD spectrum for SELENOM prepared by Sec-mediated native chemical ligation is shown in Figure 7a.

2.3 Summary of Chapter 2

This chapter presented a versatile strategy through Sec-assisted expressed protein ligation to assemble the full length, native, human SELENOM. SELENOM is an ER-residing selenoenzyme with a minimal thioredoxin fold and a catalytic CXXU motif. In the proposed method, full length SELENOM U48C was split into two fragments: N terminus with residues 25-47, and the C terminus of residues 48-125 which bears the only Sec in the sequence. The N terminus is expressed as a fusion protein with VMA intein, which through thiol cleavage, becomes an activated thioester. Meanwhile, the C terminus was expressed by misloading the cysteinyl tRNA with Sec to incorporate Sec at the 48^{th} amino acid position. Due to the low pKa of the Sec residue, it exists as the

deprotonated selenolate and initiates a nucleophilic attack of the activated thioester, then through subsequent acyl-transfer, generates the native peptide backbone. The distinct advantage of this strategy is that each fragment of SELENOM can be expressed recombinantly in *E. coli* and with high yield and Sec incorporation exceeding 95%. Typical yield of SELENOM using this method is 10 mg/mL, far exceeding those obtained with currently reported methods. In addition, no expertise in peptide synthesis or in the complex challenges of tRNA engineering is required. This strategy opens the possibility of characterizing wild type SELENOM using crystallography and NMR, methods which historically have not been possible due to the lack of sufficient yield or homogeneity of the protein sample.

This work was done in collaboration with Dr. Jun Liu.

Table 2. Two-step strategy to Sec incorporation in SELENOM^{CT}

Component	Volume (mL)
Step 1	
H_2O (autoclaved)	926
1 M $MgSO_4$	2
1000 X metals*	0.2
Vitamin mix*	1
Ampicillin 100 mg/mL	1
40 % glucose (W/V)	5
50 X M*	20
Step 2	
40 % glucose (W/V)	5
25 % aspartate (W/V)	10
17 amino acids (10 mg/mL) *	20
25 mg/mL Met	4
Selenocystine*	0.1 g

* The recipes to prepare the 1000 X metals, Vitamin mix, and 50 X M (salts mixture) and 17 amino acids are as described by Studier.[7]

** Excess selenocystine can lead to decreased yield due to toxicity. In our experience, 0.1 g L-selenocystine per L growth medium resulted in incorporation of over 90% Sec into target proteins. If incorporation levels are low, then L-selenocystine concentration should be increased to 0.15 g per L growth medium.

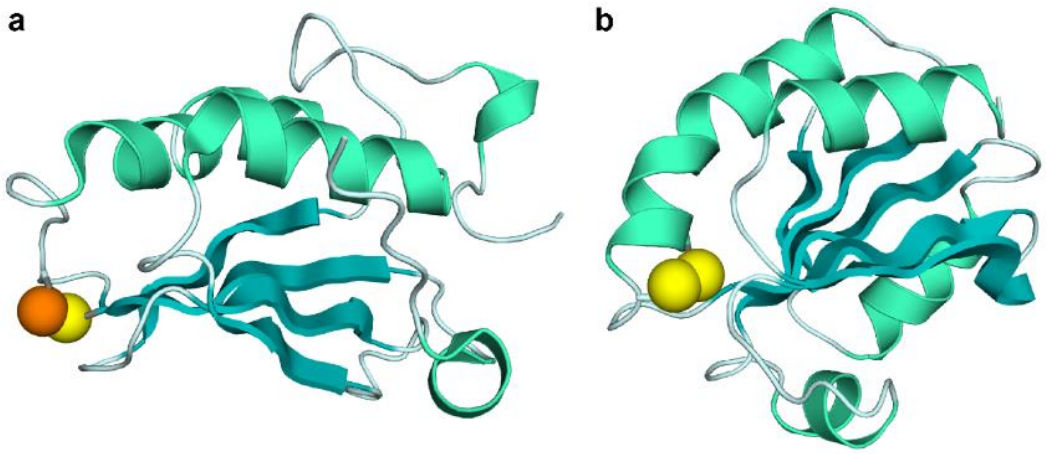


Figure 1. Selenoproteins with a minimal thioredoxin fold position their redox motif next to the first β-sheet. **(a)** The structure of *Mus musculus* SELENOM (PDB entry 2A2P) is representative of this class. **(b)** Thioredoxin and related proteins position their redox motif immediately before an α-helix. The structure of human thioredoxin (PDB entry 1ERU) exemplifies this class. The catalytic selenylsulfide bond of SELENOM and disulfide bond of Trx are depicted as spheres. Selenium is in orange and sulfur in yellow.

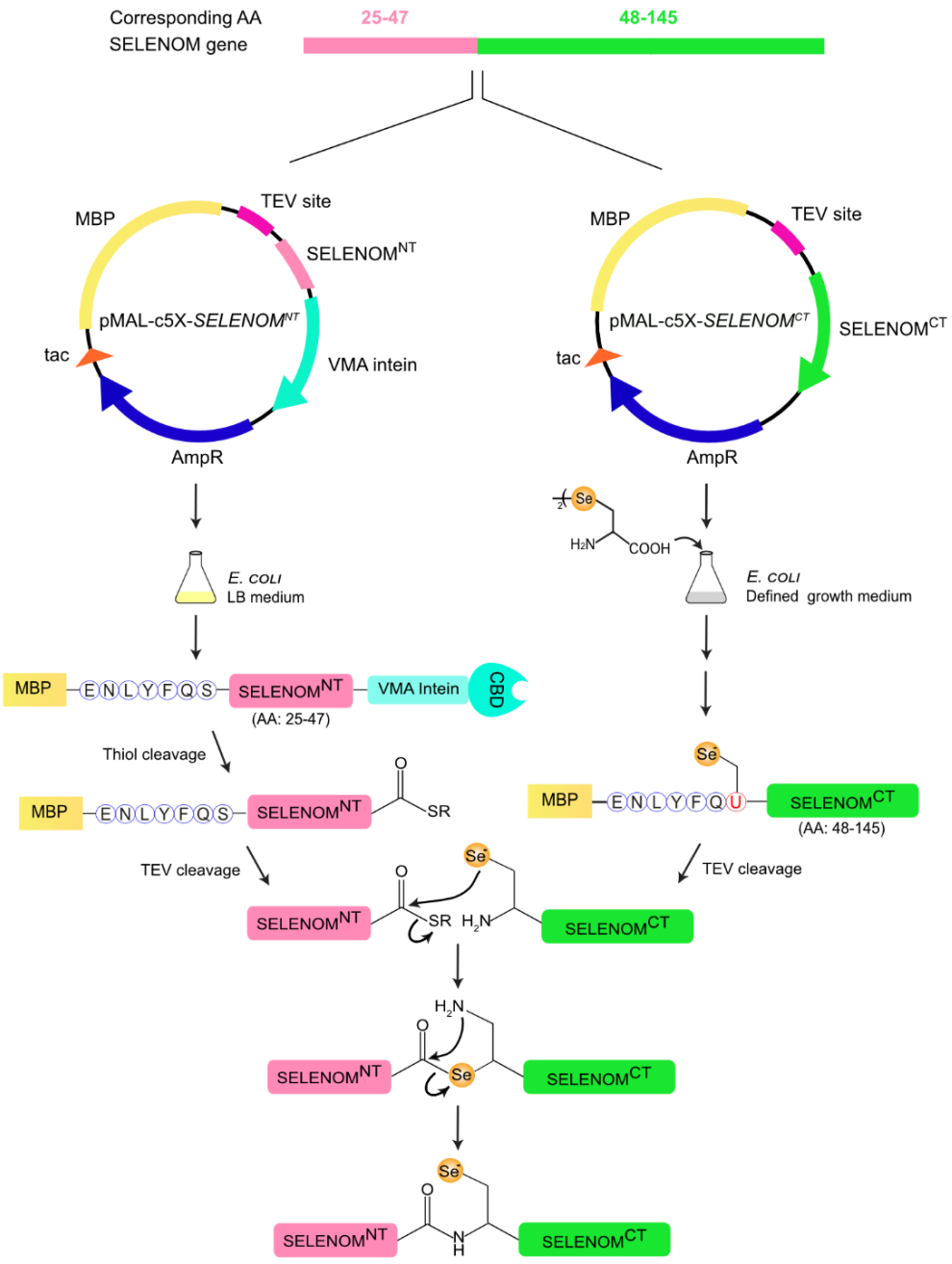

Corresponding AA
SELENOM gene
25-47
48-145
MBP
TEV site
SELENOM^NT
pMAL-c5X-SELENOM^NT
tac
VMA intein
AmpR
MBP
TEV site
pMAL-c5X-SELENOM^CT
SELENOM^CT
tac
AmpR
E. coli
LB medium
E. coli
Defined growth medium
MBP
ENLYFQS
SELENOM^NT
VMA Intein
CBD
(AA: 25-47)
Thiol cleavage
MBP
ENLYFQS
SELENOM^NT
SR
MBP
ENLYFQU
SELENOM^CT
(AA: 48-145)
TEV cleavage
TEV cleavage
SELENOM^NT
SR
SELENOM^CT
SELENOM^NT
SELENOM^CT
SELENOM^NT
SELENOM^CT

Figure 2. Scheme of Sec mediated expressed protein ligation of SELENOM. *E. coli* bearing the expression vector pMAL-c5X-MBP-SELENOM^{CT} (residues 48-145 of SELENOM) is grown in defined growth medium supplemented with L-selenocystine for the incorporation of Sec into the protein. Following TEV cleavage, the exposed selenolate of SELENOM^{CT} initiates the nucleophilic attack on SELENOM^{NT}-thioester. SELENOM^{NT}-thioester is prepared by intein technology (see Figure 3). The native peptide bond is formed through the attack of selenolate onto the thioester, followed by a Se→N shift.

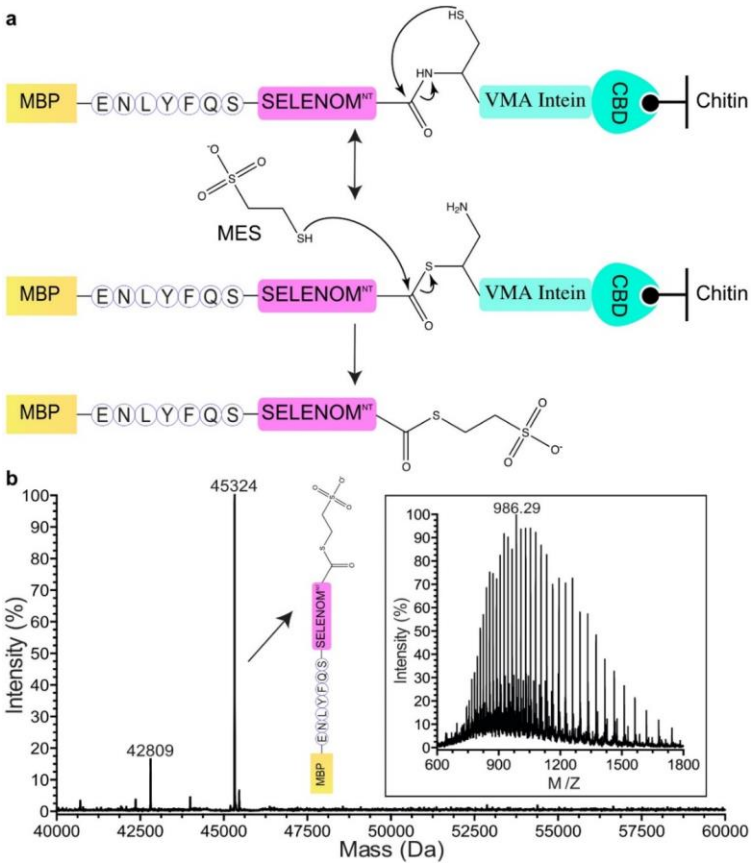


Figure 3. MBP-SELENOM^{NT} thioester characterization by mass spectrometry. **(a)** Scheme of MBP-SELENOM^{NT} thioester production. MBP-SELENOM^{NT}-VMA fusion is purified by amylose affinity chromatography and then bound to a chitin column through the chitin-binding domain fused at the C-terminal of intein. During cleavage of intein, the thioester form of MBP-SELENOM^{NT} (MBP-SELENOM^{NT}-MES) is formed and eluted from the column with the addition of 75 mM MESNA. **(b)** MBP-SELENOM^{NT} thioester detected by mass spectrometry. The theoretical mass of the MBP-SELENOM^{NT} thioester is 45322 Da, and measured as 45324 Da. The 42809 Da peak arises from a truncation. Inset shows the M/Z spectrum.

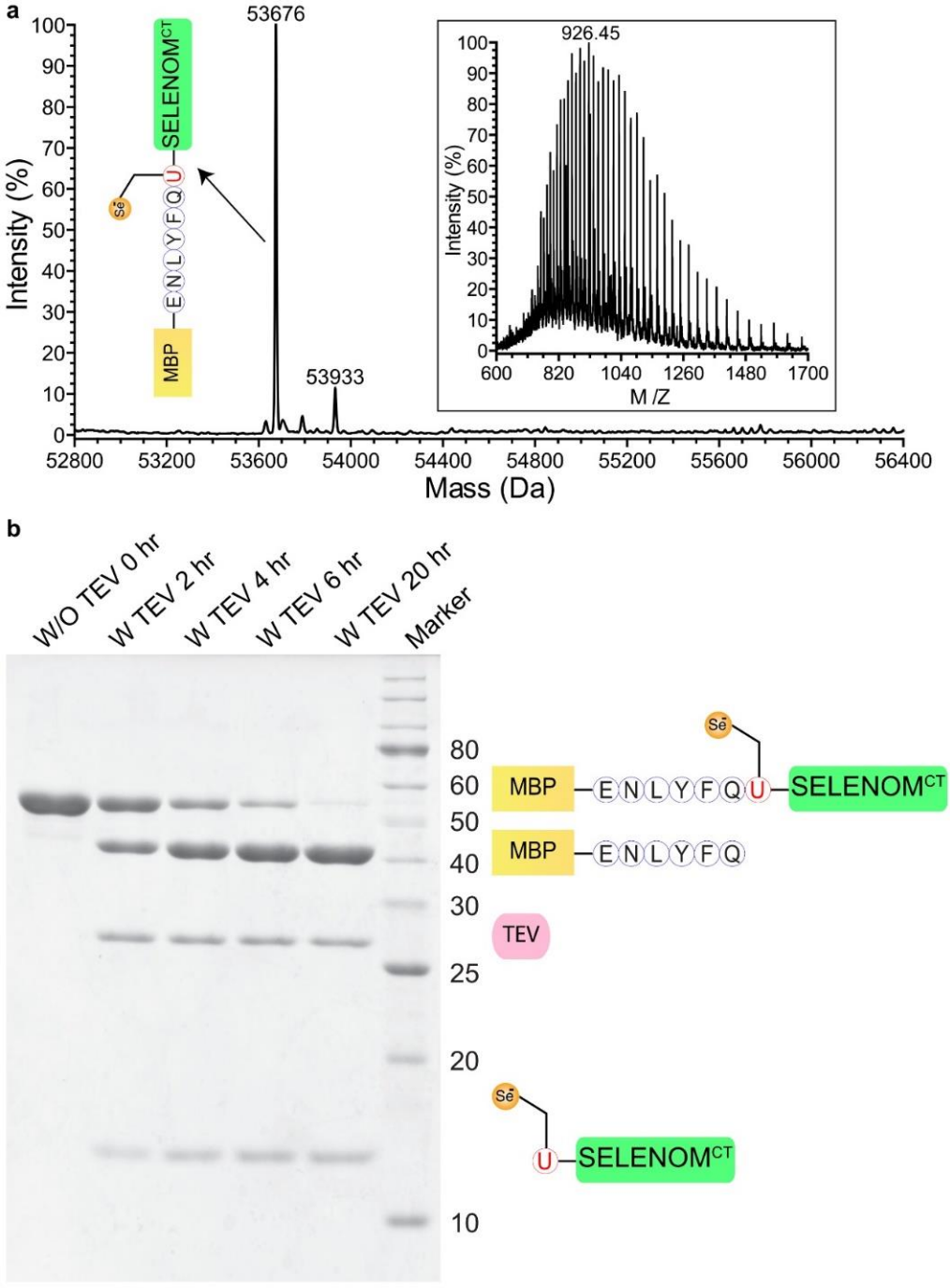

a
53676
926.45
53933
SELENOM^CT
MBP
E N L Y F Q U
Se
Intensity (%)
Mass (Da)
M /Z
b
W/O TEV 0 hr
W TEV 2 hr
W TEV 4 hr
W TEV 6 hr
W TEV 20 hr
Marker
MBP
TEV
SELENOM^CT

Figure 4. MBP-SELENOM^{CT} purification and the release of SELENOM^{CT} from MBP by TEV protease cleavage. (**a**) Deconvoluted electrospray ionization mass spectrum of reduced MBP-SELENOM^{CT}. The calculated molecular mass of MBP-SELENOM^{CT} is 53674 Da. The 53933 Da form is due to a 6-phosphogluconolactone modification, most likely on the His_6-tag at the N-terminal of MBP [36]. Inset shows the M/Z spectrum. **(b)** MBP-SELENOM^{CT} can be efficiently cleaved by TEV protease in the presence of a reductant (here DTT). Proteins are identified by 16% Tris-Glycine SDS-PAGE under reducing conditions. MBP-SELENOM^{CT} is incubated with TEV protease at a molar ratio of 5:1 in 50 mM sodium phosphate, 200 mM NaCl, 2 mM EDTA, and 10 mM DTT, pH 7.5.

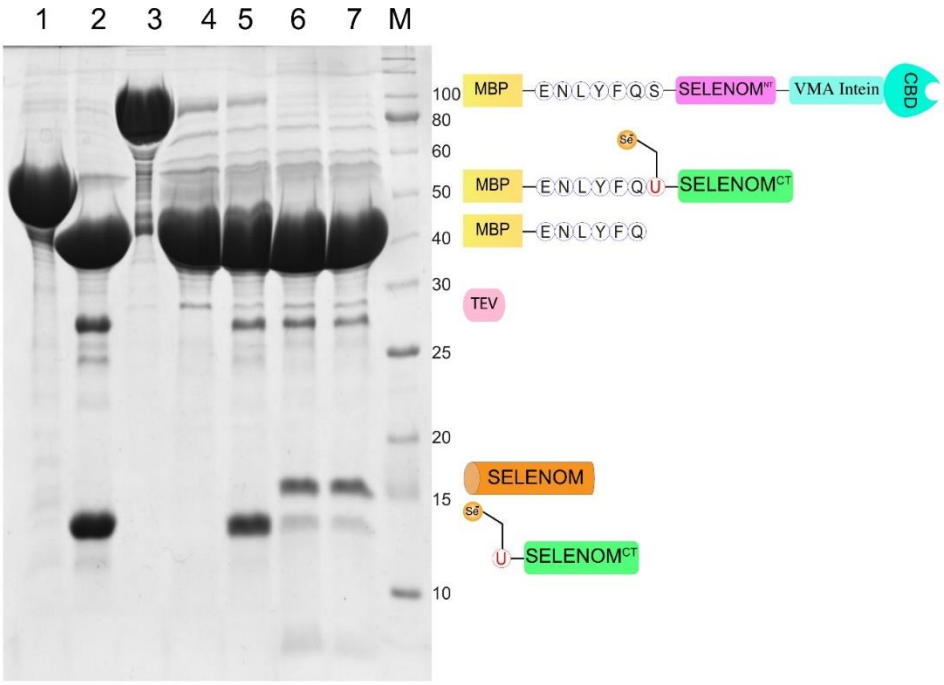


Figure 5. SELENOM ligation as monitored by SDS-PAGE: Lane 1: MBP-SELENOM^{CT}; lane 2: same as lane 1 but after TEV protease cleavage; lane 3: MBP-SELENOM^{NT}-VMA; lane 4: MBP-SELENOM^{NT} thioester; lanes 5-7: the ligation of SELENOM monitored on days 0, 1, and 2. Lane M: molecular mass standards. After TEV protease cleavage, SELENOM^{NT} thioestser cannot be detected because of its 2 kDa mass.

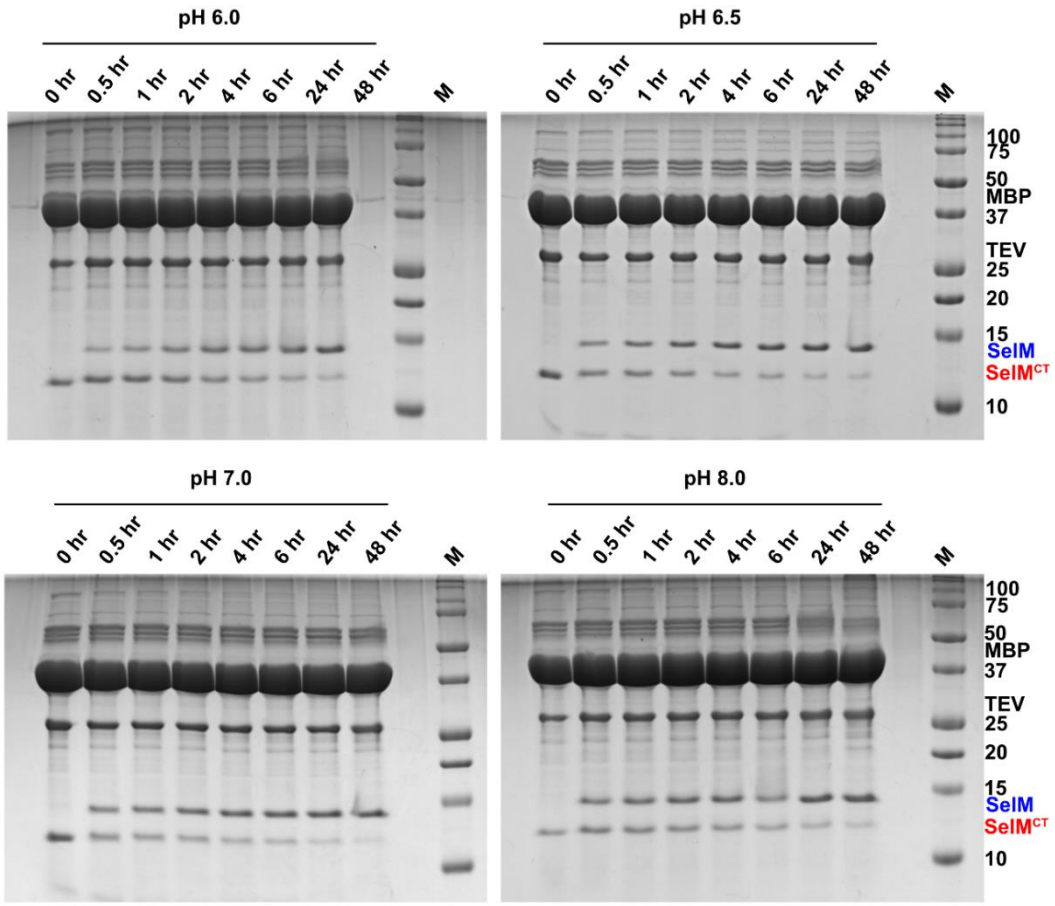


Figure 6. The rate of SELENOM ligation at 25 °C at pH 6 to 8. MBP-SELENOM^{NT} was incubated with TEV cleaved SELENOM^{CT} at a molar ratio of 5:1 in 50 mM sodium phosphate, 200 mM NaCl, and 2 mM EDTA, supplemented with 100 mM MESNA, 50 mM MPAA, and 5 mM DTT. The reaction was monitored for 48 hours. Proteins are identified on a 16% Tricine-SDS-PAGE with reducing agents. The reaction yield is higher than 90% at pH 6 and above as judged by the disappearance of SELENOM^{CT}.

