UNDERSTANDING THE STRUCTURE OF EUKARYOTIC GLUTAMINYL-TRNA SYNTHETASE: COMBINING X-RAY CRYSTALLOGRAPHY WITH STATISTICAL EVALUATIONS OF SMALL ANGLE SCATTERING DATA

By

Thomas Daniel Grant

January 2013

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Structural Biology

Dedication

I dedicate this work to God, who has been my rock and refuge, and to my family, for always keeping me grounded. Your support has encouraged me to press toward the mark.

Acknowledgments

First and foremost, I would like to thank God for giving me wisdom and guidance throughout my life and for giving me the abundance of opportunities that I have enjoyed. I thank my parents who raised me to believe that “I can do all things through Christ who strengthens me” (Philippians 4:13). I thank my brother and sister for encouraging me throughout my entire life, in that special way that siblings do.

I thank my fiancée Pei, who has been so wonderful and patient with me as we discussed much of this dissertation with a degree of detail and candor that I can enjoy with no other.

I would like to thank my advisor, Dr. Eddie Snell, for his constant guidance and for giving me the freedom to pursue my own thoughts and ideas to their conclusions, on which much of this thesis is based. I would also like to thank Joe Luft, who has always given excellent advice with kind words of support throughout my graduate career. I thank Drs. Eric Phizicky and Elizabeth Grayhack for their extensive efforts in our collaboration and the entire staff at beamline 4-2 at SSRL for their time and help. I thank each of my committee members, Drs. Bob Blessing, Richard Gillilan, Andrew Gulick, Eaton Lattman, and Wayne Schultz for their counsel and suggestions that greatly improved my thesis.

Table of Contents

Dedication iii

Acknowledgments iv

Table of Contents v

List of Tables x

List of Figures xii

List of Abbreviations xxiii

Abstract xxv

1 Introduction 1

1.1 History 1

1.2 Theory 2

1.2.1 Experimental Setup 3

1.2.2 Theory 5

1.3 High-throughput SAXS 11

1.4 Scope of This Thesis 13

1.5 References 14

2 Small Angle X-ray Scattering as a Complementary Tool for High-throughput Structural Studies 16

2.1 Introduction 16

2.2 Materials and Methods. 18

2.2.1 Samples. 18

2.2.2 Crystallization. 21

2.2.3 X-ray Crystallographic and Solution NMR Structure Determination. 21

2.2.4 SAXS data. 22

2.2.5 Comparison of structural data 24

2.3 Results. 24

2.3.1 Crystallographic and SAXS comparison 26

2.3.2 Sensitivity to different constructs. 29

2.3.3 NMR and SAXS comparison. 31

2.3.4 The Combination of Crystallography and NMR with SAXS. 34

2.3.5 Mixtures. 36

2.4 Discussion 38

2.5 References 42

3 SAXStats Automated SAXS Analysis Software Package 44

3.1 Introduction 44

3.2 Radiation Damage 45

3.2.1 Using Linear Regression T-statistic to Evaluate Radiation Damage 46

3.2.2 Detecting Changes in Scattering Profile 48

3.2.3 Detecting Changes in Maximum Particle Dimension 49

3.2.4 Detecting Changes in R_g and *I(0)* 51

3.2.5 Detecting Changes in “Foldedness” of Particle 55

3.3 Concentration Dependence 57

3.3.1 Scaling SAXS Profiles 58

3.3.2 Detecting Changes in R_g, D_{max}, and *I(0)* 61

3.3.3 Detecting Changes in Particle Volume 61

3.3.4 Evaluating Linearity in the Guinier Region 63

3.3.5 Second Virial Coefficient 64

3.4 Experimental Application of SAXStats to a Large Dataset 67

3.4.1 Description of Sample Set 67

3.4.2 Overall Success of SAXStats Software Package 68

3.4.3 Radiation Damage Statistics 69

3.4.3.1 X Statistics 72

3.4.3.2 *P(r)* Distribution Statistics 74

3.4.3.3 Guinier Approximation Statistics 77

3.4.3.4 Kratky Ratio Statistics 81

3.4.3.5 Comparing SAXS Parameters 83

3.4.4 Concentration Dependence Statistics 89

3.4.4.1 Interparticle Interaction Statistics 92

3.4.4.2 Severity of Concentration Dependence 93

3.4.4.3 Second Virial Coefficients 96

3.4.5 Summary of Results 98

3.5 References 100

4 Structural Conservation of an Ancient tRNA Sensor in Eukaryotic Glutaminyl-tRNA Synthetase 102

4.1 Introduction 102

4.2 Methods 106

4.2.1 Genetic analysis of *gln4* mutants 106

4.2.2 Protein expression and purification 108

4.2.3 tRNA purification and EMSA binding assay 108

4.2.4 *In vitro* synthesis of tRNA transcripts 109

4.2.5 Steady State Methods 109

4.2.6 Crystallization and Structure Determination 110

4.2.7 Small Angle X-ray Scattering 112

4.3 Results 113

4.3.1 Removal of the NTD impairs Gln4 function *in vivo* and *in vitro* 113

4.3.2 The Gln4 NTD is structurally similar to two subdomains in the amidotransferase that distinguish tRNA^{Gln} from tRNA^{Glu}. 118

4.3.3 The linker between the NTD subdomains is conserved and functionally important. 122

4.3.4 SAXS data reveals that the NTD structure is not distorted by crystal contacts.. 124

4.3.5 Point mutations in NTD linker impair tRNA binding integrity 125

4.4 Discussion 126

4.5 Summary 130

4.6 References 130

5 The Structure of Yeast Glutaminyl-tRNA Synthetase and Modeling Its Interaction with tRNA 134

5.1 Introduction 134

5.2 Experimental Procedures 136

5.2.1 Protein Expression and Purification 136

5.2.2 Crystallization. 137

5.2.3 Single crystal data collection and structure solution 138

5.2.4 Bioinformatics 139

5.2.5 Small Angle X-ray Scattering 139

5.2.6 Molecular Dynamics 140

5.3 Overall Structure 141

5.4 Motion of NTD in Solution is Limited in the Absence of tRNA 148

5.5 A Model of Full-length *Sc*GlnRS Bound to tRNA^{gln} Suggests a Substantial Conformational Reorientation of the NTD 152

5.6 Conformational Change in NTD Subdomains is Predicted Upon Interaction with tRNA^{gln} 154

5.7 GlnRS Phylogenetic Analysis 157

5.8 Conclusion 161

5.9 References 162

6 Summary and Future Work 167

6.1 Thesis Summary 167

6.2 Combining SAXS Data with Information from the NESG Database to Supplement the High-throughput Structural Pipeline 169

6.3 Improvements to SAXStats 170

6.4 Probing the Acceptability of SAXS Data Quality 171

6.5 Probing Long Time-scale Protein Dynamics Using SAXS Envelopes............. 173

6.6 Expanding the Structural Knowledge of Gln4 175

6.7 References 176

Appendix A. Converting t-statistic to p-value. 178

Appendix B. Script for detecting radiation damage in SAXS Profiles 179

Appendix C. Script for detecting concentration dependence in SAXS Profiles ... 196

List of Tables

Table 2.1 Samples used for the SAXS analysis are divided into four sets. The first set 1–13 contains 13 proteins, each having crystallographic structures. The second set 14–17 contains 2 proteins with two different constructs of the first having two crystallographic structures and the second a single structure. The third set 18–26 contains 9 proteins, each having an NMR structure. The fourth set 27–28 contains two proteins where both NMR and crystallographic structures are available. The sample name, ID, PDB identifier, reference, the oligomeric state in solution characterized on preparation by light scattering and gel filtration, initial concentration (mg/mL), molecular weight (Da) and number of residues are listed. The oligomeric state in solution is defined in the table as M (Monomer), D (Dimer), Tri (Trimer), T (Tetramer), Hep (Heptamer) or a combination. While all the samples have structures deposited in the PDB the majority are as yet unpublished. We are grateful to the authors in the references for the ability to use this structural data at this early stage. 20

Table 2.2 A Summary of Structural (Crystallography and NMR) and SAXS Results. The sample # refers to the identical number in Table 2.1. The number of unresolved residues in the structure (mainly crystallographic) is listed together with the Rg and Dmax (in Å) determined from the available structure. The Rg and Dmax from the SAXS data are shown together with the difference from the available structural information. The molecular weight (in Da) calculated from a Porod analysis is listed along with the ratio of this weight with that derived initially from mass spectrometry in Table 2.1. Finally the SAXS-determined oligomer, (Monomer, Dimer or Tetramer), the relationship to the available structure and the χ of the fit are listed. A special case is described below for samples 16 and 17. Further details are given in the text. 25

Table 4.1 Oligonucleotides for various Gln4 constructs... 107

Table 4.2 Yeast strains used for Gln4 expression 107

Table 4.3 Plasmids used in Gln4 study.. 107

Table 4.4 Comparison of steady state kinetic parameters for Gln4 and Gln4 variants. 115

Table 4.5 Data collection, phasing and refinement statistics for Gln4(1-187)............... 118

Table 4.6 Comparison of sequences threaded to the Gln4(1-187) NTD structure. 122

Table 5.1 Oligonucleotides for Gln4 protein constructs 136

Table 5.2 Data collection and refinement statistics for full-length Gln4 142

Table A.1 Converting t-statistic to p-value. Top row represents p-values assuming a two-tailed test. Left column refers to number of degrees of freedom, DF............. 178

List of Figures

Figure 1.1 Geometry of Scattered Radiation. The larger particle on the right scatters radiation at smaller angles than does the smaller particle on the left. λ is the wavelength and 2θ is the angle of the scattered radiation....................................... 3

Figure 1.2 Geometry of SAXS Experimental Setup. Incident X-rays corresponding to wave vector k_i with magnitude 2π/λ pass through the sample chamber and scatter with an angle of 2θ. The isotropic detector image is integrated and radially averaged to produce a 1D scattering profile of intensity (I) vs momentum transfer (q). 4

Figure 2.1 The observed SAXS data and structural fit to the observed data (continuous line) for samples with crystallographic structure. The ab initio SAXS derived envelopes overlaid with crystallographic structure are also shown illustrating the agreement between the techniques. The samples are numbered as in Table 2.1 and Table 2.2. Samples 4 and 11 contained a mixture of oligomers and are shown below. 28

Figure 2.2 The observed SAXS data and structural fit to the observed data (continuous line) for samples with crystallographic structure and multiple constructs. *Ab initio* SAXS-derived envelopes are overlaid with crystallographic structure. Sample 17 contained a mixture shown below. The figures are shown to the same approximate scale as those in Figure 2.1 and the remaining structural representations in Figure 2.3 and Figure 2.4... 31

Figure 2.3 The observed SAXS data and structural fit to the observed data (continuous line) for samples with NMR structures. The *ab initio* SAXS-derived envelopes are overlaid with the NMR structures to show the agreement between the data. The figures are shown to approximate scale as in the other structural illustrations and illustrate multiple conformations determined from the NMR data. 34

Figure 2.4 *Ab initio* SAXS-derived envelopes overlaid with NMR and crystallographic structure to show the agreement between the different structural methods. The figures are shown to approximate scale and illustrate multiple conformations determined from the NMR data. 36

Figure 2.5 Structures of oligomers based on analysis of the SAXS data and known monomer structure. The *ab initio* SAXS-derived envelopes are shown assuming a monodisperse solution. 37

Figure 3.1 Kratky Plots of Folded and Unfolded Proteins. A folded protein will exhibit a Kratky plot similar to that seen in blue, where a distinct maximum is seen and the data falls to zero at higher *q* values. Unfolded proteins show Kratky plots similar to the plot seen in red, where no characteristic peak is seen in the data. A protein that is partially folded will show a Kratky plot similar to that seen in black, where a distinct peak is found, while the data does not return to zero at higher *q* values. This figure has been adapted from (Putnam et al., 2007). 56

Figure 3.2 Frequency (Total = 284) of Multiple SAXS Parameters with Radiation Damage. While most samples experienced no radiation damage, those that did typically only had one or two parameters demonstrating the damage..................... 71

Figure 3.3 Frequency of Radiation Damage for Each SAXS Parameter. While distributed across all parameters, radiation damage is demonstrated most frequently by D_{max}. 71

Figure 3.4 Frequency of the Number of Exposures Averaged. Most samples showed no radiation damage. For damaged samples, between 4 and 6 exposures were undamaged. 72

Figure 3.5 Distribution of Average χ Values. The majority of samples exhibit χ values between 1.2 and 1.6, showing that SAXS profiles of subsequent exposures are similar. 73

Figure 3.6 Distribution of the Standard Deviation of χ. 73

Figure 3.7 Distribution of Average R_g Calculated from *P(r)*. The R_g of most particles in the sample set is between 10 and 50 Å. 75

Figure 3.8 Distribution of Standard Deviation of R_g Calculated from *P(r)*. 75% of the standard deviations of R_g are below 5 Å. 75

Figure 3.9 Distribution of Average D_{max}. The sample set shows a large range of particle sizes, with the majority less than 200 Å. 76

Figure 3.10 Distribution of Standard Deviation of D_{max}. D_{max} typically has very large standard deviations and varies widely between samples. 76

Figure 3.11 Distribution of Average R_g Calculated by the Guinier Method. The majority of R_g values is between 10 and 50 Å, similar to R_g calculated by *P(r)*. 77

Figure 3.12 Distribution of Standard Deviation of R_g Calculated by the Guinier Method. 85% of the population shows standard deviations of less than 5Å. 78

Figure 3.13 Distribution of the Difference in R_g Calculated Using Two Different Methods. The majority of samples show differences of less than 2 Å, 75% less than 5 Å, and 90% less than 10 Å. 79

Figure 3.14 Distribution of Average *I(0)*. Values vary widely with 90% less than 4000. 80

Figure 3.15 Distribution of the Standard Deviation of *I(0).* 85% of samples show standard deviations less than 100 units. 80

Figure 3.16 Distribution of the Average Kratky Ratio. Smaller values indicate a greater degree of foldedness. The distribution ranges from 0.13 to 1.0. Unfolded samples are not shown. The distribution is weighted to the left, with 56% of the samples less than 0.5. 82

Figure 3.17 Distribution of the Standard Deviation of the Kratky Ratio. The majority of samples show standard deviations less than 0.1. 82

Figure 3.18 Distribution of the Coefficient of Variation of χ. χ is relatively well determined, with most of the population having c_v values less than 0.1.................................... 84

Figure 3.19 Distribution of the Coefficient of Variation of R_g Calculated Using *P(r)*. R_g is relatively well determined, with most of the population having c_v values less than 0.1 84

Figure 3.20 Distribution of the Coefficient of Variation of D_{max}. D_{max} is not relatively well determined, with most of the population having c_v values greater than 0.1............ 85

Figure 3.21 Distribution of the Coefficient of Variation of R_g Calculated by Guinier Method. R_g is relatively well determined, with most of the population having c_v values less than 0.1. 85

Figure 3.22 Distribution of the Coefficient of Variation of *I(0)*. *I(0)* is relatively well determined, with most of the population having c_v values less than 0.1................. 86

Figure 3.23 Distribution of the Coefficient of Variation of the Kratky Ratio. Kratky Ratio is not relatively well determined, with half of the population having c_v values greater than 0.1.. 86

Figure 3.24 Coefficient of Variation of R_g as a Function of Signal-to-noise. A strong correlation exists between c_v and *I(0)*.. 87

Figure 3.25 Required Protein Concentration for Achieving a c_v Less than 0.1. The curves represent the concentration required for an 80% likelihood (black), 85% likelihood (red), 90% likelihood (green) or 95% likelihood (purple) of achieving c_v less than 0.1 for the corresponding particle molecular weight. 89

Figure 3.26 Frequency (Total = 82) of Concentration Dependence for Each Parameter. Concentration effects are seen in each parameter with varying frequency, with R_g calculated by *P(r)* demonstrating concentration dependence most frequently. 91

Figure 3.27 Frequency of Multiple SAXS Parameters Showing Concentration Dependence. Most samples did not show concentration dependence with a certainty of p<0.05, while significantly more samples demonstrated dependence with a certainty of p<0.20. 91

Figure 3.28 Frequency (Total = 246) of Interparticle Interactions. The Guinier region truncated by π/D_{max} is less sensitive to curvature than the full Guinier region. 92

Figure 3.29 Concentrations Per Sample (Total = 82) Showing Interparticle Interactions. 75% of samples show interparticle interactions in at least one concentration using a Guinier region defined by q < 1.3/R_g, while more than 85% of samples have at least one concentration not showing interparticle interactions with the same Guinier region. 93

Figure 3.30 Distribution of the Slope of R_g Calculated using Guinier Method. More than half of all samples show less than a 1 Å/mg/mL trend in R_g. 94

Figure 3.31 Distribution of the Slope of Rg Calculated using *P(r).* More than half of all samples show less than a 1 Å/mg/mL trend in R_g. 95

Figure 3.32 Distribution of the Slope of Dmax. A wide range of trends is observed with a significant number of samples showing trends greater 5 Å/mg/mL. 95

Figure 3.33 Distribution of the Slope of Porod Molecular Weight. Molecular weight changes very little in most samples, with the majority affected by less than 1 Å/mg/mL. 96

Figure 3.34 Distribution of B_{22} Calculated Using the Guinier Method. Half of the samples show B_{22} values that exist inside the so-called "crystallization slot". 97

Figure 3.35 Distribution of B_{22} Calculated Using *P(r)*. The distribution is wider than observed for that calculated using the Guinier region. 97

Figure 4.1 Crystal extraction from 1536-well crystallization screening plate. Three images show before (left), during (middle), and after (right) capillary insertion and crystal extraction. The NTD crystals can be seen spanning the horizontal width of the well. The capillary shown in the middle image can be seen as it impinges the crystal. 111

Figure 4.2 Deletion of the N-terminal domain of *GLN4* impairs function. A. Mutants bearing a *gln4* mutation in which amino acids 2-210 are deleted are defective in growth at low temperature on YP media containing glucose or glycerol as a carbon source. Serial dilutions of strains with either wild type *GLN4* or *gln4(*211-809) (marked *gln4-ΔN**) integrated at the *ade2* locus in the *gln4-ΔKanR* mutant were grown as indicated. Indicated strains carry CEN plasmids either with or without *GLN4*. B. Mutants bearing a *gln4* mutation in which amino acids 2-210 are deleted are sensitive to the glutamine synthase inhibitor L-methionine sulfoximine (MSX). 114

Figure 4.3 The N-terminal domain of Gln4 is required for specific binding to native tRNA^{Gln}(CUG). A. Gln4 variant proteins deleted for different amounts of the NTD exhibit reduced tRNA^{Gln}(CUG) binding. B. Gln4(187-809) protein exhibits detectable binding to tRNA^{Gln}(CUG) at high concentrations. 116

Figure 4.4 Gln4 binds more efficiently to native tRNA^{Gln}(CUG) than to native tRNA^{Phe} 116

Figure 4.5 Effects of glutamine, ATP and AMP-PNP on binding of Gln4 and Gln4(187-809) to tRNA^{Gln}(CUG) 117

Figure 4.6 Structure of Gln4(1-187) with comparisons to domains in *S. aureus* GatB (PDB ID: 3IP4). A. Crystallographic structure of Gln4 residues 1-187 in cartoon representation. The proposed hinge region (Gly$_{112}$Val$_{113}$Gly$_{114}$) is highlighted together with the likely interacting residue Trp$_{160}$, and shown in stick representation. B. Surface electrostatic model of Gln4 residues 1-187, shown with two orientations rotated by 90° relative to each other, with positively charged residues colored blue. C, D. Structural alignment of helical and tail domains of Gln4 NTD and *S.aureus* GatB (PDB ID: 3IP4)(Nakamura et al., 2010). C. The crystal structure of Gln4(1-

110) (red) is superposed to the helical domain of GatB(295-406) (cyan). D. The crystal structure of Gln4(119-178) (red) is superposed on the tail domains of GatB(414-475) (cyan). 121

Figure 4.7 Structural alignment of helical and tail domains of Gln4 NTD and *Thermotoga maritima* GatB (PDB ID: 3AL0). A. The crystal structure of Gln4(2-110) (red) is superposed to the helical domain of *T. maritima* GatB(303-415) (cyan). B. The crystal structure of Gln4(119-178) (red) is superposed on the tail domains of GatB (422-481) (cyan). 121

Figure 4.8 The linker between the two domains in Gln4(1-187) likely behaves as a hinge, is highly conserved and is important for tRNA binding. A. Structure of Gln4(1-187) (red) superposed on *Tm*GatB (light gray) and *Sa*GatB (dark gray) by alignment of the tail domains. B. Conservation of GlnRS NTD sequences, red-≥90%; blue-≥ 70%, with arrow at $Gln4_{187}$., aligned using Multalin (Corpet, 1988). C. Conserved residues are highlighted on Gln4(1-187) according to the color code in B with the NTD backbone shown in light grey. D. Close contacts between W160 of the Gln4 NTD and other residues. 123