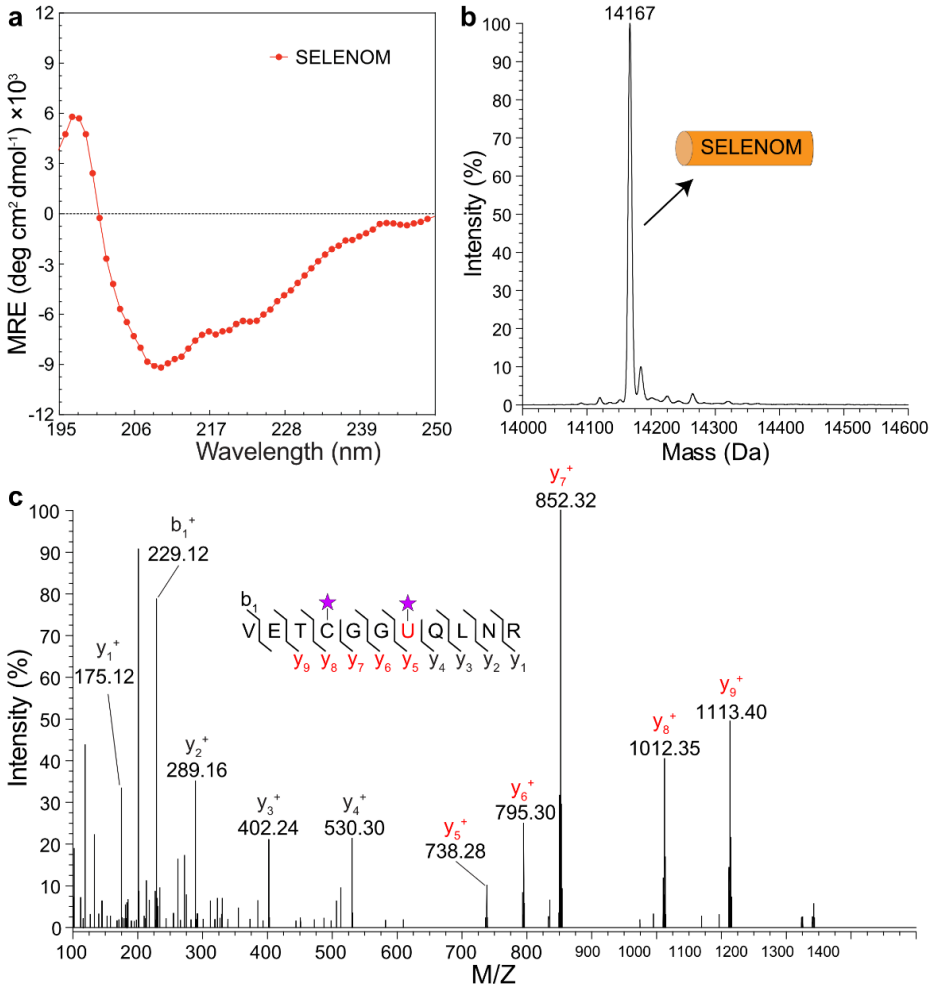


Figure 7. Characterization of SELENOM. **(a)** CD spectrum of SELENOM prepared by Sec-EPL. **(b)** Deconvoluted ESI-MS of intact SELENOM. **(c)** Tandem MS sequencing of the peptide from panel b confirms Sec's presence in SELENOM. Sec and Cys residues were alkylated with iodoacetamide (purple star). Fragment ions that contain Sec are colored red.

REFERENCES

1. Qi Y, Grishin NV (2005) Structural classification of thioredoxin-like fold proteins. Proteins: Struct Funct Bioinform 58 (2):376-388. doi:10.1002/prot.20329

2. Atkinson HJ, Babbitt PC (2009) An atlas of the thioredoxin fold class reveals the complexity of function-enabling adaptations. PLoS Comput Biol 5 (10). doi:10.1371/journal.pcbi.1000541

3. Aachmann FL, Fomenko DE, Soragni A, Gladyshev VN, Dikiy A (2007) Solution structure of selenoprotein W and NMR analysis of its interaction with 14-3-3 proteins. J Biol Chem 282 (51):37036-37044. doi:10.1074/jbc.M705410200

4. Ferguson AD, Labunskyy VM, Fomenko DE, Araç D, Chelliah Y, Amezcua CA, Rizo J, Gladyshev VN, Deisenhofer J (2006) NMR structures of the selenoproteins Sep15 and SELENOM reveal redox activity of a new thioredoxin-like family. J Biol Chem 281 (6):3536-3543. doi:10.1074/jbc.M511386200

5. Hatfield DL, Tsuji PA, Carlson BA, Gladyshev VN (2014) Selenium and Sec: Roles in cancer, health, and development. Trends Biochem Sci 39 (3):112-120. doi:10.1016/j.tibs.2013.12.007

6. Labunskyy VM, Hatfield DL, Gladyshev VN (2014) Selenoproteins: Molecular pathways and physiological roles. Physiol Rev 94 (3):739-777. doi:10.1152/physrev.00039.2013

7. Liu J, Chen Q, Rozovsky S (2017) Utilizing Sec for expressed protein ligation and bioconjugations. J Am Chem Soc. doi:10.1021/jacs.6b10991

8. Reeves MA, Bellinger FP, Berry MJ (2010) The neuroprotective functions of selenoprotein m and its role in cytosolic calcium regulation. Antioxid Redox Signal 12 (7):809-818. doi:10.1089/ars.2009.2883

9. Hwang DY, Cho JS, Oh JH, Shim SB, Jee SW, Lee SH, Seo SJ, Lee SK, Lee SH, Kim YK (2005) Differentially expressed genes in transgenic mice carrying human mutant presenilin-2 (N141I): correlation of selenoprotein M with Alzheimer's disease. Neurochem Res 30 (8):1009-1019. doi:10.1007/s11064-005-6787-6

10. Hwang DY, Sin JS, Kim MS, Yim SY, Kim YK, Kim CK, Kim BG, Shim SB, Jee SW, Lee SH, Bae CJ, Lee BC, Jang MK, Cho JS, Chae KR (2008) Overexpression of human selenoprotein M differentially regulates the concentrations of antioxidants and H2O2, the activity of antioxidant enzymes, and the composition of white blood cells in a transgenic rat. Int J Mol Med 21 (2):169-179. doi:10.3892/ijmm.21.2.169

11. Goo JS, Kim YN, Choi KM, Hwang IS, Kim JE, Lee YJ, Kwak MH, Shim SB, Jee SW, Lim CJ, Seong JK, Hwang DY (2013) Proteomic analysis of kidneys from selenoprotein M transgenic rats in response to increased bioability of selenium. Clin Proteomics 10 (1):10. doi:10.1186/1559-0275-10-10

12. Chen P, Wang RR, Ma XJ, Liu Q, Ni JZ (2013) Different Forms of Selenoprotein M Differentially Affect A beta Aggregation and ROS Generation. Int J Mol Sci 14 (3):4385-4399. doi:10.3390/ijms14034385

13. Garcia-Triana A, Gomez-Jimenez S, Peregrino-Uriarte AB, Lopez-Zavala A, Gonzalez-Aguilar G, Sotelo-Mundo RR, Valenzuela-Soto EM, Yepiz-Plascencia G (2010) Expression and silencing of selenoprotein M (SELENOM) from the white

shrimp Litopenaeus vannamei: effect on peroxidase activity and hydrogen peroxide concentration in gills and hepatopancreas. Comp Biochem Physiol 155 (2):200-204. doi:10.1016/j.cbpa.2009.10.037

14. Reeves MA, Berry MJ (2012) Selenoprotein M. Selenium: Its Molecular Biology and Role in Human Health, Third Edition. doi:10.1007/978-1-4614-1025-6_15

15. Korotkov KV, Novoselov SV, Hatfield DL, Gladyshev VN (2002) Mammalian selenoprotein in which Sec (Sec) incorporation is supported by a new form of Sec insertion sequence element. Mol Cell Biol 22 (5):1402-1411. doi:10.1128/mcb.22.5.1402-1411.2002

16. Pitts MW, Reeves MA, Hashimoto AC, Ogawa A, Kremer P, Seale LA, Berry MJ (2013) Deletion of selenoprotein M leads to obesity without cognitive deficits. J Biol Chem 288 (36):26121-26134. doi:10.1074/jbc.M113.471235

17. Ogawa-Wong AN, Berry MJ, Seale LA (2016) Selenium and metabolic disorders: An emphasis on type 2 diabetes risk. Nutrients 8 (2). doi:10.1126/SCIENCE.1083516; 10.3390/nu8020080

18. Dawson PE, Muir TW, Clark-Lewis I, Kent SB (1994) Synthesis of proteins by native chemical ligation. Science 266 (5186):776-779. doi:10.1126/science.7973629

19. Hondal RJ, Nilsson BL, Raines RT (2001) Sec in native chemical ligation and expressed protein ligation. J Am Chem Soc 123 (21):5140-5141. doi:10.1021/ja005885t

20. Quaderer R, Sewing A, Hilvert D (2001) Sec-mediated native chemical ligation. Helv Chim Acta 84 (5):1197-1206. doi:10.1002/1522-2675(20010516)84:5<1197::aid-hlca1197>3.0.co;2-#

21. GieSELENOMan MD, Xie L, van Der Donk WA (2001) Synthesis of a Sec-containing peptide by native chemical ligation. Org Lett 3 (9):1331-1334. doi:10.1021/ol015712o

22. Malins LR, Mitchell NJ, Payne RJ (2014) Peptide ligation chemistry at selenol amino acids. J Pept Sci 20 (2):64-77. doi:10.1002/psc.2581

23. McGrath NA, Raines RT (2011) Chemoselectivity in chemical biology: acyl transfer reactions with sulfur and selenium. Acc Chem Res 44 (9):752-761. doi:10.1021/ar200081s

24. Hondal RJ (2005) Incorporation of Sec into proteins using peptide ligation. Protein Pept Lett 12 (8):757-764. doi:10.2174/0929866054864319

25. Metanis N, Keinan E, Dawson PE (2006) Synthetic seleno-glutaredoxin 3 analogues are highly reducing oxidoreductases with enhanced catalytic efficiency. J Am Chem Soc 128 (51):16684-16691. doi:10.1021/ja0661414

26. Casi G, Roelfes G, Hilvert D (2008) Selenoglutaredoxin as a glutathione peroxidase mimic. ChemBioChem 9 (10):1623-1631. doi:10.1002/cbic.200700745

27. Berry SM, GieSELENOMan MD, Nilges MJ, van Der Donk WA, Lu Y (2002) An engineered azurin variant containing a Sec copper ligand. J Am Chem Soc 124 (10):2084-2085. doi:10.1021/ja0169163

28. Dery L, Reddy PS, Dery S, Mousa R, Ktorza O, Talhami A, Metanis N (2017) Accessing human selenoproteins through chemical protein synthesis. Chem Sci DOI: 10.1039/c6sc04123j. doi:10.1039/c6sc04123j

29. Kapust RB, Tozser J, Copeland TD, Waugh DS (2002) The P1' specificity of tobacco etch virus protease. Biochem Biophys Res Commun 294 (5):949-955. doi:10.1016/s0006-291x(02)00574-0

30. Muller S, Senn H, Gsell B, Vetter W, Baron C, Bock A (1994) The formation of diselenide bridges in proteins by incorporation of Sec residues - biosynthesis and characterization of (Se)(2)-thioredoxin. Biochemistry 33 (11):3404-3412. doi:10.1021/bi00177a034

31. Chong S, Williams KS, Wotkowicz C, Xu MQ (1998) Modulation of protein splicing of the Saccharomyces cerevisiae vacuolar membrane ATPase intein. J Biol Chem 273 (17):10567-10577. doi:10.1074/jbc.273.17.10567

32. Liu J, Li F, Rozovsky S (2013) The intrinsically disordered membrane protein selenoprotein S is a reductase in vitro. Biochemistry 52 (18):3051-3061. doi:10.1021/bi4001358

33. Studier FW (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif 41 (1):207-234. doi:10.1016/j.pep.2005.01.016

34. Blommel PG, Fox BG (2007) A combined approach to improving large-scale production of tobacco etch virus protease. Protein Expr Purif 55 (1):53-68. doi:10.1016/j.pep.2007.04.013

35. Ellman GL (1959) Tissue sulfhydryl groups. Arch Biochem Biophys 82 (1):70-77. doi:10.1016/0003-9861(59)90090-6

36. Geoghegan KF, Dixon HB, Rosner PJ, Hoth LR, Lanzetti AJ, Borzilleri KA, Marr ES, Pezzullo LH, Martin LB, LeMotte PK, McColl AS, Kamath AV, Stroh JG (1999)

Spontaneous alpha-N-6-phosphogluconoylation of a "His tag" in Escherichia coli: the cause of extra mass of 258 or 178 Da in fusion proteins. Anal Biochem 267 (1):169-184. doi:10.1006/abio.1998.2990

Chapter 3

PREPARATION OF SELENOMETHIONINE-CONTAINING GB1 PROTEINS

3.1 Background

Sulfur participates in a vast repertoire of biological functions. [1] Biologically relevant forms of sulfur include the canonical amino acids cysteine and methionine, but also homocysteine and taurine. Cysteine residues are incorporated into proteins as redox-active centers, which regulate cellular oxidative stress. Methionine is among the most hydrophobic of the amino acids. It is often found in the interior of the protein, or in the case of membrane proteins, they are found to interact with the lipid bilayer. However, surface exposed methionine residues are regarded as antioxidants to proteins. [2] In addition, there exists an endogenous pool of sulfide compounds including hydrogen sulfide and polysulfide and the cysteine-containing enzymes that produce them, which are involved in sulfur metabolism and regulatory functions. [1]

NMR is a spectroscopic tool that analyzes the changes in the electronic structure surrounding an atom. This technique can be used to identify a chemical species, or to follow the reactions that the species undergoes by monitoring perturbations at the nucleus of interest. This makes NMR an attractive method for studying sulfur chemistry. As a sulfur-containing species becomes chemically transformed, its local electronic environment will be changed as atoms are added or taken away, bonds formed or

broken. However, the sulfur nucleus itself is not amenable to biological NMR. [3-4] This is because the only NMR active nucleus of sulfur, ^{33}S is a quadrupolar nucleus with low natural abundance on par with that of ^{13}C, and broad linewidth making detection and data interpretation a challenging task. [3] Sharp spectroscopic lines are only attainable for high symmetric sulfur species, rendering this technique ineffective for studying sulfur species in any biological samples. [5]

Since sulfur cannot be directly probed using NMR, selenium was found to be a viable surrogate to sulfur. ^{77}Se is a spin 1/2 nucleus with 7.63% natural abundance, roughly 10 times that of ^{33}S. It has good NMR sensitivity and can be detected by conventional hardware and experiments. [3-9] In addition, sulfur and selenium share many physiochemical properties such as comparable electronegativity and the same redox states. [6] The difference in sulfur and selenium's atomic radius and the length of covalent bonds with carbon is about 10%. ^{77}Se chemical shifts in biological systems spans over 3000 ppm for different chemical species in selenium-containing amino acids, that is the biologically relevant Se-, Se-Se, SeH, Se-OH, Se-O_2H, SeO_3H, CH_2-Se-CH_3, CH_2-Se(O)-CH_3 and CH_2-Se(O_2)-CH_3. [4]

The feasibility of biological selenium NMR has been demonstrated nearly 40 years ago. [10-11] Like most heavy atoms, selenium has large chemical shielding response which translates to efficient relaxation routes resulting in short transverse relaxation rates and broad spectroscopic lines. [3,4,8,12] In addition, currently there is sparse research on biological selenium NMR which makes data interpretation difficult. Unlike for other established NMR nuclei such as ^{1}H, ^{13}C, ^{15}N for which there exist

comprehensive chemical shift libraries such as the BMRB Biochemical Reference, there is no such systematic for selenium. Therefore, this study aims to establish a model system which can be used to compile ^{77}Se chemical shift tensors in proteins to understand how the chemical environment in biological molecules influences ^{77}Se spectra.

The model protein selected for this task is the well-characterized immunoglobulin - binding B1 domain of Streptococcal Protein G (GB1). This is a highly compact and stable small protein unit, with its melting point approximately 20 °C above that of the average protein. [13-14] Furthermore, GB1's fold contains all secondary structural elements: helix, beta sheets both parallel and antiparallel, as well as loops. (Figure 8) The entire protein is 56 residues long and the core of the protein is solvent inaccessible, consisting of residues Leu5, Leu7, Ala26, Phe30, Ala34, Trp43, Phe52 and Val54. The sole Trp at position 43 is found on β-strand 3, and its sidechain is in van der Waals contact with Phe30 and Phe52, as well as aliphatic residues Leu5, Ala34 and Val54. Phe30 and Phe52 are arranged so that the aromatic sidechains are perpendicular to each other, with Trp43, Tyr33 and Tyr45 packed loosely in the vicinity. With this configuration, 95% of GB1's hydrophobic residues are shielded from the solvent environment, which is a significant contributor to its high thermodynamic stability.

Within the protein there is also an extensive hydrogen bonding network. As much as 80% of the residues in GB1 are hydrogen bonded (H-bonded). The first pair of anti-parallel β-ribbons is fully H-bonded, with Gln2 and Asp22 sharing two H-bonds between them. In the second anti-parallel β-ribbon pair, there are three H-bonds between

residues Glu42 and Thr55, Thr44 and Thr53, and Asp46 and Thr51. Between the central parallel β-strands 1 and 4 there are six H-bonds on the peptide backbone. In addition, there are four hydrogen bonds between the sidechains of Thr49 and Thr51, Thr44 and Thr53, Glu56 and Lys10, as well as Glu56 and Asp40.

Extensive biochemical and biophysical studies are available on GB1, and its structure is well documented using X-ray crystallography, solution- and solid-state NMR with over one hundred structures deposited in the Protein Data Bank. Stability of the GB1 protein is also well-documented, multiple studies reported mutations to the primary amino acid sequence of GB1 in order to investigate the impact on the protein's structure and stability. [13-30] Taken together, these studies suggest that GB1 has a stable fold, where mutations of up to 14% of the sequence do not significantly alter protein structure. [19,25,31]

In the present study, six dissimilar locations were selected for mutagenesis to selenomethionine: Leu5, Ile6, Val29, Ala34, Val39 and Val54. Of these locations, Ile6 and Val29 are located on the solvent exposed sides of the first β-strand and the α-helix, respectively. Because neither are involved in the hydrophobic packing of the protein core, these mutations are expected to not alter the protein structure. [16,31-34] Similar mutations to Thr16 and Thr18 posed no structural disruptions to β-sheet propensity, so it can be postulated that Ile6 mutation will yield the same result. In the case of Val29, it has been identified as a "boundary position" on the GB1 structure, where its sidechain can be solvent exposed to shielded based on its local interactions. [32-33] For example, V29F is stabilizing to the global structure by 1.7 kJ/mol, while V29K is destabilizing

by 2.2 kJ/mol. This destabilization can also be rescued by placing the protein in high salt environment, where up to 2M of NaCl shows destabilization by 1.5 kJ/mol. The authors hypothesized that it is due to Lys29 forming unfavorable electrostatic interactions with the solvent which becomes partially diminished in a high salt environment. In the wild type GB1, a packing defect has been identified involving the sidechains of Val29, Phe30 and Tyr33, where a void in space was created by these three residues and which could be stabilized by substituting Val29 to larger aromatic residues. Bulky sidechains at the 29th position can potentially turn inward to enhance packing and stability in this area. [31,33] In general, it has been reported that mutations which fill hydrophobic pockets are generally stabilizing. [34] Therefore, since the SeM sidechain is larger than the wild type Val, it is postulated that Val to SeM mutation will be stabilizing to the global protein structure.

The contiguous core of GB1 where the remaining four mutations are selected from, is formed by residues at positions 3, 5, 7, 20, 43, 52 and 54 on the β-strand, positions 26, 30, 34 on the α-helix and position 39 on the loop. [34] It has been shown that when mutations at these core positions disrupt the natural packing of the protein, it is compensated by the protein burying additional sidechains in the hydrophobic core and to reduce the volume of the core. Thus the size of the core residues tends to increase for mutations that present a stabilizing force to the structure. [35] Mutation at Leu5 had no significant impact on overall structure but increasing the size of the sidechains at positions 34 and 54 have both demonstrated to drastically alter the native structure. [19,35] *Thoms et al.* showed that with A34L and V54F mutations, both of which

increase the size of the residue from the wild type, displace Trp43 and partially expose its sidechain to the solvent environment. As a result, new hydrophobic driving forces may emerge to compensate for the loss of local stability. When Trp43 is displaced, the resulting structure shows increased disorder in this region, as competing forces attempt to reestablish favorable interactions in the protein core. [35] However, it is worth noting that these mutations still produce a stable protein structure, and it may be interesting to investigate the new interactions that arise from mild core-destabilizing mutations.

Overall, the six SeM mutations identified for this study are expected to produce stable proteins whose global structural features resemble those of the wild type GB1 (Table 3). This chapter discusses the relevant strategies in preparing these SeM-containing GB1 variants for both solution and solid-state NMR. Data from solution NMR compiles the characteristic isotropic chemical shifts of the selenium atom. However due to rapid molecular tumbling in a liquid medium, individual configurations of the protein cannot be resolve and only an averaged conformation can be obtained. In the case of solid-state NMR, although in many cases similar to a solution NMR experiment, contains a wider range of information including the means to determine molecular orientation.

Detection of SeM in GB1 by solution NMR can be achieved using natural abundance selenium. [3] However, for efficient data acquisition using solid-state NMR, GB1 proteins enriched with ^{77}Se were prepared.

3.2 Materials and Methods

3.2.1 Gene, Plasmids and Molecular Cloning

The gene of the immunoglobulin-binding B1 domain of streptococcal protein G (GB1) was codon optimized for expression in E. coli and the gene synthesized by DNA2.0. The synthetic gene of N-terminal hexahistidine-tag tagged GB1 was cloned into the bacterial expression vector pJexpress414 (DNA2.0). The hexahistidine-tag included the TEV protease cleavage sequence ENLYFQG to allow for affinity tag removal of the hexahistidine-tag. The tag-free GB1 has Gly at position 1 instead of Met so that a sole Met residue can be inserted at other locations of the protein. In addition, the engineered GB1 sequence following cleavage with TEV protease also contained a substitution of Gln for Thr at position 2 since the majority of studies on GB1 utilize this mutation to improve sample homogeneity. [16] This vector was subsequently used as a template to introduce all the Met variants using PCR-based mutagenesis.

3.2.2 Protein Expression and Purification for Solution NMR

Proteins samples for solution NMR were prepared by utilizing the direct incorporation of the SeM amino acid into the protein chain, which results in samples containing natural abundance ^{77}Se at 7.63%.

To begin, the appropriate Met variant was transformed into *E. coli* BL21(DE3) cells. To incorporate the expressed protein with SeM, a defined media was used that supplied no Cys or Met sources by the time protein expression was induced. In brief, an overnight starter culture in low salt LB broth (10 g of tryptone, 5 g of yeast extract, 5 g NaCl per L), supplemented with 100 µg/ml ampicillin was inoculated and grown at 37 °C with good aeration.

Next day, 10 ml of starter culture was used to inoculate 1 L of LB broth, supplemented with 100 µg/ml ampicillin. When optical density (OD) at 600 nm reaches 0.7, the cells were pelleted by centrifugation for 10 min at 4°C and 5000 rpm in a Sorvall RC5B floor centrifuge equipped with a SLA3000 rotor. The supernatant was discarded, and the cell pellet was washed with ice cold sterilized water. The cells were pelleted again and resuspended in modified MDAG media from *Studier* (Table 4). [36] The temperature was lowered to 18 °C, and the cells were allowed to recover for 15 – 30 min. L-SeM (Acros Organics) was supplied at 0.1 g/L with no additional Met supplementation, and after an additional 10 min, protein expression was induced with 0.5 mM Isopropyl *β-D*-1-thiogalactopyranoside (IPTG). The cells were grown for 20 hours at 18 °C.

The cells were pelleted as described and resuspended in 50 mM sodium phosphate, 200 mM NaCl, 10 mM imidazole, pH 8.0, supplemented with 1 mM phenylmethylsulfonyl fluoride (PMSF) and 1 mM benzamidine. Cells were disrupted by homogenization and the cell debris was removed by centrifugation at 17000 rpm for 1 hr. The supernatant was pooled and loaded onto a 15 ml immobilized metal affinity

chromatography (IMAC) column that was pre-equilibrated with 5 CV (Column Volume) of IMAC equilibration buffer (50 mM sodium phosphate, 200 mM NaCl, 10 mM imidazole). The column was washed with 15 CV of buffer supplemented with 30 mM imidazole and His_6-$GB1^{SeM}$ protein was eluted with a final imidazole concentration of 500 mM.

The yield of His_6-$GB1^{SeM}$ is on average 50 mg per L of defined media. The ratio of SeM incorporation into GB1 was assessed by mass spectrometry after removal of the polyhistidine tag due to its post-translational modification which interferes with accurate mass calculation. Using this method, SeM incorporation typically reaches > 95% with the Met variant undetectable in some preps. (Figure 10a) His_6-$GB1^{SeM}$ variants were cleaved by TEV protease at 1:3 molar ratio in 50 mM sodium phosphate, 200 mM NaCl, pH 8.0, supplemented with 5 mM of β-mercaptoethanol (BME). The progress of the cleavage was monitored by 16% tris-tricine SDS-PAGE.

After TEV cleavage, the protein mixture was loaded onto a second IMAC column in order to remove the TEV protease. The flow through containing the tag-free GB1 was pooled and concentrated using an Amicon Ultra-15 centrifugal filter unit with MWCO of 3 kDa. The concentrated protein was loaded onto an appropriate size exclusion column suitable for small proteins (< 20 kDa) such as Superdex™10/300 GL or Sephacryl™ S-100, pre-equilibrated with gel filtration buffer such as 50 mM sodium phosphate, pH 5.5. Fractions containing GB1 was collected and visualized on 16% Tris-tricine SDS-PAGE. Protein concentration was determined by UV-Vis spectroscopy with an extinction coefficient of 9970 M^{-1} cm^{-1}, or 1.617 g L^{-1} when appropriate.

3.2.3 Sample Preparation for Solution NMR

GB1 SeM variants containing natural abundance SeM were used for solution NMR experiments. This is a viable strategy due to the comparably high natural abundance of the ^{77}Se nucleus (7.63%). Purified GB1 in 50 mM sodium phosphate buffer pH 5.5 and 10% D_2O was concentrated to approximately 10 mM in 500 uL using Amicon Ultra-15 centrifugal filter unit with MWCO of 3 kDa. Such a high sample concentration is required in order to observe the ^{77}Se signal under natural abundance conditions.

3.2.4 Protein Expression and Purification for Solid-State NMR

Proteins samples for solid-state NMR were prepared based on the method reported by *Schaefer et al.* to incorporate elemental ^{77}Se in order to isotopically enrich the sample. [3]

To begin, the appropriate Met variant was transformed into *E. coli* BL21(DE3) cells. A 10 mL starter culture supplemented with 5 mM Na_2SO_4 and 100 μg/mL of ampicillin was grown at 37 °C for 9 hours. The starter culture contained 50 mM Na_2HPO_4, 50mM KH_2PO_4, 10 mM NaCl, 50 mM NH_4Cl, 2 mM $MgCl_2$, 0.2x metals, 1x vitamins, 0.4% glucose and 200 μM $CaCl_2$. The 1x trace metal solution contained 50 μM $FeCl_3$, 20 μM $CaCl_2$, 10 μM $MnCl_2$, 10 μM $ZnCl_2$, 2 μM $CoCl_2$, 2 μM $CuCl_2$, 2 μM $NiCl_2$, 2 μM Na_2MoO_4 and 2 μM H_3BO_3. The recipe for 1000x vitamin solution was as

detailed by *Studier*. [36] It is worth noting that the components used in this media including the antibiotic selection were kept sulfur-free except for Na_2SO_4 which was the sole source of sulfur.

When the optical density of the starter culture reached approximately 1.2, 1 ml of the starter culture was used to inoculate 1 L of growth media supplemented with 50 μM Na_2SO_4 and 100 μM/ml of ampicillin in 2.8 L baffled flasks. Cells were grown at 37 °C with good aeration until sulfur depletion, which was indicated by a plateau in OD600nm measurements at about 0.95 that was typically reached after 15 hours. At this point, the cells were provided with a mixture of 20 μM Na_2SO_4/70 μM Na_2SeO_3. Protein expression was induced with 0.8 mM IPTG. Cells were grown for an additional 12 hours at 37 °C then harvested by centrifugation as previously described. (Figure 10b)

When ^{77}Se isotope was to be used, an appropriate amount of elemental ^{77}Se powder was carefully oxidized to ^{77}selenite using nitric acid. No more than 100 μM of nitric acid can be present in the cell culture to minimize cellular toxicity.

Purification procedure for the isotopically labeled protein is exactly as described for the natural abundance version. However, when using gel filtration to polish the purified GB1, the column was pre-equilibrated with buffer suitable for solid-state NMR experiments such as 25 mM sodium acetate, pH 5.5. Fractions containing GB1 was collected and their purity was evaluated using 16% Tris-tricine SDS-PAGE. Protein concentration was determined by UV-Vis spectroscopy with an extinction coefficient of 9970 M^{-1} cm^{-1}, or 1.617 g L^{-1} when appropriate. Recombinant expression strategies of GB1 are robust and consistently yield 50 mg of pure protein per each liter of rich

media. The yield does not suffer significantly when using minimal expression media for the production of isotopically labeled samples, with average final yield of 40 mg per liter of medium, and average isotopic labeling of 80% as confirmed by mass spectrometry (Figure 11).

3.2.5 Sample Preparation for Solid-State NMR

Microcrystalline protein samples for solid-state NMR experiments were prepared by batch crystallization on ice, similar to the conditions previously reported by *Schmidt et al*. [22] In brief, a solution of GB1 protein at 50 mg/ml in 25 mM sodium acetate buffer pH 5.5 was added by increments of the precipitant of 2-methyl-2,4-pentanediol (MPD) to isopropyl alcohol (IPA) in 2:1 v/v ratio. After the addition of each increment, the mixture was gently tapped to mix. Small increments of the precipitant were continuously added in this fashion until microcrystals begin to form. This was evident by the appearance of a birefringent cloudiness that appears in the mixture and did not disappear upon tapping. The mixture was then left undisturbed at 4 °C overnight. Crystal growth typically fully formed in 2 days. Crystals were packed by gentle centrifugation into a 3.2 mm Bruker MAS thin-wall rotor, immersed in excess mother liquor. Typical experiments utilized ~45 mg of GB1 of which up to 80% was enriched with ^{77}Se.

3.3 Summary of chapter 3

This study aims to establish a model protein system which can be used to compile ^{77}Se chemical shift tensors to readily identify specific types of interactions. As a viable surrogate to sulfur, selenium NMR will help to shed light on the chemistry of sulfur and sulfur reaction mechanisms in biological processes. To this end, six SeM-containing GB1 variants were expressed and purified, each to be used in a series of solution and solid-state NMR experiments. Solution NMR samples were prepared using protein containing natural abundance selenium, while solid-state NMR samples were prepared via batch crystallization of isotopically enriched GB1.

Table 3. Local interactions of six engineered GB1-SeM variants

Mutation	Location	Interaction
L5M	Hydrophobic; close to Phe30 Trp43	Sulfur – Aromatic
I6M	Solvent exposed near Lys4 and Glu15	London Dispersion
V29M	Surface exposed on helix	London Dispersion
A34M	Hydrophobic; close to Phe30 Trp43	Sulfur – Aromatic
V39M	Loop, oriented into core	Flexible
V54M	Hydrophobic; close to Trp43	Sulfur – Aromatic

Table 4. Defined growth medium for natural abundance SeM-incorporation

Components	Volume (ml)
Sterilized water	940
1 M Mg_2SO_4	2
1000x metals*	0.2
1000x vitamins*	1
40% w/v glucose	10
50 XM*	20
25% aspartate	10
17 amino acids (10 mg/mL)*	20
L-SeM	0.1 g/L

*The recipes for the preparation of 1000x metals, 1000x vitamins, 50XM salts and 17 amino acids are as described by *Studier* [36].

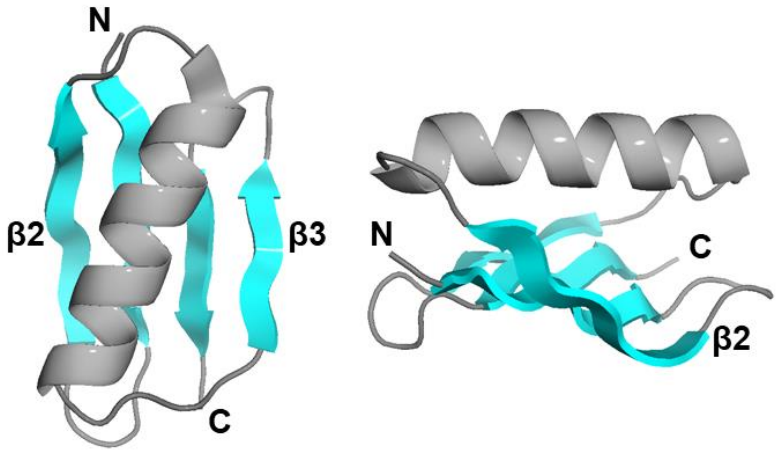


Figure 8. Global structural features of the wild type GB1 domain obtained through X-ray diffraction (PDB Accession Code: 2QMT). Left: front view of the GB1 structure, comprises a four-stranded β-sheet and a long central helix connecting strands 2 and 3. Right: side view of the same structure showing packing of the α-helix against the β-strands. [15]

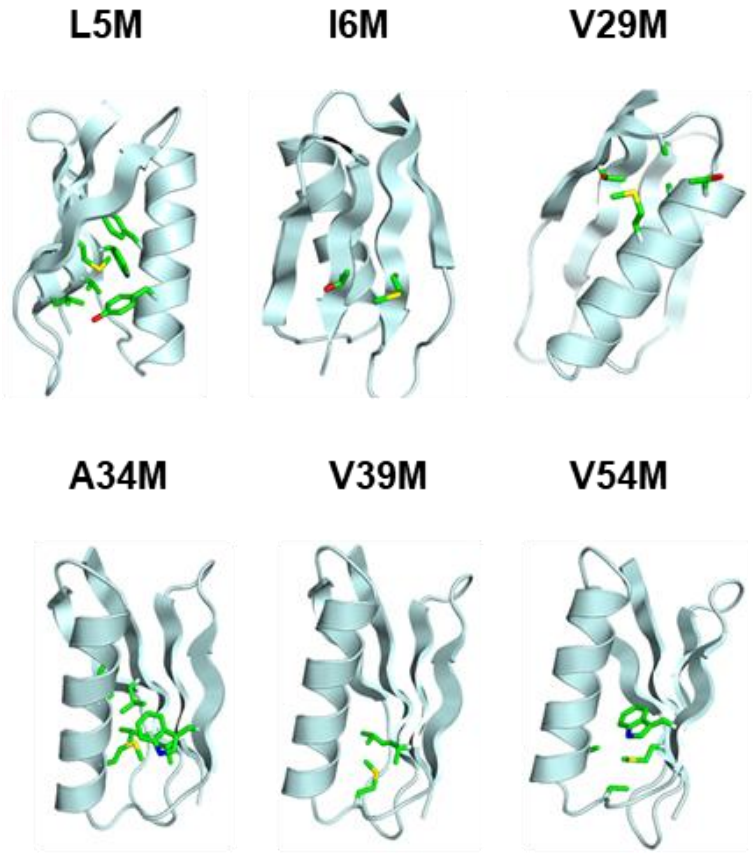


Figure 9. Energy minimized structures of the six GB1 Met variants. The sulfur atom of Met is colored yellow. The structures were minimized using the Met instead of SeM as Force fields for SeM were not yet reported. Residues within van der Waals contact to each Met sidechain are also shown in sticks. Energy minimization and figure courtesy of Shiping Xu.

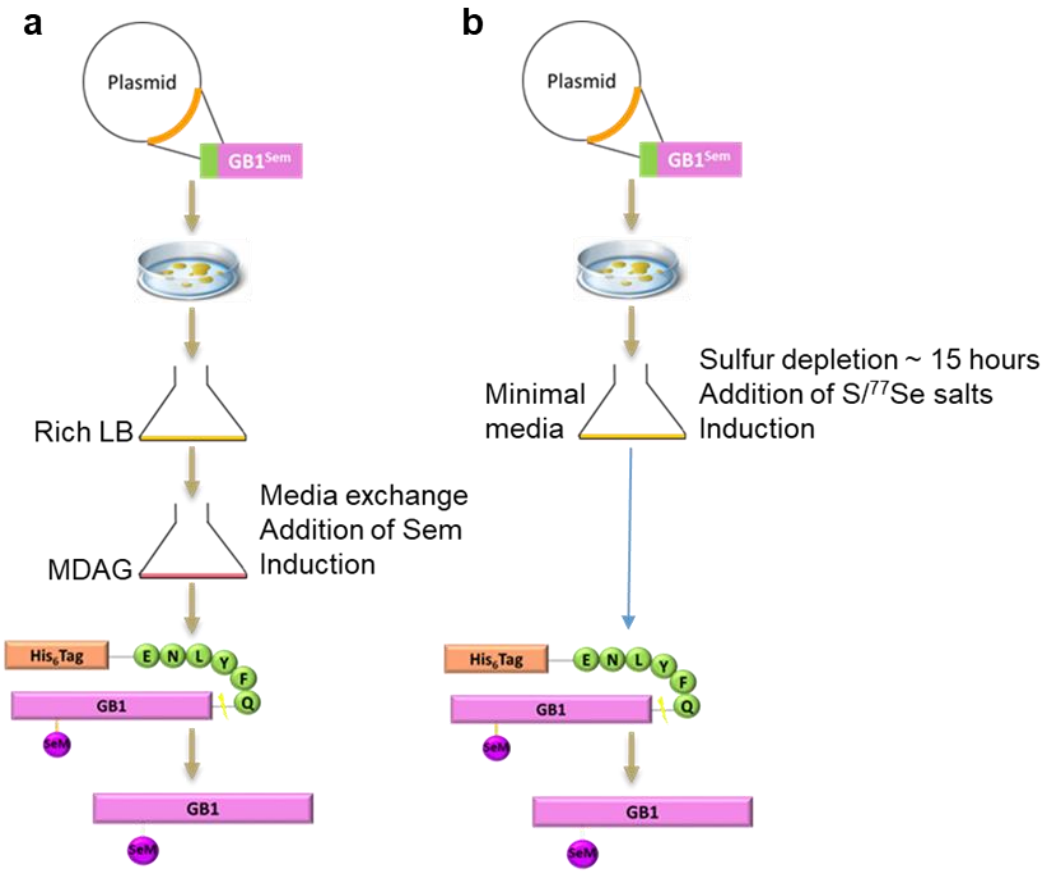


Figure 10. Schematic diagrams of the expression and purification of GB1 SeM variants. A) protein sample for solution NMR was expressed by direct incorporation of natural abundance SeM amino acid; B) protein sample for solid-state NMR was prepared via sulfur depletion and the incorporation of elemental ^{77}Se isotope in the form of selenite.

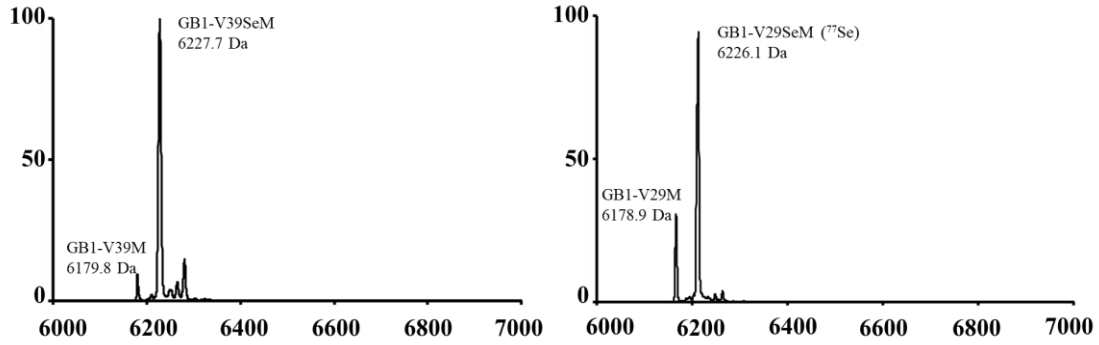


Figure 11. Representative deconvoluted intact-mass electrospray ionization mass spectrum of GB1 proteins. A) GB1 V39SeM prepared by the incorporation of SeM amino acid at natural abundance ratio during protein expression. B) GB1 V29SeM prepared by the random incorporation of ^{77}Se isotope. Met variant is present at 30%.

REFERENCES

1. Nagahara N and Wróbel M. (2017) Atomic sulfur: An element for adaptation to an oxidative environment. Molecules 22 (11): 1821-1827.

doi: :10.3390/molecules22111821

2. Brosnan JT and Brosnan ME. (2006) The Sulfur-Containing Amino Acids: An Overview. J Nutr 136 (6-Suppl): 1636S-1640S. PMID: 16702333

3. Schaefer SA, Dong M, Rubenstein RP, Wilkie WA, Bahnson BJ, Thorpe C and Rozovsky A. (2013) ^{77}Se enrichment of proteins expands the biological NMR toolbox. J Mol Biol 425: 222-231. doi: 10.1016/j.jmb.2012.11.011

4. Rozovsky S. (2012) "^{77}Se NMR spectroscopy of selenoproteins." (2013) In Biochalcogen chemistry: the biological chemistry of fulfur, selenium, and tellurium. ACS Press. doi: 10.1021/bk-2013-1152.ch006

5. Barbarella G. (1993) Sulfur-33 NMR. Prog Nucl Magn Reson Spectrosc 25 (4): 317-343. doi: 10.1016/0079-6565(93)80001-A

6. . Johansson L, Gafvelin G and Arnér ESJ. (2005) Selenocysteine in proteins—properties and biotechnological use. Biochim Biophys Act 1726: 1-13. doi: 10.1016/j.bbagen.2005.05.010

7. Duddeck H. (2004) Se-77 NMR spectroscopy and its applications in chemistry. Annu Rep NMR Spectrosc 52: 105-166. doi: 10.1016/S0066-4103(04)52003-3

8. Liu J and Rozovsky S.(2016) ^{77}Se NMR spectroscopy of selenoproteins. In Selenium its molecular biology and role in human health, with Hatfield DL, Gladyshev VN,

Schweizer U, and Tsuji PA Eds. Springer: 187-198. doi: 10.1007/978-3-319-41283-2_15

9. Wessjohann LA, Schneider A, Abbas M and Brandt W. (2007) Selenium in chemistry and biochemistry in comparison to sulfur. Biol Chem 388 (10): 997-1006. doi: 10.1515/BC.2007.138

10. Luthra NP, Costello RC, Odom JD and Dunlap RB. (1982) Demonstration of the feasibility of observing nuclear magnetic resonance signals of 77Se covalently attached to proteins. J Biol Chem 257 (3): 1142-1144.

11. Dunlap RB and Odom JD. (1988) Magnetic resonance studies of selenium containing compounds. Phosphorus Sulfur Silicon Relat Elem 38 (3-4): 217-229. doi: 10.1080/03086648808079718

12. Gettins P and Wardlaw SA. (1991) NMR relaxation properties of ^{77}Se-labeled proteins. J Biol Chem 266 (6): 3422-3426.

13. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT and Clore GM. (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253: 657-661. doi: 10.1126/science.1871600

14. Alexander P, Fahnestock S, Lee T, Orban J and Bryan P. (1992) Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domain B1 and B2: Why small proteins tend to have high denaturation temperatures? Biochemistry 31: 3597-3603.

15. Schmidt HL, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, Wilson SR and Rienstra CM. (2007) Crystal Polymorphism of Protein GB1 Examined by Solid-State

NMR Spectroscopy and X-ray Diffraction. J Phys Chem B 111: 14362-14369. doi: 10.1021/jp075531p

16. Gallagher T, Alexander P, Bryan P and Gilliland GL. (1994) Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 33: 4721-4729.

17. Smith CK, Withka JM and Regan L. (1994) A thermodynamic scale for the β-sheet forming tendencies of the amino acids. Biochemistry 33: 5510-5517.

18. Gronenborn AM, Frank MK and Clore GM. (1996) Core mutants of the immunoglobulin binding domain of streptococcal protein G: stability and structural integrity. FEBS Lett 398: 312-316. doi: 10.1016/S0014-5793(96)01262-8

19. Frank MK, Dyda F, Dobrodumov A and Gronenborn AM. (2002) Core mutations switch monomeric protein GB1 into an intertwined tetramer. Nat Struct Biol 9 (11): 877-885. doi: 10.1038/nsb854

20. Byeon IL, Louis JM and Gronenborn AM. (2003) A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol 333: 141-152. doi:10.1016/S0022-2836(03)00928-8

21. Franks WT, Zhou D, Wylie BJ, Money BG, Graesser DT, Frericks HL, Sahota G and Rienstra CM. (2005) Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): ^{15}N and ^{13}C chemical shift assignments and conformational analysis. J Am Chem Soc 127: 12291-12305. doi: 10.1021/ja044497e

22. Frericks Schmidt HL, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, Wilson SR and Rienstra CM. (2007) Crystal polymorphism of protein GB1 examined by solid-state NMR spectroscopy and X-ray diffraction. J Phys Chem 111: 14362-14369. doi: 10.1021/jp075531p

23. Jee JG, Byeon IL, Louis JM and Gronenborn AM. (2008) The point mutation A34F causes dimerization of GB1. Proteins 71: 1420-1431. doi: 10.1002/prot.21831

24. Thorns S, Max KEA, Wunderlich M, Jacso T, Lilie H, Reif B, Heinemann U and Schmid FX. (2009) DImer formation of a stabilized Gβ1 variant: a structural and energetic analysis. J Mol Biol 391: 918-932. doi: 10.1016/j.jmb.2009.06.031

25. Strop P, Marinescu Am and Mayo SL. (2000) Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. Protein Sci 9: 1391-1394. doi: 10.1110/ps.9.7.1391

26. Cunningham TF, Pornsuwan S, Horne WS and Saxena S. (2016) Rotameric preferences of a protein spin label at edge-strand beta-sheet sites. Protein Sci 25: 1049-1060. doi: 10.1002/pro.2918

27. Cunningham TF, McGoff MS, Sengupta I, Jaroniec CP, Horne WS and Saxena S. (2012) High-resolution structure of a protein spin-label in a solvent-exposed beta-sheet and comparison with DEER spectroscopy. Biochemistry 51: 6350-6359. doi: 10.1021/bi300328w

28. Cunningham TF, Putterman MR, Desai A, Horne WS and Saxena S. (2015) The double-histidine Cu^{2+}-binding motif: a highly rigid, site-specific spin probe for electron

spin resonance distance measurements. Angew Chem Int Ed Engl 54: 6330-6334. doi: 10.1002/anie.201501968

29. Reinert ZE, Lengyel GA and Horne WS. (2013) Protein-like Tertiary Folding Behavior from Heterogeneous Backbones. J Am Chem Soc 135: 12528-12531. doi: 10.1021/ja405422v

30. Davey JA, Damry Am, Goto NKand Chica RA. (2017) Rational design of proteins that exchange on functional timescales. Nat Chem Biol 13: 1280-1285. doi: 10.1038/nchembio.2503

31. Wunderlich M, Mas KEA, Roske Y, Mueller U, Heinemann U and Schmid F. (2007) Optimization of the Gβ1 domain by computational design and by in vitro evolution: structural and energetic basis of stabilization. J Mol Biol 373: 775-784. doi: 10.1016/j.jmb.2007.08.004

32. Dahiyat BI and Mayo SL. (1997) Probing the role of packing specificity in protein design. Proc Natl Acad Sci 94 (19): 10172-10177. doi: 10.1073/pnas.94.19.10172

33. Malakauskas SM and Mayo SL. (1998). Design, structure and stability of a hyperthermophilic protein variant. Nature Struct Biol 5: 470–475. doi: 10.1038/nsb0698-470

34. Saito M, Kono H, Morii H, Uedaira H, Tahirov TH, Ogata K and Sarai A. (2000). Cavity-filling mutations enhance protein stability by lowering the free energy of native state. J Phys Chem ser B: 104: 3705–3711. doi: 10.1021/jp991717f

35. Thoms S, Mas KEA, Wunderlich M, Jasco T, Lilie H, Reif B, Heinemann U and Schmid FX. (2009) Dimer formation of a stablized GB1 variant: a structural and energetic analysis. J Mol Biol 391: 918-932. doi: 10.1016/j.jmb.2009.06.031

36. Studier FW. (2005) Protein production by auto-induction in high-density shaking cultures. Protein Expression Purif 41: 207-234. doi: 10.1016/j.pep.2005.01.016

Chapter 4

STRUCTURES OF GB1 SELENOMETHIONINE VARIANTS BY X-RAY DIFFRACTION

4.1 Background

4.1.1 Structure Determination through X-ray Diffraction

Structure determination through crystallography began at the dawn of the 20th century with the discovery of X-rays by German scientist Wilhem Conrad Röntgen. The next 60 years saw tremendous advances in both the theory and experimentation in the analysis of crystal structures using X-rays, resulting in a number of Nobel Prizes in the fields of physics, chemistry and medicine. [1-2] In 1958, the first high resolution protein structure was determined for myoglobin and soon after for hemoglobin through X-ray crystallography by John Kendrew and Max Perutz. [3-5] These high resolution structures revealed vastly different shapes and features of each protein, and sparked intense interest in studying protein structural diversity that has lasted to the present day. For their seminal work in macromolecular crystallography, Kendrew and Perutz shared the 1962 Novel Prize in chemistry. [2]

The latest survey of the Protein Data Bank shows over 1.3 million structures deposited and 90% of them were obtained using X-ray crystallography. This is a powerful biophysical technique that allows direct visualization of the atomic makeup of

a target. With this method, the arrangement of atoms in proteins and nucleic acids can be determined and often at the atomic level with advances in building more powerful radiation sources. The ability to visualize biological macromolecules in 3-dimensional space and near-atomic resolution allows us to explore enzymatic mechanisms or protein-ligand interactions. These discoveries bridge the understanding between structure to function, and the results have often found use in rational therapeutic designs. [5-7]

The "structures" obtained through an X-ray diffraction experiment, or more precisely, the atomic coordinates determined by any structural biology technique, represent both a temporally and spatially averaged set of molecular properties. Specifically, the structure obtained using X-ray diffraction is an averaged representation of all the molecules periodically arranged in the protein crystal lattice. [8-10] During data collection, a single protein crystal containing millions of copies of the protein molecule is subjected to a beam of X-ray for minutes or hours. [10,11] Thermal energy excites the molecules in the crystal allowing them to sample their conformational energy landscape. X-ray diffraction captures the protein structure averaged across all the copies of molecules in the crystal lattice across the entire length of data acquisition time, thus providing a thorough sampling of the energy landscape available to each molecule. The electron density map that is produced represents this averaging across all molecules and the energy landscapes available to each, and can be said as "blurred" by the combination of motions the protein experiences. Therefore, while it is possible to extract motional information from an X-ray diffraction experiment, the information about the rates at

which such motions occur is lost due to this averaging. Also, static vs dynamic disorders cannot be distinguished by X-ray diffraction. Where motion does occur in a crystal structure, whether it is static (spatial disorder) or via site exchange (temporal disorder) by rapid jumping between different conformations, these phenomenon results in structures with fractional populations. [19] This is a key disadvantage to the X-ray diffraction technique, where differences between equivalent sites remain undetected and where exchange rates information cannot be extracted. [8,12-15]

A complementary structural biology technique to X-ray diffraction that has emerged in the last 40 years is NMR. There is a synergy between X-ray crystallography and NMR spectroscopy [5] and these techniques can be used together in trying to solve the structure of a biological macromolecule. In fact, these two techniques have long been recognized as complementary methods in structural biology. [7-9]

When an NMR experiment is performed on a protein molecule, the resulting spectrum is informative of several aspects of the protein's properties, such as the identity, quantity, and relative location within the protein molecule that the nucleus of interest can be found. A solution NMR experiment provides a structure that is averaged over the random orientations of the molecule tumbling in a liquid medium. In solid-state NMR, since the samples can be either powder materials or crystalline, this technique permits the direct visualization of molecular orientations. [16] The complementarity of NMR to X-ray diffraction arises from the experimental time scales that can be accessed by NMR. Unlike X-ray diffraction which takes place over seconds to hours, NMR time scales are typically micro- to millisecond range, where many types

of protein motions occur. [17-19] In situations where structural disorders are present, instead of taking an average over the entire crystal, NMR can potentially reveal individual configurations if the exchange rate between the configurations are slower than the NMR experiment timescale. [19] For example, if a given sidechain residue samples two conformations, the local environment between the two conformations are likely to be different as well. As the local electronic environment changes, the frequency at which the nucleus absorbs radio frequency will change, thereby changing the observed NMR spectrum. When the rate of exchange between the two conformations is slow, it is possible to observe distinct peaks representing each of the conformations. However, when exchange rate between the two conformations happens on the same time scale as the molecular motion that induces the exchange between the two conformations, the observed peaks may coalesce, and become one broader peak. Therefore, in NMR experiments peak width can be indicative of molecular motions experienced by the nucleus of interest. [19] The possibility to determine individual configurations coupled with the fact that exchange rate information can also be extracted make NMR a powerful approach for full structural characterization of crystalline protein molecules.