Figure 4.9 SAXS data shows that the NTD crystal structure is similar to that found in solution. A. Simulated scattering profiles calculated by CRYSOL for the Gln4 NTD (red), *Tm*GatB (green), and *Sa*GatB (blue) are shown overlaid on top of experimental SAXS data from the Gln4 NTD in solution. Goodness of fit values (χ) are given in parentheses. B. The *ab initio* envelope reconstructed from the experimental scattering profile of the Gln4 NTD is shown superimposed onto the crystal structures of the Gln4 NTD (red), *Tm*GatB (green), and *Sa*GatB (blue). The orientation of these structures is similar to that presented in Figure 4.8, A. 125

Figure 4.10 Mutations in conserved amino acids in the putative hinge of the NTD affect the interaction of Gln4 with native $tRNA^{Gln(CUG)}$. A., B. EMSA wild type and mutant

Gln4 proteins (23 nM to 2,017 nM). C. Binding as a function of Gln4 protein concentration. 126

Figure 4.11 Phylogenetic tree indicating the relationships between the organisms from which GlnRS NTDs were compared in Figure 4.7B. The scale bar indicates the number of amino acid substitutions per site. 128

Figure 4.12 Phylogenetic distribution of GlnRSs with and without N-terminal appended domains. Eukaryotic and bacterial GlnRS sequences were aligned using ClustalW and a phylogenetic tree was constructed using MEGA5, with 200 bootstraps carried out to test statistical relevance. In addition to bacteria, a set of Euglenozoa protists lacks the appended domain.. 129

Figure 5.1 X-ray Crystal Structure of *Sc*GlnRS. Cartoon representation of the *Sc*GlnRS structure is shown color-coded according to domains and insertions relative to *E. coli* GlnRS. Residues 1-214 are missing in the electron density. Domains are labeled with corresponding amino acid numbers.. 142

Figure 5.2 Disorder Prediction Analysis of the Primary Sequence of *Sc*GlnRS. The probability of disorder is shown on the y-axis and the residue number is shown on the x-axis. The linker connecting the N-terminal and C-terminal domains extends from residue 188 to 214. Disorder probability was calculated using DISOPRED2. 144

Figure 5.3 Dissolved *Sc*GlnRS crystals show only full-length *Sc*GlnRS protein. A. SDS PAGE gel showing dissolved Gln4 protein crystals is shown in the left lane, and the molecular weight ladder is shown in the right lane. Labels for the full-length protein, and both the NTD and CTD fragments are given. The presence of full-length Gln4 and absence of NTD and CTD fragments indicates that only the full-length protein is present in the crystal. B. Western blot using an anti-His antibody for crystals containing both His-tagged (left-most lane) and non-His-tagged (right-most lane) Gln4 protein. The molecular weight ladder is shown in the middle lane. 144

Figure 5.4 Crystallographic Packing Arrangement of *Sc*GlnRS. Shown oriented with the z-axis perpendicular to the page. Minimum and maximum diameters of solvent channels are labeled with dashed lines. Arg215 is labeled. 145

Figure 5.5 Domain Architecture of GlnRS. Domains for GlnRS are shown for Fungi, Eukaryotes (other than Fungi), and Bacteria and Archaeal GluRS. NTD = N-terminal Domain. Major domains are labeled NTD, Catalytic Domain and Anticodon Binding Domain. Insertion 1, Insertion 2, Insertion 3, and Loop 1 are labeled. 146

Figure 5.6 Zinc Finger Motif in *Sc*GlnRS. A. Location of zinc ion shown as gray sphere. B. Zinc binding motif residues shown coordinating zinc ion. 147

Figure 5.7 SAXS Data of Full-length *Sc*GlnRS. For clarity, all plots have been scaled. A. The log of the scattering intensity plotted as a function of momentum transfer of full-length *Sc*GlnRS (red) and the CTD (blue). B. Guinier plots of full-length *Sc*GlnRS (red) and the CTD (blue). The dashed line shows linearity in Guinier region. C. Porod-Debye plots of full-length *Sc*GlnRS (red) and the CTD (blue). The linear fits are based on the corresponding Porod region. Full-length *Sc*GlnRS has a Porod exponent of 3.4 and the CTD has a Porod exponent of 4.0. D. Pair distribution functions of full-length *Sc*GlnRS (red) and the CTD (blue). 149

Figure 5.8 SAXS Rigid Body Model of Full-length *Sc*GlnRS. A. Two orientations rotated 90° relative to each other are shown in cartoon representation and colored according to secondary structure and superposed onto the *ab initio* envelope shown in gray. B. The calculated scattering of the rigid body model (solid black line) fit to the experimental SAXS data (red circles). 151

Figure 5.9 R_g Distribution of Ensembles Calculated by the Ensemble Optimization Method. The R_g distribution of the pool of 10,000 random conformers (red) is shown with the distribution of the ensemble calculated using experimental SAXS data (purple) and using the simulated scattering profile of the rigid body model (blue). 151

Figure 5.10 Homology Model of Full-length *Sc*GlnRS Bound to tRNA^{gln}. A. Full-length *Sc*GlnRS shown bound to tRNA^{gln}. B. Enlarged and rotated model showing gap between NTD helical subdomain and tRNA molecule. 153

Figure 5.11 Comparison of NTD Position Before and After tRNA Binding. A. *Sc*GlnRS is shown in cartoon representation. The NTD position prior to tRNA binding determined by SAXS rigid body modeling is shown in red. The NTD position upon binding tRNA predicted by homology modeling is shown in blue. The CTD is shown in green. B. The calculated scattering of the rigid body model (red line) and the homology model (blue line) are shown fitted to the experimental SAXS data (circles). 154

Figure 5.12 Molecular Dynamics Simulation of *Sc*GlnRS bound to tRNA^{gln}. A. Plot of backbone RMSD of molecular dynamics trajectory as a function of time fit to the structure at time t=0 ns and t=10 ns. B. *Sc*GlnRS bound to tRNA^{gln} after 70 nanoseconds of molecular dynamics simulation colored according to secondary structure. 156

Figure 5.13 NTD undergoes conformational change after binding to tRNA. A. Plot of backbone RMS fluctuation as a function of residue. The RMS has been calculated as the deviation from the starting structure. B. Structure of *Sc*GlnRS before (red) and after (blue) molecular dynamics simulation. The tRNA molecule has been removed for clarity. 156

Figure 5.14 Comparison of NTD Structure Before and After Molecular Dynamics Simulation. The position of the NTD before the simulation is shown in gray and after the simulation in cyan. The solid black arrow shows the degree and direction of angular motion calculated by DynDom. 157

Figure 5.15 Conserved Structural Motifs of GlxRS Family of Enzymes. Sections of residues are shown to demonstrate the conservation of select motifs identified through PROMALS3D sequence and structure alignment. Structures of *Mt*GluRS, *Sc*GlnRS and *Ec*GlnRS were used to aid sequence alignment and identify motifs. Motifs are boxed and labeled as described in the text. Zinc coordinating residues of the zinc-binding motif are highlighted in red. 158

Figure 5.16 GlxRS Evolutionary Pathway. Boxed numbers refer to steps in the pathway. Unboxed numbers are ClustalW scores for adjacent protein sequences. Further details are given in the text. Steps: 1. The last universal common ancestor containing only the GluRS catalytic domain splits and adds an alpha helical anticodon binding domain in bacteria and a beta barrel anticoding binding domain in archaea. 2. The archaeal GluRS evolves in early eukaryotes by adding the P-loop in the anticodon binding domain. 3a. The early eukaryotic GluRS evolves into the modern eukaryotic GluRS by adding an NTD. 3b. The early eukaryotic GluRS evolves into the first dedicated GlnRS in early eukaryotes by altering the catalytic domain, and adding the G-loop and the NTD. 4a. The early eukaryotic GlnRS experiences little evolutionary change in modern eukaryotes. 4b. Only in fungi, Loop 1 is added with the zinc-binding motif. 4c. The early eukaryotic GlnRS evolves into the bacterial GlnRS by losing the appended NTD and gaining Loop 1. The dashed line represents the less likely route of GlnRS horizontal gene transfer to prokaryotes. 159

List of Abbreviations

1D	One Dimensional
3D	Three Dimensional
ADP	Adenosine Diphosphate
AMP	Adenosine Monophosphate
ATP	Adenosine Triphosphate
BSA	Bovine Serum Albumin
CD	Circular Dichroism
CTD	C-terminal Domain
Da	Daltons
DF	Degrees of Freedom
D_{max}	Maximum Dimension
DNA	Deoxyribonucleic Acid
EcGlnRS	*Eschericia coli* Glutaminyl-tRNA Synthetase
EOM	Ensemble Optimization Method
FEA	Finite Element Analysis
FFEA	Fluctuating Finite Element Analysis
GatCAB	Glutamine Amidotransferase CAB
Gln	Glutamine
GlnRS	Glutaminyl-tRNA Synthetase
Glu	Glutamate
GluRS	Glutaminyl-tRNA Synthetase
GlxRS	Glutamyl- or Glutaminyl-tRNA Synthetase
kDa	Kilodaltons
MD	Molecular Dynamics
MtGluRS	*Methanothermobacter thermoautotrophicum* Glutamyl-tRNA Synthetase

NESG	Northeast Structural Genomics
NgGlnRS	*Naegleria gruberi* Glutaminyl-tRNA Synthetase
NgGluRS	*Naegleria gruberi* Glutamyl-tRNA Synthetase
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect Spectroscopy
NSD	Normalized Spatial Discrepancy
NTD	N-terminal Domain
PDB	Protein Data Bank
R_g	Radius of Gyration
RMSD	Root-mean-square Deviation
RNA	Ribonucleic Acid
SaGatB	*Staphylococcus aureus* Glutamine Amidotransferase Subunit B
SANS	Small Angle Neutron Scattering
SAS	Small Angle Scattering
SAXS	Small Angle X-ray Scattering
ScGlnRS	*Saccharomyces cerevisiae* Glutaminyl-tRNA Synthetase
ScGluRS	*Saccharomyces cerevisiae* Glutamyl-tRNA Synthetase
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
TmGatB	*Thermotoga maritima* Glutamine Amidotransferase Subunit B
tRNA	Transfer Ribonucleic Acid

Abstract

Proteins are the machinery of the cell, carrying out critical functions in living organisms. Amino acyl tRNA synthetases perform the vital cellular function of attaching amino acids to their cognate tRNA molecule for their use in protein synthesis. In essence, these enzymes function as the “codebook” of life by translating information inscribed in codons into amino acid sequence. While a dedicated tRNA synthetase is used to attach most amino acids to their cognate tRNA molecules, glutaminyl-tRNA synthetase (GlnRS) is absent in most prokaryotes. These organisms instead use an indirect method of attaching glutamine to tRNA by first misacylating $tRNA^{gln}$ with glutamic acid. Glutamic acid is then converted to glutamine by an amidotransferase enzyme, GatCAB. Eukaryotes, however, do not employ this indirect route, but encode a dedicated GlnRS to directly attach glutamine to $tRNA^{gln}$. Additionally, eukaryotic GlnRS has an appended N-terminal domain that is absent from its prokaryotic homolog, whose function is currently unknown. To date, no eukaryotic GlnRS structure is known.

This work describes a complementary approach to understanding the structure and function of yeast GlnRS, Gln4. To determine the structure and function of Gln4, we used molecular biology, X-ray crystallography, small angle X-ray scattering (SAXS), and bioinformatics. To objectively evaluate SAXS data, we have developed statistical methods using data from high-throughput structural genomics initiatives. Using SAXS data and the crystal structures of both the N and C-terminal domains of Gln4, we present a model of the first full-length eukaryotic GlnRS in solution. Our results describe a previously unknown structural homology between the appended N-terminal domain of Gln4 and the B subunit of GatCAB. Using this structural homology, coupled with the known structure of *E. coli* GlnRS bound to tRNA and molecular dynamics simulations, we present the first model of a full-length eukaryotic GlnRS bound to tRNA and a mechanism of binding to tRNA.

1 Introduction

Structural biology is used to uncover the three dimensional structures of biological macromolecules to understand the relationship between structure and function. X-ray crystallography and nuclear magnetic resonance (NMR) are the predominant techniques used to determine structural details at atomic resolution (Berman et al., 2000). These techniques, however, are limited in their application due to difficulties in creating single, diffraction-quality crystals or due to limitations imposed by protein size. Small angle X-ray scattering (SAXS) is a solution technique that can yield low-resolution structural information about size, shape, and protein flexibility (Putnam et al., 2007). Since SAXS is a solution technique, it does not require forming a crystal, and additionally is unrestricted by protein size. SAXS is particularly useful as a complementary technique, for example in determining the organization of a macromolecular complex from high-resolution structures of subunits or for visualizing regions of protein structure unresolved by X-ray crystallography (Putnam et al., 2007; Svergun and Koch, 2003). The versatility of SAXS and the ease of preparing solutions and performing experiments have caused the technique to become an increasingly important tool for the structural biologist to gain a complete understanding of a biological system.

1.1 History

In 1939 André Guinier published the now famous paper discussing the relationship between the SAXS profile and particle size (Guinier, 1939). Guinier realized this relationship rather fortuitously when he was attempting to correct for parasitic scatter during his studies on diffuse scatter. He developed the Guinier Camera to study weak diffuse scatter and found that scattering at the smallest angles was only present for heterogeneous solutions. He also found that the X-ray intensity was strongest at these angles for fine grains 10 to 100 nm in size and determined a method, known as the Guinier approximation, to calculate the sizes of the particles from the small angle scattering (Guinier and Foumet, 1955).

SAXS began being used on biological macromolecules in the 1960s as a method to gain low-resolution structural information in the absence of crystals (Svergun and Koch, 2003). The advent of synchrotron radiation in the 1970s saw a breakthrough in the ease of SAXS data collection. The introduction of high-flux neutron sources enabled contrast variation studies using small angle neutron scattering (SANS) of perdeuterated solutions (Engelman and Moore, 1975; Ibel, 1975).

Until the 1990s, only parameters about shape and size could be extracted from SAXS data including radius of gyration and particle volume, and information about the 3D structure of a particle was limited to modeling estimations using simple geometrical bodies such as ellipsoids. Advances in computer algorithms brought about the ability to obtain reliable *ab initio* 3D shape reconstructions from the 1D intensity profiles and rigid body modeling (Svergun and Koch, 2003). Further developments in instrumentation and third generation synchrotron radiation sources enabled high-throughput as well as sub-millisecond time-resolved SAXS (Grant et al., 2011; Hura et al., 2009; Pollack et al., 1999). With these advances in methodology, the applicability of small angle scattering has grown, as has the user community.

1.2 Theory

Elastic scattering of X-rays occurs when electrons in a medium resonate with the frequency of the incident photon and emit secondary waves that interfere. The interference of these waves is governed by the geometrical properties of the source of the scattering. Figure 1.1 shows the basic geometry of scattered radiation from a spherical particle with uniform density. Coherent scattering between two electrons inside the particle only occurs when the scattered radiation from each is in phase. The path difference for each of the smaller and larger spheres is one wavelength, λ, however the angle, 2θ, at which this occurs is much smaller for the larger particle than it is for the smaller particle. Larger particles will therefore exhibit narrower intensity profiles focused at smaller angles than the profiles of smaller particles. It follows from this simple geometrical picture that the scattering of any particle shape can be calculated, even for anisotropic particles, by averaging the scattering from every possible orientation.

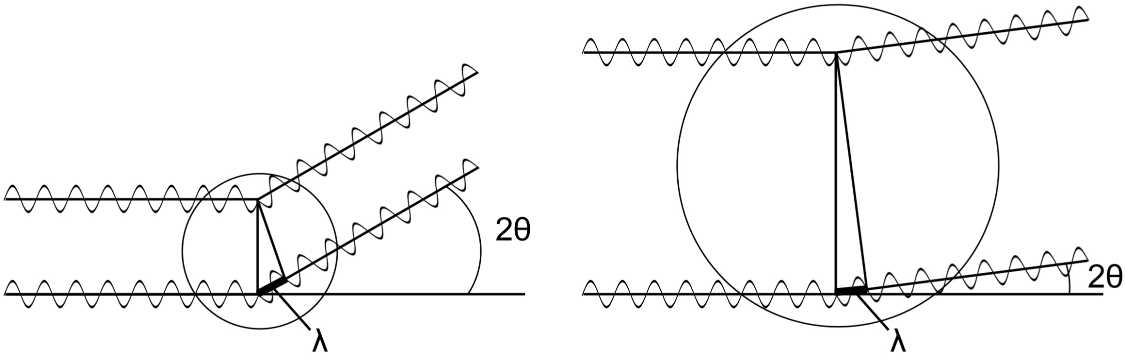


Figure 1.1 Geometry of Scattered Radiation. The larger particle on the right scatters radiation at smaller angles than does the smaller particle on the left. λ is the wavelength and 2θ is the angle of the scattered radiation.

1.2.1 Experimental Setup

A typical SAXS experiment consists of an X-ray source, optics, a sample chamber, and a detector. Laboratory setups often use X-ray tubes or rotating anodes to produce X-ray beams that are then collimated to a small diameter to pass through the sample chamber where the beam is scattered and collected on a detector. In a synchrotron setup the primary X-rays are polychromatic and must pass through a monochromator before passing through the sample chamber. A simple geometric setup for a SAXS experiment is shown in Figure 1.2. As the incident beam passes through the sample chamber, X-rays interact with solution and scatter at an angle dependent upon the interatomic pair distances between the scattering electrons within a particle. The momentum transfer, q, is related to the scattering angle 2θ by

$$q = \frac{4\pi \sin\theta}{\lambda} \qquad (1.1.)$$

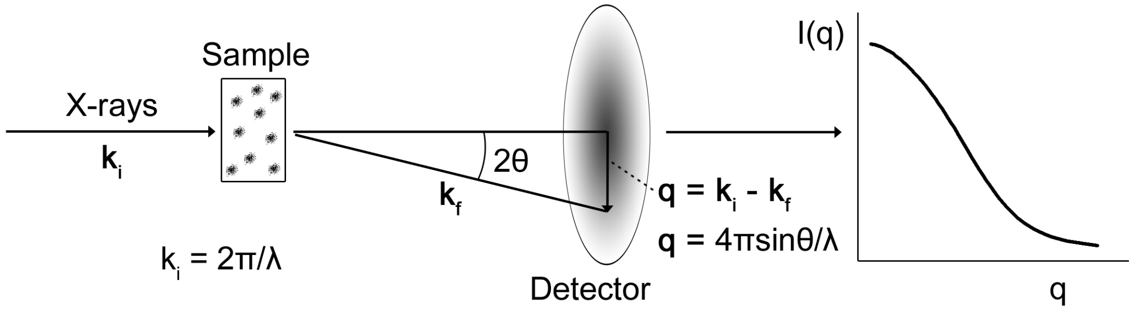


Figure 1.2 Geometry of SAXS Experimental Setup. Incident X-rays corresponding to wave vector k_i with magnitude 2π/λ pass through the sample chamber and scatter with an angle of 2θ. The isotropic detector image is integrated and radially averaged to produce a 1D scattering profile of intensity (I) vs momentum transfer (q).

While data from higher concentrations are collected to increase signal to noise, data from lower concentrations are also collected to ensure no concentration dependent effects are occurring. Volumes required vary between experimental set ups, typically ranging from 10 μL to 60 μL for each concentration (Hura et al., 2009; Smolsky et al., 2007). SAXS data from both the protein solution as well as an identical buffer blank are collected and the scattering from the buffer is subtracted from the total protein solution scattering resulting in an intensity profile corresponding to the protein of interest. The sample to detector distance is an important parameter to define prior to data collection. If particularly large particles are being studied, then the sample to detector distance should be increased to ensure scattering at the lowest angles is collected.

In addition to increasing concentration, signal to noise can be improved by increasing the time of exposure. Increasing the time of exposure, however, can result in radiation damage (Kuwamoto et al., 2004). Therefore data collection often involves multiple exposures for comparison. Laboratory systems typically require exposure times up to one hour whereas synchrotron sources can provide sufficient signal to noise with exposures as short a one second.

1.2.2 Theory

The scattering from a single particle is governed by the electron density function in real space, *ρ(**r**)*. *ρ(**r**)* is related to the scattered intensity by the Fourier transform

$$I(q) = \int \bar{\rho}^2(\boldsymbol{r}) \cdot e^{-iqr} \ dV \tag{1.2.}$$

where ρ^2 is the autocorrelation function of the electron density, commonly known as the Patterson function, *dV* is a volume element located at position ***r*** and the integral is over the total volume (Glatter and Kratky, 1982). The problem of small angle scattering can be made simpler by assuming the following two restrictions are met: 1) the system is isotropic, as freely tumbling particles in solution, and 2) long range order is negligible, *i.e.* particles do not interact. The first assumption allows the use of the Debye approximation (Debye, 1915) for averaging over all ***r***

$$\langle e^{-iqr} \rangle = \frac{\sin qr}{qr} \tag{1.3.}$$

where *r* is the magnitude of position ***r***. Combining equations 1.2 and 1.3 yields

$$I(q) = \int 4\pi r^2 \cdot \bar{\rho}^2(r) \cdot \frac{\sin qr}{qr} \ dr \tag{1.4.}$$

Equation 1.3 is given by the spherical average over all ***r***, which leads to ambiguity in the SAXS profile. The result is that SAXS data can be fit equally well by multiple models, leading to a problem of uniqueness (Volkov and Svergun, 2003). Additionally, the averaging of all possible orientations of the particle as it tumbles in solution yields an effective spatial resolution that is not directly related to the corresponding Bragg resolution of *d = 2π/q* due to the loss of information caused by the spherical averaging.

The second assumption says that at sufficiently large *r* no interactions exist and the electron density tends towards an average value, $\bar{\rho}$. This background value is subtracted from the particle density, $\rho - \bar{\rho}$. This is an important point that warrants further discussion. Given that the background density of the solvent is subtracted from the particle electron density, SAXS is inherently a contrast method. The electron density of water is 0.33 $e^{-}/Å^{3}$, whereas for protein the average electron density 0.44 $e^{-}/Å^{3}$ (Putnam et al., 2007). The scattering intensity is therefore approximately 5% of what it would be if in a vacuum (Das and Doniach, 2006). It is from this difference in electron density that small angle scattering is detected. Solvent conditions are often not simply water, but salt solutions sometimes containing several different molecules of varying electron densities. In formulating an experiment it is important to prepare a solvent with as low of an electron density as possible, to increase the contrast between solute and solvent. For example, high molar salt concentrations may result in very small differences between protein and solvent density, resulting in intensities with low signal to noise and poor quality data. Nucleic acids, which contain many electron-rich phosphorous atoms, have very high electron densities compared to protein, and can give very high signal to noise even at low concentrations (Lipfert and Doniach, 2007).

This contrast property of small angle scattering has been exploited in neutron scattering as well, in a technique known as contrast matching (Jacques and Trewhella, 2010). This is a very useful tool in particular for macromolecular complexes containing mixtures of proteins, nucleic acids and lipids. Using varying mixtures of H_2O, D_2O and perdeuterated solutes, the scattering from single or multiple components of a complex can be isolated and studied while the scattering of the remaining components matches the solvent and is effectively zero. While contrast matching can be performed using X-rays, the method is difficult and typically yields very little signal, since most elements in biological solutions have similar electron densities and therefore similar scattering cross sections. However, in neutron scattering the scattering cross section is not dependent on the number of electrons in the atom, but on the properties of the nucleus. The deuterium isotope of hydrogen has a significantly different cross section than that of hydrogen, and this difference is what is often exploited in neutron scattering contrast matching.

The autocorrelation function of the difference between the particle and solvent electron density is termed *γ(r)*. The histogram of all interatomic pair distances is known as the pair distribution function, *P(r)*, which is related to *γ(r)* according to

$$P(r) = r^2 \, \gamma(r) \tag{1.5.}$$

This can now be combined with equation 1.4 to yield the relationship between the scattered intensity and the pair distribution function:

$$I(q) = 4\pi \int_0^{D_{max}} P(r) \cdot \frac{\sin qr}{qr} \, dr \tag{1.6.}$$

where the limits of the integral range from zero to the maximum dimension of the particle (D_{max}) (Feigen and Svergun, 1987; Glatter and Kratky, 1982). When analyzing the collected intensity data, the pair distribution function can be obtained similarly by the inverse Fourier transform

$$P(r) = \frac{r^2}{2\pi^2} \int_0^{\infty} I(q) \cdot \frac{\sin qr}{qr} \, dq \tag{1.7.}$$

Knowing the pair distribution function we can now uncover information about the shape and size of a particle from the SAXS profile.

One of the first parameters often extracted from the data is the radius of gyration, *R_g*. *R_g* is defined as the root mean square distance of all atoms in a particle from the center of mass. The Guinier approximation (Guinier, 1939; Guinier and Foumet, 1955) can be used to obtain *R_g*, which can be derived as follows. The Maclaurin series of sin(*qr*)/*qr* is given by

$$\frac{\sin qr}{qr} = 1 - \frac{(qr)^2}{3!} + \frac{(qr)^4}{5!} - \cdots \tag{1.8.}$$

Considering only very small angles, we will approximate by ignoring terms of order greater than two. Equation 1.8 can then be combined with equation 1.6 to yield

$$I(q) = 4\pi \int_0^{D_{max}} P(r)\ dr - \frac{4\pi}{6} \int_0^{D_{max}} P(r) \cdot q^2 r^2\ dr \tag{1.9.}$$

By defining *I(0)* and *R_g* as follows

$$I(0) = 4\pi \int_0^{D_{max}} P(r)\ dr \tag{1.10.}$$

$$R_g^2 = \frac{\int_0^{D_{max}} P(r) \cdot r^2\ dr}{2\int_0^{D_{max}} P(r)\ dr} \tag{1.11.}$$

equation 1.9 becomes

$$I(q) = I(0)\left(1 - \frac{q^2 R_g^2}{3}\right) \tag{1.12.}$$

Guinier recognized this as the first two terms in the Maclaurin series of *e^x* such that

$$I(q) = I(0)\, e^{-\frac{q^2 R_g^2}{3}} \qquad (1.13.)$$

for sufficiently small *q*. Thus, by taking the natural logarithm of both sides we obtain

$$\ln I(q) = \ln I(0) - \frac{q^2 R_g^2}{3} \qquad (1.14.)$$

which is in the familiar form of the linear equation *y* = *mx*+*b*. Therefore, by plotting the natural log of intensity versus the square of the momentum transfer we can obtain R_g via the slope of the line such that

$$R_g = \sqrt{-\frac{slope}{3}} \qquad (1.15.)$$

Checking linearity in the low angle Guinier region is an important quality control step, since deviation from linearity suggests polydispersity is present and therefore the scattering intensity will reflect a population-weighted average of all particles present in solution.

As can be readily seen from the derivation, approximations resulting from truncation of the infinite series require *q* to be very small. Due to the truncation, the smaller the angle is, the smaller the contribution from higher order terms are and the more accurate the R_g estimation is. However, in practice, the number of data points at low angles, the inability to collect data approaching $q = 0$, and the degree of noise in the data can cause extremely small angles to result in inaccurate approximations and unreliable estimations of linearity (Feigen and Svergun, 1987). Therefore a balance has been struck such that for globular particles, the Guinier approximation is valid for angles where $qR_g < 1.3$ (Guinier and Foumet, 1955).

In addition to R_g, the forward scattering, *I(0)*, can also be estimated from equation 1.14. The actual forward scattering is not collected in a SAXS experiment, since it is in line with the primary beam, which is masked by the beamstop. However, by extrapolating the linear fit of the Guinier region, *I(0)* can be estimated. *I(0)* is an important parameter as it is directly related to molecular mass (Mylonas and Svergun, 2007).

The Guinier region for a small particle will be larger than the Guinier region for a large particle. For particularly large particles this can present a problem since very few data points may be in the Guinier region, making it difficult to accurately estimate R_g and *I(0)*. However, these parameters can also be calculated using the pair distance distribution function. Equations 1.10 and 1.11 describe how R_g and *I(0)* can be calculated by integrating over *P(r)*. This has the advantage that all points in the experimental profile are used in the calculation of these parameters, which greatly exceeds the number of points in the Guinier region. Since the entire available *q*-range is used in the calculation, this method also has the advantage of being less sensitive to aggregates that may be distort the Guinier region estimation (Putnam et al., 2007).

In the general case, the intensity profile is the relationship between the form factor, describing the shape and size of a particle, and the structure factor, describing the interactions between neighboring particles in solution. This relationship is given by

$$I(q) = F(q) \times S(q) \qquad (1.16.)$$

where *F(q)* is the form factor and *S(q)* is the structure factor (Putnam et al., 2007; Svergun and Koch, 2003). According to Restriction 2, interparticle interactions must not exist in solution for the prior derivation to be valid, resulting in a structure factor of one. Therefore, the scattered intensity is assumed equivalent to the form factor. If attraction between particles exists in solution, then the structure factor will cause the data to trend upwards as *q* approaches zero. If repulsion exists, then the data will trend downwards. These trends in the data will cause nonlinearity in the Guinier region and will distort the R_g and *I(0)* determined from the Guinier approximation. This is generally not desired and therefore protein concentrations are often diluted to alleviate these effects. If these distortions are mild, extrapolation of multiple concentrations can be used to obtain values from a theoretically infinite dilution (Konarev et al., 2003).

1.3 High-throughput SAXS

The rate at which genomic sequence data is becoming available has rapidly increased providing biologists with a vast surplus of biologically diverse macromolecules to study the relationship between structure and function. The Protein Structure Initiative (PSI) has been established to determine the structures of a broad range of macromolecules pertaining to biological and biomedical problems. The Northeast Structural Genomics Consortium (NESG) is one of four large-scale centers funded by the National Institutes of Health (NIH) as part of the PSI. Addressing the challenge of understanding the structure-function relationship for the increasingly large quantity of unique proteins sequences requires the ability to characterize not only the structures of individual macromolecules, but also their complex assemblies and conformations in solution. X-ray crystallography and NMR have proven to be highly effective methods to uncover the high-resolution structures of many of these molecules, however limitations of each technique have greatly restricted their applicability to only a small fraction of macromolecules. Only 12% of soluble, purified proteins have resulted in structures deposited into the Protein Data Bank (PDB) (Chen et al., 2004). To keep pace with the rapidly growing database of genomic sequence data, a greater capacity of high-throughput structural biology approaches must be developed.

SAXS offers many advantages in high-throughput structural analysis, such as ease in sample preparation, low volumes required, structural characterization in solution, applicability to a very wide range of molecular sizes, and the speed and efficiency with which data can be collected. The ability to perform high-throughput SAXS experiments at several beam lines now exists, generating vast quantities of data that require analysis. Some semi-automated software exists to quickly provide users with parameters, such as R_g, *I(0)*, and D_{max}, which is necessary to enable rapid characterization of SAXS data (Franke et al., 2012). However, subjective interpretation of data is still required to fully assess data quality and ensure that conclusions that are drawn are not erroneous. Historically, expert analysis has been required to accurately evaluate data quality. While X-ray crystallography and NMR have quantitative standards by which data and model quality can be compared, such as R factors, SAXS has no equivalent of such metrics. A set of parameters that serves as a guideline for publication quality SAXS data has been published (Jacques et al., 2012), however this still only provides a qualitative measure of SAXS data quality that is largely dependent on the expertise of the user. Since the SAXS user community has experienced extensive growth recently, SAXS is no longer a technique exclusive to scientists specially trained in the field, and therefore the risk of conclusions drawn from unreliable data has also increased. Therefore there is a growing need for objective, statistically significant measures of SAXS data quality that can be employed by users of varying degrees of expertise. Furthermore, software procedures that can provide these objective measures remove the need for time-consuming and laborious manual analysis, which is required for efficient, high-throughput structural pipelines.

1.4 Scope of This Thesis

In this thesis SAXS is introduced and the utility of SAXS discussed as a complementary tool used to fully understand biological systems. Chapter 2 describes a study of 28 proteins, carried out as part of a high-throughput structural pipeline, for which SAXS data were collected, whose high-resolution structures are known by X-ray crystallography, NMR, or both. Analysis using manual data quality metrics coupled with semi-automated software programs demonstrate that resulting SAXS parameters including R_g, D_{max}, molecular weight, and even low-resolution molecular envelopes agree well with high-resolution structural data. Moreover, in several cases analyzed, SAXS yields additional information such as oligomeric state and visualization of regions of structure unresolved by X-ray crystallography providing a more complete understanding of the biological system.

Chapter 3 presents a new automated approach, called SAXStats, to evaluate data quality utilizing objective, statistical measures. SAXStats was applied to a set of 100 proteins in a high-throughput manner and structural information for each protein sample was collected including, but not limited to, R_g, D_{max}, and molecular weight. By consolidating this large quantity of data and evaluating trends, a relationship has been found between the precision with which SAXS parameters can be determined and the signal to noise of SAXS intensity. This provides the user community with a valuable resource, which can be used to increase the likelihood of successful SAXS experiments by ensuring protein solutions are prepared with sufficient concentration prior to data collection.

Chapters 4 and 5 demonstrate the usefulness of SAXS as a complementary tool not only in a high-throughput environment, but also in a specific biological case, that of the yeast glutaminyl-tRNA synthetase, Gln4. In this study a suite of biochemical, structural and bioinformatics tools has been used to study Gln4. The combination of these techniques yielded the first structural characterization of any eukaryotic glutaminyl-tRNA synthetase. SAXS data, coupled with objective evaluation provided by SAXStats, proved to be integral in understanding the complete biological system which otherwise would not have been possible.

1.5 References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J. (2004). TargetDB: a target registration database for structural genomics projects. Bioinformatics *20*, 2860-2862.

Das, R., and Doniach, S. (2006). Structural studies of proteins and nucleic acids in solution using small angle x-ray scattering (SAXS) (Berlin, Springer).

Debye, P. (1915). Zerstreuung von Röntgenstrahlen. Annalen der Physik *351*, 809-823.

Engelman, D.M., and Moore, P.B. (1975). Determination of quaternary structure by small angle neutron scattering. Annu Rev Biophys Bioeng *4*, 219-241.

Feigen, V., and Svergun, D. (1987). Structure analysis by small-angle X-ray and neutron scattering., Vol XIII (New York/London, Plenum Press).

Franke, D., Kikhney, A.G., and Svergun, D.I. (2012). Automated acquisition and analysis of small angle X-ray scattering data. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment *689*, 52-59.

Glatter, O., and Kratky, O. (1982). Small angle x-ray scattering (London ; New York :, Academic Press).

Grant, T.D., Luft, J.R., Wolfley, J.R., Tsuruta, H., Martel, A., Montelione, G.T., and Snell, E.H. (2011). Small angle X-ray scattering as a complementary tool for high-throughput structural studies. Biopolymers *95*, 517-530.

Guinier, A. (1939). La diffraction des rayons X aux tres petits angles; application a l'etude de phenomenes ultramicroscopiques. Ann Phys (Paris) *12*, 161-237.

Guinier, A., and Foumet, F. (1955). Small Angle Scattering of X-rays (New York, Wiley Interscience).

Hura, G.L., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., 2nd, Tsutakawa, S.E., Jenney, F.E., Jr., Classen, S., Frankel, K.A., Hopkins, R.C.*, et al.* (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods *6*, 606-612.

Ibel, K. (1975). Comparison of neutron and X-ray scattering of dilute myoglobin solutions. J Mol Biol *93*, 255-265.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta crystallographica Section D, Biological crystallography *68*, 620-626.

Jacques, D.A., and Trewhella, J. (2010). Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls. Protein Sci *19*, 642-657.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr *36*, 1277-1282.

Kuwamoto, S., Akiyama, S., and Fujisawa, T. (2004). Radiation damage to a protein solution, detected by synchrotron X-ray small-angle scattering: dose-related considerations and suppression by cryoprotectants. J Synchrotron Radiat *11*, 462-468.

Lipfert, J., and Doniach, S. (2007). Small-angle X-ray scattering from RNA, proteins, and protein complexes. Annu Rev Biophys Biomol Struct *36*, 307-327.

Mylonas, E., and Svergun, D.I. (2007). Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. J Appl Crystallogr *40*, s245-s249.

Pollack, L., Tate, M.W., Darnton, N.C., Knight, J.B., Gruner, S.M., Eaton, W.A., and Austin, R.H. (1999). Compactness of the denatured state of a fast-folding protein measured by submillisecond small-angle x-ray scattering. Proc Natl Acad Sci U S A *96*, 10115-10117.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys *40*, 191-285.

Smolsky, I.L., Liu, P., Niebuhr, M., Ito, K., Weiss, T.M., and Tsuruta, H. (2007). Biological small-angle x-ray scattering facility at the Stanford synchrotron radiation laboratory. J Appl Crystallogr *40*, S453-S458.

Svergun, D.I., and Koch, M.H.J. (2003). Small-angle scattering studies of biological macromolecules in solution. Reports on Progress in Physics *66*, 1735.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.

2 Small Angle X-ray Scattering as a Complementary Tool for High-throughput Structural Studies

2.1 Introduction

Structural biology aims to understand life on an atomic scale by using structural information to discern a molecule's functional attributes. To date over 86,000 macromolecular structures have been deposited in the Protein Data Bank (PDB) (Berman et al., 2000). Of these approximately 86% were determined with single crystal X-ray diffraction methods. The importance of this method is reflected in the number of Nobel prizes that have been associated with it including the determination of the structure of DNA (Watson and Crick, 1953), the structure of vitamin B_{12} (Hodgkin et al., 1957), the structure of the photosynthetic reaction center (Deisenhofer et al., 1984), the enzymatic mechanism underlying the synthesis of adenosine triphosphate (Abrahams et al., 1994), the structure of potassium channels (Doyle et al., 1998), the molecular basis of eukaryotic transcription (Cramer et al., 2001) and most recently the ribosome (Schlunzen et al., 1995) and G-protein coupled receptors (Rasmussen et al., 2007). Unfortunately, most proteins do not readily produce diffraction-quality crystals. Where failure as well as success has been rigorously tracked only 34% of expressed and purified targets provide a crystal and only 12% result in a structure deposited in the PDB (Chen et al., 2004). Crystallographic structures require crystals, while crystallization remains fundamentally a hit-or-miss proposition.

The Hauptman-Woodward Medical Research Institute provides a High-throughput crystallization screening service to the structural genomics and biological crystallography community. Macromolecular samples are screened against 1536 chemically diverse cocktails (Luft et al., 2003) using the microbatch-under-oil technique (Chayen, 1996). This service has been in operation for over 10 years and to date has screened 12,500 proteins for over 1,000 laboratories worldwide. The screening laboratory has worked in close collaboration with the North East Structural Genomics (NESG) consortium, screening their samples for crystallization leads. In this effort approximately 50% of soluble proteins that enter the screening laboratory provide promising crystallization lead conditions; ~45% of these have been successfully optimized by NESG resulting in a PDB deposition. While this success rate is relatively good in the structural genomics field (providing evidence of good initial sample preparation and crystallization methods), this means that 78% of the soluble, purified proteins do not result in crystallographic structures. NMR techniques can provide structural information for samples recalcitrant to crystallization. In the NESG case, approximately 44% of the structural depositions result from NMR methods alone. For the Protein Structure Initiative (PSI) as a whole, only ~10% of the soluble purified targets make it to a PDB deposition, 31% of which were determined by NMR. To put this into perspective, there are more than 30,000 soluble, purified samples from the US PSI that failed to provide structures; this number is almost half of the current structural information in the PDB. Even low-resolution structural information from these samples would significantly enhance the understanding of the biological world.

Small Angle X-ray Scattering (SAXS) is a technique that can provide low-resolution structural information, a molecular envelope from a solution of the sample (Putnam et al., 2007); a crystal is not required. In this recently published study (Grant et al., 2011), high-throughput SAXS has emerged as a technique that is complementary to crystallography and NMR. The remainder of solutions from samples provided by NESG for high-throughput crystallization screening have been used for SAXS analysis. To date, this has been carried out for over five hundred samples. In this chapter the information obtained from SAXS on a subset of 28 samples where either crystallographic, NMR, or a combination of both of these structures are available is described. We demonstrate how information from SAXS can complement and enhance high-resolution structural information. Based upon these observations it is proposed that SAXS should be adopted as a routine, complementary analysis technique in structural biology that can be used to resolve and improve the interpretation of biological function from structural information.

This chapter has been published in (Grant et al., 2011). My role in this work involved the collection and analysis of SAXS data and the comparison with high-resolution structural data.