In summary, X-ray diffraction and NMR are complementary techniques that together become informative of the structure of a protein molecular and the rate at which the molecule samples different conformations available to it. Using X-ray crystallography, the relative population of different conformational states along with their respective orientation information can be obtained, including the determination of the favorable energy landscapes for each of the conformations. NMR experiments are

informative of the rates at which exchanges occur between these available conformational states, but provide little information on the distance or direction of these conformations.

4.1.2 Motivation for obtaining GB1 crystal structures

As previously discussed, this research project seeks to establish a comprehensive library of selenium NMR chemical shift parameters. Information about selenium's chemical shielding tensors will be sought after using both solution and solid-state NMR strategies.

Because selenium exhibits high sensitivity to its local environment, it is expected that the isotropic chemical shifts of the SeM sidechain will vary between each GB1 SeM variant, as they each experience a different set of chemical environment. Such differences in the chemical shifts are indicative of the differences in the local electronic environment around the selenium nucleus. Preliminary solid-state CP/MAS NMR data using ^{77}Se isotopically labeled V29SeM showed two major peaks and two minor peaks. This data suggests that up to four conformations are possibly available to the SeM sidechain in this structure (data not shown). Methods of quantum chemistry, such as DFT calculations, can be used to calculate the chemical shift tensors, and relate theoretical predictions to experimentally observed NMR spectra. [20]

DFT calculations can be performed using an optimized or predicted structure of the protein molecule, or in the case of GB1, structures obtained by X-ray

crystallography. The structure of the wild type GB1 is well documented, and the conditions to grow diffraction-quality crystals can be easily reproduced, as already discussed in the preceding chapter. Crystals grown from GB1 proteins frequently diffract to atomic resolution, where backbone trace and most sidechains can be clearly observed. At this resolution, it is possible to determine if multiple conformations exist for the same sidechain. This is especially critical when determining if any variations exist for the SeM local environment. Recent advances in X-ray crystallography software have nearly fully automated structure solution. [21-23] With a number of high resolution GB1 structures to use as search templates for molecular replacement, obtaining a fully refined structure of a variant GB1 is an easily accessible task.

Applying quantum chemical calculations of the chemical shift tensor is crucial to understanding the relationship between structural characteristics and the observed chemical shifts as results of these structures. Obtaining high resolution, accurate X-ray crystal structures of the GB1 SeM variants provides the structural factors that can be used in said quantum chemical calculations. Taken together, the experimental chemical shielding tensor, and the quantum chemical calculations using structural information from the same protein crystals used to obtain experimental NMR data, allow the interpretation and prediction of the local environment of the selenium nucleus in a given protein. Once a match between experimental data and theoretical calculations is established, theoretical studies can be expanded to systematically study the environment of SeM in proteins and to understand ligand binding, protein-protein interactions and protein dynamics in isolation and in protein complexes. [5]

4.2 Materials and Methods

4.2.1 Crystallization Screening

Conditions for single crystal growth were screened based on the original conditions reported by *Gallagher et al.* [24] and *Schmidt et al.* [25], but with the exception of zero salt concentration to facilitate the applicability of the successful condition in subsequent NMR experiments. Each GB1 variant was screened using the identical set of conditions, as the structural differences between the SeM variant and the wild type were minimal. Initial screening was conducted using the hanging-drop vapor-diffusion method in 48-well plates at 283.15 K. Equal volumes of protein and reservoir solutions of 0.5 µl were mixed together and equilibrated against 0.25 ml of the reservoir solution. Single crystals formed in 2 to 3 days and finished growing within 1 week. Preliminary crystals were obtained for each type of variant from the broad screening range of 45-50% v/v of MPD (2-methyl-2,4-pentanediol), 20% v/v of IPA (isopropyl alcohol), in 25 mM sodium acetate buffer with pH ranging from 4.5 to 4.9. The conditions which yielded the most promising morphologies were optimized by varying the precipitant gradient, buffer pH, and protein concentration. The optimized conditions were then prepared with freshly purified protein in duplicates by multiple operators to ensure reproducibility as well as to maximize the chances for harvesting the highest quality of crystals.

Due to the high concentration of MPD in the crystallization condition, no additional cryoprotectants were required during the freezing process. Instead, a single crystal was picked up in a cryoloop and directly flash-cooled in liquid nitrogen. X-ray diffraction data were collected on beamline 5.0.1 at the Advanced Light Source (ALS) at the Lawrence Berkeley Laboratory equipped with a Pilatus 6M detector. Each data set was indexed and integrated with XDS. [26] Data reduction was done using AIMLESS, all of which are parts of the CCP4 program suite. [22-23,27]

4.2.2 Structure Refinement

The structure of wild type GB1 reported by *Schmidt et al.* [25] (PDB ID: 2QMT) was used as the initial search model in molecular replacement for each GB1 SeM variant. [28] The six GB1 variant structures were refined independently and without reference to each other. Refinement was performed iteratively using phenix.refine and then with COOT to visualize the model after each cycle of refinement [29-30]. Each model was initially refined with rigid body and simulated annealing. Hydrogens were refined as riding hydrogens. The fit of the model to the electron density map was analyzed for each amino acid residue, and manual adjustments such as alternate sidechain conformations or chirality centers, were made using COOT with the decisions guided by the density map. Ligands were identified both with LigandFit [31] as part of the PHENIX program suite, as well as manually by inspecting the map. Placement of the ligands was verified by monitoring improvements in the map and the refinement

statistics. The density map of the entire structure was inspected after each round of refinement to determine the appropriateness of the fit. The SeM sidechain was refined with the help of the anomalous difference signal derived from the selenium atom and the anomalous maps calculated by PHENIX. [32] Occupancy of the selenium atom was refined iteratively. X-ray verses stereochemistry weights, as well as X-ray verses atomic displacement parameters (ADP) were refined at the end of the refinement process. Refinement progress was monitored by the analysis of the results using POLYGON and MolProbidity. [33] A randomly selected subset of the data which was omitted before the start of refinement was used to calculate the free R value which was used to cross-validate the model. [34]

A simulated annealing omit map calculated from PHENIX was used to validate each of the final refined models and to reduce potential bias that can arise from the probe model structure. [35] In each case, the omit map was virtually identical with the $2F_0$- F_c electron density map and for each variant the omit map showed continuous density for all amino acid residues.

4.3 Crystal Structures of GB1 SeM Variants

4.3.1 GB1 SeM Variant Crystal Morphologies

Thin rod-shaped crystals were obtained for five of the six SeM variants: L5SeM, I6SeM, V29SeM, A34SeM and V39SeM. The dimensions for these crystals averaged

800 μm in length, 50 μm in width and approximately 10-20 μm in depth. Crystal morphologies for each variant largely resembled thick needles, often growing out of a single nucleation point in the hanging drop. Such "dendrite" morphology is a common phenomenon in protein crystallization (Figure 12). When a nucleus is formed through heterogeneous epitaxy, which happens via attachment to a mechanical surface such as particles in the protein drop or the glass slide, a protein crystal begins to develop. However, one face of the crystal cannot develop through restriction by the mechanical barrier, while the other face extends unhindered. This one-dimensional growth introduces tremendous stress to the crystal lattice. In order to relieve the strain, the crystal splinters and initiates multiple separate growths in all other unrestricted directions. The end result of this process is a dendritic crystal often referred to as a "crystal bouquet". [36]

No diffraction quality crystals were obtained for the V54SeM over several attempts. The crystals that grew from V54SeM lacked visible birefringence and had rounded edges. Lack of sharp crystal facets is often an indication that the growth rate is too fast and the resulting crystals tend not to diffract well or at all. [37-38] Multiple protein concentrations and temperatures were attempted but without improvement to the morphology of these crystals. The location of the V54SeM mutation is completely shielded from solvent on the last β-strand, though may be located near the crystal contact edges, which can disrupt crystal packing. Therefore, for this particular variant, wider screening conditions are still required in the future in order to produce diffraction quality crystals.

From these crystals, the structures of the five GB1 SeM variants have been solved at atomic resolution and deposited into the Protein Data Bank (PDB). Overall structures of the variants resemble that of the wild type protein (Figure 13). The PDB accession codes for the GB1 variants are: L5SeM (6CNE), I6SeM (6CPZ), V29SeM (6C9O), A34SeM (6CHE) and V39SeM (6CTE). A sample of the electron density map for A34SeM is shown in Figure 14.

4.3.2 Crystal Contacts

Crystal contact for protein crystals can be complex to analyze, especially if more than one copy of the molecule is found in the asymmetric unit (AU), which adds interactions between the two copies in the asymmetric unit on top of the interactions formed at the crystal packing interface. An AU is the smallest unit which contains all of the structural information. From the AU and by application of the symmetry operators the unit cell can be constructed. The content of the AU depends on both the crystallized protein's position and conformation within the unit cell. Two generalized scenarios can occur for the AU depending on the crystallization conditions and local packing: 1) each copy of the protein within a unit cell has identical conformation and occupies symmetry-related positions. In this scenario, the biologically-relevant assembly can be a monomeric protein chain or two or more symmetry-related chains coming together to form a larger complex; or in case 2) copies of the protein in the AU each have unique positions and different conformations from each other. As a result, each copy of the

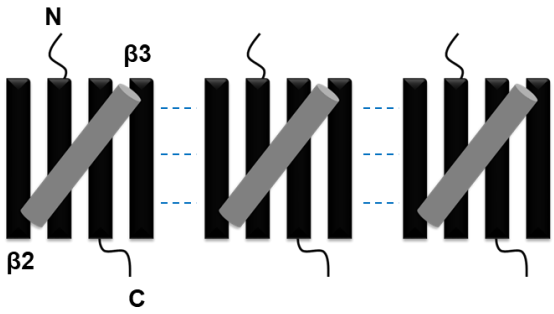
protein in the unit cell may represent structurally similar but functionally diverse states of the biological assembly. [39-41]

All five GB1 variants crystallized in the monoclinic P1 21 1 space group, the same as the wild type structure. [24,25,41,50] In fact, from a survey of nearly 10,000 nonredundant crystal structures deposited in the Protein Data Bank, more than 75% of them crystallized in one of the primitive lattice types. [42] The monoclinic crystal system is one of seven crystal systems in crystallography. Each crystal system is described by three vectors and three angles which the vectors intersect. For the monoclinic, the vectors of the crystal are of unequal lengths to each other, and form a rectangular prism with a parallelogram as its base. Therefore, through this morphology, the angles can also be determined: two vectors intersect at 90° to each other, while the third vector meets the other two at a non-right angle.

The preference for proteins to adopt some symmetries over others cannot be fully explained yet. [42] *Wukovitz and Yeates* developed an entropic model which describes the nonuniformity using rigid-body degrees of freedom (D). [43] The model suggests that space groups with higher D-values occur more frequently and in the case of the series of GB1 SeM variants, D was calculated to be 6 using the entropic model, which placed the P1 symmetry group as the sixth most frequently observed for protein crystals. A number of similar studies attempted to also address the preference for certain symmetries in protein crystals. [42-46] An interesting observation was that lower symmetries are associated with lower solvent content, which is why triclinic, monoclinic (such as GB1 SeM variants), and orthorhombic systems are found to have

10% lower average solvent content compared to cubic crystals. Solvent content of GB1 crystals are L5SeM 25.18%, I6SeM 24.80 %, V29SeM 28.87%, A34SeM 22.56% and V39SeM 23.92% (Table 5). Average solvent content of a protein crystal is 43%. [47] However, even when available studies are taken collectively, the overall frequency distribution of crystal symmetries cannot be adequately explained by solvent content, or the theory of entropy, or symmetry operators that promote closer packing alone. It is indeed possible that distribution of space-group frequencies is simply the result of errors in the reported crystal symmetry: either error in the space group but correct crystal system, or errors in both the space group and the crystal system. [48-49]

An intrinsic property of the GB1 protein is that it favors association via stretches of unsatisfied hydrogen bonds along its edge β-strands 2 and 3. [41] In dilute solutions individual GB1 molecules will not spontaneously associate due to energetic penalties in decreasing entropy of the system. However, at high protein concentrations the formation of edge hydrogen bonds becomes favored and is the stabilizing force to drive crystal growth. In fact, the crystal contact surface for GB1 variants are found along these edge strands. This preferred site of association also explains the tendency of these GB1 variants to form needle-shaped crystals. As growth and elongation is favored only along one face of the structure, the crystal propagates by aligning itself along these β-strands as shown in the schematic graph below, where blue dashes indicate hydrogen bonds that can be satisfied upon association.



The crystal contact interface for each GB1 SeM variant are formed between the second β-strand of one molecule and the third β-strand of another (Figure 15). Strand 2 includes residues TLKGETT (amino acids 11 to 17) and strand 3 includes residues GEWTYDD (amino acids 41 to 47), which are positioned within 3 Å anti-parallel to each other. Nine polar contacts are found within this stretch: five of which are formed between backbone amide protons and carbonyl groups (Glu15 to Trp43, Lys13 and Tyr45, amide proton of Asp47 and carbonyl of Thr11), the other four of which involve sidechain interactions (amide proton of Thr17 to Glu42 sidechain as well as to hydroxyl of Thr16, indole ring proton of Trp43 to Glu15 carboxyl sidechain, hydroxyl of Tyr45 to carboxyl sidechain of Asp47). However, it is worth noting that in order to promote crystallization, the pH of the system is generally poised close to the theoretical pI of the protein, making the charge state of the amino acids in a crystal structure ambiguous. [51-52]

Interactions involved in crystal packing are generally considered to be non-specific, so long as the interactions favor the formation and growth of crystals. This generalization is based on the wide polymorphism exhibited by protein molecules and the crystals which are derived from them. [53] *Carugo et al.* systematically analyzed 78

non-redundant crystal structures deposited in the PDB and determined that the composition of amino acids that form the crystal contact region is statistically indistinguishable from that which form the solvent exposed surfaces of the protein. [53-54] The residues that participate in forming the crystal packing interface tend to participate in hydrogen bonding in order to form stable interactions and also quite counter-intuitively, have higher atomic thermal factors compared to their solvent-shielded counterparts, suggesting they share the flexibility generally associated with protein surfaces. This conclusion was later verified by *Luo et al.* using a database of 773 protein structures, who concluded that there exist no significant geometric or physiochemical properties in the amino acids that form crystal packing interfaces. [56]

4.3.3 Description of the Global Features of GB1 SeM Crystals

In each GB1 SeM variant, the initiating Met residue of the wild type protein was mutated to Gly so that the protein sequence was devoid of any other sulfur containing amino acids except the engineered Met residue which was expressed as SeM. In addition, the Thr at position 2 in the wild type was substituted for Gln in order to improve the homogeneity of the purified protein. [57] These modifications to the N-terminal sequence is not expected to have a significant impact on the global structure of the protein. SeM substitutions located on the solvent exposed surface of GB1 also should not destabilize the protein such as for I6SeM and V29SeM, however a detailed

comparison of local structural perturbations may be required for the residues whose SeM substitution is on the interior of the protein (L5SeM, A34SeM and V39SeM).

Root-mean-square distance, or RMSD is a standard measurement of the similarity between two superimposed proteins through pairs of their equivalent Cα atoms. The expression for RMSD calculation is as follows: [58-59]

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2}$$

In the expression above, *n* represents the number of pairs of equivalent atoms, d_i is the distance between the two atoms in any given *i*-th pair. From the expression above it is clear there are several precautions that should be taken into consideration when RMSD is used to compare protein structures.

The main disadvantage of RMSD lies in the fact that it is easily biased by the amplitude of errors. [58] If two structures are similar but possess regional differences through conformational change or the presence of flexible terminal loops, their RMSD value could be artificially high due to these localized structural discrepancies even though the two protein molecules are virtually identical. In other words, RMSD cannot effectively distinguish between two pairs of structures where one pair is indistinguishable from each other except for a significant regional deviation, and the other pair where a multitude of smaller scale rearrangements occur. For the series of SeM variants, it will be particularly important to examine the stretches of amino acids around the SeM residue mutation, as that is likely where the highest degree of structural

perturbations is found. An additional feature of the RMSD measurement is its dependence on resolution. *Caruugo et al.* demonstrated that smaller RMSD values are associated with structures refined at higher resolution, and that it tends to increase for two structures under comparison that have dissimilar resolutions. [59]

GB1 Wild Type The structure of the GB1 protein has been extensively characterized by a large number of studies and covered in detail in the preceding section. [60-64] Briefly, the domain is made up of four strands of β-sheets and one α-helix which crosses over them. The central two β-strands are parallel to each other, with residues 1 to 8 making up of strand 1 and residues 50 to 56 forming the second strand. The two outer strands, β2 and β3, are anti-parallel to β1 and β4, respectively. The β2 strand consists of residues from 13 to 20 and β3 strand encompasses residues 42 to 47. Residues 9 to 12 form a type I connecting loop between the first and second strand, while the third and fourth strands are connected by an unusual six residue turn formed by residues 46 to 51. [60]

Overview of GB1 SeM Variants The structures of five GB1 SeM variants obtained by X-ray diffraction is shown in Figure 16. Each horizontal panel of the figure depicts the overall structure of the variant, indicating the relative position of the SeM sidechain to the body of the protein. In addition, residues found within 4.5 Å of the selenium atom are illustrated in middle and bottom panels. The SeM sidechains are clearly observed for each variant; in the case of I6SeM, two conformations were modeled as was justified by the $2F_o$-F_c electron density map as well as anomalous diffraction map of selenium.

L5SeM L5SeM is a solvent-inaccessible site on the first β strand. Despite having the SeM mutation oriented into the core of the protein, RMSDs between chains A and B of L5SeM to 2QMT are 0.140 and 0.180, respectively, which indicates no significant structural difference from the introduction of the mutation. This could be explained by Leu and SeM having similar sized sidechains so that the substitution of Leu does not affect packing. Both amino acids contain Cα, Cβ, Cγ and Cδ (which is substituted for selenium in SeM), with Leu having an extra branched methyl group and SeM instead having a linear methyl group. This finding is consistent with previous reports that L5V does not alter GB1 structure. [66-67]

Six amino acids are found within 4.5 Å of SeM5: Leu7, Phe30, Trp43 and Val54 interact more strongly with selenium while Thr16 and Ala34 have weaker interactions. From the orientation of each interacting sidechain, they are engaged in non-polar interactions with the SeM residue. Only one of the terminal branched methyl groups of Leu7 and Val54 is within van der Waals distance to the selenium: $C_{\delta}2$ of Leu7 is 3.96 Å from selenium, while Cγ2 of Val54 is found 4.14 Å away. The other methyl group in each case is > 5 Å from the selenium atom and pointed away therefore unlikely to interact. Two aromatic residues, Phe30 and Trp43 are found 4 Å away from the selenium. The phenyl ring of Phe30 is staggered over SeM5, placing the Cβ of SeM5 over the ring face but not the selenium atom itself. Instead, the selenium atom approaches the edge of the Trp43 phenyl ring at 3.8 Å (Figure 17). These distances are too long to be hydrogen bonds, which typically range from 2.3 to 2.7 Å with selenium being the acceptor group. [68] However, energetically favorable sulfur – aromatic

interactions have been identified from as much as 5 to 7 Å apart. [69] Nonetheless, the interactions surrounding the SeM sidechain locked the sidechain into one distinct conformation (Figure 18A). In addition, anomalous data from selenium was used to further confirm the localization of the selenium atom in the sidechain (Figure 18B).

While the interactions described above are for subunit A of the two copies of the molecule in the asymmetric unit, the configurations in subunit B are similar. One notable difference is that Thr16 in subunit A is modeled with two alternate conformations, whereas in subunit B only one conformation can be clearly observed. In either case, the selenium atom does not form any interactions with the Thr sidechain, rather it is the Cε of SeM5 that approaches the Thr hydroxyl group at 3.5 Å in both subunits.

I6SeM GB1 I6SeM contains a SeM located on the exterior of the first β-strand of the protein. Because it is solvent accessible, it is not expected to cause significant structural perturbations compared to the native protein. Structural alignment showed little global variation between them as expected, with alignment RMSD of 0.202 between 2QMT to chain A and 0.152 to chain B – the two copies of I6SeM molecule in the AU. SeM incorporation did not alter the backbone of the protein, and the Cα of the residues surrounding the mutation site for both chains exhibit an exact match to the model structure.

Solvent exposed sidechains exhibit more flexibility compared to the buried ones, and several residues in this structure have been modeled with alternate conformations: Gln2, Lys10 and SeM6 of chain A, and SeM and Thr16 of chain B. Gln2 occupies two

conformations, one of which the amide proton of the sidechain interacts via hydrogen bonding (2.8 Å) with the carboxylic acid sidechain of Glu19. Gln is capable of hydrogen bonding through its two amine groups, the backbone amine cation NH_3^+ and the amine on the sidechain NH_2. [65] Thus Gln can hydrogen bond through donation from NH_3^+ and NH_2 to surrounding water molecules or other eligible sidechains. Lys10 also exhibits two conformations, one of which extends into the solvent and engages a water molecule while the other forms a salt bridge with Glu56. Based on the refined occupancy of each conformation for Gln2 and Lys10, there is no preference for the sidechain to adopt one conformation over another. The two alternate conformations of Thr16 on chain B do not form specific interactions with nearby residues however in the more preferable conformation with occupancy refined to 70%, the hydroxyl group is found within polar contact distance to a nearby water molecule.