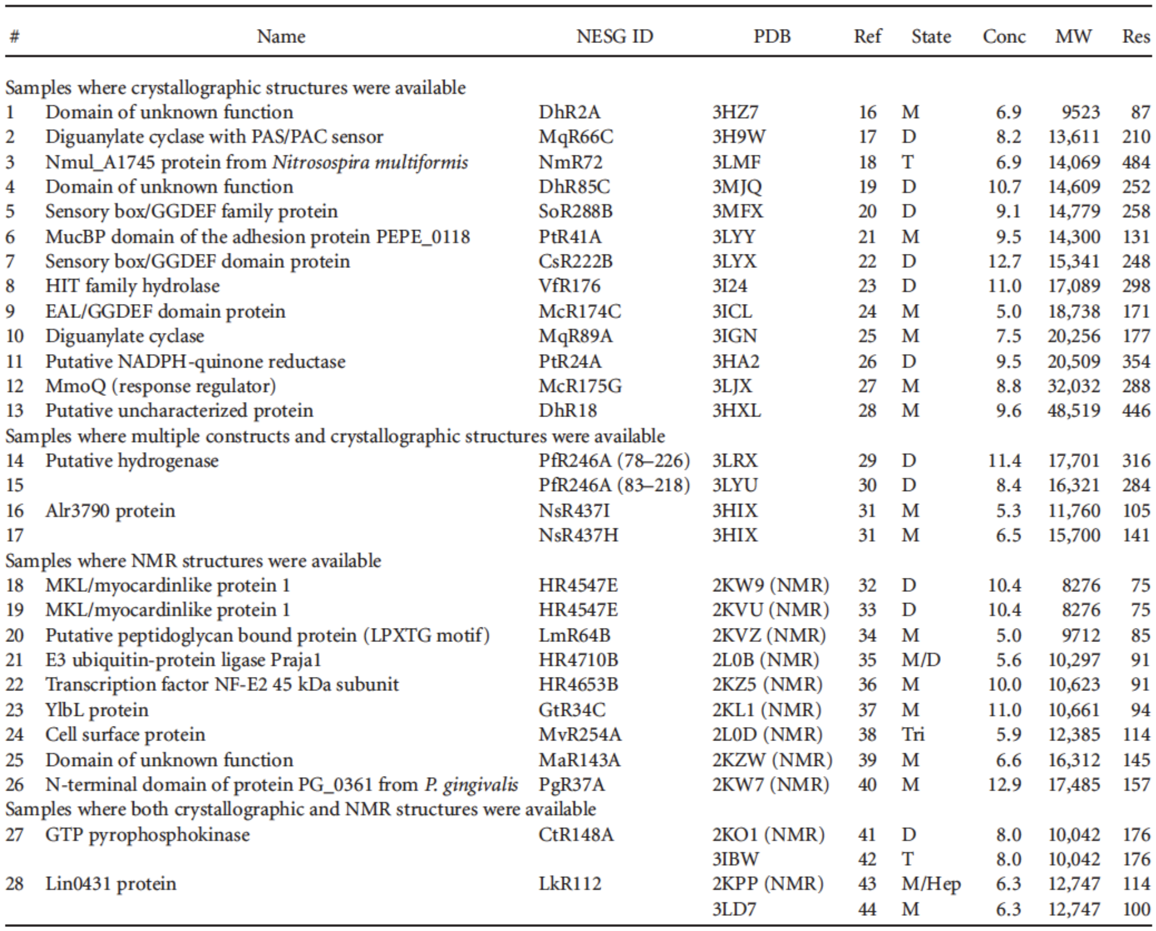
2.2 Materials and Methods.

2.2.1 Samples.

In this study complementary SAXS data were collected from 28 different protein samples where structural information is available through crystallography, NMR or a combination of both techniques. The protein samples are NESG targets that represent large protein domain families, biomedical themes, and targets nominated by the biomedical community. The NESG biomedical themes focus on eukaryotic proteins, particularly human proteins involved in cancer biology, protein-protein interaction networks, specific biochemical pathways, and proteins implicated in other human diseases. The protocols for selection, cloning, expression, purification and crystallization of each sample is described elsewhere (Acton et al., 2005; Acton et al., 2011; Xiao et al., 2010). After purification each sample is concentrated to between 5-10 mg/mL using an Amicon (Millipore, Billerica, MA) centrifugal filtration unit with a 5 kDa molecular weight cutoff membrane. SDS-PAGE and mass spectrometry analysis are used to confirm purity and molecular weight respectively. Analytical gel filtration with static light scattering detection is used to screen for aggregation and determine the oligomeric state of each sample.

Table 2.1 summarizes information about these 28 proteins, which are divided into four sets. The first set encompasses 13 proteins where a crystallographic structure is available. These range in molecular weight from 9.5 kDa to 48.5 kDa. The second set consists of two proteins where two constructs were studied for each protein target. Two crystallographic structures were available from different constructs for the first and a single crystallographic structure for the second. The third set consists of nine proteins for which there is an NMR structure; the fourth set includes two protein targets where both NMR and crystallographic structures are available.

Table 2.1 Samples used for the SAXS analysis are divided into four sets. The first set 1–13 contains 13 proteins, each having crystallographic structures. The second set 14–17 contains 2 proteins with two different constructs of the first having two crystallographic structures and the second a single structure. The third set 18–26 contains 9 proteins, each having an NMR structure. The fourth set 27–28 contains two proteins where both NMR and crystallographic structures are available. The sample name, ID, PDB identifier, reference, the oligomeric state in solution characterized on preparation by light scattering and gel filtration, initial concentration (mg/mL), molecular weight (Da) and number of residues are listed. The oligomeric state in solution is defined in the table as M (Monomer), D (Dimer), Tri (Trimer), T (Tetramer), Hep (Heptamer) or a combination. While all the samples have structures deposited in the PDB the majority are as yet unpublished. We are grateful to the authors in the references for the ability to use this structural data at this early stage.



#	Name	NESG ID	PDB	Ref	State	Conc	MW	Res
Samples where crystallographic structures were available								
1	Domain of unknown function	DhR2A	3HZ7	16	M	6.9	9523	87
2	Diguanylate cyclase with PAS/PAC sensor	MqR66C	3H9W	17	D	8.2	13,611	210
3	Nmul_A1745 protein from *Nitrosospira multiformis*	NmR72	3LMF	18	T	6.9	14,069	484
4	Domain of unknown function	DhR85C	3MJQ	19	D	10.7	14,609	252
5	Sensory box/GGDEF family protein	SoR288B	3MFX	20	D	9.1	14,779	258
6	MucBP domain of the adhesion protein PEPE_0118	PtR41A	3LYY	21	M	9.5	14,300	131
7	Sensory box/GGDEF domain protein	CsR222B	3LYX	22	D	12.7	15,341	248
8	HIT family hydrolase	VfR176	3I24	23	D	11.0	17,089	298
9	EAL/GGDEF domain protein	McR174C	3ICL	24	M	5.0	18,738	171
10	Diguanylate cyclase	MqR89A	3IGN	25	M	7.5	20,256	177
11	Putative NADPH-quinone reductase	PtR24A	3HA2	26	D	9.5	20,509	354
12	MmoQ (response regulator)	McR175G	3LJX	27	M	8.8	32,032	288
13	Putative uncharacterized protein	DhR18	3HXL	28	M	9.6	48,519	446
Samples where multiple constructs and crystallographic structures were available								
14	Putative hydrogenase	PfR246A (78–226)	3LRX	29	D	11.4	17,701	316
15		PfR246A (83–218)	3LYU	30	D	8.4	16,321	284
16	Alr3790 protein	NsR437I	3HIX	31	M	5.3	11,760	105
17		NsR437H	3HIX	31	M	6.5	15,700	141
Samples where NMR structures were available								
18	MKL/myocardinlike protein 1	HR4547E	2KW9 (NMR)	32	D	10.4	8276	75
19	MKL/myocardinlike protein 1	HR4547E	2KVU (NMR)	33	D	10.4	8276	75
20	Putative peptidoglycan bound protein (LPXTG motif)	LmR64B	2KVZ (NMR)	34	M	5.0	9712	85
21	E3 ubiquitin-protein ligase Praja1	HR4710B	2L0B (NMR)	35	M/D	5.6	10,297	91
22	Transcription factor NF-E2 45 kDa subunit	HR4653B	2KZ5 (NMR)	36	M	10.0	10,623	91
23	YlbL protein	GtR34C	2KL1 (NMR)	37	M	11.0	10,661	94
24	Cell surface protein	MvR254A	2L0D (NMR)	38	Tri	5.9	12,385	114
25	Domain of unknown function	MaR143A	2KZW (NMR)	39	M	6.6	16,312	145
26	N-terminal domain of protein PG_0361 from *P. gingivalis*	PgR37A	2KW7 (NMR)	40	M	12.9	17,485	157
Samples where both crystallographic and NMR structures were available								
27	GTP pyrophosphokinase	CtR148A	2KO1 (NMR)	41	D	8.0	10,042	176
			3IBW	42	T	8.0	10,042	176
28	Lin0431 protein	LkR112	2KPP (NMR)	43	M/Hep	6.3	12,747	114
			3LD7	44	M	6.3	12,747	100

There is a high percentage of crystallographic structures that have residues missing in the crystallographic structure when compared to the total number of residues in the protein sequence. Indeed, it is estimated that conformational flexibility results in unstructured regions of 40 amino acids or more in length in 50% of eukaryotic proteins (Vucetic et al., 2003). Although some efforts were made in construct design to eliminate large disordered N- and C-terminal segments (Xiao et al., 2010), in many cases disordered ends and disordered internal loops are observed in these protein structures. Since dynamics and conformational changes are crucial for the function of many macromolecular complexes and enzymes (Boehr et al., 2006), even low resolution information on these residues is useful.

2.2.2 Crystallization.

Each sample (450 µL at ~5-13 mg/mL concentration) was shipped to the Hauptman-Woodward Medical Research Institute's high-throughput screening laboratory on dry ice, thawed upon arrival (typically within one day of receipt), and set up in 1536 crystallization plates (Luft et al., 2003). Each of the 1536 experiments was imaged immediately after the sample was added, and then in weekly intervals for six weeks (as well as a control imaging before the sample was added to the cocktails). For all of the NESG samples, each image was manually inspected and classified as 'crystal' or 'no crystal'. These classifications and the images were then communicated to NESG scientists for crystallization optimization and structural data collection.

2.2.3 X-ray Crystallographic and Solution NMR Structure Determination.

The crystallographic and NMR structures used in this study were all solved by NESG staff scientists, and have been deposited in the Protein Data Bank (Berman et al., 2000). The crystallographic asymmetric unit is not necessarily the biological oligomer. This oligomer was predicted using a theoretical analysis of binding energy and entropy of dissociation with the Protein Interfaces, Surfaces and Assemblies (PISA) service at the European Bioinformatics Institute (Krissinel and Henrick, 2007).

2.2.4 SAXS data.

SAXS data were collected at beamline 4-2 (Smolsky et al., 2007) of the Stanford Synchrotron Radiation Lightsource (SSRL). The SAXS experiments used samples that were re-frozen after crystallization screening; all of the SAXS samples underwent two freeze/thaw cycles. Typically, a minimum sample volume of 60 μL was used. The sample was diluted with its matching sample buffer to prepare 3 solutions of known concentrations. At the beamline an automated sample loader (manuscript in preparation), compatible with the PCR tubes, was used to collect data on as many as 96 experiments without user intervention, or the need to open the hutch. A wavelength of 1.3 Å was used for eight consecutive two-second exposures collected at each of the 3 sample concentrations. Each sample was oscillated back and forth in a quartz capillary cell during data collection to minimize radiation damage effects. All of these samples had an identical matching buffer. The samples were loaded in 8 well PCR strips such that a buffer blank was recorded followed by three concentrations of each of the two samples and then a final buffer blank with a wash cycle between each. The original concentration was diluted in 2:1, 1:2 and 1:5 ratios of sample and buffer blank. Using the 96 well capacity of the beamline sample loader, a series of 24 proteins was studied in each automated run. Typical time for a single sample concentration series was approximately 15 minutes with the majority of that time spent on liquid handling, *e.g.* sample loading and washing the fluid apparatus between each concentration. The data were processed and azimuthally integrated with SASTool (manuscript in preparation) and then visually examined with Primus (Konarev et al., 2003). All eight exposures were compared for similarity to ensure no radiation damage took place and were averaged using SASTool to increase the signal to noise ratio. The SAXS data for different protein concentrations were assessed with Kratky plots and screened for aggregation using Guinier plots

(Guinier and Foumet, 1955). Guinier regions and Radius of gyration (R_g) estimates were derived by the Guinier approximation $I(q) = I(0) \exp(-q^2 R_g^2/3)$ with $qR_g<1.3$ using the AutoRg function of Primus where $q = 4\pi \sin \theta / \lambda$. The highest quality estimate as determined by AutoRg was used to select which of the three concentrations would go on to further processing. Zero extrapolated curves were not used because the examination of the concentration series showed no evidence of aggregation or repulsion in the higher concentration, stronger signal data. AutoGNOM (Svergun, 1992) was used to compute the pair distribution functions, *P(r)*, for each sample and to determine the maximum particle dimension, D_{max}, and these values were compared with those determined manually by GNOM to ensure consistency. A molecular weight was estimated from the program AutoPOROD of the ATSAS package (Petoukhov et al., 2012). Five *ab initio* shape reconstructions (molecular envelopes) were generated by DAMMIF (Franke and Svergun, 2009) and averaged with DAMAVER (Volkov and Svergun, 2003). The program CRYSOL (Svergun et al., 1995) was used to calculate the scattering intensity from deposited crystallographic and NMR structures and estimate an R_g and fit the data by minimizing the discrepancy, χ, according to:

$$\chi^2(r_0, \delta_\rho) = \frac{1}{N_p}\sum_{i=1}^{N_p}\frac{\left[I_e(q_i) - cI_c(q_i, r_0, \delta_\rho)\right]^2}{\sigma_i^2} \qquad 2.1$$

where I_e is the experimental scattering, I_c is the calculated scattering, and σ is the experimental error as determined by SASTool (Smolsky et al., 2007). Other variables are given elsewhere (Svergun et al., 1995). In our case the experimental errors are underestimated as the detector is treated as an ideal photon counter; χ values here should therefore be regarded as a relative indicator of goodness of fit. Volume fractions for cases of oligomeric mixtures were estimated using the program OLIGOMER

(Konarev et al., 2003). In these cases estimates for χ and the convoluted R_g of the mixture in solution and were taken from OLIGOMER.

2.2.5 Comparison of structural data

R_g and D_{max} were calculated from the crystallographic and NMR structures using the program CRYSOL (Svergun et al., 1995). These values were compared with those derived from the SAXS data with constant subtraction enabled. For visualization purposes envelopes produced by the SAXS data were automatically overlaid with structures derived from X-ray crystallographic and NMR techniques using the program SUPCOMB, and were followed by manual adjustments using the program PyMOL.

2.3 Results.

Table 2.2 summarizes and compares the values calculated from crystallographic and NMR structural information with those measured from the SAXS data. In general, experimentally determined R_g values are consistent with those calculated from the structural information. In most cases the SAXS calculated D_{max} are somewhat larger than those for the crystallographic cases, but smaller than those for the NMR structures. This might be expected due to missing residues in the crystallographic case. On the other hand, these NMR structure ensembles may overestimate the breadth of the true conformational distribution, since the set of 20 conformers deposited in the PDB does not account for the population distribution across the ensemble. The Porod calculated molecular weights are, again for the most part, integer multiples of the measured molecular weight. The SAXS determined oligomer is shown along with the relationship to that seen in the crystallographic structure.

Table 2.2 A Summary of Structural (Crystallography and NMR) and SAXS Results. The sample # refers to the identical number in Table 2.1. The number of unresolved residues in the structure (mainly crystallographic) is listed together with the R_g and D_{max} (in Å) determined from the available structure. The R_g and D_{max} from the SAXS data are shown together with the difference from the available structural information. The molecular weight (in Da) calculated from a Porod analysis is listed along with the ratio of this weight with that derived initially from mass spectrometry in Table 2.1. Finally the SAXS-determined oligomer, (Monomer, Dimer or Tetramer), the relationship to the available structure and the χ of the fit are listed. A special case is described below for samples 16 and 17. Further details are given in the text.

#	Residues Observed	# Res Missing	R_g Structure	D_{max} Structure	R_g SAXS	ΔR_g	D_{max} SAXS	Δ d_{max}	Porod MW	MW Ratio	SAXS Oligomer	Oligomer Assign.	SAXS Fit (χ)
Samples where crystallographic structures were available													
1	74	13	13.7	42.0	14.9	1.2	53.2	11.2	7827	0.8	M		4.2
2	198	12	16.6	67.0	19.8	3.2	67.4	0.4	24,555	1.8	D	sym	2.6
3	436	48	22.4	62.3	23.2	0.8	75.3	13.0	50,064	3.6	T	sym	1.6
4	214	38	23.3	81.2	23.6	0.3	82.7	1.5	37,348	2.6	D/T[a]	PDB	2.6
5	224	34	19.9	57.6	19.8	−0.1	64.2	6.6	28,828	2.0	D	PDB	2.2
6	107	24	19.6	76.3	21.5	1.9	82.0	5.7	11,085	0.8	M		6.1
7	236	12	21.4	64.7	22.2	0.8	76.8	12.1	31,410	2.0	D	PDB	3.8
8	286	12	20.5	63.1	21.1	0.6	71.4	8.3	34,786	2.0	D	PDB	2.0
9	162	9	17.6	54.0	18.7	1.1	65.5	11.5	20,468	1.1	M		3.7
10	165	12	17.5	58.0	18.5	1.0	65.8	7.8	19,069	0.9	M		4.2
11	336	18	26.1	80.8	26.0	−0.1	89.7	8.9	59,937	2.9	D/T[a]	PDB/sym	1.4
12	252	36	21.3	61.5	22.5	1.2	81.9	20.4	37,254	1.2	M		2.9
13	416	30	28.5	95.0	27.6	−0.9	98.5	3.5	40,027	0.8	M		1.4
Samples where multiple constructs and crystallographic structures were available													
14	272	44	20.8	59.6	21.1	0.3	69.2	9.6	30,670	1.9	D	PDB	1.9
15	258	26	21.1	61.8	22.0	0.9	79.7	17.9	32,657	2.0	D	PDB	1.8
16	93	12	18.0	59.5	18.2	0.2	64.7	5.2	15,875	1.3	D2	PDB	1.7
17	93	48	20.4	75.0	20.8	0.4	73.0	−2.0	15,920	1.0	D1	PDB	2.5
Samples where NMR structures were available													
18	75	0	22.5	122.4	16.8	−0.9	58.4	−64.0	6771	0.8	M		4.7
19	75	0	17.7	94.4	16.5	−1.2	58.4	−36.0	6771	0.8	M		1.4
20	85	0	19.0	80.8	18.7	−0.3	68.0	−12.8	9724	1.0	M		1.7
21	91	0	16.4	71.0	15.9	−0.5	59.6	−11.4	7862	0.8	M		1.5
22	91	0	22.3	123.1	19.6	−2.7	68.0	−55.1	10,762	1.0	M		1.6
23	87	7	14.3	55.8	14.5	0.2	49.7	−6.1	8479	0.8	M		1.4
24	114	0	16.5	67.8	19.6	3.1	66.6	−1.2	12,609	1.0	M		5.9
25	145	0	49.0	325.5	26.6	−22.4	94.7	−230.8	15,386	0.9	M		7.4
26	157	0	19.8	67.5	17.5	−2.3	60.6	−6.9	15,238	0.9	M		2.1
Samples where both crystallographic and NMR structures were available													
27[a]	176	0	18.0	66.7	19.1	1.1	68.3	1.6	22,589	2.2	D	PDB	2.5
	158	18	18.1	52.5	19.0	0.9	68.3	15.8				PDB	2.4
28[a]	114	0	18.5	104.4	18.5	0.0	68.2	−36.2	10,721	0.8	M		2.3
	87	13	14.8	44.1	18.4	3.6	68.2	24.1					7.4

The SAXS data are recorded from specimens diluted from the initial sample preparation and the crystallographic structures are necessarily determined under different biochemical conditions. Details for each group, and in particular deviations from the known crystallographic structure, are described below. The observed data and structural fit to the observed data (continuous line) are shown with the structures and *ab initio* envelopes calculated for each group in Figure 2.1 to Figure 2.4 and for those where a mixture was observed, Figure 2.5. In the majority of cases the fit to the experimental data is good.

2.3.1 Crystallographic and SAXS comparison

For the 13 samples in the set of crystallographically determined structures, there was relatively good agreement between the R_g of the model and that calculated from the SAXS data with an average deviation of < 1 Å, Table 2.2. The difference in D_{max} between the crystallographic and SAXS envelope is greater, having no correlation with the percentages of missing residues in the crystallographic structure. This is not surprising given that missing residues may contribute to the D_{max} if missing from the longest axis or may have little to no contribution if predominately missing from a shorter axis.