In each subunit, the SeM6 sidechain was modeled with two alternate conformations, guided by the $2F_o$-F_c as well as the anomalous map density. Of the two conformations, the $2F_o$-F_c density can be observed for Cε of one of them which enables an exact measurement of the C-Se-C angle, as well as the C-C-Se-C χ3 torsion angle surrounding the selenium atom (Table 6). This is the dominant conformation at 80% of occupancy. The relevant angles for the less populated conformation was measured on the final refined structure, however since density around Cε cannot be observed, the angles also cannot be accurately measured (Figure 19). [71]

The dominant conformation in each subunit is surrounded by Cγ of Thr51 and Thr53 at 4 Å, as well as the C5 methyl group of MPD which is 4.7 Å away. There are

no specific interactions for the less populated conformation of SeM6 in subunit A, however in subunit B, the selenium atom is within 5 Å of the salt bridge formed between Lys4 and Glu15 (Figure 20). At 1.12 Å resolution, the assignment of the sidechains is unambiguous. Therefore, it can be clearly distinguished that the interacting functional groups with the selenium atom are the aliphatic methyl groups of Thr and MPD, instead of the hydroxyl groups. Water molecules are found within hydrogen bonding distances of the hydroxyl groups of each Thr sidechain, further confirming the assignment of the sidechain (Figure 21). [72] Sulfur-aliphatic interactions are moderately stabilizing, which may be the underlying explanation that electron density of SeM6 Cε methyl group is observed for the dominant conformation located near three aliphatic groups. [73]

V29SeM V29SeM is the only other GB1 variant with SeM mutation designed to be solvent exposed. Here, SeM29 is positioned in the helical surface of the protein. RMSD between chain A of V29SeM and 2QMT is 0.173, and with chain B is 0.210 indicating no major structural perturbation caused by the mutation. Minor conformational variations for surfaced exposed residues exist however for most residues located in the core of the protein, their conformations match exactly with 2QMT. One distinct positional difference in the protein core is seen for Tyr33 whose hydroxyl group is shifted by 1.5 Å compared to 2QMT in chain A and 1.3 Å in chain B. Thr15 of chain B has two sidechain conformations with one configuration rotated 90° from the other. Interestingly, while Val29 in the wild type protein is surface exposed, the sidechain of SeM29 extends back into the core of the protein and is not surface accessible though the peptide backbone still remains exposed (Figure 22, Upper Panel). This is more

pronounced for SeM in chain A whose terminal methyl is within 4 Å Thr18, Ala20 and Ala26, the latter amino acids being solvent-shielded residues. In chain B, the methyl group of SeM extends outward towards the solvent, while the Se atom is angled inward. The closest residues near Se are the sidechains of Ala20 and Thr25. In the case of Thr25, it is the methyl group which faces the Se.

Multiple conformations of SeM29 sidechain exist in both subunit A and B in this structure (Figure 22, Lower Panel). Anomalous diffraction map acquired at the selenium absorption edge shows a continuum which is in distinct contrast to that of SeM5 and SeM6 previously discussed. Each of the anomalous diffraction map exhibits a "peanut" shape, where two positions appear to be more favored compared to the conformation that exists in the pinched region of the peanut. Nonetheless, the selenium atom cannot be accurately located in this variant. In addition, electron density surrounding the terminal methyl group (Cε) is not observed. Therefore, bond angles for each sidechain cannot be accurately measured. The following residues are located 4.5 Å away from the selenium atom in subunit A: Thr18, Glu19, Ala20, Thr25, Ala26. One MPD molecule is 4.39 Å away from the Cε methyl group of SeM29, and > 5 Å away from the selenium atom. The distances in subunit B are similar.

A34SeM This is the only GB1 SeM variant which crystallized with one copy of the molecule in the AU. In addition, this is also the only structure that displays the largest degree of deviation from the model structure of 2QMT. RMSD between this variant and the wild type is 0.446, the highest of all five crystal structures. Unlike L5SeM which a structurally similar SeM was substituted for Leu, in this case a much

smaller Ala residue located inside the protein core was replaced by larger SeM. Located at the C-terminal end of the helix, Ala34 sidechain points up towards strand β4 and inward to the core of the protein. [67] Placement of SeM forces the protein core in the immediate vicinity to expand, and as a result the N-terminal of the β3 strand is displaced by 1.0 Å outwards compared to the wild type GB1. Residues found in the immediate surroundings of the mutation also exhibited displaced positions. Lys31, Glu42, Trp43 side chains are in new positions relative to the wild type GB in order to accommodate the larger SeM sidechain, with the exception of Asp40 which shifted inward by as much as 1.6 Å (Figure 23). In this conformation Asp40 possibly interacts via hydrogen bond to Lys31, and together they effectively form a "cap" at the end of the protein, in order to compensate for the loss of hydrophobicity within the core due to the structural displacements. The sidechain from Val39, the central residue of the hinge-loop, twists 90° inward possibly to complete the formation of the "cap" (Figure 23 Inset). The C-terminal of the neighboring strand 4 also expanded outward by 1.1 Å, however there are no sidechain displacements on this strand, except for Thr55 being shifted outward but without any conformational differences on the sidechain.

The conformation of Trp43 presents an interesting phenomenon in protein structure stabilization (Figure 24). In the wild type protein, Ala34 is directly across from the plane of the phenyl ring of Trp43, and the distal end of Lys31 is perpendicular to its ring face. Both Ala34 and Lys31 are 3.7 Å away. The sidechain projects inward with the plane of the rings perpendicular to the α-helix across the central channel. In A34SeM variant, if the position of Trp43 were to remain unchanged, the distance from the

selenium atom to the phenyl ring of Trp would be 1.7 Å, leading to steric clashes in the region. In order to accommodate the SeM sidechain, the aromatic ring of Trp43 is dislocated from being completely solvent shielded to partially solvent exposed. The sidechain of Lys31 is displaced further away from Trp. The distance between the amide group of Lys to the ring face of Trp increases from 3.7 Å to 6.9 Å. This conformation may become destabilizing to the highly hydrophobic protein core, however the crystal structure clearly reveals a bound imidazole ligand placed halfway between Lys31 and Trp43. GB1 SeM variants were prepared via IMAC therefore during purification imidazole was used to elute the protein and as an additive to prevent nonspecific binding to the column. It is likely at this point the imidazole was incorporated into the protein chain in order to stabilize the structure. At pH 4.7, both imidazole nitrogens should be protonated to give the cationic imidazolium, and one of the charged nitrogens is stacked directly over the center of the Trp43 phenyl ring approximately 4 Å away. At the same time, the imidazolium is 3.9 Å away from the selenium atom of SeM34, the center of the ring placed over the selenium atom.

The crystal structure also revealed less populated conformations of SeM34. It is possible that the SeM sidechain is populated across a sweep of locations (Figure 25). In its alternate conformations, the carbons at Cα, Cβ and Cγ remain stationary, and motion predominantly came from the selenium atom and the terminal methyl group. The distance between the terminal methyl groups of conformations A and C is 3.0 Å and conformation B is sandwiched in between and 1.5 Å away from each of the other two. The configuration of B and C are shifted downward where they may be able to engage

the nearby Leu7 and Val54 in nonpolar interactions. It also moves the SeM away from the imidazole and Trp43 possible to minimize clashing in the region.

V39SeM The SeM39 sidechain is located on the loop connecting the α helix to the third β-strand. Residues found on loops in a protein structure typically exhibit more flexibility However, as the electron density shows, residues from 37 to 41 have well-defined conformations, and the data does not suggest the presence of any alternate conformations, including the SeM39. Comparison between the variant and wild type yielded RMSD of 0.177, indicating no structural perturbations as a result of this mutation. Val39 and SeM39 both project into the core of the protein. However, as the residues are located on a loop, there is sufficient space to accommodate the larger SeM sidechain (Figure 26A).

The residues located within 4.5 Å of the selenium atom are: Leu7, Leu12, Ala34 and Val54. No polar groups are found to interact with selenium, except the backbone carbonyl oxygen of Ala34 which is located 4.4 Å away and therefore may be weakly interacting. No ligands were located nearby. This is the case for both subunits in the asymmetric unit. These interactions suggest that the SeM39 sidechain is fully encased in a nonpolar environment, surrounded by aliphatic amino acids. The SeM sidechain in each subunit is best fitted by one conformation, as guided by the $2F_o$-F_c density map and the anomalous diffraction map acquired of the selenium atom (Figure 26BC).

V54SeM No diffraction quality crystals were obtained with the V54SeM variant. It is possible to postulate based on existing data the primary reason behind this. Val54 is located on the last β strand inside the protein core, shielded from solvent. Trp43

is located directly across Val54 on strand 3, the distance between them is 3.5 Å. This distance is even smaller compared to the distance between Trp43 and Ala34, which is 4.1 Å (Figure 27). It is possible that a Val to SeM mutation at this position will significantly disrupt the protein-protein interactions responsible for crystallization. Trp43 can become dislodged even further from its original position as compared to the A34SeM variant, due to the close distance between the residues. If that were the case, and the crystal packing surfaces for the other five GB1 variants were all found along the β2, β3 strand, this can in turn disrupt the crystal packing surface, leading to severe defects in the crystals that cannot diffract. If the displacement of Trp43 is indeed the underlying cause in obtaining diffraction-quality crystals, co-crystallization with high imidazole, similar to that of A34SeM may be a viable strategy to stabilize any flexibility at the crystal contact region and produce usable crystals.

4.4 Discussion

4.4.1 Conformational Flexibility in SeM Sidechain

X-ray crystallography measures the intensities of X-rays diffracted by the atoms in a molecule. The electron density map constructed as a result is used to locate the positions of these atoms and in turn construct a complete model of the protein of interest. Electron density distributions are Gaussian, and structural models are built by placing

the atoms into the highest peak of the distribution. [75] The lower electron density regions fall off from the peak position and can be modeled either isotropically, or at sufficiently high resolution (typically < 1.2 Å), can be described by anisotropic B-factors. Within these local minima electron densities, there may be alternate positions of the atoms that contain vital information about alternate conformations of the residue. In addition, even in a rigid crystal lattice, and at cryogenic temperatures, the protein can still adopt multiple conformations from either static or dynamic positional disorders. Such conformational flexibilities should be carefully inspected, since they may be of functional importance especially in biologically relevant systems (e.g. enzyme catalysis mechanisms).

The assignment of alternate conformations must be guided by more than the common iso-surface wire-frame representation of the $2F_o$-F_c density maps. Difference maps, or F_o-F_c maps, are preferable in identifying conformational heterogeneity in weak, irregular election densities. Contour levels of these maps are often enhanced in the model building stage, and also during refinement to aid visual inspection of the resulting protein models. Additional tools used to modify and enhance map feasures include local feature enhancement, maximum entropy principles, and B-factor sharpening. [77-79] Even more complex is to plot electron density distribution as a function of dihedral angle. [80] Most amino acid sidechains adopt preferential rotameric torsion angles, which can be used to assign alternate conformations especially where electron densities are weak. In the case of GB1, most electron densities are

unambiguous, and complex fitting programs were not necessary to build or refine the structures.

In a recent survey of sidechain conformational flexibilities, *Miao et al.* summarized four types of conformational variations that exist in protein crystal structures. Type I are described as fixed conformers, such as most buried residues, whose atomic coordinates are definite. Type II describes discrete conformations, where different conformations can be definitively built into the electron density. Conformational variations described by Type III and Type IV are different degrees of conformational flexibility. In Type III, the so-called cloud conformation, the locations of the sidechain cover a limited but continuous region. In Type IV, the flexible conformation, the electron density of the sidechain cannot be captured in the diffraction experiment, the resulting sidechain conformation is therefore a "prediction" of the crystallographer. [76]

In this series of GB1 SeM variants, SeM5 and SeM39 sidechains are classified as Type I conformers. SeM6 is a Type II conformer, its sidechain clearly described by two discrete conformations due to discrete locations of the Se_{ε} position. SeM29 and SeM34 are both cloud conformers, as the positions of their selenium atoms cannot be pinpointed, but rather exist over a continuous region (Figure 30). So far no SeM sidechain falls into Type IV conformers.

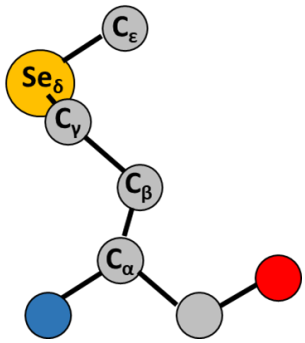
According to *Miao et al.*, the likelihood of a residue to adopt alternate conformations is related to the degree of freedom (number of χ dihedral angles) available to its sidechain. Therefore, residues with longer sidechains such as Arg, Lys,

Glu, Gln and Met are more likely to have alternate conformations than residues such as Cys with one dihedral angle. In the case of GB1, both solvent-exposed SeM sidechains (SeM6, SeM29) exhibit conformational flexibility, which agrees with the general assumption that exposed residues are less restrained and more likely to include such structural variations. However, it is worth noting that, at least for the case of GB1, most solvent-exposed residues are still clearly modeled by a single conformation. The reason for this could be that their locations are constrained by interaction partners, and the relative small size of GB1 may also contribute to structural rigidity.

4.4.2 Conformational Characteristics of the SeM sidechain

Bond angles and torsion angles for each SeM sidechain is summarized in Table 6. A total of 11 individual SeM sidechains have been modeled for five GB1 SeM structures, including alternate conformations and multiple copies in the asymmetric unit. Average bond angle about the selenium atom (*C-Se-C*) is 99.5° and torsion angle χ_3 clusters around 60° and 300°. These measurements are consistent with the reports by *Virrueta et al.* from a survey of the Dunbrack 1.7 Å and small molecule databases, which cited for selenomethionine average *C-Se-C* bond angle of 98.3° (±2.2°) and sidechain dihedral distribution of P(χ_3) at 60° (41%) and 300° (42%), the remaining 17% represents a third smaller cluster at 180°. [71] This bimodal distribution is characteristic of methionine and selenomethionine, both of which incorporating a chalcogen at the sidechain δ position. The distribution is significantly different for norleucine (Nle),

where Se_δ is replaced by C_δ in the sidechain, and for whom nearly 80% of the χ_3 dihedral angle adopts 180° conformation, and the remaining 20% is evenly distributed between the 60° and 300° bins. In all three cases, Met, SeM and Nle, the minima angles χ_3 at 120° and 240° are caused by clashes between the Cε and the hydrogen atoms on Cγ, as well as hydrogens on Cε and Cγ. [71]



Computational prediction of the dihedral angles for about half of all canonical amino acids have been successful at recapitulating the distributions of most of the torsion angles, and in the case of Met, χ_1, χ_2 were successfully calculated but χ_3 with sulfur or selenium have not been accurately captured. [71] The presence of the chalcogen complicates computational efforts for reasons not yet fully elucidated. However, using the hard sphere dipeptide model, *Vittueta et al*. proposed the addition of attractive forces between the hydrogen atoms in the amino acids which yields a closer prediction of the distribution of the χ_3 angle in Met and SeM and without significant perturbations to the calculations of other torsion angles. [71]

4.4.3 Future Implications

The structures of five GB1 SeM variants were obtained via X-ray crystallography all at atomic resolution (≤ 1.2 Å). Each variant was designed to carry one SeM amino acid residue at a structurally distinct location on the polypeptide sequence. Of the final 11 sidechains modeled for SeM, including alternate conformations, the majority display strong enough electron density around the terminal methyl group to unambiguously model SeM conformation. Confidence in the conformation of the SeM is essential for subsequent theoretical studies, as the structures obtained in this work will be used as the starting place for any computational calculations downstream. Anomalous data acquired at the selenium absorption edge was also used to assign the specific location of each selenium atom. In addition, most water molecules and co-crystallized solvent molecules such as MPD, acetate and phosphate ions were observed which is especially important if the solvent ions are found near the selenium and could influence its local environment. Taken together, these structures serve as the road map for the compilation of SeM sidechain interactions and of the selenium atom itself. This set of data represents a systematic approach to catalog the interactions of SeM, which is complementary to the crystal structures deposited in the Protein Data Bank, where SeM is found on different proteins with different resolution ranges and levels of refinement appropriate for that resolution only.

An exciting discovery from this study was the co-crystallization of imidazole in the A34SeM variant. In this structure, the imidazole molecule is located close to Trp43

whose position became dislocated with the Ala-to-SeM mutation. Therefore, it is highly plausible that imidazole plays a stabilizing role in the protein structure and could have been the deciding factor in obtaining atomic resolution X-ray diffraction data.

The same logic was applied to the sixth GB1 SeM variant, V54SeM. Val54 is an immediate neighbor to Trp43, where Trp43 is found on the outer strand 3 at the crystal contact interface, and Val54 is found on one of the inner strands number 4. Mutation of the smaller valine to a larger selenomethionine could disrupt the crystal interface, which contribute to the poor quality of the crystals grown from this variant. Several attempts yielded crystals that lacked birefringence, almond-shaped and most importantly, did not diffract any X-ray. After obtaining the A34SeM structure, additional efforts were directed to grow V54SeM crystals by applying conditions with imidazole. At the time this dissertation is being finalized, V54SeM crystals have successfully diffracted to 2.2 Å. Efforts to obtain a fully refined crystal structure is still on-going. As an added benefit, the effect of point mutation on protein structure can be investigated, following the footsteps of a myriad of work in this field, such as that of *Smith et al. 1994*. [57]

4.5 Summary of Chapter 4

In this chapter, the five crystal structures of GB1 SeM variants were examined in detail. Each crystallized in the monoclinic space group P 1 21 1, and exhibited the same crystal contact surface along the edges of the second and third β-strand. The

structures were obtained at atomic resolution. The final refined models verified by simulated annealing OMIT maps for both the entire molecule and also each SeM sidechain and bound ligands (Figure 28, 29). The region of crystal contact is highly hydrogen bonded which promotes crystal growth. The sequence of each SeM variant is analogous to the wild type, and as expected there are minimal structural differences between them. The largest structural perturbation was observed for the A34SeM variant, as the substitution of SeM34 dislodges Trp43 and other nearby residues from their native positions, and as a result an imidazole molecule was bound to the protein core. Within the vicinity of SeM are aliphatic residues, predominately Leu and Val. These interactions are cataloged and will be crucial in understanding Met interactions in proteins. In summary, through these five crystal structures one can obtain a direct visualization of the local electronic environment of the SeM residue. The structural information will be used in theoretical calculations to predict the chemical shift of the selenium nucleus in the specific protein environment that is captured by the crystal structures. A relationship between predicted and observed chemical shifts of the selenium assigns the NMR behavior to a given electronic structure, providing the information needed to build a database of ^{77}Se selenium NMR.

Table 5. Single crystal X-ray diffraction data for GB1 SeM variants

	L5Sem	*I6Sem*	*V29Sem*	*A34Sem*	*V39Sem*
Wavelength	0.9795	0.9795	0.9795	0.9795	0.9795
Resolution range	25.74 - 1.2 (1.243 - 1.2)	27.57 - 1.12 (1.16 - 1.12)	29.1 - 1.2 (1.243 - 1.2)	26.96 - 1.1 (1.139 - 1.1)	27.17 - 1.2 (1.243 - 1.2)
Space group	P 1 21 1	P 1 21 1	P 1 21 1	P 1 21 1	P 1 21 1
Unit cell	27.844 36.43 48.963 90 99.273 90	27.8898 36.489 48.9398 90 98.6744 90	27.506 36.497 48.843 90 99.237 90	22.965 37.048 28.719 90 110.164 90	27.592 36.552 48.978 90 100.027 90
Total reflections	58904 (5615)	71575 (6899)	58077 (5491)	36693 (3551)	57403 (5335)
Unique reflections	29585 (2844)	36641 (3554)	29169 (2768)	18407 (1792)	30116 (2940)
Multiplicity	2.0 (2.0)	2.0 (1.9)	2.0 (2.0)	2.0 (2.0)	1.9 (1.8)
Completeness (%)	96.71 (94.28)	97.43 (94.78)	97.03 (93.82)	99.79 (98.79)	99.67 (99.02)
Mean I/sigma(I)	16.85 (5.61)	12.98 (6.89)	9.36 (3.94)	11.55 (7.46)	8.92 (3.34)
Wilson B-factor	7.74	7.95	9.47	6.66	8.45
R-merge	0.05208 (0.4714)	0.01784 (0.05328)	0.02561 (0.1201)	0.02873 (0.05434)	0.02903 (0.1944)
R-meas	0.07366 (0.6666)	0.02524 (0.07535)	0.03622 (0.1698)	0.04063 (0.07685)	0.04105 (0.2749)
R-pim	0.05208 (0.4714)	0.01784 (0.05328)	0.02561 (0.1201)	0.02873 (0.05434)	0.02903 (0.1944)
CC1/2	0.992 (0.429)	0.999 (0.996)	0.997 (0.979)	0.998 (0.99)	0.999 (0.963)
*CC**	0.998 (0.775)	1 (0.999)	0.999 (0.995)	0.999 (0.997)	1 (0.99)
Reflections used in refinement	29431 (2835)	36580 (3540)	29132 (2763)	18402 (1792)	30074 (2933)
Reflections used for R-free	996 (95)	1996 (194)	1283 (122)	1581 (153)	1332 (131)
R-work	0.1930 (0.2419)	0.1453 (0.1253)	0.1616 (0.1926)	0.1746 (0.1624)	0.1629 (0.2218)
R-free	0.2160 (0.2704)	0.1615 (0.1591)	0.1869 (0.2338)	0.1853 (0.1632)	0.1793 (0.2328)
CC(work)	0.961 (0.903)	0.967 (0.985)	0.970 (0.967)	0.962 (0.962)	0.971 (0.956)
CC(free)	0.941 (0.882)	0.961 (0.984)	0.944 (0.904)	0.947 (0.961)	0.962 (0.950)
Number of non-hydrogen atoms	1096	1129	1077	555	1090
macromolecules	878	898	879	437	875
ligands	42	52	40	21	70
solvent	176	179	158	97	145
Protein residues	112	112	112	56	112
RMS(bonds)	0.010	0.009	0.007	0.007	0.006
RMS(angles)	1.44	1.48	1.15	1.27	1.23

Ramachandran favored (%)	98.15	98.15	98.15	98.15	97.22
Ramachandran allowed (%)	1.85	1.85	1.85	1.85	2.78
Ramachandran outliers (%)	0.00	0.00	0.00	0.00	0.00
Rotamer outliers (%)	0.00	2.08	0.00	0.00	0.00
Clashscore	5.52	5.84	4.37	5.50	3.81
Average B-factor	14.25	13.88	15.67	11.98	14.61
macromolecules	11.30	10.53	12.61	8.96	11.45
ligands	30.29	34.12	30.86	25.97	34.89
solvent	25.18	24.80	28.87	22.56	23.92

Table 6. Bond angle and torsion angles of the SeM sidechain in each GB1 variant. The bond angles are identified separately for each chain. For A34SeM, only one chain crystallized in the asymmetric unit. For I6SeM, two conformations of SeM were modeled in each chain, resulting in two angle measurements per chain.

	C-Se-C (A)	*C-Se-C (B)*	*χ_1 (A)*	*χ_2 (A)*	*χ_3 (A)*	*χ_1 (B)*	*χ_2 (B)*	*χ_3 (B)*
L5SeM	100.7	99.9	169.4	176.1	52.1	176.5	180.0	49.4
I6SeM	99.3 99.1	100.0 98.9	297.7 309.6	179.2 292.4	69.5 74.4	295.5 310.9	180.7 291.9	68.1 282.7
V29SeM	104.6	94.3	296.7	174.4	314.9	295.7	166.1	86.5
A34SeM	97.9	-	290.1	162.3	53.2	-	-	-
V39SeM	100.5	99.9	304.6	184.2	85.5	306.8	182.8	86.4

Table 7. Interactions within approximately 4.5 Å of SeM sidechain in each GB1 SeM variant. In each interaction, the functional group of the sidechain or peptide backbone that is within the interaction distance to the selenium atom is identified in parentheses, followed by the interaction distance. Two entries are made for I6SeM, as the SeM sidechain was modeled with two conformations on each chain. The dominant conformation is indicated by an asterisk (*). (BB: backbone)

	Chain A		Chain B	
	Interactions with Sidechains	*Interactions with Backbone or Ligands*	*Interactions with Sidechains*	*Interactions with Backbone or Ligands*
L5SeM	Leu7 (Cδ2) – 3.96 Leu7 (Cγ) – 4.34 Phe30 (Cβ) – 3.78 Phe30 (Cγ) – 4.11 Trp43 (CH2) – 3.70 Trp43 (CZ3) – 3.84 Val54 (Cγ1) – 4.14 Ala34 (Cβ) – 4.82	None	Leu7 (Cδ2) – 4.05 Leu7 (Cγ) – 4.38 Phe30 (Cβ) – 3.80 Phe30 (Cγ) – 4.11 Trp43 (CH2) – 3.65 Trp43 (CZ3) – 3.84 Val54 (Cγ1) – 4.13 Ala34 (Cβ) – 4.78	None
I6SeM Conf. 1	Thr51 (Cγ2) – 4.00 Thr53 (Cγ2) – 4.04 Lys4 (NZ) – 4.62	MPD (C5) – 4.69 H_2O – 2.15	Thr51 (Cγ2) – 4.11 Thr53 (Cγ2) – 4.06 Lys4 (NZ) – 4.80	MPD (C5) – 4.65 No water
I6SeM Conf. 2	Lys4 (Cε) – 3.60 Lys4 (NZ) – 3.87 Thr51 (Cγ2) – 4.68 Glu15 (Oε2) – 3.18	Gly14 (BB CO) – 4.77	Thr51 (Cγ2) – 4.62 Glu15 (Oε2) – 3.30	None
V29SeM	Ala20 (Cβ) – 4.75 Thr18 (Oγ1) – 4.48 Thr18 (Cγ2) – 4.79 Thr25 (Cγ2) – 4.30 Ala26 (Cα) – 4.87	Glu19 (BB CO) – 4.49 Thr25 (BB CO) – 4.09 MPD (O4) – 4.39	Ala20 (Cβ) – 4.15 Thr18 (Oγ1) – 4.06 Thr18 (Cγ2) – 4.54 Thr25 (Cγ2) – 3.98 Ala26 (Cα) – 4.41	Glu19 (BB CO) – 3.99 Thr25 (BB CO) – 3.77 H_2O – 3.85
A34SeM	Leu5 (Cδ1) – 4.62 Leu5 (Cδ2) – 4.48 Phe30 (Cβ) – 4.83 Lys31 (Cα) – 4.36 Val54 (Cγ1) – 4.65 Val54 (Cγ2) – 3.97	Phe30 (BB CO) – 3.71 Val39 (BB CO) – 4.92 Imidazole – 4.10	-	-

	Leu7 (Cδ1) – 4.95			
V39SeM	Leu7 (Cδ2) – 3.71 Leu12 (Cβ) – 4.71 Leu12 (Cγ) – 4.70 Leu12 (Cδ2) – 4.05 Ala34 (Cα) – 4.05 Ala34 (Cβ) – 3.98 Val54 (Cγ1) – 4.10 Val54 (Cγ2) – 4.78	Ala34 (BB CO) – 4.44 H_2O – 3.86	Leu7 (Cδ2) – 3.67 Leu12 (Cβ) – 4.77 Leu12 (Cγ) – 4.63 Leu12 (Cδ2) – 4.01 Ala34 (Cα) – 4.08 Ala34 (Cβ) – 3.93 Val54 (Cγ1) – 4.15 Val54 (Cγ2) – 4.87	Ala34 (BB CO) – 4.35 H_2O – 4.93

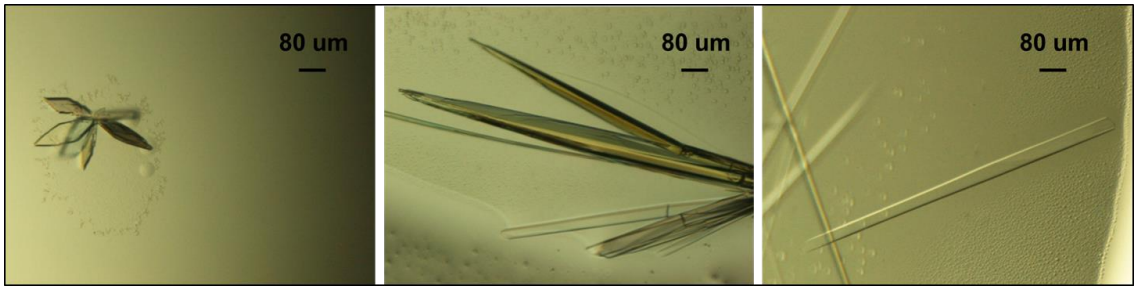


Figure 12. Representative crystals grown of the GB1 V39SeM variants. The crystals display dendritic morphology as a result of growing from a common nucleation point. Diffraction-quality crystals were harvested from regions of single crystal growth, taking care to avoid areas of crystal splinter where the lattice may be twinned.

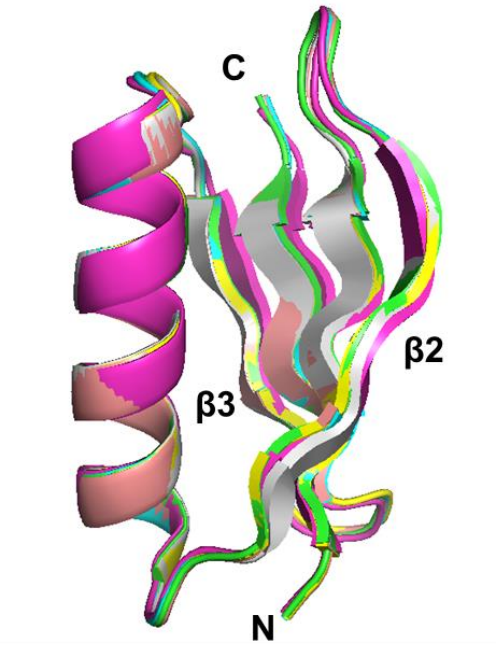


Figure 13. Cartoon overlay of GB1 wild type protein (PDB Accession Code: 2QMT) with five SeM variants: L5SeM, I6SeM, V29SeM, A34SeM and V39SeM. Global structure of the variants close resemble that of the wild type with RMSD of 0.232.

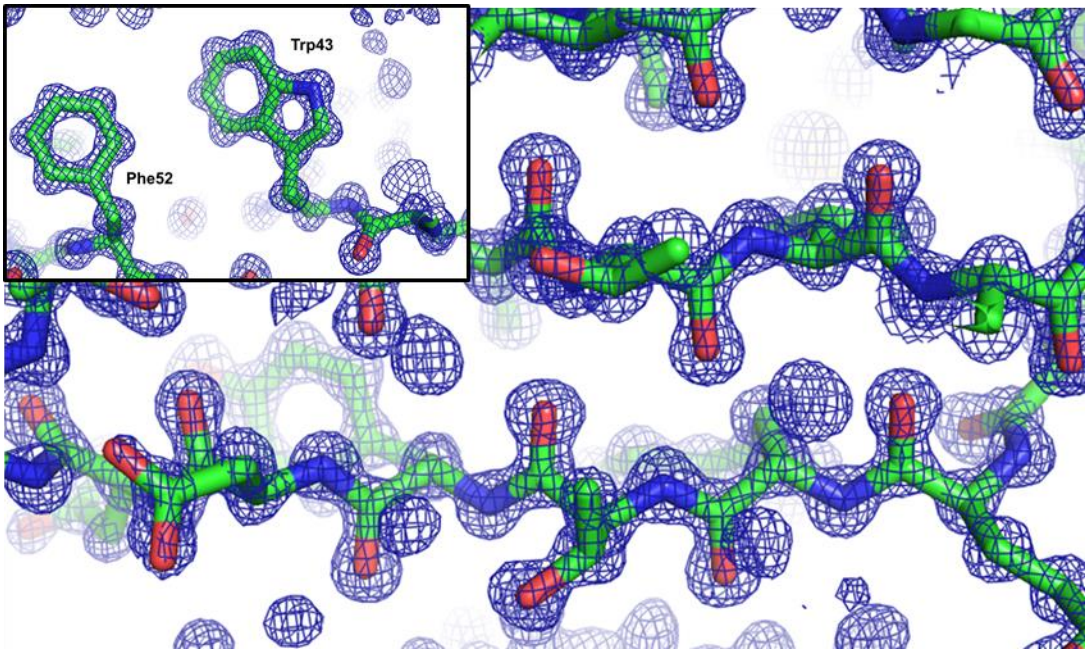


Figure 14. Representative $2F_o - F_c$ electron density map at 1.10 Å for GB1 A34SeM variant, calculated from the iteratively refined atomic model (Phenix map calculation). For clarity water molecules are not shown. **Inset:** local electron density map of aromatic residues at atomic resolution. The holes in the center of Trp43 indole and phenyl rings and the phenyl ring of Phe52 are both clearly visible at this resolution.

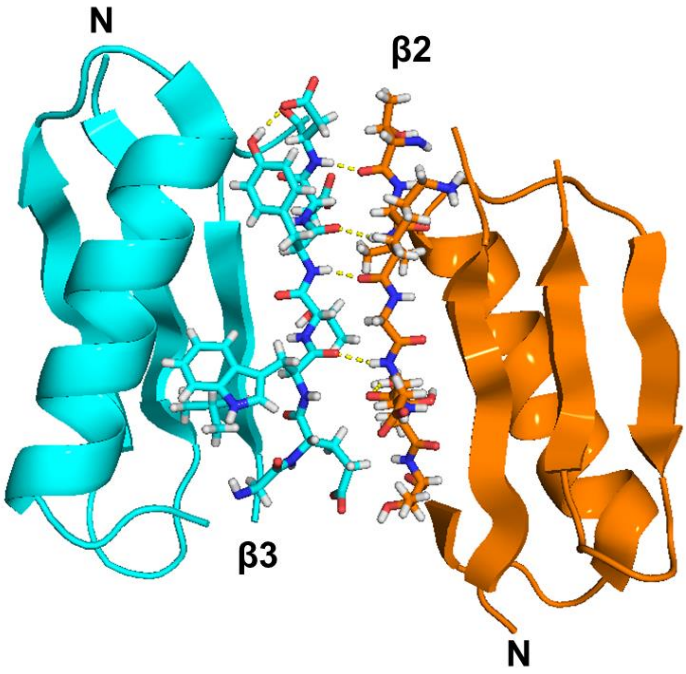


Figure 15. Representative crystal contact surfaces shown of L5SeM formed between strand 2 of one copy of the molecule (cyan) and strand 3 of its neighboring molecule (orange) in the crystal lattice. The crystal contact region is shown in sticks Strand 2 includes residues TLKGETT (amino acids 11 to 17) and strand 3 includes residues GEWTYDD (amino acids 41 to 47). Polar contacts are shown in yellow dashes.

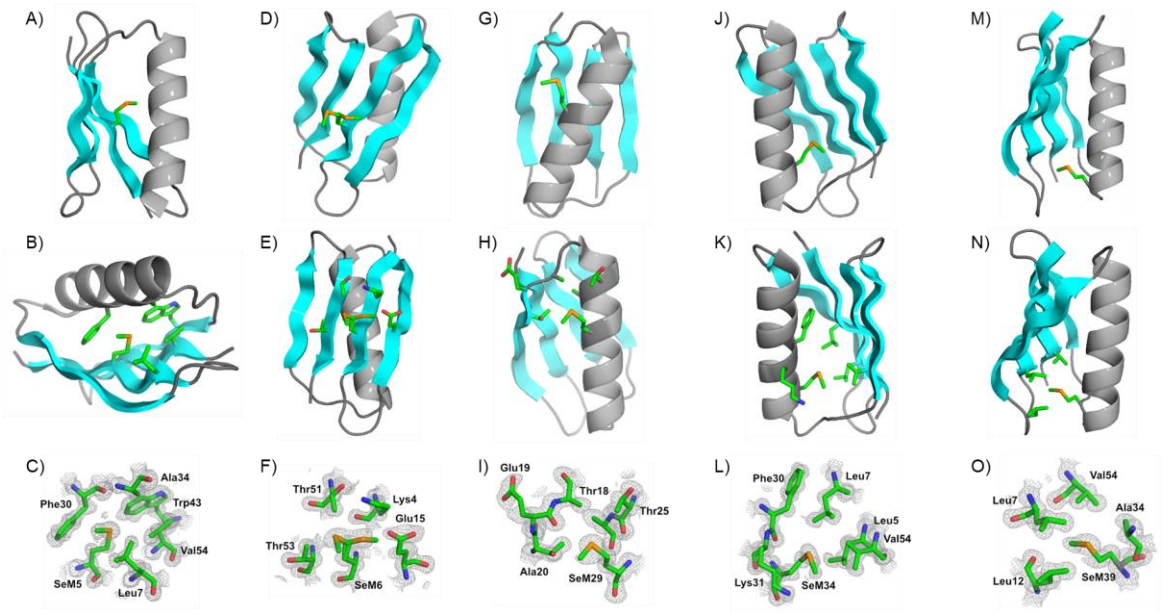


Figure 16. Overall crystal structures of five GB1 SeM variants. A) L5SeM overall structure showing location of SeM within the protein for subunit A. The SeM sidechain is shown in sticks, the selenium atom is colored orange; B) SeM5 sidechain and residues located within 4.5 Å of the selenium atom; C) SeM5 siddechain and residues within 4.5 Å of the selenium atom with $2F_o$-F_c electron density map. Map is contoured at 1.0 σ. Each residue is shown in sticks, with oxygen in red, nitrogen in blue, carbon in green, and selenium in orange; D-F) I6SeM showing the structure of subunit B; G-I) V29SeM subunit A; J-L) A34SeM; M-O) V39SeM subunit A. SeM6 sidechain is modeled with two alternate conformations.

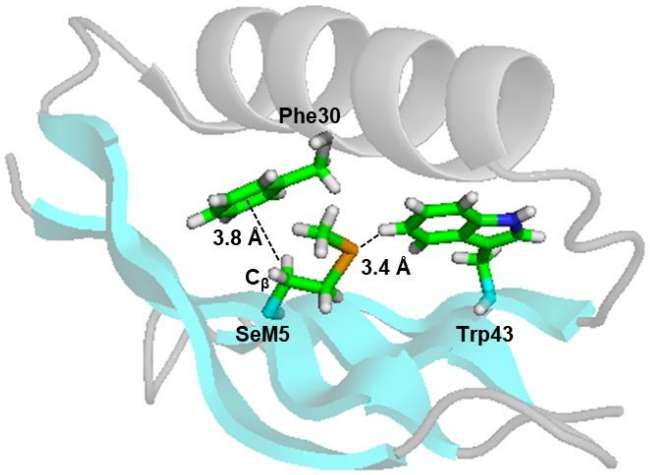


Figure 17. SeM5 sidechain interaction with nearby aromatic resiues in the GB1 L5SeM variant. The selenium atom approaches the edge of Trp43 at 3.4 Å, unlikely to have hydrogen bonding characteristics. The C-Se-C bond angle is 100.65° for the SeM sidechain. Typical bond angle reported for C-S-C angle is between 95° to 100°. [70] This value is comparable to the selenium angle reported here.

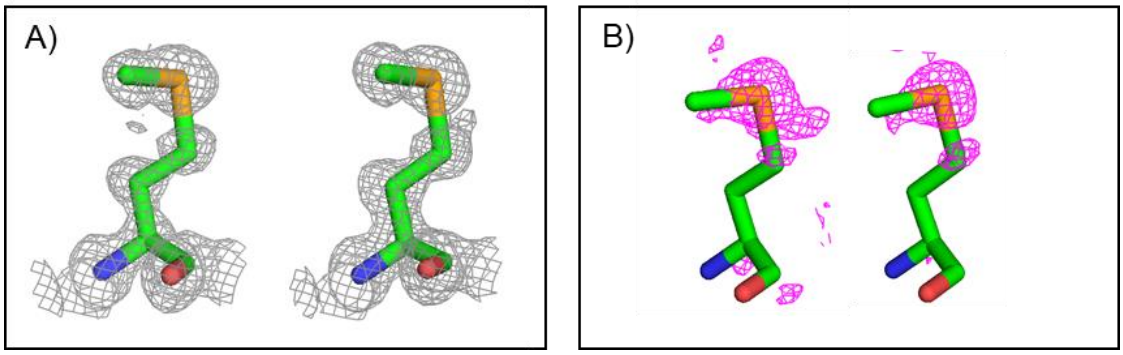


Figure 18. Electron density surrounding SeM5 sidechain. A) $2F_o$-F_c map shown in grey surrounding SeM5. One conformation of the sidechain is clearly assigned for both subunits A (Left) and B (Right) of the asymmetric unit. Electron density of the terminal methyl group is also obsevred, which enables the measurement of the C-Se-C angle, as well as the C-C-Se-C χ_3 torsion angle around the selenium atom. B) Placement of the selenium atom was aided by the anomalous diffraction map. Electron density shown in magenta is that acquired of the heavy atom at 0.9795 Å, for subunits A (Left) and B (Right) in the asymmetric unit. The selenium atom is colored orange.

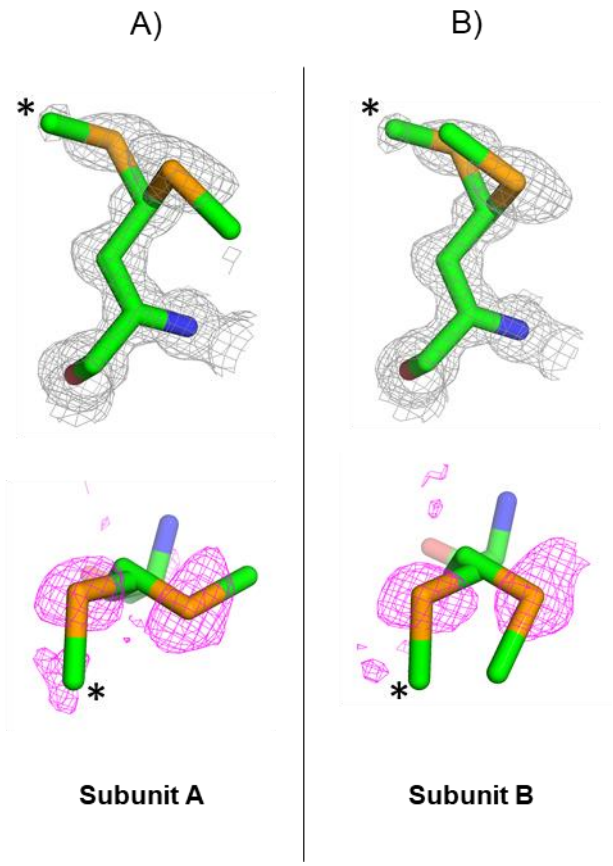


Figure 19. *2F$_o$-F$_c$* (Grey) and anomalous diffraction (Magenta) electron density map for SeM6 sidechain in subunits A (Left) and B (Right) of I6SeM variant. The dominant conformation is marked with an asterisk (*). The *2F$_o$-F$_c$* electron density surrounding Cε is observed for the dominant conformation, which allows accurate measurement of the bond angles surrounding the selenium atom. Assignment of two conformations for each SeM6 sidechain is guided by both the *2F$_o$-F$_c$* map, as well as the anomalous map which specifically locates the selenium.

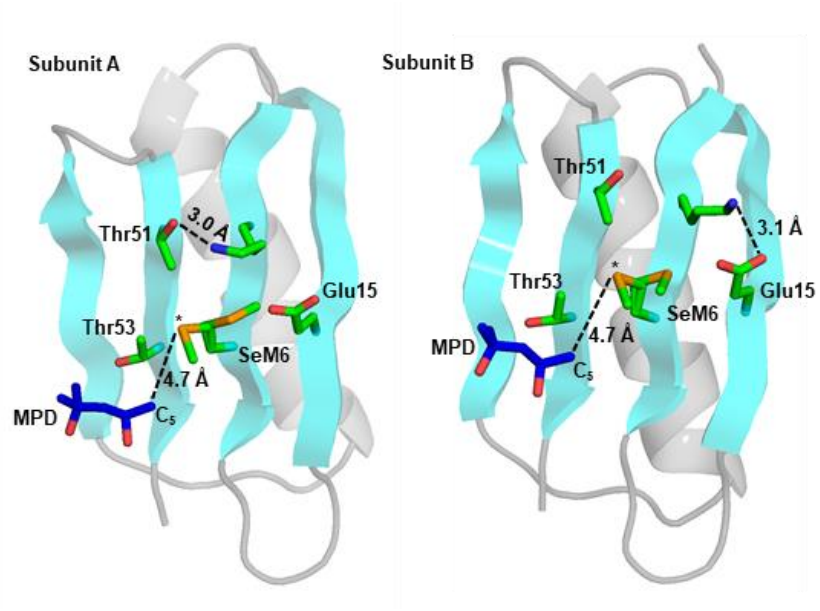


Figure 20. Interactions within 4.5 Å of the selenium atom in Sem6 sidechain. Each interacting residue is shown in sticks. Carbon is colored green, oxygen in red, nitrogen in blue, and selenium in orange. The interacting ligand, MPD, is colored deep blue, with oxygens in red. In each subunit, two conformationsn of SeM6 sidechain were modelled. The dominant conformation is marked with an asterisk (*). This conformation is within 5 Å of the Cγ methyl groups of Thr51 and Thr53, as well as C_5 methyl group of nearby MPD. Lys4 forms a salt bridge with the hydroxyl group of Thr51 in subunit A, and with the carboxyl group of Glu15 in subunit B. However, in each case the electron density surrounding the distal atoms of Lys4 are weak, therefore its location cannot be accurately placed. This flexibility of Lys4 to form two favorable electrostatic interactions may contribute to its weak electron density.

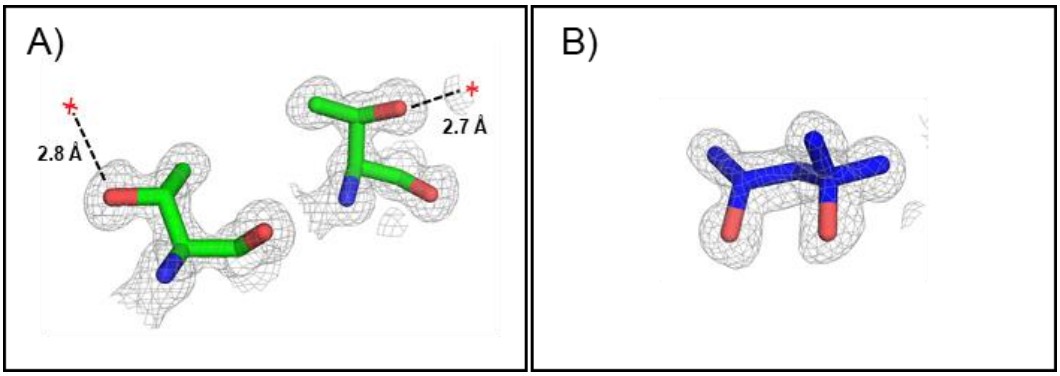


Figure 21. Representative $2F_o$-F_c electron density map for A) Thr51, Thr53 sidechains and B) MPD precipitant used in crystallization cocktail which co-crystallized with the protein. The crystal structure is refined at 1.12 Å. The density map is contoured at 1.0 σ. The electron density shows distinct differences surrounding the methyl and hydroxyl groups which facilitates the assignment of each functional group in the Thr sidechain. In addition, water molecules located within hydrogen bond distance confirms the assignment of the hydroxyl functional group. [72] Differences in electron density facilitates assignment of the functional groups in MPD, however the nearest water molecule is located 4 Å away and cannot reliably be used for placement of the functional group (not included).

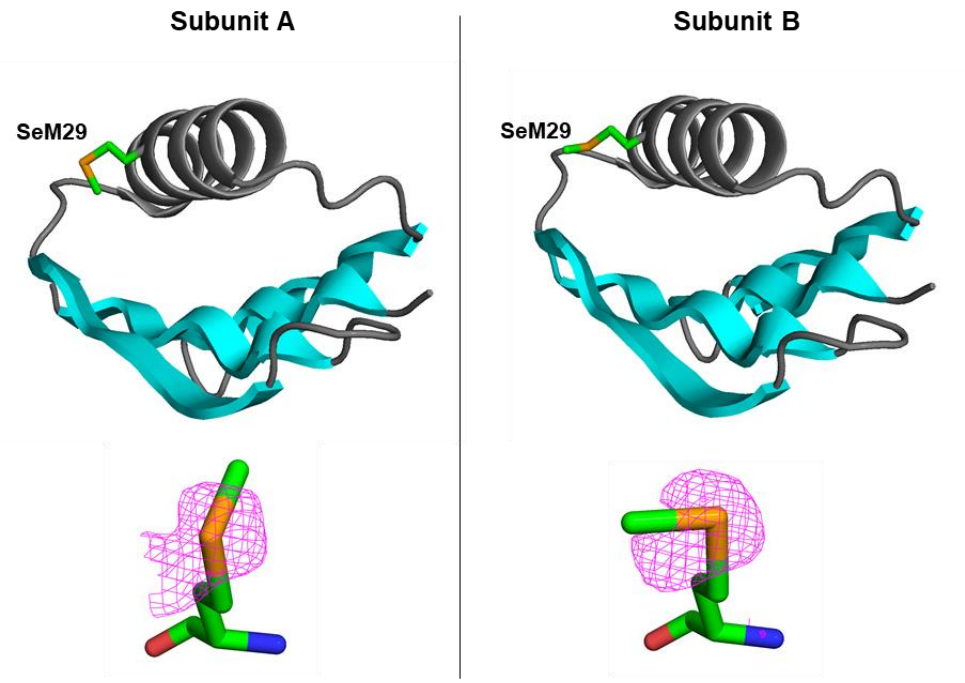


Figure 22. Orientations of the SeM29 sidechains in GB1 V29SeM variant subunit A (Left) and B (Right). The SeM sidechain is flexible at this location, which is further confirmed by the anomalous diffraction map (magenta). The selenium atom cannot be accurately located, but rather occupies multiple positions in the crystal.