The observed data and structural fit to the observed data (continuous line) with crystallographic structures and *ab initio* molecular envelopes are shown in Figure 2.1 with the exception of samples 4 and 11 where a mixture of oligomers was noted (see below). Outliers clearly visible by eye occur in samples 1 and 6; however there is a good correlation with all the envelopes and the known structure. The known crystallographic structure is represented in a ribbon form for clarity but in reality occupies more space when side chains are taken into account. For a number of cases the molecular envelope clearly extends beyond the known structure, extending further than can be explained by side chains on the backbone. These instances are consistently located in areas with residues missing from the crystallographic structure, but could also be attributed to slight undetected aggregation artificially enhancing the calculated D_{max}. The highest χ values are observed for samples 6 (χ=6.1), 1 (χ=4.2), and 10 (χ=4.2). The crystallographic structures for these samples are missing 22%, 18% and 7% of the residues respectively which could contribute to this, although better χ values are observed for samples 4 and 12 which are both missing 14% of their residues in the crystallographic structures. In sample 1, a total of 13 residues are unresolved in the crystallographic structure. The molecular envelope reconstruction suggests evidence of these on the left hand side of the envelope. Sample 2 is a dimer in solution and 12 residues are unresolved in the crystallographic structure. When reconstructing the molecular envelope it is possible to add known symmetry information to the reconstruction and averaging (as in the case of an oligomer) but in our case *ab initio* modeling and averaging without symmetry constraints were used. A similar effect is seen for sample 7 where 12 residues were missing from the crystallographic structure. The molecular envelope accounts for missing residues in samples 8, 9 and 12, with 12, 9 and 36 unresolved residues, respectively. In each case the portion of envelope unexplained by the available crystallographic structure is positioned adjacent to the point where residues become unresolved in the structure. In sample 5 there are 34 residues missing but in this case it is not clear where those residues reside. These samples are structurally diverse yet in all the cases, the molecular envelopes show good agreement (at the resolution of the technique) with the known structures.

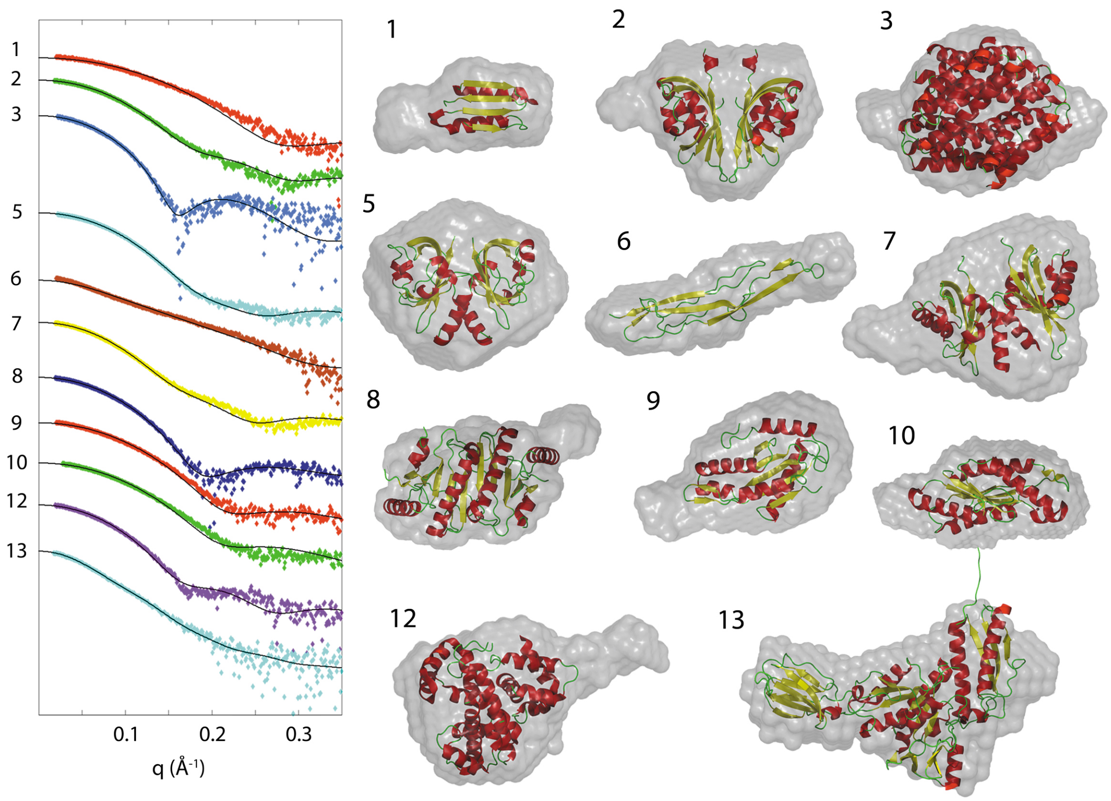


Figure 2.1 The observed SAXS data and structural fit to the observed data (continuous line) for samples with crystallographic structure. The ab initio SAXS derived envelopes overlaid with crystallographic structure are also shown illustrating the agreement between the techniques. The samples are numbered as in Table 2.1 and Table 2.2. Samples 4 and 11 contained a mixture of oligomers and are shown below.

Where the molecular weight calculated from the Porod volume indicated an oligomer, different oligomers were compared with the experimental scattering. In Table 2.2 the oligomer assignment is noted as either “PDB”, where the chosen oligomer is present within the asymmetric structure provided by the PDB, or as “sym” where the oligomer in solution is not present in the PDB, but is chosen based on the crystal symmetry operator. In two cases, the oligomer seen in solution was not the oligomer indicated by the asymmetric unit in the PDB, but was a smaller unit present within it. In case 5, the asymmetric unit contains a trimer, whereas the SAXS data not only favored a dimer, but also was able to clearly distinguish which dimer from the two possibilities. In case 9, the PDB oligomer is a dimer, whereas the SAXS selects the monomer. In both cases, the SAXS-selected oligomer agrees with the oligomer found via gel filtration. For the majority of cases, with the exception of samples 6 and 9, the PISA prediction was in good agreement with the SAXS derived oligomer. In sample 6, PISA predicted an elongated dimer and in sample 9 three dimers in solution were predicted. Neither of these cases were supported by the SAXS data. In certain cases no oligomer provided by the PDB or via symmetry operation was found to be consistent with the SAXS data when comparing the R_g, D_{max}, and overall fit to the curve. In this study, since the atomic structure is already known the results can be analyzed as a mixture of oligomers. Samples 4 and 11 showed clear evidence of oligomer mixtures with sample 4 consisting of 63% dimer and 37% tetramer and sample 11 consisting of 47% dimer and 53% tetramer. These are discussed below.

2.3.2 Sensitivity to different constructs.

Alternate constructs were available for two samples. The first sample, a putative hydrogenase, had crystallographic structures for both constructs where SAXS data was collected, samples 14 and 15 with 316 and 290 residues, respectively. Interestingly, for sample 15, a significantly larger D_{max} (79.7 Å) is found for the construct with fewer residues compared to the D_{max} (69.2 Å) for the construct with more residues. The corresponding crystallographic structure for sample 15 shows a dimer in the PDB where one monomer has two fewer residues in the electron density than the adjacent monomer. In the adjacent monomer, these two residues appear to form a beta strand secondary structural element while in the adjacent monomer, lacking these two residues, this element is not present. This may reflect a higher level of disorder for these and neighboring residues in solution and subsequently for the five additional residues absent from this terminus. The SAXS envelope for this sample fits well to the overall crystallographic structure with the exception of an additional region present on only one side of the dimer. The disordered residues present in one monomer may be occupying this area. However, SAXS is a technique sensitive to aggregation and this extension of the SAXS envelope by an additional 10 Å compared to the similar construct may result from minor levels of aggregation present in solution that has escaped detection via static light scattering and Guinier analysis of multiple concentrations. Without further data it is not possible to distinguish the source of this difference.

The second example of multiple constructs, the protein Alr3790, has a single crystallographic structure (PDB ID 3HIX) for the two constructs, whereas SAXS data for each construct, samples 16 and 17, are clearly different. These constructs had 105 residues (providing the crystallographic structure) and 141 residues respectively (out of 151 in the protein). The 3HIX structural model shows a trimer in the asymmetric unit. This trimer did not fit the SAXS data for either construct. Breaking the trimer into two separate dimers, D1 and D2, showed that each construct forms a structurally distinct dimer in solution. Sample 16 contains 36 fewer residues than sample 17 and these extra residues are located precisely at the D1 dimer interface. A possible explanation for the two solution states is that these extra residues impede D1 dimer formation in sample 17, but not being present in sample 16, allow the formation of dimer D2. The comparison of the calculated scattering for each possible dimer configuration with the experimental SAXS data clearly distinguishes the correct dimer formation for each construct. The envelope for sample 17 appears to underestimate the volume of the entire D1 dimer. Given that analytical gel filtration data and Porod molecular weight indicated a monomer in solution, it is possible that the monomer form may exist for a significant population in solution. A mixture analysis using both monomer and dimer only gave marginal improvements to the fit (χ = 2.1 to 2.5 respectively), and no improvement to the size parameters. If a monomer population is also present at a low concentration it does not appear to greatly affect the SAXS curve. The PISA analysis predicted a stable hexamer consisting of a trimer of dimers for samples 14 and 15. While the hexamer is not shown to exist from the SAXS data, the dimer is present. For samples 16 and 17, the PISA analysis predicts both dimers to be equally stable. The SAXS data for sample 16 shows one dimer, while for sample 17 the SAXS data shows the other. For both examples the observed SAXS data, structural fit to the observed data and the envelopes with overlaid crystallographic structure are shown in Figure 2.2. Again the globular region of these proteins is well represented by the SAXS-derived *ab initio* molecular envelope.

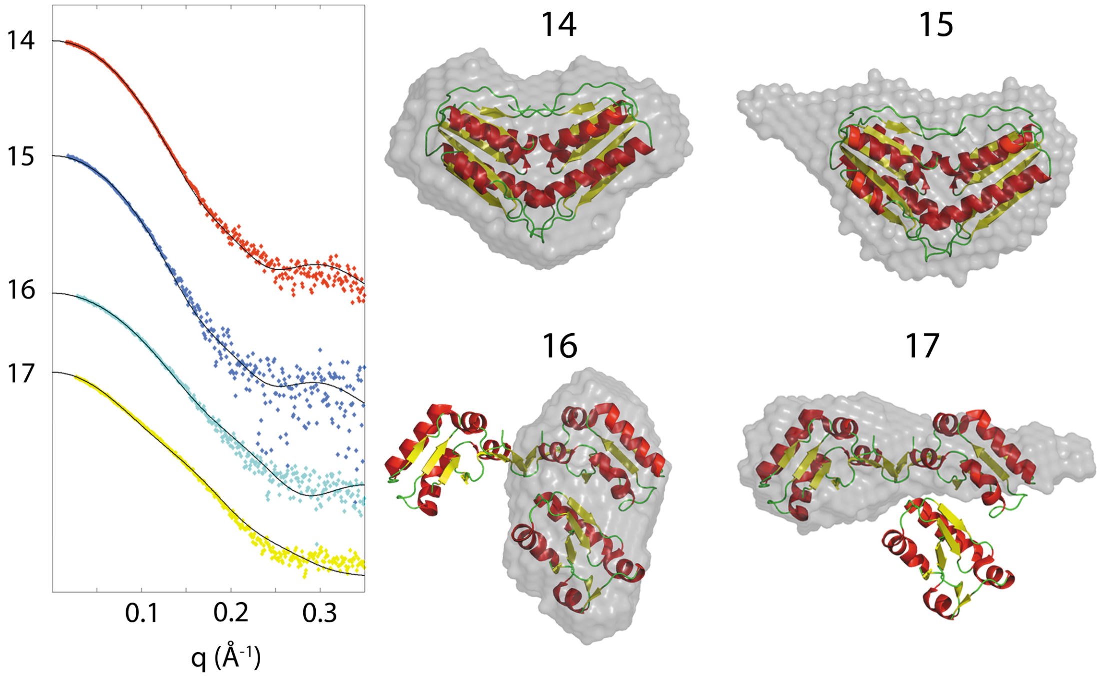


Figure 2.2 The observed SAXS data and structural fit to the observed data (continuous line) for samples with crystallographic structure and multiple constructs. *Ab initio* SAXS-derived envelopes are overlaid with crystallographic structure. Sample 17 contained a mixture shown below. The figures are shown to the same approximate scale as those in Figure 2.1 and the remaining structural representations in Figure 2.3 and Figure 2.4.

2.3.3 NMR and SAXS comparison.

In the NMR case the SAXS D_{max} was consistently smaller than that derived from the NMR structure. In each case the NMR structural data consists of the 20 lowest energy conformers from 100 that were calculated. The D_{max} in the NMR case is calculated from the maximum dimension of the total envelope of all 20 conformers. As such it can lead to an overestimate of D_{max} as it measures the extremes within this set of conformers and does not take into account relative populations or dynamics. Although it is possible to obtain such population distributions from NMR studies, this information is not available from these NMR structural ensembles. For all the samples except sample 25, the calculated and measured R_g's are similar. The R_g is defined as the root mean square distance of the atoms in the molecule from their common center of gravity. As such it is less sensitive to extremes within the population of conformers derived from the NMR data. Of note are samples 18 and 19, where the same SAXS data is compared against two NMR structures. The first, sample 18, was compared to a structure with no residual dipolar coupling information and the second, sample 19, was compared to a structure making use of residual dipolar coupling. Although these two NMR structures are similar, with backbone rmsd between the mean coordinates of each ensemble of 4.6 Å (1.5 Å for the well defined residues, 20-75), the fit of the SAXS data is significantly better to the latter. Figure 2.3 shows the SAXS data, structural fit to the data, and the NMR structures overlaid on each SAXS envelope. For samples 20 through 24, results similar to those observed in comparing SAXS data to the crystallographic structures are seen. The SAXS envelope accurately contains the globular portion of the NMR model; where the SAXS and NMR model diverge is consistent with the expected location of disordered residues. An exception to this appears to occur in the case of sample 21, where the NMR model indicates disordered structure extending away from the top right of the ordered portion but the SAXS envelope indicates that structural envelope is predominately to the right of the ordered portion of the molecule. Samples 22 and 25 both show large structurally disordered regions. SAXS is a technique that is sensitive to the time- and ensemble-averaged volume occupied by a protein, but there will be a case where the amount of time a protein molecule is in a particular position or the percentage of molecules in that position is too small to produce a signal that can be interpreted as the envelope and not noise. Though NMR can be used to characterize distributions of conformations in disordered regions by interpreting the data as arising from ensemble averaging, these methods were not used for these NMR structures and the distributions of conformations in disordered regions cannot be interpreted as representative of the true conformational distributions in solution. In this case limitations of each technique must be realized and a balance needs to be made between the limitations of both techniques.

The SAXS data shows that samples 18, 19 and 24 are in the monomeric state, which is in disagreement with the oligomeric state determined by analytical gel filtration. NMR 1D N_{15} T1/T2 measurements are a higher fidelity technique than gel filtration for oligomer determination. Applying this NMR technique to samples 18, 19 and 24 confirms the monomeric state is in good agreement with the SAXS data.

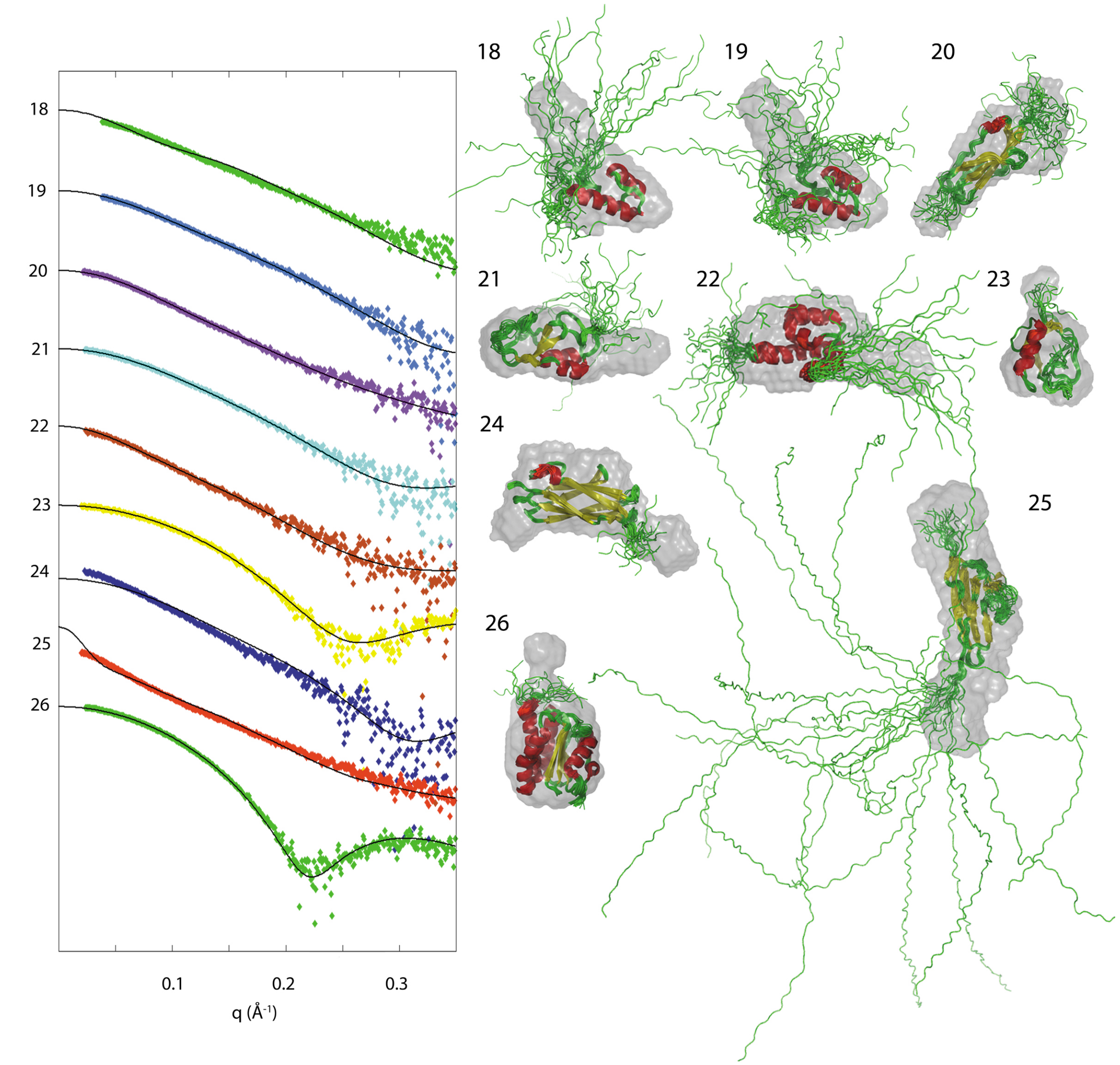


Figure 2.3 The observed SAXS data and structural fit to the observed data (continuous line) for samples with NMR structures. The *ab initio* SAXS-derived envelopes are overlaid with the NMR structures to show the agreement between the data. The figures are shown to approximate scale as in the other structural illustrations and illustrate multiple conformations determined from the NMR data.

2.3.4 The Combination of Crystallography and NMR with SAXS.

For two samples, 27 and 28, both crystallographic and NMR structures were available. In each of these cases, the SAXS envelopes were in good agreement with the crystallographic and NMR structures. For sample 27, the R_g and D_{max} from the SAXS data were each within ~1 Å of the NMR structure. The R_g was within ~1 Å of the crystallographic structure, but the D_{max} measures 15.8 Å greater when SAXS data is compared to the crystallographic structure. This is consistent with the crystallographic structure having 18 missing residues, ~10% of the structure, while NMR accounted for all of the residues. For sample 28, the R_g for the NMR data was in exact agreement with the SAXS data but the D_{max} from the SAXS data was ~36 Å less. The crystallographic structure had 13 missing residues, 13% of the structure, which accounts for a smaller R_g and D_{max} when compared to the SAXS values. The difference in D_{max} from the NMR structure and SAXS data is discussed previously.

The observed SAXS data and structural fit to the observed data, along with *ab initio* envelopes with structures overlaid are shown in Figure 2.4. The NMR and crystallographic structures are similar and fit well into the SAXS-derived molecular envelopes. In the case of sample 27, no residues are missing from the NMR structure (a) and 18 are missing from the crystallographic structure (b). For the NMR structure, regions of structural disorder are consistent with envelope regions otherwise not explained by the NMR structure. Similarly, in the X-ray structural case, missing residues are represented by the molecular envelope density consistent with the position and number of those residues. Sample 28 is missing 13 residues in the crystallographic structure (a), which are positioned by making use of the SAXS envelope. The NMR data (b), while fitting the experimental SAXS data better than the crystallographic (accounting for these missing residues), appears to place the bulk of the disordered region in a different location than the SAXS envelope suggests. When comparing these two examples, 18 missing residues can make little difference to the overall calculated curve in cases such as sample 27, while a similar number of missing residues can have a great impact on the calculated curve, as seen in sample 28. This may be due to location of the missing residues as well as their size compared to the size of the particle as a whole. The PISA analysis predicted the same dimer organization for sample 27 as determined by the SAXS data. However, a dimer predicted by PISA for sample 28 was not seen in the SAXS data.