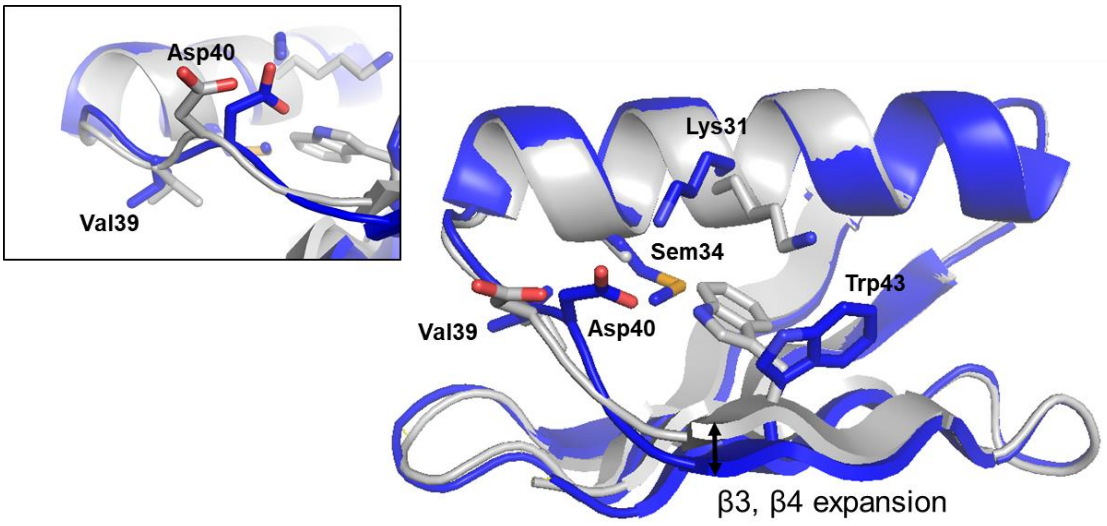


Figure 23. Core expansion caused by Ala-to-SeM mutation at position 34. Silver: structure of 2QMT showing the wild type positions of residues immediately surrounding the mutation site. Blue: structure of GB1 A34SeM showing altered positions of the residues as a result of the Ala-to-SeM substitution. Residues Lys31 and Trp43 are pushed outward from their wild type positons, while Asp40 located on the flexible loop bends inward possibly to compensate for the expansion of the core. The start of β-strand 3 is expanded outward by 1 Å compared to the wild type position. The strand gradually returns to wild type configuration as the strand proceeds. Similarly, the end of β-strand 4 shows similar displacement by 1.1 Å, as this is also close to the site of mutation. **Inset:** The sidechain of Val39 and Asp40 are both twisted inward to form an end "cap" to shield the core from expansion.

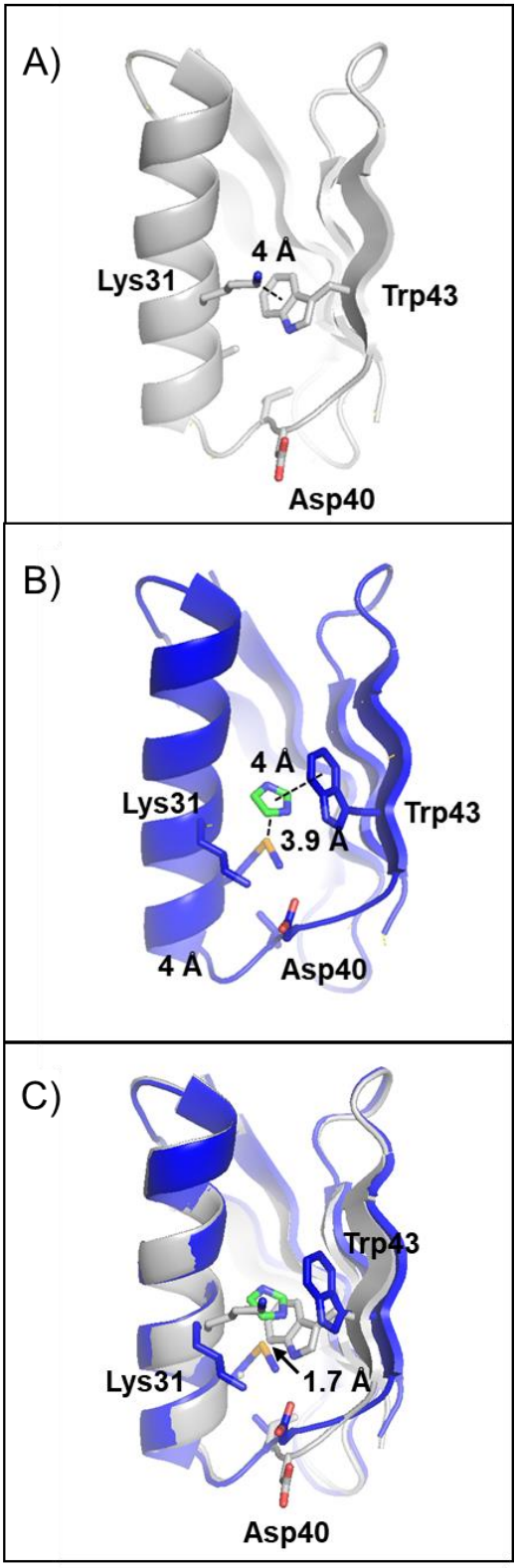


Figure 24. Conformaion rearrangement in the core of GB1 A34SeM variant compared to the wild type structure.

A) Silver: Trp43 in the wild-type structure 2QMT. The center of the Trp phenyl ring is directy across Lys31. Trp43 is completely solvent shielded in the wild type.

B) Blue: Lys31, Trp43 sidechains pushed outward by SeM34 mutation. An imidazolium molecule is found approximately halfway between SeM34 and Trp43, presumably acting as a stabilizing force to replace the displaced Trp sidechain. In this conformation, Trp43 becomes partially exposed to the solvent. Expansion of the protein core likely stabilized by Asp40 turning inward.

C) Overlay of 2QMT with GB1 A34SeM variant. Arrow points to the steric clash if the position of Trp43 remains unchanged. Conformational change of Asp40 is also depicted.

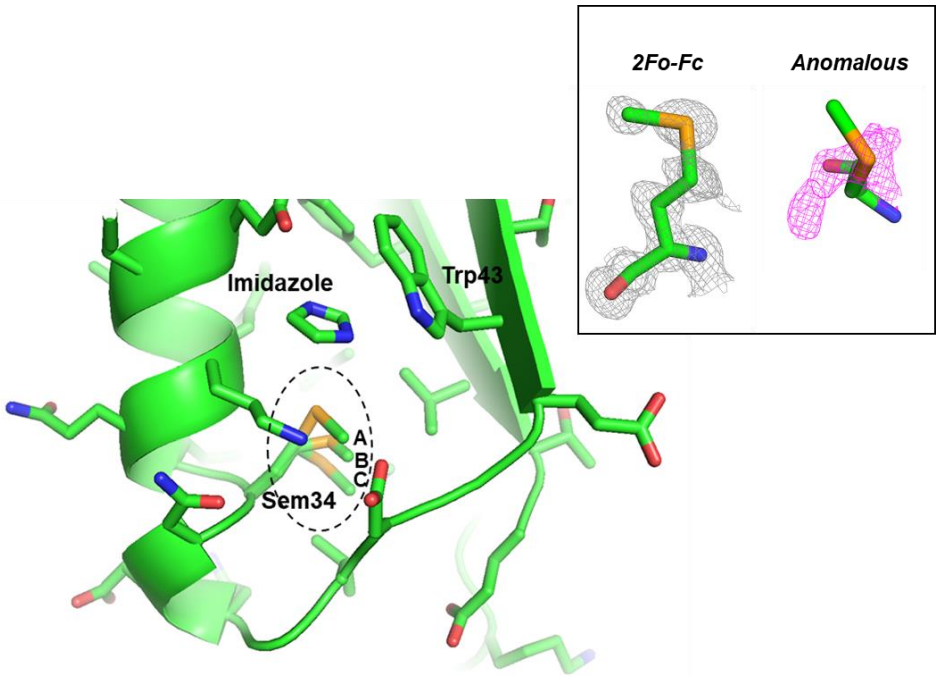


Figure 25. GB1 A34SeM variant showing alternate conformations of the SeM34 sidechain. The three conformations from top to bottom are A, B and C. Conformation A is the most populated species. In B and C, the sidechain of SeM34 is shifted downward and away from Trp43 to within 4 Å of both Leu7 and Val54. **Inset:** (Left) $2F_o$-F_c electron density map of the most populated conformation A. Electron density around Cε of the SeM sidechain is clearly observed. (Right) anomalous diffraction of the selenium atom shown in aerial view. More than one conformation may be accessible to the SeM sidechain. However, in the final refined model, only the most chemically feasible conformation was modeled.

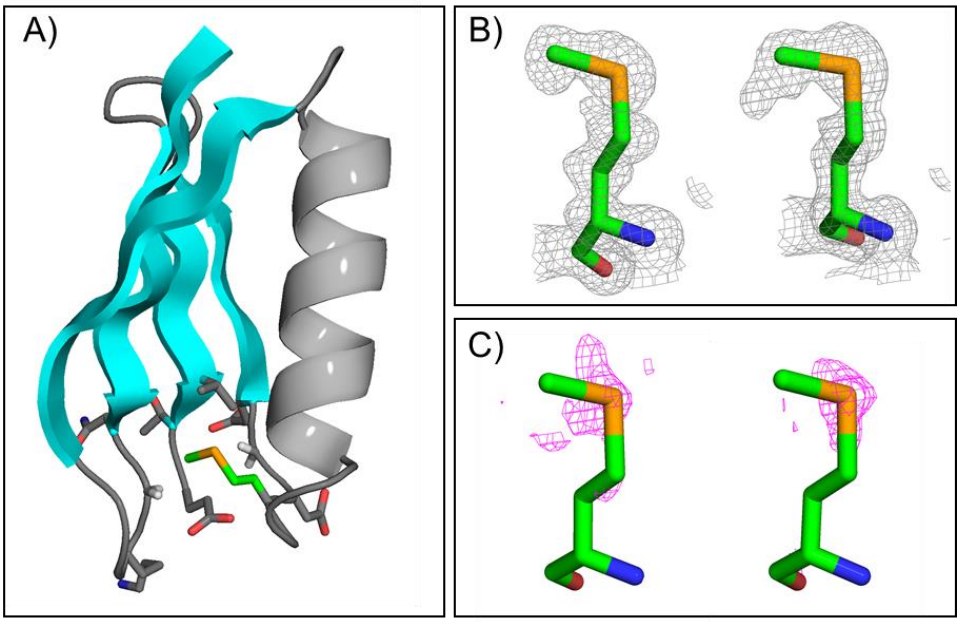


Figure 26. GB1 V39SeM variant. A) Immediate vincinity surrounding the SeM39 residue. The SeM sidechain is located on the loop positioned in the protein core. The mutation is accomondated spatially by nearby residues and does not alter the overall structure of the protein. The SeM residue is colored green, the selenium atom is colored orange. Neighboring atoms are colored according to the corresponding secondary structure. B) The SeM sidechain is modeled with one conformation, guided by the density maps shown in grey. C) Anomalous diffraction map acquired of the selenium atom further confirms the conformation of the sidechain in each subunit.

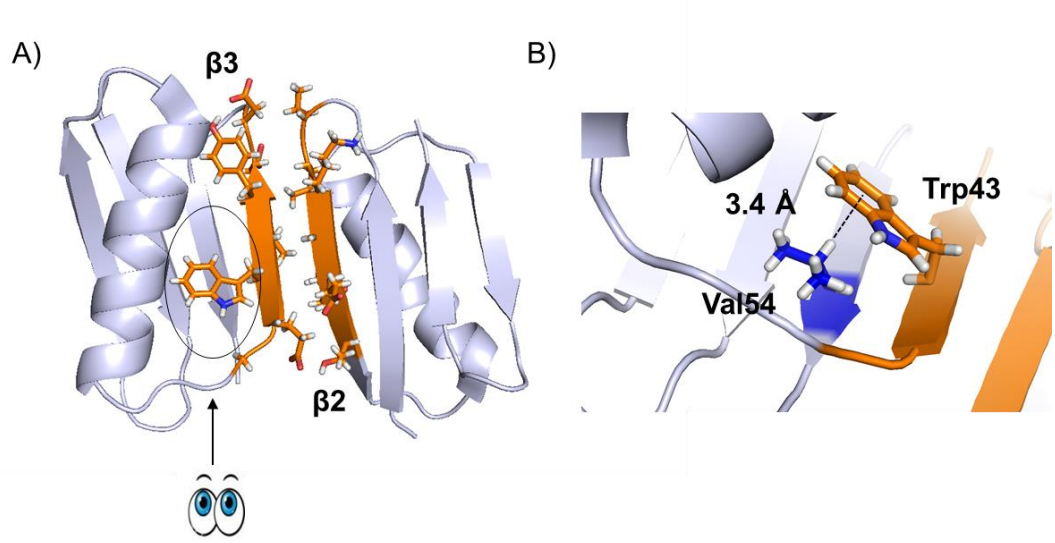


Figure 27. The structure of 2QMT showing the immediate vicinity of Val54. A) Crystal contact for current five GB1 SeM variants. B) Placement of Val54 with respect to Trp43. Val54 is a solvent inaccessible residue located 3.5 Å away from Trp43. By mutating Val to SeM, it could disrupt the core of the protein and displace the position of Trp43. Trp43 is located on the β strand resoponsible for forming the crystal contact in these SeM muants. Such positional alterations may be the cause that no diffraction quality crystals of V54SeM were obtained.

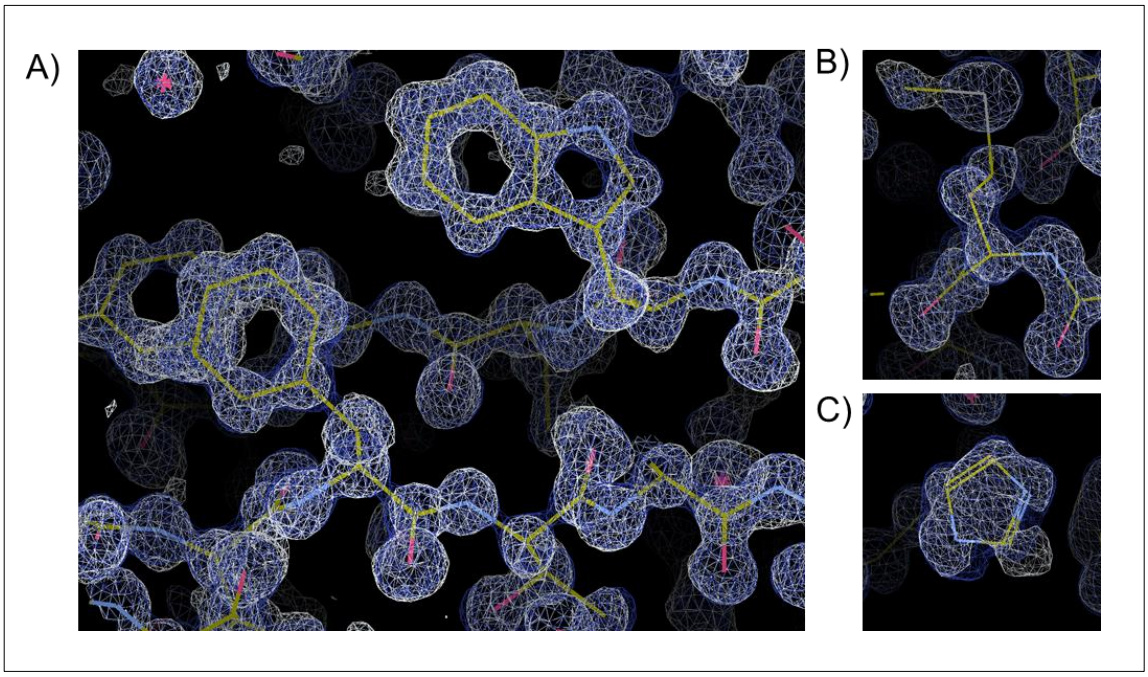


Figure 28. Representative simulated annealing OMIT maps (white) compared to the experimental calculated $2F_o$-F_c maps for A) Whole structure, showing the electron density holes in both tryptophan and phenylalanine residues are clearly visible. B) SeM39 sidechain in subunit A, and C) imidazole bound to A34SeM variant.

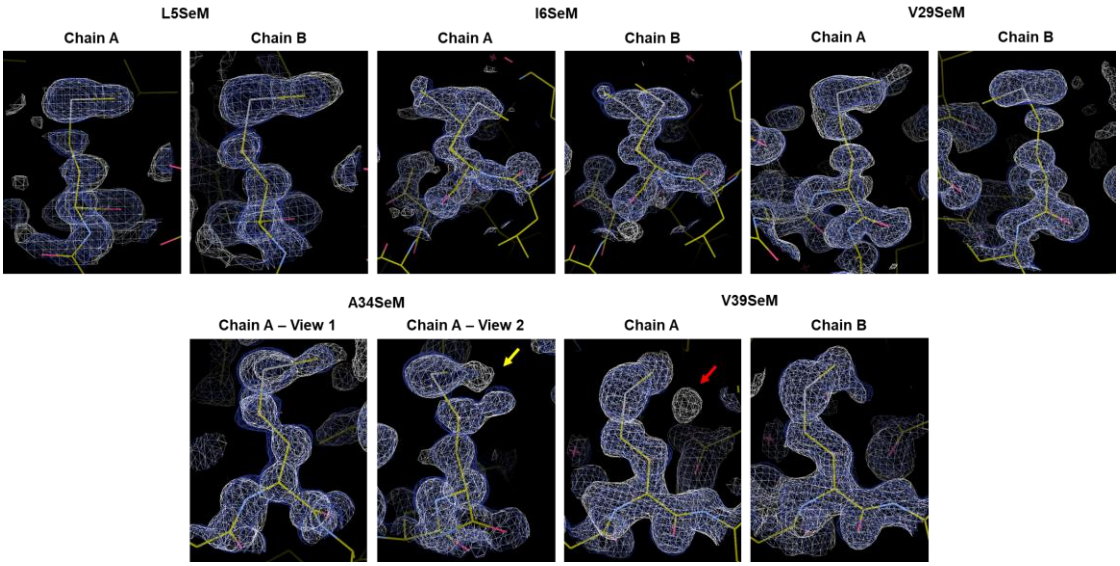


Figure 29. Composite OMIT maps with simulated annealing for the SeM residue in five GB1 SeM variants. Blue: experimental $2F_o$-F_c electron density maps. White: Simulated OMIT electron density maps. The SeM residue for A34SeM variant shows two views of the same residue, as A34SeM contains only one chain in the asymmetric unit. In view B, the presence of at least one other conformation of SeM is confirmed by the OMIT map (yellow arrow). In V39SeM, the possibility of a bound water molecule or ligand is confirmed by the OMIT map (red arrow). Attempts were made to model this ligand in the protein structure. Neither the placement of water nor sodium ion was justified through refinement.

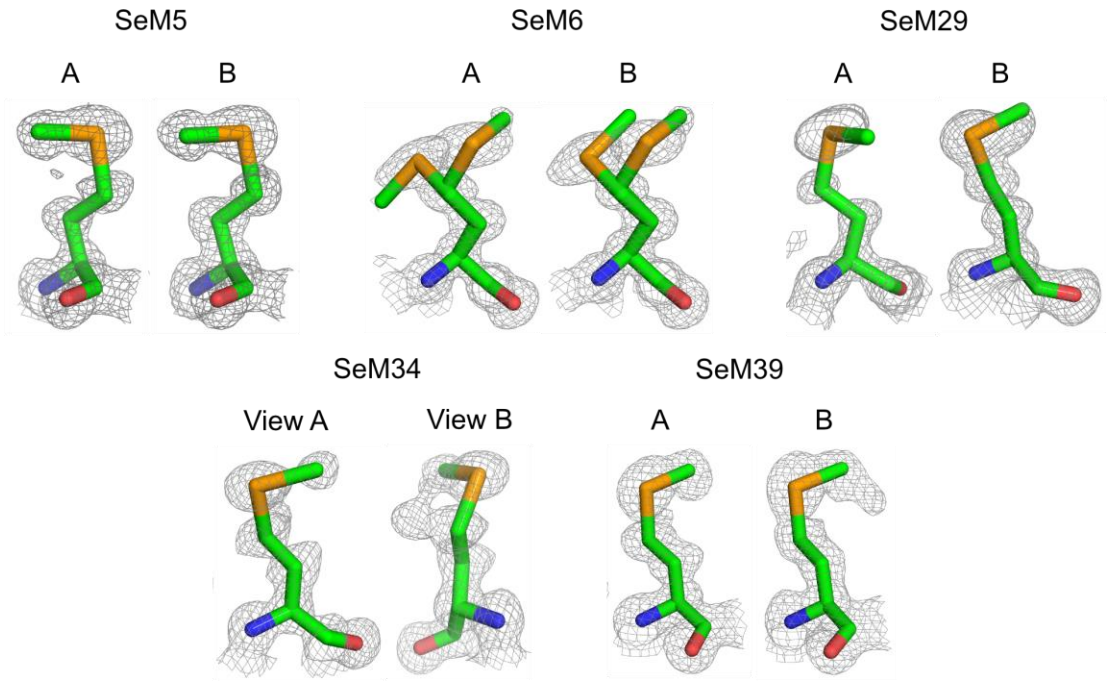


Figure 30. Sidechain conformational variations of SeM sidechains with local $2F_o$-F_c electron density maps. SeM5 and SeM39 in L5SeM and V39SeM variants each has one definite location, resolution 1.2 Å. SeM6 has two alternate locations per subunit. In each subunit, the dominant state (shown on right) has 0.80 occupancy. In the less occupied state, the electron density at the terminal methyl group is missing. Therefore, the exact conformation of this state cannot be determined. The same methyl group electron density is missing in SeM29. In addition, selenium in each SeM29 occupies a continuous range of locations. Same cloud conformation is observed for SeM34. In View B, other possible conformations of SeM34 is clearly shown by the electron density.

REFERENCES

1. Hauptman HA. (1990) History of X-ray crystallography. Struct Chem 1 (6): 617-620. doi: 10.1007/978-3-319-19827-9_2

2. Jaskolski M, Dauter Z and Wlodawer A. (2014) A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. 281 (18): 33985-4009. doi: 10.1111/febs.12796

3. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H & Phillips DC (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181, 662–666. doi: 10.1038/181662a0

4. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G & North ACT (1960) Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. Nature 185, 416–421. doi: 10.1038/185416a0

5. Strandberg B. (2009) Chapter 1: building the ground for the first two protein structures: myoglobin and haemoglobin. J Mol Biol 392 (1) 2-10. doi: 10.1016/j.jmb.2009.05.087

6. Lovette MA and Doherty MF. (2012) Needle-shaped crystals: causality and solvent selection guidance based on periodic bond chains. Cryst Growth Des 13: 3341-3352. doi: dx.doi.org/10.1021/cg301830u

7. Yang LW, Eyal E, Chennubhotta C, Jee JG, Gronenborn AM and Bahar I. (2007) Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions. Structure 15: 741-749. doi: 10.1016/j.str.2007.04.014

8. Brünger AT. (1997) X-ray crystallography and NMR reveal complementary views of structure and dynamics. Nat Struct Biol Suppl: 862-865. PMID:9377160

9. Yee AA, Savchenko A, Ignachenko A, Lukin J, Xu X, Skarina T, Evdokimova E, Liu CS, Semesi A, Guido V, Edwards AM and Arrowsmith CH. (2005) NMR and X-ray crystallography, complementary tools in structural proteomics of small proteins. J Am Chem Soc 127 (47): 16512-16517. doi: 10.1021/ja053565+

10. Shi Y. (2014) A Glimpse of structural biology through X-Ray Crystallography. Cell 159: 995-1014. doi: 10.1016/j.cell.2014.10.051

11. Phillips WC, Li Y, Stanton M, Xie J, O'Mara D, and Kalata K. (1993) A CCD-based area detector for X-ray crystallography using synchrotron and laboratory sources. Nucl Instrum Methods Phys Res A 334, 621–630. doi: 10.1016/0168-9002(93)90830-B

12. Gros P, van Gunsteren WF, Hol WG. (1990) Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science 249 (4973): 1149-1152. doi: 10.1126/science.2396108

13. Schmidt A and Lamzin VS. (2010) Internal motion in protein crystal structures. Protein Sci 19 (5): 944-953. doi: 10.1002/pro.371

14. Ma P, Xue Y, Coquelle, N, Haller JD, Yuwen T, Ayala I, Mikhailovskii O, Willbold D, Colletier JP, Skrynnikov NR and Schanda P. (2015) Observing the overall rocking motion of a protein in a crystal. Nat. Commun 6: 1-10. doi: 10.1038/ncomms9361

15. Kurauskas V, Izmailov SA, Rogacheva ON, Hessel A, Ayala I, Woodhouse J, Shilova A, Xue Y, Yuwen T, Coquelle N, Colletier JP, Skrynnikov NR, Schanda P.

(2017) Slow conformational exchange and overall rocking motion in ubiquitin protein crystals. Nat Commun 8 (1) 145. doi: 10.1038/s41467-017-00165-8

16. Robustelli P, Cavalli A and Vendruscolo. (2008) Determination of protein structures in the solid state from NMR chemical shifts. Structure 16: 1764-1769. doi: 10.1016/j.str.2008.10.016

17. Ishima R and Torchia DA. (2000) Protein dynamics from NMR. Nat Struct Biol 7: 740-743. doi: 10.1038/78963

18. Kleckner IR and Forster MP. (2011) A introduction to NMR-based approaches for measuring protein dynamics. Biochim Biophys Acta 1814 (8): 942-968. doi: 10.1016/j.bbapap.2010.10.012

19. Fenwick RB, van den Bedem H, Fraser JS and Wright PE. (2014) Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. Proc Natl Acad Sci USA 111 (4) E445-E454. doi: 10.1073/pnas.1323440111

20. Victora A, Moller HM and Exner T. (2014) Accurate *ab initio* prediction of NMR chemical shifts of nucleic acids and nucleic acids/protein complexes. Nucleic Acids Res 42 (22): e173. doi: 10.1093/nar/gku1006

21. Adams PD, Afonine PV, Chen V, Echols N, Headd JJ, Hung LW, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Read RJ, Richardson DC, Richardson JS, Terwilliger TC and Zwart PH. (2010) PHENIX - A Comprehensive Python-based System for Macromolecular Structure Solution. Acta Cryst. D66, 213-221.

doi: 10.1107/S0907444909052925

22. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A and Wilson KS. (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr 67 (Pt4): 235-242. doi: 10.1107/S0907444910045749

23. Potterton L, Agirre J, Ballard C, Cowtan K, Dodson E, Evans PR, Jenkins HT, Keegan R, Krissinel E, Stevenson K, Lebedev A, McNicholas SJ, Nicholls RA, Noble M, Pannu NS, Roth C, Sheldrick G, Skubak P, Turkenburg J, Uski V, von Delft F, Waterman D, Wilson K, Winn M and Wojdyr M. (2018) CCP4i2: the new graphical user interface to the CCP4 program suite. Acta Crystallogr D Struct Biol 74(Pt 2):68-84. doi: 10.1107/S2059798317016035

24. Gallagher T, Alexander P, Bryan P and Gilliland GL. (1994) Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 33: 4721-4729.

25. Schmidt HL, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, Wilson SR and Rienstra CM. (2007) Crystal Polymorphism of Protein GB1 Examined by Solid-State NMR Spectroscopy and X-ray Diffraction. J Phys Chem B 111: 14362-14369.
doi: 10.1021/jp075531p

26. Kabsch, W. (2010) XDS. Acta Cryst D66: 125-132.
doi: 10.1107/S0907444909047337

27. Evans PR and Murshudov GN. (2013) How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr 69 (Pt7): 1204-1214. doi: 10.1107/S0907444913000061

28. Bunkóczi G, Echols N, McCoy AJ, Oeffner RD, Adams PD and Read RJ. (2013) Phaser.MRage: automated molecular replacement. Acta Crystallogr D Biol Crystallogr 69, 2276-86. doi: 10.1107/S0907444913022750

29. Emsley P, Lohkamp B, Scott WG and Cowtan K. (2010) Features and development of Coot. Acta Cryst D66: 486-501. doi: 10.1107/S0907444910007493

30. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH and Adams PD. (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr 68: 352-67. doi: 10.1107/S0907444912001308

31. Klei HE, Moriarty NW, Echols N, Terwilliger TC, Baldwin ET, Pokross M, Posy S and Adams PD. (2014) Ligand placement based on prior structures: the guided ligand-replacement method. Acta Crystallogr D Biol Crystallogr 70 (Pt 1): 134-143. doi: 10.1107/S1399004713030071

32. Merritt EA. (2012) X-ray anomalous scattering. http://www.bmsc.washington.edu/scatter.

33. Chen VB, Arendall WB III, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS and Richardson DC. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66 (Pt 1): 12-21. doi: 10.1107/S0907444909042073

34. Brünger AT. (1992) Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355 (6359): 472-475. PMID: 18481394.

35. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Adams PD, Read RJ, Zwart PH and Hung LW. (2008) Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. Acta Crystallogr D Biol Crystallogr 64 (Pt 5): 515-524. doi: 10.1107/S0907444908004319.

36. Frey M, Genovesio-Taverne JC and Fontecilla-Camps JC. (1991) Molecular packing and morphology of protein crystals. J Phys D: Appl Phys 24: 105-110.

37. Nayhouse M, Kwon JS, Christofides PD and Orkoulas G. (2013) Crystal shape modeling and control in protein crystal growth. Chem Eng Sci 87: 216-223. doi: 10.1016/j.ces.2012.10.020

38. Nanev C N. (2015) Bond selection during protein crystallization: crystal shapes. Cryst Rez Technol 50 (6): 451-457. doi: 10.1002/crat.201500013

39. Krissinel E and Henrick K. (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774-797. doi: 10.1016/j.jmb.2007.05.022

40. Lawson CL, Dutta S, Westbrook JD, Henrick K and Berman HM. (2008) Representation of viruses in the remediated PDB archive. Acta Cryst D Biol Crystallogr 64: 874-882. doi: 10.1107/S0907444908017393

41. Thoms S, Mas KEA, Wunderlich M, Jasco T, Lilie H, Reif B, Heinemann U and Schmid FX. (2009) Dimer Formation of a Stabilized Gβ1 Variant: A Structural and Energetic Analysis. J Mol Biol 391: 918-932. doi: 10.1016/j.jmb.2009.06.031

42. Chruszcz M, Potrzebowski W, Zimmerman MD, Grabowski M, Zheng H, Lasota P and Minor W. (2008) Analysis of solvent content and oligomeric states in protein crystals—does symmetry matter? Protein Sci 17 (4): 623-632.
doi: 10.1110/ps.073360508

43. Wukovitz SW, Yeates TO. (1995) Why protein crystals favour some space-groups over others. Nat Struct Biol 2 (12): 1062-7. doi: 10.1038/nsb1295-1062

44. Padmaja N, Ramakumar S, Viswamitra MA. (1990) Space-group frequencies of proteins and of organic compounds with more than one formula unit in the asymmetric unit. Acta Crystallogr A 46:725–730. doi: 10.1107/S0108767390004512

45. Pellegrini M, Wukovitz SW, Yeates TO. (1997) Simulation of protein crystal nucleation. Proteins 28 (4): 515-21.
doi: 10.1002/(SICI)1097-0134(199708)28:4<515:;AID-PROT5>3.0.CO;2-8

46. Andersson KM, Hovmöller S. (2000) The protein content in crystals and packing coefficients in different space groups. Acta Crystallogr D Biol Crystallogr 56 (Pt 7): 789-790. doi: 10.1107/S0907444900005163

47. Matthews BW. (1968) Solvent content of protein crystals. J Mol Biol 33 (2): 491-497. doi: 10.1016/0022-2836(68)90205-2

48. Marsh RE. (1995) Some thoughts on choosing the correct space group. Acta Crystallogr B 51: 897-907. doi: 10.1107/S0108768195008901

49. Marsh RE, Bernal I. (1995) More space-group changes. Acta Crystallogr B: 51: 300–307. doi: 10.1107/S0108768194011857

50. Franks WT, Wylie BJ, Stellfox SA and Rienstra CM. (2006) Backbone conformational constraints in a microcrystalline U-^{15}N-labeled protein by 3D dipolar-shift solid-state NMR spectroscopy. J Am Chem Soc 128: 3154-3155. doi: 10.1021/ja058292x

51. Mollica L, Baias M, Lewandowski JR, Wylie BJ, Sperling LJ, Rienstra CM, Emsley L and Blackledge M. (2012) Atomic-Resolution Structural Dynamics in Crystalline Proteins from NMR and Molecular Simulation. J Phys Chem Lett 3 (23): 3657-3662. doi: 10.1021/jz3016233

52. Varma S, Chiu SW and Jakobsson E. (2006) The Influence of Amino Acid Protonation States on Molecular Dynamics Simulations of the Bacterial Porin OmpF. Biophys J 90 (1): 112-123. doi: 10.1529/biophysj.105.059329

53. Carugo O and Argos P. (2008) Protein—protein crystal-packing contacts. Protein Sci 6 (10): 2261-2263. doi: 10.1002/pro.5560061021

54. Fusco D, Headd JJ, De Simone A, Wang J and Charbonneau P. (2014) Characterizing protein crystal contacts and their role in crystallization: rubredoxin as a case study. Soft Matter 10 (2): 290-302. doi: 10.1039/C3SM52175C

55. Capitani G, Duarte JM, Baskaran K, Bliven S and Somody JC. (2015) Understanding the fabric of protein crystals: computational classification of biological interfaces and crystal contacts. Bioinformatics 32 (4): 481-489. doi: 10.1093/bioinformatics/btv622

56. Luo J, Liu Z, Guo Y and Li M. (2015) A structural dissection of large protein-protein crystal packing contacts. Sci Rep 5: 14214. doi: 10.1038/srep14214

57. Smith CK, Withka JM and Regan L. (1994) A Thermodynamic scale for the beta.-Sheet forming tendencies of the amino acids. Biochemistry 33 (18): 5510-5517. doi: 10.1021/bi00184a020

58. Kufareva I and Abagyan R. (2012) Methods of protein structure comparison. Methods Mol Biol 857: 231-257. doi: 10.1007/978-1-61779-588-6_10

59. Carugo O. (2007) Statistical validation of the root-mean-square-distance, a measure of protein structural proximity. Protein Eng Des Sel (PEDS) 20 (1): 33-37. doi: 10.1093/protein/gzl051

60. Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253 (5020): 657-661. doi: 10.1126/science.1871600

61. Argos P, Rossmann MG, Grau UM, Zuber H, Frank G, and Duri Tratschin J. (1979) Thermal stability and protein structure. Biochemistry 18 (25): 5698-5703. doi: 10.1021/bi00592a028

62. Alexander P, Fahnestock S, Lee T, Orban J and Bryan P. (1992) Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. Biochemistry 31 (14): 3597-3603. doi: 10.1021/bi00129a007

63. Gallagher T, Alexander P, Bryan P and Gilliland GL. (1994) Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 33 (15): 4721-4729. doi: 10.1021/bi00181a032

64. Sheinerman FB and Brooks CL. (1998) Molecular picture of folding of a small α/β protein. Proc Natl Acad Sci USA 95 (4): 1562-1567.

65. Rhys NH, Soper AK and Dougan L. (2012) The hydrogen-bonding ability of the amino acid glutamine revealed by neutron diffraction experiments. J Phys Chem B 116: 13308-13319. doi: 10.1021/jp307442f

66. Frank MK, Dyda F, Dobrodumov A and Gronenborn AM. (2002) Core mutations switch monomeric protein GB1 into an intertwined tetramer. Nature Struct Biol 9: 877–885. doi: 10.1038/nsb854

67. Byeon IL, Louis JM and Gronenborn AM. (2003) A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J Biol Mol 333: 141-152. doi: 10.1016/S0022-2836(03)00928-8

68. Mishra KK, Singh SK, Ghoshh P, Ghosh D and Das A. (2017) The nature of selenium hydrogen bonding: gas phase spectroscopy and quantum chemistry calculations. Phys Chem Chem Phys 19: 24179-24187. doi: 10.1039/c7cp05265k

69. Valley CC, Cembran A, Perimutter JD, Lewis AK, Labello NP, Gao J and Sachs JN. (2012) The methionine-aromatic motif plays a unique role in stabilizing protein structure. J Biol Chem 287 (42): 34979-34991. doi: 10.1074/jbc.M112.374504

70. Forbes CR, Sinha SK, Ganuly HK, Bai S, Yao GPA, Patel S and Zondlo NJ. (2017) Insights into thiol−aromatic interactions: A stereoelectronic basis for S−H/π interactions. J Am Chem Soc 139: 1842-1855. doi: 10.1021/jacs.6b08415

71. Virrueta A, O'Hern CS and Regan L. (2016) Understanding the physical basis for the side-chain conformational preferences of methionine. Proteins 84 (7): 900-911.

doi: 10.1002/prot.25026

72. Nittinger E, Schneider N, Lange G and Rarey M. (2015) Evidence of water molecules -a statistical evaluation of water molecules based on electron density. J Chem Inf Model 55: 771-783. doi: 10.1021/ci500662d

73. Gómez-Tamayo JC, Cordomí A, Olivella M, Mayo E, Fourmy D and Pardo L. (2016) Analysis of the interactions of sulfur-containing amino acids in membrane proteins. Protein Sci 25 (8): 1517-1524. doi: 10.1002/pro.2955

74. Thoms S, Mas KEA, Wunderlich M, Jasco T, Lilie H, Reif B, Heinemann U and Schmid FX. (2009) Dimer formation of a stablized GB1 variant: a structural and energetic analysis. J Mol Biol 391: 918-932. doi: 10.1016/j.jmb.2009.06.031

75. Woldeyes RA, Sivak DA and Fraser JS. (2014) E pluribus unum, no more: from one crystal, many conformations. Curr Opin Struct Biol 28: 56-62.
doi: 10.1016/j.sbi.2014.07.005

76. Miao Z and Cao Y. (2016) Quantifying side-chain conformational variations in protein structure. Sci Rep 6: 37024. doi: 10.1038/srep37024

77. Afonine PV, Moriaty NW, Mustyakimob M, Sobolev OV, Terwilliger TC, Turk D, Urzhumtsev A and Adams PD. (2013) FEM: Feature enhanced map. Acta Crystallogr D Biol Crystallogr 4(2):28–29. doi: 10.1107/S1399004714028132

78. Glykos NM and Kokkinidis M. (2000) GraphEnt: a maximum-entropy program with graphics capabilities. J App Crystallogr 33 (3): 982–985.
doi: 10.1107/S0021889800004246

79. Liu C and Xiong Y. (2014) Electron density sharpening as a general technique in crystallographic studies. J Mol Biol 426 (4): 980–993. doi: 10.1016/j.jmb.2013.11.014

80. Shapovalov MV and Dunbrack RL Jr. (2007) Statistical and conformational analysis of the electron density of protein side chains. Proteins 66 (2): 279–303. doi: 10.1002/prot.21150

Chapter 5

UNDERSTANDING SELENIUM CHEMISTRY IN HUMAN SELENOPROTEINS

5.1 Background

There are two naturally occurring amino acids which contain sulfur, Met and Cys. Each one also has a selenium-containing analog: SeM which was discussed in these texts, and also Sec which has not been covered here in detail. Research into the physiological roles of selenium suggest that Sec, recognized as the 21st natural amino acid in the genetic code, is largely responsible for the health benefits of selenium. [1]

Sec is found in 25 human proteins, collectively they make up the human selenoproteome. In most cases, the Sec residue is found in the enzymatic active site, where they are responsible for carrying out redox reactions and maintaining cellular redox balance. In the past decade, there have been significant progress made in understanding the biological functions of human selenoproteins. Comprehensive research is available for the GPx family and the TrxR family. However, out of the 25 human selenoproteins, at least 13 of which whose physiological roles are not well understood. These include selenoproteins W, T, H, V, I, M, K, S, O, N, P and the 15 kDa selenoprotein more commonly referred to as Sep15. Regardless, these less-characterized selenoproteins may participate in a wide range of physiological functions such as to facilitate protein folding in the ER (SELENOK, SELENOS), and links to cognitive function as well as obesity (SELENOM). [1]

Our understanding of human selenoproteins has been hindered by the lack of an efficient recombinant scheme to produce Sec-containing proteins. [2] The biosynthesis of Sec requires the cotranslational incorporation of the Sec amino acid residue by in-

frame UGA codon found in selenoprotein mRNAs. The UGA codon normally signals for translational termination. However, under the influence of Sec-incorporating protein factors, the designated Sec-tRNA, and the *cis*-acting Sec-insertion sequence (SECIS element), translation is allowed to proceed and the UGA codon instead is decoded as Sec. Sec-containing proteins are found in all three branches of life, and variations exist between their respective biosynthetic pathways. Due to this feature, and the inherent complexity of Sec-incorporation, either traditional recombinant methods or sophisticated bioengineering have not identified a robust technique to produce adequate amounts of selenoproteins for biochemical characterization.

The expressed selenoprotein ligation method developed by *Liu et al.* can circumvent the issue of low protein production. [2] For the model selenoproteins that were tested with this method, the overall yield of SELENOM is 10 mg and that of SELENOW is 2 mg per each production cycle. This quantity makes the study of selenoproteins by NMR a possibility, where typical sample concentrations vary between 0.5 to 1 mM.

Selenium NMR is still in its nascent stage of development. While the common benefits and challenges for SeM, such as high sensitivity but broad lines also exist for Sec, in the case of Sec the problem becomes much more pronounced. For example, the chemical shift tensor for L-SeM spans roughly 580 ppm and can be detected with ^{77}Se at natural abundance, the resonance widths for Sec have been reported to be between 500-900 ppm for diselenide bonds, and can be even broader for the naturally occurring, redox sensitive selenylsulfide moieties. [3] Such broad lines compromise detection

sensitivity, necessitating the need for isotopically labeled samples, longer acquisition times, and results that may be difficult to interpret especially due to the paucity of previously available research. To date, the only experimental chemical shift tensor information available on Sec-containing systems is from *Struppe et al.*, where the authors examined the chemical shift tensors of crystallized L-selenocystine and its temperature dependence. [4] It was concluded that the environmental effect, such as the bonding environment, even crystal packing each generates significant effects in ^{77}Se NMR properties, but the sidechain dihedral angle was the most critical parameter. Combining theoretically calculation with experimental data, the studied determined that in L-Sec crystals, it only takes less than 2.5 kcal/mol of energy to alter the dihedral angle of the Sec by as much as 30° around the energy minimum conformation. However, this study was carried out on a small molecule system consisting of one diselenide bond. It provides a starting point for the analysis of Sec using NMR, but may not be easily extrapolated to complex biological systems. Again, the sensitivity of ^{77}Se NMR means that the Sec residue can also be applied as a reporter to probe the details of molecular structures. For instance, the geometry of the Sec sidechain, or that of the selenylsulfide bond, can directly impact the protein's ability to perform oxidoreductase reactions. To understand the details between the electronic environment of Sec and its reactivity, can help to elucidate the enzymatic functions of many selenoproteins. This information can also aid in engineering novel selenoproteins by taking advantage of the more stable and more reactive diselenide bond, or perhaps generating stable protein therapeutics for biomedical applications. [4]

5.2 Building a Sec repertoire for GB1

The same model system using protein GB1 was selected for its high protein stability and overall high yield. A number of Cys and Sec variants of GB1 are made available through the mutagenesis procedures as discussed in *Section 3.2.1*. The details of each variant are summarized in Table 7.

The location of Sec in each variant follows the similar rationale of selecting a wide range of local interactions of Sec, either in its monomeric form, or as part of a diselenide bond. Lys4 is a solvent-exposed residue on the first β-strand. It is in close proximity to both Thr44 and Thr51. It may also be able to interact with Thr17 depending on their respective conformations. Cys5 is a solvent-shielded residue on the first β-strand. It is located in the interior of the protein, surrounded by three aromatic residues Phe30, Tyr33 and Trp43, and it is within van der Waals contact to both Phe30 and Trp43. It can potentially also sense Leu7 and Val54, both of which are located 4 Å away. Lys28 is positions on the only helix of the protein structure. It is completely solvent exposed at the outer edge of the helical turn. Cys34 maybe in π interaction with Trp43 which is located approximately 5 Å away. In the SeM variant this particular location is rather challenging experimentally due to its extremely broad lines. Thr44 and Thr53 are found on the solvent exposed sides of β-strands 3 and 4 and within van der Waals contact of each other.

Since the chemical shift tensors of Sec are typically broad, preliminary NMR experiments using solution samples indicate that the detection of ^{77}Se through natural

abundance sample is not feasible. Therefore, ^{77}Se-isotopically enriched samples must be used for both solution- and solid-state NMR experiments. The initial procedure for enrichment is as discussed in *Section 3.2.4*. Once expressed, the protein can be purified following the affinity chromatography method discussed for GB1 SeM variants discussed in Chapter 3.

5.3 Concluding Remarks

Significant progress has been made in understanding the roles of selenium and various selenoproteins in human physiology over the past decade, and the entire human selenoproteome consisting of 25 selenoproteins has been identified. The functions of several families of selenoproteins, including the glutathione peroxidases family, the thyroid hormone deodinases family, and the thioredoxin reductases family are well established. Each member of these protein families is responsible for a wide range of activities from hydrogen peroxide signaling, thyroid hormone regulation and cellular redox homeostasis. However, there still remains a large number of these selenoproteins whose functions are putative or completely unknown. Based on protein structure, these members of the family likely participate in redox-related activities, but it has come to light that selenoproteins do not function solely as antioxidants. [1] Over the course of history of selenium since its discovery, the role of selenium has been at various times linked to cancer, diabetes, obesity, inflammation, neurodegenerative diseases and

cardiovascular diseases. Therefore, research into selenium and selenoproteins may provide insights for the development of new therapeutics for a number of different diseases that have been linked to these proteins.

5.4 Acknowledgement

The work described in this thesis has been made possible through collaborations with many talented mentors and colleagues. Dr. Jun Liu led the research on selenocysteine-mediated expressed protein ligation of SELENOM. Dr. Xingyu Lu and Shiping Xu led the efforts in ^{77}Se-NMR data acquisition using solid-state NMR samples. Dr. Xingyu Lu and Michael Boeri acquired data using solution-state NMR samples. Elizabeth Diaz was instrumental in the screening efforts to obtain the highest diffraction quality crystals. Dr. Brian J Bahnson provided invaluable insight on the topic of protein X-ray crystallography, as well as technical advice on growing, harvesting and freezing protein crystals. Dr. Marc Allaire collected atomic resolution diffraction data on all GB1 SeM crystals at the Advanced Light Source beamline 5.0.2. I also would like to thank Dr. Henry Bellamy, Dr. Pavel Affonine and Dr. Karl Schmitz for helpful discussions on data acquisition and analysis.

Table 8. Selenocysteine-variants of GB1. [5]

Mutation	Location	Interaction
K4C	Solvent exposed near Glu15 and Lys4	Dispersion
L5C	In contact with Phe30	Sulfur – Aromatic
K28C	Solvent exposed on helix	Dispersion
A34C	Hydrophobic, near Trp43	Sulfur – Aromatic
T44C	Solvent exposed near Thr53	Dispersion
T53C	Solvent exposed near Thr53	Dispersion
T44C T53C	Across each other on strands 3 & 4	Disulfide/Diselenide
K4C T17C	Across each other on strands 1 & 2	Disulfide/Diselenide

REFERENCES

1. Labunskyy VM, Hatfield DL, and Gladyshev VN. (2014) Selenoproteins: molecular pathways and physiological roles. Physiol Rev 94 (3): 739-777.
doi: 10.1152/physrev.00039.2013

2. Liu J, Chen Q and Rozovsky S. (2017) Utilizing selenocysteine for expressed protein ligation and bioconjugations. J Am Chem Soc 139 (9): 3430-3437. doi: 10.1021/jacs.6b10991

3. Schaefer SA, Dong M, Rubenstein RP, Wilkie WA, Bahnson BJ, Thorpe C and Rozovsky S. (2013) ^{77}Se Enrichment of proteins expands the biological NMR toolbox. J Mol Biol Comm 425: 222-231. doi10.1016/j.jmb.2012.11.011

4. Struppe J, Zhang Y and Rozovsky S. (2015) ^{77}Se Chemical shift tensor of L-selenocystine: experimental NMR measurements and quantum chemical investigations of structural effects. J Phys Chem B 119: 3643-3650.
doi: 10.1021/jp510857s

5. Jaroniec C. (2016) Structural studies of proteins by paramagnetic solid-state NMR spectroscopy. J Magn Reson 253: 50-59. doi: 10.1016/j.jmr.2014.12.017