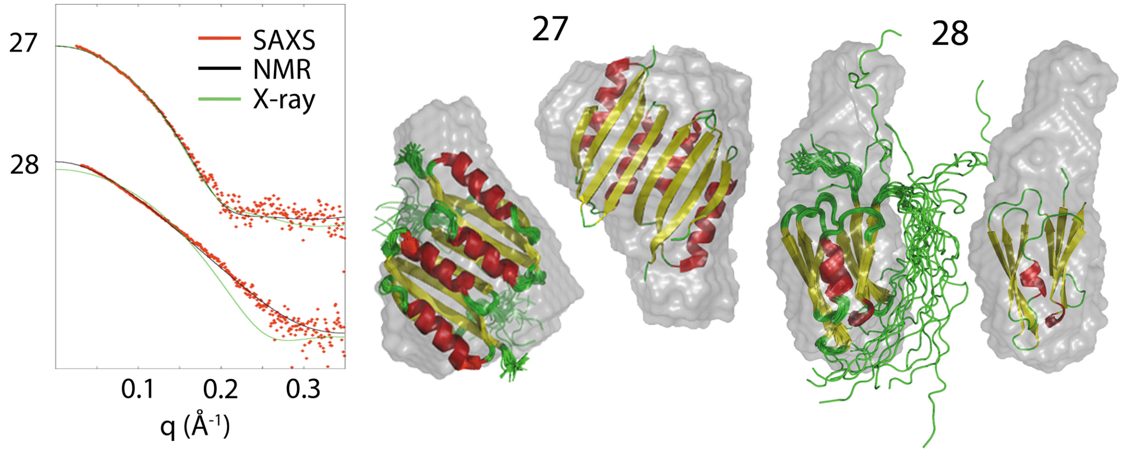


Figure 2.4 *Ab initio* SAXS-derived envelopes overlaid with NMR and crystallographic structure to show the agreement between the different structural methods. The figures are shown to approximate scale and illustrate multiple conformations determined from the NMR data.

2.3.5 Mixtures.

This study has been used to determine how well SAXS *ab initio* molecular envelope reconstructions represent known structures and from this gain an idea of the accuracy of cases where no structural information is present. However, having this structural information also allows us to analyze samples as mixtures. Samples 4 and 11 were determined to be mixtures of oligomers from the SAXS data, Figure 2.5.

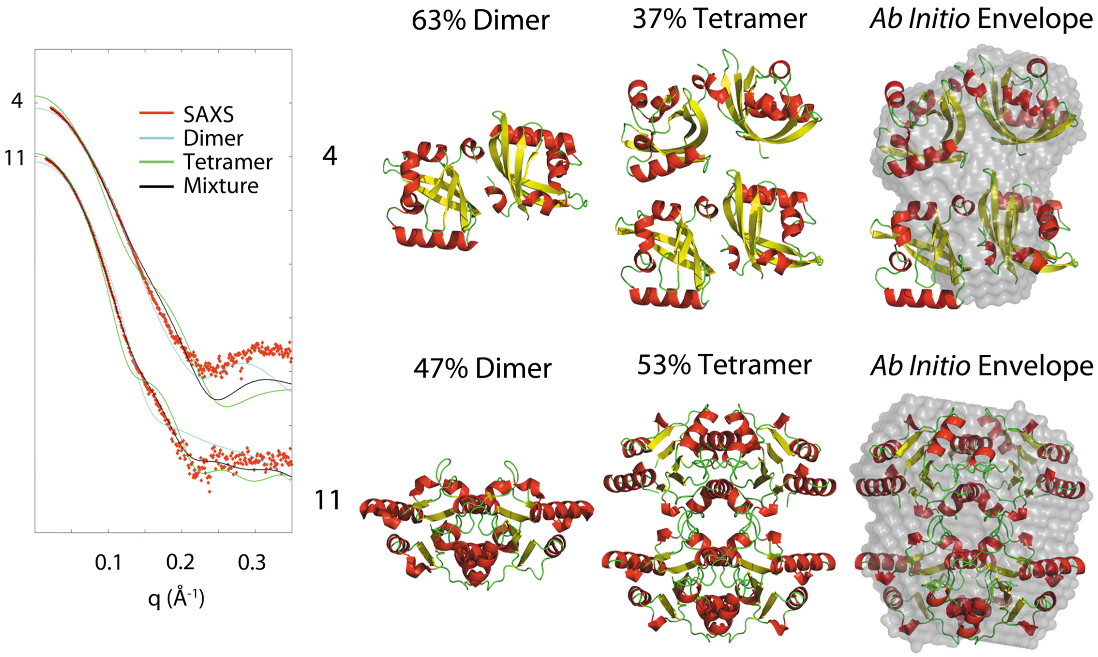


Figure 2.5 Structures of oligomers based on analysis of the SAXS data and known monomer structure. The *ab initio* SAXS-derived envelopes are shown assuming a monodisperse solution.

In these examples, scattering from each oligomer was calculated and estimates of volume fractions present in solution were carried out. For sample 4, the fit to the curve improved from χ = 7.8 for the dimer (not shown) to χ = 2.6 for the dimer-tetramer mixture. This is seen primarily in the improvement of the low q-region of the curve, corresponding to the overall size of particles in solution. The D_{max} values reflect this as well, 58.7 Å for the dimer, 81.2 Å for the tetramer, in good agreement with the SAXS estimated value of 82.7 Å. The 28 residues of the dimer, and 56 residues of the tetramer that were missing in the crystallographic structure, may explain the poor fit beyond about 0.13 $Å^{-1}$ for the mixture. For sample 11, the fit improved dramatically from χ = 13.5 for the dimer and χ = 10.4 for the tetramer to χ = 1.4 for the mixture. Similarly, the D_{max} for the dimer is only 71.0 Å, but for the tetramer it increases to 80.8 Å, closer to the SAXS derived D_{max} of 89.7 Å. The PISA analysis indicated that both dimers were in stable oligomeric states but did not identify either tetramer.

Without knowledge of the structure, one is unable to determine volume fractions of oligomers in solution. *Ab initio* reconstructions for mixtures should not be carried out because most algorithms, including that used in the DAMMIF reconstructions in this study, assume a monodisperse solution and are not suited for polydisperse mixtures (Franke and Svergun, 2009). Attempts at reconstructing *ab initio* envelopes when samples are known to be polydisperse are shown in Figure 2.5. It is readily seen that in some cases, *i.e.* sample 4, the envelope is a poor representation of either oligomer in solution, whereas in sample 11, the envelope appears to be able to accommodate most of the tetramer. This illustrates that while an *ab initio* model may be constructed, it is not *reliable* for either oligomer if the solution is polydisperse. In a polydisperse solution containing multiple oligomers of the same basic quaternary unit the intramolecular distances within the basic quaternary unit will be similar for each oligomer and thus contribute similarly to the intensity profile as a monodisperse solution. Only the additional intramolecular distances present in the larger oligomer that are not present in the basic quaternary unit will contribute to the scattering differently than the monodisperse solution. This highlights the importance of oligomer screening prior to SAXS data collection, or the use of biophysical or biochemical separation techniques to ensure a single oligomer population exists within the sample, especially when no other structural information is available. If prior structural information exists, this illustrates the strength of the application of SAXS for mixed-oligomer analysis to characterize solutions containing mixtures of quaternary structures.

2.4 Discussion

SAXS is not a new biophysical technique but it has only recently been applied to high-throughput structural biology (Hura et al., 2009). In this paper, SAXS has been approached from a different perspective, that of a high-throughput crystallization screening laboratory. SAXS has been used to characterize remnants of samples that remained after crystallization screening. Over five hundred different proteins from this group have been characterized to date. From these samples we have presented a subset of cases where crystallographic and/or NMR structural information was available. In some cases this was known prior to SAXS, in other cases it became known subsequently. In all cases, SAXS studies using minimal amounts of sample at multiple concentrations but a single buffer condition, produced molecular envelopes that were consistent with crystallographic and NMR based structural knowledge. We acknowledge the limitations of SAXS; for example, disordered regions may be averaged to a single area that is not representative of the actual molecular structure. Similarly, the SAXS envelope may not be completely sensitive to highly dynamic regions of a structure and in extreme cases could insufficiently sample and subsequently incorrectly represent the volume occupied by the flexible portion of the molecule. SAXS experiments can be performed on all of the greater than 30,000 soluble, purified samples produced by the US PSI. SAXS could be used to structurally characterize the majority of these samples. We have demonstrated that these envelopes appear to be highly consistent with known structural information. If these samples could be characterized structurally, albeit at low resolution, they would significantly increase the amount of structural knowledge that is currently available.

The fact that our envelopes are in good agreement with known structures does not imply that envelopes for samples recalcitrant to crystallization will necessarily be representative of the structures of these samples. There could be significant biochemical, biophysical, or structural reasons for failure to crystallize. However, SAXS is a powerful technique for characterizing samples in solution. It can distinguish between natively unfolded samples, those with flexible disordered regions and those that may have multiple globular regions with flexible linkers. We can identify these problem cases and limit our analysis to those samples that are well behaved. In doing so we can have reasonable confidence that the molecular envelope produced from SAXS data reflects the molecular structure. However, reasonably confident is not completely confident. Without complementary structural, or biochemical knowledge we can never be 100% certain of the accuracy of the envelope. We have to remain wary and have to settle for the fact that most of what we see from envelope reconstructions is correct, but this is not always going to be the case.

An important note in this study is the observation of two cases of mixtures. We have a limited sample set that has been well characterized on preparation but then cycled through freeze thaw cycles both before and after crystallization trials prior to SAXS analysis. Samples should be as fresh as possible and homogeneous. One approach that is clearly recommended is the use of size-exclusion chromatography and light scattering techniques immediately before SAXS data collection to monitor monodispersity (Rambo and Tainer, 2010).

SAXS is clearly complementary to high-resolution structural techniques such as crystallography and NMR spectroscopy. We have demonstrated that it provides unique quaternary structural information from the solution state that can be leveraged into biological knowledge that is not determined using independent methodologies. This is exemplified by the identification of oligomer organizations for samples 2, 3, 4, 5, 6, 9, 11, 16 and 17 that are alternatives to those seen in the crystallographic structures. Binding energy and entropy of dissociation can be calculated where structural information is present enabling prediction of the biological oligomer with services such as PISA (Krissinel and Henrick, 2007). This has been shown to be successful in 80-90% of cases, a similar success rate seen with our data. However, SAXS can directly identify these oligomers removing any uncertainty. In the case of NMR, special data collection and analysis methods are required to determine the correct representation of highly disordered regions. In such disordered regions, SAXS data indicate a more compact structure than that indicated by the reported conformational ensemble. To some extent, this is an issue with the calculation of D_{max} from an ensemble of conformers, but there are clear cases where this alone does not fully explain difference in D_{max} values. Specific modeling of SAXS sensitivity is needed to resolve this case. There are methods to treat molecules or parts of molecules as ensembles of conformers within the SAXS analysis. The Ensemble Optimization Method (EOM) (Bernado et al., 2007) randomly generates conformers, bins them to create ensembles and using a genetic algorithm, optimizes the ensembles by comparing the average scattering profile of their conformers to the experimental data. Using an increasing number of conformers per ensemble, and an analysis of the deviation of experimental data from predicted data, SAXS analysis can be used to study dynamic structural regions. In this study we have not made use of these methods due to the number of samples examined and the computational resources required for each case. On the other hand, the NMR methods used by the NESG consortium are not aimed at accurate representation of conformational distributions in disordered regions, which requires special methods and considerations.

The structural and biochemical data used in this study are publically available. We are happy to provide the SAXS data associated with this study to groups that may use it for further development. We have used SAXS to complement high-throughput crystallization screening and are uniquely positioned with the availability of a large number of well-behaved and well-characterized samples courtesy of the NESG efforts. We have presented a top-level overview of our initial results on a subset of samples where structural information was already available. SAXS data have been useful and provided additional information in these cases.

The strength of SAXS shown by our results causes us to echo the conclusions of Hura *et al.* (Hura et al., 2009) in adopting the method for high-throughput structural genomic studies and to go one step further in suggesting that it is in fact essential. While X-ray crystallography and NMR are clearly powerful structural techniques, when SAXS analysis is added, the synergistic relationship between the techniques provides a far greater understanding of the biological system as a whole.

2.5 References

Abrahams, J.P., Leslie, A.G., Lutter, R., and Walker, J.E. (1994). Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Nature *370*, 621-628.

Acton, T.B., Gunsalus, K.C., Xiao, R., Ma, L.C., Aramini, J., Baran, M.C., Chiang, Y.W., Climent, T., Cooper, B., Denissova, N.G., *et al.* (2005). Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. Methods in enzymology *394*, 210-243.

Acton, T.B., Xiao, R., Anderson, S., Aramini, J., Buchwald, W.A., Ciccosanti, C., Conover, K., Everett, J., Hamilton, K., Huang, Y.J., *et al.* (2011). Preparation of protein samples for NMR structure, function, and small-molecule screening studies. Methods in enzymology *493*, 21-60.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc *129*, 5656-5664.

Boehr, D.D., Dyson, H.J., and Wright, P.E. (2006). An NMR Perspective on Enzyme Dynamics. Chemical Reviews *106*, 3055-3079.

Chayen, N.E. (1996). A novel technique for containerless protein crystallization. Protein Engineering *9*, 927-929.

Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J. (2004). TargetDB: a target registration database for structural genomics projects. Bioinformatics *20*, 2860-2862.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Deisenhofer, J., Epp, O., Miki, K., Huber, R., and Michel, H. (1984). X-ray structure analysis of a membrane protein complex. Electron density map at 3 A resolution and a model of the chromophores of the photosynthetic reaction center from Rhodopseudomonas viridis. J Mol Biol *180*, 385-398.

Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T., and MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science *280*, 69-77.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr *42*, 342-346.

Grant, T.D., Luft, J.R., Wolfley, J.R., Tsuruta, H., Martel, A., Montelione, G.T., and Snell, E.H. (2011). Small angle X-ray scattering as a complementary tool for high-throughput structural studies. Biopolymers *95*, 517-530.

Guinier, A., and Foumet, F. (1955). Small Angle Scattering of X-rays (New York, Wiley Interscience).

Hodgkin, D.C., Kamper, J., Lindsey, J., MacKay, M., Pickworth, J., Robertson, J.H., Shoemaker, C.B., White, J.G., Prosen, R.J., and Trueblood, K.N. (1957). The Structure of Vitamin B12 I. An Outline of the Crystallographic Investigation of Vitamin B12.

Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences *242*, 228-263.

Hura, G.L., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., 2nd, Tsutakawa, S.E., Jenney, F.E., Jr., Classen, S., Frankel, K.A., Hopkins, R.C.*, et al.* (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods *6*, 606-612.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr *36*, 1277-1282.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*, 774-797.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol *142*, 170-179.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr *45*, 342-350.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys *40*, 191-285.

Rambo, R.P., and Tainer, J.A. (2010). Improving small-angle X-ray scattering data for structural analyses of the RNA world. Rna *16*, 638-646.

Rasmussen, S.G., Choi, H.J., Rosenbaum, D.M., Kobilka, T.S., Thian, F.S., Edwards, P.C., Burghammer, M., Ratnala, V.R., Sanishvili, R., Fischetti, R.F.*, et al.* (2007). Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature *450*, 383-387.

Schlunzen, F., Hansen, H.A., Thygesen, J., Bennett, W.S., Volkmann, N., Levin, I., Harms, J., Bartels, H., Zaytzev-Bashan, A., Berkovitch-Yellin, Z.*, et al.* (1995). A milestone in ribosomal crystallography: the construction of preliminary electron density maps at intermediate resolution. Biochem Cell Biol *73*, 739-749.

Smolsky, I.L., Liu, P., Niebuhr, M., Ito, K., Weiss, T.M., and Tsuruta, H. (2007). Biological small-angle x-ray scattering facility at the Stanford synchrotron radiation laboratory. J Appl Crystallogr *40*, S453-S458.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr *28*, 768-773.

Svergun, D.I. (1992). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. J Appl Crystallogr *25*, 495-503.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.

Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. (2003). Flavors of protein disorder. Proteins *52*, 573-584.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Xiao, R., Anderson, S., Aramini, J., Belote, R., Buchwald, W.A., Ciccosanti, C., Conover, K., Everett, J.K., Hamilton, K., Huang, Y.J.*, et al.* (2010). The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. Journal of structural biology *172*, 21-33.

3 SAXStats Automated SAXS Analysis Software Package

3.1 Introduction

The use of SAXS in structural biology studies is growing rapidly. Third generation synchrotron sources coupled with detectors that have low noise and high dynamic range and the development of computational power to exploit new algorithms have led to a resurgence of the technique (Grossmann, 2007). Unfortunately, the analysis of SAXS data is deceptively simple compared to crystallography and subtle effects may belie more serious problems that invalidate the analysis (Jacques and Trewhella, 2010). As the technique grows, it takes time to build the expertise necessary to spot these problems and avoid misinterpretation of the data. An objective system of data analysis yielding quantitative evaluation of data quality would greatly reduce the likelihood of problematic SAXS data being used to make erroneous conclusions.

As described in section 1.2.2, the intensity data collected by small angle scattering experiments is given by:

$$I(q) = F(q) \times S(q) \qquad (3.1.)$$

where *I* is intensity, *F* is the form factor of the particle in solution, *S* is the structure factor and *q* is the momentum transfer. Ideally, particles in dilute solution conditions act independently of one another, exhibiting no interparticle effects, resulting in a structure factor of one. However, oftentimes particles in SAXS experiments will interact in solution. Interparticle interactions will cause the structure factor to deviate from unity, causing the intensity profile to inaccurately reflect the form factor of the particle, which includes the desired information about size and shape. Most modeling software presumes monodispersity and such interparticle interactions may result in incorrect modeling and data interpretation.

To aid the user in objective, statistically significant measurements of SAXS parameters, we have developed a process implemented through a series of computer scripts to perform statistical analyses on SAXS data and identify data that should be treated with caution. This series of scripts, called SAXStats, tests for two major occurrences in SAXS data that can negatively distort SAXS profiles, namely radiation damage and interparticle interactions. Unwanted trends in SAXS data resulting from radiation or interparticle interactions are measured and the significance of these trends is examined using the linear regression t-test. Since the detection of radiation damage and interparticle interactions relies on statistical significance, the identification of problematic data is objective and therefore does not require expert analysis.

3.2 Radiation Damage

SAXS experimental data acquisition often utilizes high-flux synchrotron X-ray sources to irradiate the sample solution to obtain the signal-to-noise ratio required for accurate form factor evaluation. Occasionally proteins can form high molecular weight oligomers due to the generation of inter-protein crosslinking reactions, formation of disulfide bonds, hydrophobic interactions, or electrostatic interactions that can result from ionizing radiation (Davies and Delsignore, 1987; Le Maire et al., 1990). These radiation damage effects manifest as aggregation in the Guinier plot and increases in radius of gyration (R_g), maximum particle dimension (D_{max}), and forward scattering intensity ($I(0)$), as well as the destruction of bonds resulting in protein denaturation, detected using the Kratky plot. By monitoring these parameters as a function of exposure time, radiation damage resulting in the formation of aggregates or protein unfolding can be evaluated.

3.2.1 Using Linear Regression T-statistic to Evaluate Radiation Damage

To quantitatively evaluate the likelihood of radiation damage, we have applied statistical methods to assess various aspects of the integrated intensity profiles, including R_g, *I(0)*, D_{max}, and the similarity of each exposure to the first exposure, χ^2. By plotting each of these parameters as a function of exposure, we calculated a linear regression from which we obtained a t-statistic (Kenney, 1962). In linear regression analysis, the t-statistic is a value that describes the likelihood that a slope is significant. By calculating the t-statistic for these plots, we can determine whether or not the trends in SAXS parameters as a function of radiation are significant, and therefore whether or not radiation damage is present.

To determine the t-statistic, the parameter of interest determined for each exposure is plotted against the exposure number. The calculation of the t-statistic requires the determination of the slope of the regression and the associated standard errors. To best fit the data points, the method of ordinary least squares is used to minimize the sum of the square residuals of the linear regression model. The equation of the linear regression model is:

$$y = ax + b \qquad (3.2.)$$

where *y* is the dependent variable, *i.e.* the SAXS parameter of interest, *x* is the independent variable, *i.e.* the exposure number, *a* is the slope and *b* is the *y*-intercept of the regression. First, the following summations are calculated for variables *y* and *x*:

$$S_x = \sum x_i \qquad (3.3.)$$

$$S_y = \sum y_i \tag{3.4.}$$

$$S_{xx} = \sum x_i^2 \tag{3.5.}$$

$$S_{xy} = \sum x_i y_i \tag{3.6.}$$

$$S_{yy} = \sum y_i^2 \tag{3.7.}$$

Second, the slope is calculated as

$$a = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2} \tag{3.8.}$$

where *n* is the number of exposures and the *y*-intercept is calculated as

$$b = \frac{1}{n} S_y - a \frac{1}{n} S_x \tag{3.9.}$$

which, inserting equation 3.8, simplifies to

$$b = \frac{S_{xx} S_y - S_x S_{xy}}{nS_{xx} - S_x^2} \tag{3.10.}$$

Now that we have a simple linear regression model for the SAXS parameters as a function of radiation, we can determine whether or not the slope of this line differs significantly from zero, indicating that there is a linear dependence and that radiation damage is present. For simple linear regression, the t-statistic is equal to

$$t = \frac{a}{s_a} \tag{3.11.}$$

where s_a is the standard error in the estimate of the slope and t has n – 2 degrees of freedom. The standard error of the slope coefficient takes the form

$$s_a^2 = \frac{ns_e^2}{nS_{xx} - S_x^2} \tag{3.12.}$$

where s_e is the standard error of the residuals

$$s_e^2 = \frac{1}{n(n-2)}\left(nS_{yy} - S_y^2 - a^2(nS_{xx} - S_x^2)\right) \tag{3.13.}$$

The t-statistic can be converted to a p-value to determine the statistical significance of radiation damage independent of the number of degrees of freedom, *i.e.* the number of exposures. To convert the t-statistic to a p-value we use a two-tailed t-table organized by the number of degrees of freedom in Appendix A (Goulden, 1956). Radiation damage is assumed to be present if the p-value is less than 0.05, a threshold commonly chosen to indicate statistical significance.

3.2.2 Detecting Changes in Scattering Profile

Typical scattering profiles cover structural information ranging from a resolution of hundreds of angstroms to as high as ten angstroms. Two scattering profiles can be directly compared for overall similarity through the use of the reduced χ^2 statistic employed in the software program DATCMP (Petoukhov et al., 2012). χ^2 is defined as

$$\chi^2 = \frac{1}{\nu}\sum_{i=1}^{n} \frac{\left[I_2(q_i) - I_1(q_i)\right]^2}{\sigma_{2i}^2} \tag{3.14.}$$

where *n* is the number of data points *i*, and $I_1(q_i)$ and $I_2(q_i)$ are the intensities of the scattering profiles of interest at q_i with error σ_i and *v* is the number of degrees of freedom. For two identical scattering profiles, the χ^2 will equal zero, while two similar profiles will approximately equal one, and two dissimilar profiles will be much greater than one. While in general the comparison of two scattering profiles is a non-trivial task, this simple discrepancy criterion can be used since we are comparing two profiles on the same scale. Since error bars are included in the determination of χ^2 it is important to compare only two scattering profiles with similar degrees of noise, as is the case for multiple identical exposures of the same sample.

In order to determine whether or not radiation damage results in changes in the overall scattering profile, each scattering profile for subsequent exposures was compared against the first exposure and the χ^2 is calculated. The first exposure is used for comparison since it will have experienced the least amount of ionizing radiation. To determine if radiation damage is present that results in changes in the overall scattering profile, the χ^2 is plotted as a function of exposure number and the data is best fit according to the method described above. The t-statistic is then calculated and then converted to a p-value based on the number of exposures. If the p-value is less than 0.05, exposures that are more than two standard deviations from the y-intercept of the linear regression equation are rejected and the remaining exposures are averaged together using the software program DATAVER (Petoukhov et al., 2012), which averages the intensities of each data point from all accepted exposures.

3.2.3 Detecting Changes in Maximum Particle Dimension

Oftentimes radiation damage can result in the formation of aggregates in solution, which may alter the maximum particle dimension (D_{max}) if the growth in size is along the largest axis of the particle. SAXS data can be used to estimate D_{max} from the pair distribution function (*P(r)*) using the software program GNOM (Semenyuk and Svergun, 1991). GNOM uses an indirect Fourier transform to evaluate *P(r)* from *I(q)* according to the following equation

$$P(r) = \frac{r}{2\pi^2}\int_0^\infty I(q)\, q\ \sin(qr)\, dq \qquad (3.15.)$$

In practice the user is required to select D_{max} such that the resulting *P(r)* decays smoothly to zero without significant oscillations or systematic deviations in the curve. Typically the user will begin with a predicted D_{max} of around 3.5*R_g and increase or decrease D_{max} until a suitable value is found. In addition to calculating *P(r)*, GNOM also calculates the inverse Fourier transform to examine how well the resulting *I(q)* fits the scattering profile. This is an important step to ensuring the most accurate D_{max} and *P(r)* are determined. In the program DATGNOM (Petoukhov et al., 2012) this process has been automated, and a series of perceptual criteria such as oscillation, stability, and deviation of the fitted versus the experimental *I(q)* are used to select the best D_{max}.

To evaluate the likelihood of radiation damage resulting in increases in D_{max}, the linear regression t-statistic described in section 3.2.1 is used. While errors in estimating D_{max} can be significant, approaching five to ten per cent of D_{max}, these errors are likely to be distributed randomly among the series of sequential exposures, resulting in a t-statistic that will only reflect changes due to radiation damage. Once the t-statistic is determined it is converted to a p-value. If radiation damage is present ($p < 0.05$) then any exposures that are more than two standard deviations from the y-intercept of the linear regression are rejected from averaging.

3.2.4 Detecting Changes in R_g and *I(0)*

Increases in the average size of particles in solution can also manifest as an increase in the measured R_g and *I(0)*. To detect whether or not radiation damage results in changes in R_g and *I(0)* we plot the R_g and *I(0)* as a function of exposure number and proceed to determine the linear regression. The t-statistic is calculated from this regression and converted to a p-value. If radiation damage is present ($p < 0.05$) then any exposures that are more than two standard deviations from the y-intercept of the linear regression are rejected from averaging.

R_g can be calculated from scattering data using two independent methods. The first method is the most commonly recognized one, and calculates R_g from the Guinier plot. To determine R_g using this method the log of the intensity is plotted as a function of q^2. For a monodisperse solution the Guinier plot is linear in the low resolution regime where $q < 1.3 / R_g$ (Guinier and Foumet, 1955). In this region the slope of the line passing through the data is related to R_g according to the following equation:

$$slope = -\frac{R_g^2}{3} \qquad (3.16.)$$

It is important to note that while the R_g is calculated from the slope of the line through the data points in the Guinier region, the Guinier region is dependent upon the R_g. Typically the R_g and Guinier region are determined through an iterative cycle of calculating R_g and adjusting the Guinier region accordingly followed by recalculating the R_g and so on. The final determination of the accepted Guinier region and calculated R_g is then arrived at through a somewhat subjective interpretation of what the user finds to be an acceptable linear region. This becomes particularly difficult when particles are large, resulting in very few data points and a heightened sensitivity to the estimation of R_g when varying the Guinier region by as little as one data point. This procedure is automated in the software program AutoRg (Petoukhov et al., 2012) which attempts to determine the Guinier region by fitting several slightly different regions, calculating the R_g for each region, evaluating how the R_g changes as a function of additional data points, and accepting the region that minimizes the variance in R_g. While AutoRg works well for many samples, cases that contain few points in the Guinier region as a result of a large R_g, or data that is particularly noisy, may cause the software to incorrectly estimate R_g or fail to find one entirely. In a high-throughput setting, where the focus is to characterize samples rapidly rather than collect optimal data from each, many samples may suffer from low signal-to-noise, or from having too few data points in the Guinier region. Therefore it is important to have as high of a success rate as possible in accurately estimating the Guinier region and R_g.

To remove the inherent subjective nature of estimating the Guinier region, and to ensure the highest possible rates of successfully evaluating the Guinier region and R_g, we have employed an independent method of determining the R_g and subsequently the Guinier region. In the previous section we described the determination of the maximum particle dimension and the pair distribution function calculated using the software program DATGNOM. Using *P(r)* we can calculate the R_g of the particle using an independent formalism from that of the Guinier estimation, shown by the following equation:

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 \, P(r)dr}{2\int_0^{D_{max}} P(r)dr} \tag{3.17.}$$

where D_{max} is the maximum particle dimension, *r* is the interatomic distance, and *P(r)* is the pair distribution function (Putnam et al., 2007). While the determination of *P(r)* still requires an estimation of D_{max}, there is an advantage to calculating R_g using this method.

In the Guinier approximation, even slight modifications to the Guinier region by as little as a few data points can result in a significantly different R_g. Using equation 3.17, however, the R_g is estimated from the pair distribution function, which in turn has been calculated using all available data points in *I(q)*, greatly exceeding the number of data points in a Guinier plot and incorporating information from all regions of reciprocal space. While small errors in the estimation of D_{max} may alter *P(r)* slightly, they have little effect on integration over all *r*, thus providing us with a robust calculation of R_g (Jacques and Trewhella, 2010). In addition to estimating D_{max}, DATGNOM reports the R_g calculated from the pair distribution function. We can now use this new method of calculating R_g to determine the limit of the Guinier region.

Until now we have considered only the upper limit of the Guinier region to be $1.3/R_g$. Occasionally, data at very low resolution, close to the beam stop, can be influenced by external factors such as parasitic scatter and divergence in the beam (Li et al., 2012; Wignall et al., 1990). To alleviate the adverse affects of such factors, it is advantageous to select a minimum cutoff for *q* such that the desired information about shape and size is not lost or distorted. The minimum *q* value required to accurately restore the size and shape information present in the form factor is given by the Shannon sampling theorem which states that the information content in the continuous function *I(q)* can be represented by its values on a discrete set of points, termed Shannon channels (Svergun and Koch, 2003). A measure of the information content is given by Shannon's sampling theorem, such that

$$q\,I(q) = \sum_{i=1}^{\infty} q_k\,I_k(q_k)\left[\frac{\sin D_{max}(q-q_k)}{D_{max}(q-q_k)} - \frac{\sin D_{max}(q+q_k)}{D_{max}(q+q_k)}\right] \qquad (3.18.)$$

where $q_k = k\pi/D_{max}$. The number of parameters required to represent *I(q)* on an interval $[q_{min}, q_{max}]$ is given by the number of Shannon channels

$$N_S = \frac{D_{max}\,(q_{max} - q_{min})}{\pi} \tag{3.19.}$$

where *q_{max}* here refers to the highest resolution collected in the experiment. This provides a lower bound on q_{min} such that its value does not exceed the first Shannon channel, *i.e.* that

$$q_{min} < \frac{\pi}{D_{max}} \tag{3.20.}$$

By utilizing this boundary on q_{min}, and our previously described boundary of q_{max} for the Guinier region, we can limit the Guinier region to the interval

$$[q_{min}, q_{max}] = \left[\frac{\pi}{D_{max}}, \frac{1.3}{R_g}\right] \tag{3.21.}$$

After determining the new Guinier region, we proceed to calculate the R_g using the Guinier method, which provides us with a second independent measure of R_g that can be compared against that estimated from *P(r)* for consistency (Putnam et al., 2007). If the R_g from the Guinier approximation differs significantly from the R_g calculated using the pair distribution function, then this is an indication that there may be additional interparticle interactions affecting only the data at low resolution that would be most immediately influenced by these interactions. Additionally, while radiation damage may not result in changes in the R_g estimated using the pair distribution function, it may affect the R_g estimated from the Guinier approximation that in turn may influence the overall intensity profile.

To calculate the R_g using the Guinier approximation, we have implemented a similar approach to fit the data points using the least squares minimization method as that described previously using equations 3.2 through 3.10. The slope calculated in equation 3.8 can then be used to calculate the R_g according to equation 3.16. To test if radiation damage is present, we again use the linear regression t-statistic.

Additionally, from equation 3.10 we can obtain the intensity extrapolated to q = 0. This extrapolated intensity, *I(0)*, is a very useful quantity as it is directly proportional to the square of the number of electrons in the particle, *i.e.* the molecular weight (Putnam et al., 2007). If radiation damage results in interparticle interactions, *I(0)* will be affected. The linear regression t-statistic is used to reject any exposures that suffer from radiation damage.

3.2.5 Detecting Changes in "Foldedness" of Particle

Not only can radiation damage result in the formation of aggregates, it can also result in protein unfolding (Garrison, 1987). SAXS data can qualitatively assess the degree of foldedness in protein structure using Kratky plots (Glatter and Kratky, 1982). Typical Kratky plots for well-folded or unfolded proteins are shown in Figure 3.1 (adapted from (Putnam et al., 2007)). Globular proteins with a well-defined surface will follow Porod's law in the high-q region as *I(q)* decays proportional to q^{-4} (Glatter and Kratky, 1982; Porod, 1951). In a Kratky plot ($I(q)*q^2$ vs. q) these particles exhibit a characteristic bell curve with a maximum whose position is roughly related to R_g (Receveur-Brechot and Durand, 2012). This maximum is sometimes followed by a minimum in the higher resolution region due to the breakdown of the Porod law resulting from the influence of shape and internal structure on the scattering curve (Rambo and Tainer, 2011). Unfolded proteins, however, do not decay as q^{-4} and instead more closely resemble worm-like chains that decay as q^{-2} (Putnam et al., 2007). Unfolded proteins therefore yield a Kratky plot lacking the bell curve and instead show a continuous increase in $I*q^2$. To obtain a quantity describing foldedness, we derive a new term, called the "Kratky ratio", calculated as the ratio of the height of the minimum of the Kratky plot to the height of the maximum. For a well-folded protein the Kratky ratio will be close to zero while partially folded proteins will have values between zero and one, whereas for an unfolded protein there will be no minimum or maximum. By assessing changes in the Kratky ratio as a function of exposure we can monitor radiation damage effects resulting in protein unfolding.

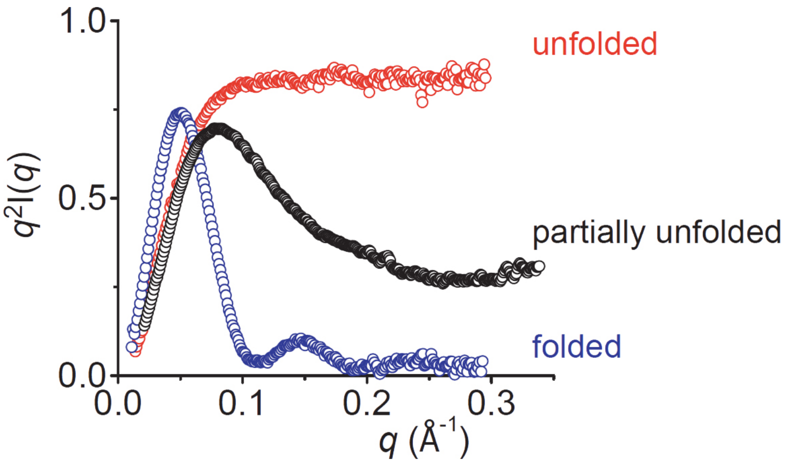


Figure 3.1 Kratky Plots of Folded and Unfolded Proteins. A folded protein will exhibit a Kratky plot similar to that seen in blue, where a distinct maximum is seen and the data falls to zero at higher *q* values. Unfolded proteins show Kratky plots similar to the plot seen in red, where no characteristic peak is seen in the data. A protein that is partially folded will show a Kratky plot similar to that seen in black, where a distinct peak is found, while the data does not return to zero at higher *q* values. This figure has been adapted from (Putnam et al., 2007).

It is important to note that an absolute measure of protein foldedness is difficult to obtain due to the varying levels of noise that different data sets may exhibit which are exacerbated in a Kratky plot caused by scaling the intensity by q^2. In many cases, low signal-to-noise ratios result in the appearance of unfoldedness and increasing the signal-to-noise ratio either by increasing concentration of the solute or by increasing the time of exposure will reveal that the particle is in fact folded. Here we do not seek to determine an absolute degree of foldedness, but only to compare similar data sets separated only by X-ray dose that share similar levels of signal-to-noise.

In order to alleviate some of the problems in determining the Kratky ratio due to low signal-to-noise we have decided to evaluate the fit to the scattering data rather than the raw data itself. While this method does introduce a degree of modeling, since we are utilizing the t-statistic to evaluate changes in foldedness only, and since any discrepancies resulting from the fitting procedure are not likely to vary from profile to profile in a dose dependent manner, calculating the Kratky ratio using the fit to the data will not skew the final result. The fit to the scattering profile used for this analysis is provided by the output from DATGNOM.

The maximum of the Kratky plot is determined through the use of moving averages. For each data point $I*q^2$ is evaluated and averaged over a block of ten data points. The block is then shifted forward five data points and the procedure is repeated. If the second block yields a higher average than the first block, then the block is shifted another five data points. This procedure is repeated until the average of the current block of points falls below the previous block of points, indicating that a maximum has occurred. Once the maximum has been reached the remaining points are divided into blocks of ten points, each shifted five points from the previous block, and the average $I*q^2$ is evaluated for each. This list of averages is then sorted and the minimum average is selected as the global minimum, following the previously determined maximum. The Kratky ratio is then calculated as the minimum divided by the maximum. If no maximum is found, then the program reports to the user that the particle is unfolded. To assess whether radiation damage results in protein unfolding, the linear regression t-test is used and if radiation damage is present ($p < 0.05$), then exposures that yield Kratky ratios more than two standard deviations from the y-intercept of the linear regression are excluded from averaging.

3.3 Concentration Dependence

While increasing the time of exposure is one way to increase the signal-to-noise in a scattering profile, radiation damage limits the total dose that can be placed on the sample and therefore limits the maximum signal-to-noise that can be achieved. However, another way to increase signal-to-noise is to increase the concentration of the protein in solution. This can significantly increase the signal-to-noise ratio. However, as the concentration of the protein in solution is increased, the average distance between individual particles decreases, and therefore the likelihood of their interaction increases. Variations in electrostatic charge or hydrophobic regions distributed across the surface of the particle can result in either attractive or repulsive forces between neighboring particles when concentration is increased beyond a particular threshold. These interparticle interactions directly affect the structure factor in equation 3.1, causing it to deviate from unity. This results in a breakdown of the assumption that *I(q)* collected in a SAXS experiment can be treated as the form factor, which contains the size and shape information desired. Therefore, it is very important that no interparticle interactions are present in the course of a SAXS experiment. One way to monitor for the presence of interparticle interactions is to evaluate SAXS parameters as a function of concentration. By plotting each SAXS parameter as a function of increasing concentration, it can be determined whether or not concentration dependent effects are distorting the intensity profile. A minimum of three concentrations is required for this analysis to calculate the linear regression. If the t-statistic shows a significant trend, this suggests that interparticle interactions exist due to increasing concentration. If interparticle interactions do exist, these may be alleviated using more dilute solution conditions, or by modifying the buffer conditions such that interparticle interactions do not occur.

3.3.1 Scaling SAXS Profiles

An important step in assessing the linear regression for a series of data points is determining the independent variable. When discussing radiation damage this step was trivial, since each exposure is identical to every other exposure, and since we are comparing serial exposures, the independent variable is simply taken to be the exposure number. However, in the case of testing for concentration dependence, this step is non-trivial. The independent variable is no longer exposure number, but is concentration of the particle in solution. Oftentimes concentration series are not on a linear scale, so one cannot simply plot the independent variable as the sample number in the concentration series. Users often prepare multiple concentrations of a sample by performing serial dilutions, which aids in creating a linearly dependent set of concentrations. However, in practice it is often the case that errors in estimated concentration occur, either from human error in pipetting small μL volumes of liquid, or from varying rates and times of exposure to air resulting in different levels of evaporation of solvent causing unpredictable changes in concentration. This problem can be alleviated through the use of a UV spectrometer, which can be used to measure the absorbance of ultraviolet light at 280 nm. Using the known extinction coefficient for the protein the concentration can then be calculated at the time of data collection. While more accurate, since the extinction coefficient is many times predicted from primary amino acid sequence, changes in absorption due to the tertiary structure of the protein or a lack of tryptophan in the primary sequence can result in errors in concentration estimation. Additionally, many samples that have not had this information recorded at the time of data collection would be precluded from this analysis, instead relying on predicted concentrations from serial dilutions. Even if accurate protein concentrations are known, in a high-throughput setting the manual input of several concentrations for possibly hundreds or thousands of samples may prove to be cumbersome and intractable, though that option is available in the SAXStats scripts.

To determine the correct abscissa for each concentration we have chosen to evaluate each concentration in a series on a relative scale, *i.e.* one that is not dependent on knowing the absolute solution concentration in mg/mL. One way to approach this would be to simply divide the *I(0)* of each concentration by the *I(0)* of the lowest concentration, since *I(0)* is dependent on the number of particles in the illuminated volume. However, this simple approach has the flaw that *I(0)* is also dependent upon the molecular weight of the particle in solution, and if interparticle interactions are occurring as a result of increasing concentration, then the relative scaling factors will be skewed. Another possible approach is to scale the scattering profiles using the full *q*-range. Since most small changes in interparticle interaction will be manifest at only the lowest resolution data points, the scaling is likely to be more accurate and is in fact the method used to scale data sets in the software program PRIMUS. However this method also suffers from inaccuracies often due to errors in the high-resolution data points. These errors can result from improper background subtraction or high levels of noise that systematically skew the scaling factor.

In order to alleviate the possible errors in scaling resulting from both the high and low-resolution regions of the scattering profile, we have chosen to select from the data a region of one hundred data points beginning at q = 0.07 Å^{-1} as this region should not experience distortion from small changes in interparticle interactions for most proteins while still being at a low enough resolution to avoid the region most sensitive to low signal-to-noise. To determine the proper scaling factor between concentrations the following procedure is performed. First, each data point in the scaling region for each concentration is divided by the corresponding data point in the first concentration, yielding a list of ratios. Second, this list is sorted from least to greatest and the median ratio is selected as the scale factor for that concentration. Lastly, the first concentration is given an abscissa of one while each additional concentration is given an abscissa equal to its corresponding scale factor. These abscissae are now used as the regressors in the linear regression analysis to detect changes as a function of concentration.

3.3.2 Detecting Changes in R_g, D_{max}, and *I(0)*

Changes in particle size as a function of concentration may manifest as changes in R_g and D_{max}. To calculate R_g we have employed two methods described in section 3.3.3. First, the pair distribution function is used to calculate R_g and D_{max} using the software program DATGNOM. These values are subsequently used to determine the Guinier region according to equation 3.21. Second, equation 3.16 is used to calculate R_g according to the Guinier approximation after using the least squares method described in equations 3.2 to 3.10 to best fit the data. Using the previously described linear regression t-test, each of these three parameters, R_g (Guinier), R_g (*P(r)*), and D_{max} are tested for their dependence on concentration. The p-value of each is reported to the user.

In addition to changes in the apparent size and length of the particle, interparticle interactions may also result in changes in *I(0)*, which depends on molecular weight. Similar to R_g, *I(0)* can also be determined from *P(r)* according to the following equation:

$$I(0) = 4\pi \int_{0}^{D_{max}} P(r)dr \qquad (3.22.)$$

For each concentration (after scaling) the *I(0)* is calculated using both equation 3.22 and the y-intercept of the linear regression in the Guinier approximation. Both estimates are used for detecting dependence on concentration and are reported to the user for comparison.

3.3.3 Detecting Changes in Particle Volume

Occasionally, due to the shape of a particle, increases in particle size may not significantly alter R_g or the D_{max} and thus may remain undetected under the current analysis. However, another measure of particle size is the excluded particle volume. This value, also known as the Porod volume, is based on the observation by Porod that globular particles that have a sharp interface between the surface and the solvent display a decay in the high resolution region that goes as q^{-4}. Porod found that the volume of the particle could be calculated according to the following equation:

$$V = \frac{2\pi^2 \, I(0)}{Q} \tag{3.23.}$$

where V is volume and Q is the Porod invariant such that

$$Q = \int_0^\infty [I(q) \, - \, k] \, q^2 dq \tag{3.24.}$$

where k is a constant subtracted to ensure the asymptotical intensity decay proportional to q^{-4}. This calculation is provided by the software program DATPOROD (Petoukhov et al., 2012) and requires the output from DATGNOM, already calculated in the previous section. The Porod volume is directly proportional to the molecular weight of the protein assuming a typical particle density of 1.37 g/cm^3 according to the following equation:

$$MW \; (Da) = \frac{V}{1.66} \tag{3.25.}$$

where *MW* is the molecular weight in Daltons (Rambo and Tainer, 2011). After performing a linear regression analysis on Porod volume as a function of concentration, the p-value is reported to the user. Additionally the molecular weight for each protein is reported and the average is calculated and can be compared with the predicted molecular weight for consistency and to check if oligomers are present.

3.3.4 Evaluating Linearity in the Guinier Region

In section 3.3.3 we described in detail the method for estimating the correct Guinier region for calculating R_g. While this is an effective method for determining R_g, it does so regardless of the linearity of the data in the Guinier region. Linearity in the Guinier region is an important prerequisite to ensure monodispersity (Jacques and Trewhella, 2010). If data in the Guinier region are nonlinear, this suggests that either interparticle interactions or polydispersity are present in the sample. While the software program AutoRg possesses a function to test for aggregation in a sample, it suffers from similar problems discussed earlier in calculating R_g, *i.e.* that sufficiently noisy data or Guinier regions that are particularly sparse in data points can cause the program to fail. In order to evaluate whether or not the data in the Guinier region is linear, we have developed a method utilizing the previously described linear regression t-statistic to calculate the significance of nonlinearity.

After determining the two intervals for the Guinier region, *i.e.* $q < 1.3/R_g$ and $[\pi/D_{max}, 1.3/R_g]$, the method of least squares is applied to fit each block of three consecutive data points, the minimum required to calculate a linear regression. The slope of the line through these three points is calculated and the block of points is shifted one data point and the procedure is repeated. Next, a linear regression is calculated for the slopes as a function of the data point. If the region is linear, then the slope of each consecutive block of three points should not be dependent upon where in the Guinier region the block is. Therefore, using the t-statistic, we can calculate a p-value for the likelihood that a trend is significant, and therefore that linearity in the Guinier region is not upheld. While typically a p value of 0.05 is recognized as significant, since nonlinearity in the Guinier region can greatly affect further analyses of the SAXS profile, we have relaxed the p-value threshold to 0.20. The slope of the linear regression is used to determine whether the interparticle interactions are attractive or repulsive. If the slope of the regression is positive, this suggests that the interactions are attractive. If the slope of the regression is negative, the interactions are repulsive. To ensure the highest degree of data quality, both previously described intervals of $q < 1.3/R_g$ and $[\pi/D_{max}, 1.3/R_g]$ are used for examining nonlinearity in the Guinier region and are reported to the user. If very slight levels of repulsion appear to be occurring, then the user can either merge data from high and low concentrations or use zero extrapolation to generate a curve with sufficient signal-to-noise while alleviating the influence of interparticle interactions (Petoukhov et al., 2012).

3.3.5 Second Virial Coefficient

Until now we have discussed how interparticle interactions can be detrimental to obtaining information about the shape and size of a particle in solution from a SAXS profile. As mentioned, this is due to the introduction of a structure factor not equal to one, which results in *I(q)* not equal to the form factor. While interparticle interactions may preclude the analysis of shape and size information, these interactions themselves can provide useful information about the effect of the solution conditions on the particle.

One quantity used to measure interparticle interactions is the second virial coefficient, or B_{22}. B_{22} is the second coefficient in the virial expansion of the many-particle system that provides corrections to the ideal gas law. B_{22} is typically the most significant correction to the ideal gas law since it depends only on the pair interactions between particles, whereas higher order coefficients depend on multi-body interactions that are usually rare, and in fact are usually dropped from the expansion to obtain an accurate approximation.

It has been shown that protein solutions resulting in crystal formation have B_{22} values that reside in a narrow window, termed the “crystallization slot”, which ranges from ~ -8 x 10^{-4} to -1 x 10^{-4} mol*mL/g^2 (George and Wilson, 1994). It was also shown that protein solutions that failed to promote crystallization yielded B_{22} values well outside this window. Knowledge of B_{22} could theoretically provide insight on directing buffer conditions such that the solution properties cause B_{22} to fall into the crystallization slot, and may result in higher success rates for crystallization. However, in practice, the large solution volumes and low efficiency of determining B_{22} have hindered the wide adoption of this method.

SAXS data from a concentration series of a protein can be used to calculate B_{22}. With the advent of high-throughput robotics on SAXS beamlines and the low µL volumes of liquid required to perform the experiment, B_{22} values can be obtained with much greater speed and efficiency. B_{22} can be obtained from a concentration series of scattering profiles by utilizing the structure factor according to the following equation:

$$\frac{F(q=0)}{I(c,q=0)} = \frac{1}{S(c,q=0)} = 1 + (2MB_{22})c \tag{3.26.}$$

where *F(q=0)* is the form factor evaluated at *q*=0, *I(c,q=0)* is the forward scattering intensity for each concentration, *S(c, q=0)* is the structure factor evaluated at *q*=0 for each concentration, *M* is the molecular weight of the particle, and *c* is the concentration (Bonnete and Vivares, 2002; Pollack, 2011). To solve this equation, the forward intensity is determined for each concentration after scaling. Ideally one requires the form factor of the particle, which can be calculated if the atomic structure is known. However, for many SAXS experiments the atomic structure is unknown. In order to estimate *F(q=0)* without the knowledge of atomic structure, the lowest concentration is used and assumed to be equivalent to the form factor, since it will experience the least degree of interparticle interactions among the series of concentrations. The estimated forward scattering is then determined from the lowest concentration using the *P(r)* method according to equation 3.22, since this method is least sensitive to interparticle interactions. However, when determining the forward scattering intensity for the various concentrations, the Guinier approximation is used since it is most sensitive to interparticle interactions, which is of interest when discussing second virial coefficient. While this method is preferred, the forward scattering for each concentration using the *P(r)* method is also determined for comparison. After determining *F(q=0)* and *I(c, q=0)* the molecular weight is estimated from each concentration using the Porod volume and equation 3.27. If the user knows the molecular weight from primary sequence analysis or experimental methods such as mass spectrometry, that value can also be used. If the user knows the concentration it is recommended to give the concentration in mg/mL of the first concentration of the series. If the user does not know the concentration, it may be estimated from a protein standard such as bovine serum albumin, lysozyme, xylose isomerase, or another well characterized protein standard. The concentration will be calculated for the first concentration in the series according to the following equation:

$$\frac{MW_{protein} \times C_{protein}}{I_{protein}(0)} = \frac{MW_{standard} \times C_{standard}}{I_{standard}(0)} \qquad (3.27.)$$

where $MW_{protein}$ is the molecular weight of the protein taken from the user input if given or estimated from the Porod volume, $I_{protein}(0)$ is the forward scattering intensity of the protein, $MW_{standard}$ is the molecular weight of the protein standard, $C_{standard}$ is the concentration of the protein standard in mg/mL, $I_{standard}(0)$ is the forward scattering intensity of the protein standard, and the equation is solved for the concentration of the protein, $C_{protein}$. From this value calculated for the first concentration in the series, the remaining concentrations are estimated using the scale factor determined in section 3.4.1. Next, *F(q=0)/I(c,q=0)* is plotted as a function of concentration with the addition of the point (0,1) according to the *y*-intercept of equation 3.26. A linear regression is calculated for these points and from the slope of this line and the estimated or user-defined molecular weight, B_{22} is calculated according to equation 3.26. When B_{22} is negative the particle interactions are attractive, when B_{22} is positive the particle interactions are repulsive, and when B_{22} is zero no interparticle interactions exist.

3.4 Experimental Application of SAXStats to a Large Dataset

In section 2 it was shown for a set of 28 proteins that SAXS data can provide structural information that agrees well with known high-resolution structural data. Each of these samples was analyzed with a combination of manual inspection and automated software procedures. Whereas numerical measurements such as R_g and D_{max} were collected using automated software, trends in measurements as a function of either radiation or concentration were analyzed manually. In a high-throughput setting, where hundreds or even thousands of scattering curves need to be analyzed, manual analysis is likely not feasible. Additionally, at a beam line where informing the user about failed experiments can lead to changing solution conditions to increase success, fast, objective analysis is required due to typically brief shifts at beam lines.

3.4.1 Description of Sample Set

To test the success of the SAXStats package, 100 protein samples supplied by the Northeast Structural Genomics Consortium (NESG) were taken to beam line 4-2 at SSRL (Smolsky et al., 2007; Xiao et al., 2010). SAXS data were collected for each protein sample at three different concentrations according to the methods described in section 2.2. The NESG is one of four large-scale NIH funded structural genomics centers of the Protein Structure Initiative. Targeted proteins are typically representatives from large protein domain families or biomedical themes, or have been selected as targets whose known structure would be significant to the biomedical community (Wunderlich et al., 2004). Most target proteins are full-length polypeptide chains shorter than 340 amino acids selected from domain sequence clusters (Liu et al., 2004; Liu and Rost, 2004), which are organized in the PEP/CLUP database (Carter et al., 2003). Each protein cluster corresponds to putative structural domains whose 3D structure is not known nor can be accurately modeled through homology. Target taxa range from bacteria and archaea to eukaryotes, with a focus on human proteins.

Each target has a series of biochemical experiments performed including analytical gel filtration, static light scattering, mass spectrometry, NMR spectroscopy for determining rotational correlation time and, if possible, high resolution structural data, and, if crystals can be formed, X-ray crystallography (Bertone et al., 2001; Goh et al., 2003). The Hauptman-Woodward Institute houses the High-throughput Screening (HTS) laboratory for crystallization (Luft et al., 2003), where NESG sends many soluble, purified targets for crystallization. The HTS requires approximately 420 µL of protein solution for screening, leaving approximately 60 µL on average left over. The remaining volume of protein solution is used for SAXS analysis. It is important to note that each of these 100 samples has experienced at least two freeze-thaw cycles; primarily once when the protein is shipped to the HTS lab, and a second time when the remaining protein is shipped to SSRL. For many proteins, a freeze-thaw cycle can prove to be detrimental to solution conditions, causing the protein to aggregate or precipitate, and two freeze-thaw cycles increases the likelihood of this occurring. As part of the standard protein solution preparation that NESG does to ensure efficiency in its high-throughput structural genomics procedure, each protein target is prepared in identical buffer solution conditions consisting of 100 mM NaCl, 5 mM DTT, and 10 mM Tris at a pH of 7.5.

3.4.2 Overall Success of SAXStats Software Package

Robustness of the SAXStats package was assessed based on whether or not every statistic was calculated and yielded a non-zero value. In every case that not all statistics were calculated, R_g calculated from the Guinier approximation yielded a value of zero angstroms, and therefore this was used to detect failure of the program. Of the 100 total protein samples used in this study, SAXStats successfully calculated each statistic for all three concentrations for 88 samples. For the remaining 12 samples, 9 samples had two of the three concentrations that were successful, 2 samples had one concentration that was successful, and 1 sample had no concentrations that were successful. Of the 16 concentrations that failed, 14 failed because all the data points were negative. Inspection of the raw X-ray images showed no X-rays struck the detector during the exposure time. This may be due to an instrumentation error at the beam line that resulted in data not being collected for those concentrations, possibly that the shutter wasn't open during the time of collection or that the beam dumped and no X-rays were entering the hutch. While this would not necessarily result in negative data points in the total scattering curve, the integration procedure employed in SASTool also subtracts the signal from the buffer blank taken prior to the protein solution scattering experiment, which would result in all data points being negative. Of the remaining 2 concentrations that yielded an R_g of zero for only the Guinier approximation, the failure was due to particle sizes determined from *P(r)* that were too large, *i.e.* that resulted in Guinier regions that existed outside of the *q* range that was collected. Therefore, it has been shown that SAXStats is highly robust in calculating the described statistics in a high-throughput manner.

3.4.3 Radiation Damage Statistics

In total, six different SAXS parameters were tested for radiation damage, including χ (reported as the square root of χ^2), R_g calculated from the Guinier approximation, R_g calculated from *P(r)*, D_{max}, *I(0)*, and the Kratky Ratio. Of the 284 concentrations for which SAXStats successfully calculated statistics, 66 concentrations experienced radiation damage with $p < 0.05$ for at least one SAXS parameter. Some concentrations showed radiation damage for multiple SAXS parameters, shown in Figure 3.2. By analyzing both the average values of the parameters and the associated standard deviations, we can investigate the precision that parameters are determined. Knowledge of the precision of these parameters for 284 SAXS profiles allows us to determine a minimum signal-to-noise required to obtain well-defined values for future SAXS experiments. Using known standards to calibrate for the specific beam line and sample characteristics then allows us to calculate the minimum concentration needed for proteins of varying molecular weight required to achieve precise values for SAXS parameters.

Among the six different SAXS parameters, χ detected the likely presence of radiation damage least often (10 occurrences) and D_{max} most often (22 occurrences) (Figure 3.3). While radiation damage may distort SAXS profiles of the 66 concentrations where damage occurred, not all exposures suffered from radiation damage. Frames that did not suffer from radiation damage were averaged to produce one SAXS profile for each concentration. The distribution of the number of exposures not suffering from radiation damage, and thus used in averaging, is shown in Figure 3.4. On average, when radiation damage occurred in any of the 8 total exposures collected, it usually affected between 2 and 4 exposures, leaving between 4 and 6 exposures for averaging.

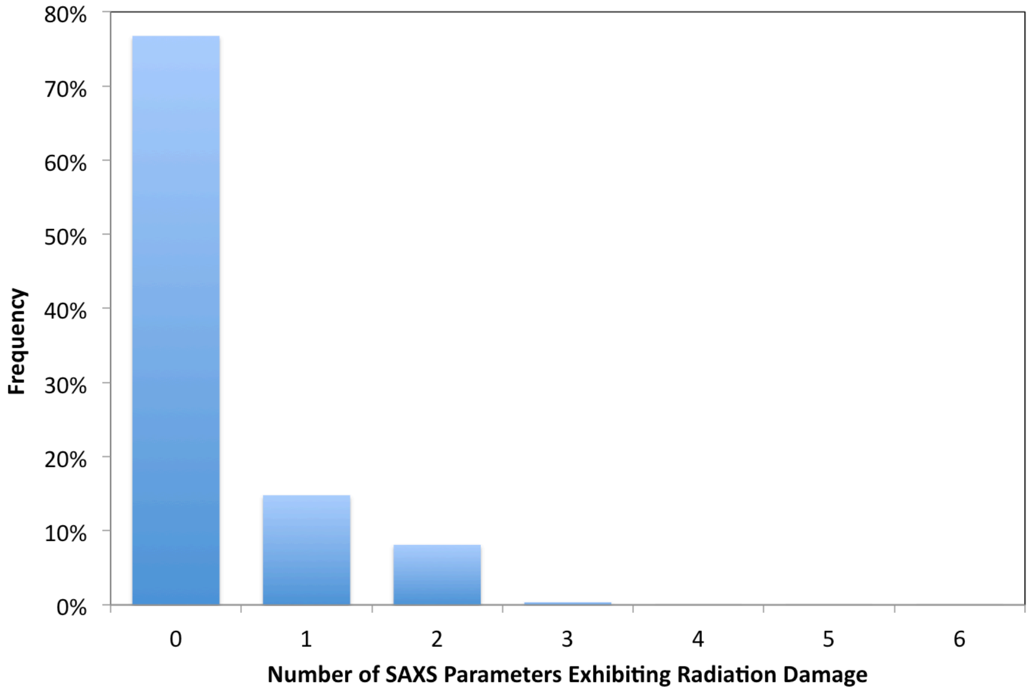


Figure 3.2 Frequency (Total = 284) of Multiple SAXS Parameters with Radiation Damage. While most samples experienced no radiation damage, those that did typically only had one or two parameters demonstrating the damage.

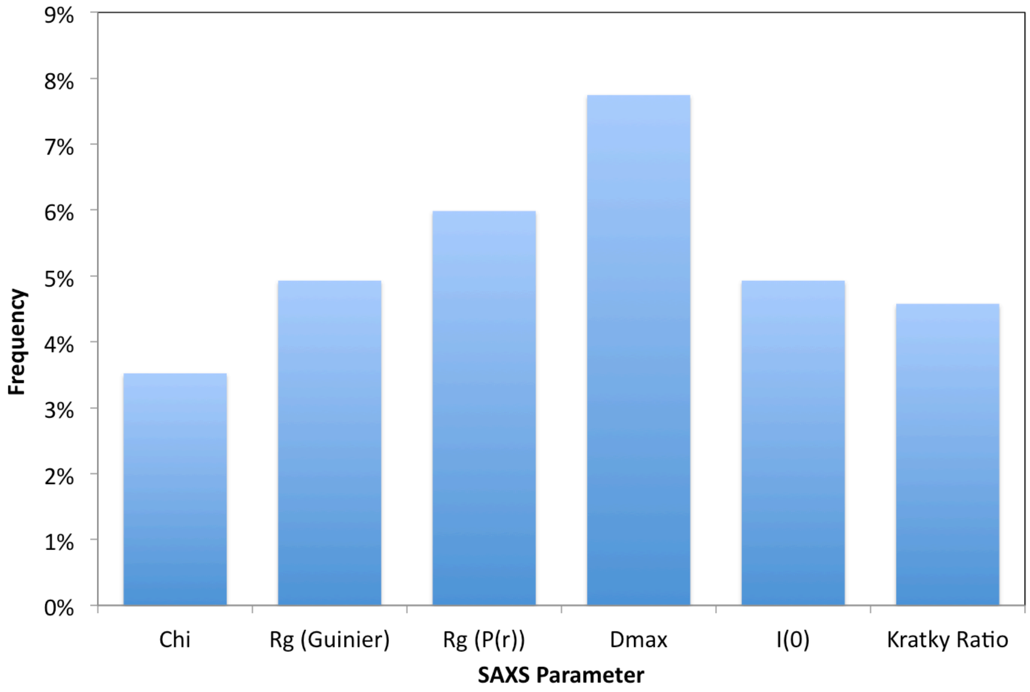


Figure 3.3 Frequency of Radiation Damage for Each SAXS Parameter. While distributed across all parameters, radiation damage is demonstrated most frequently by D_{max}.

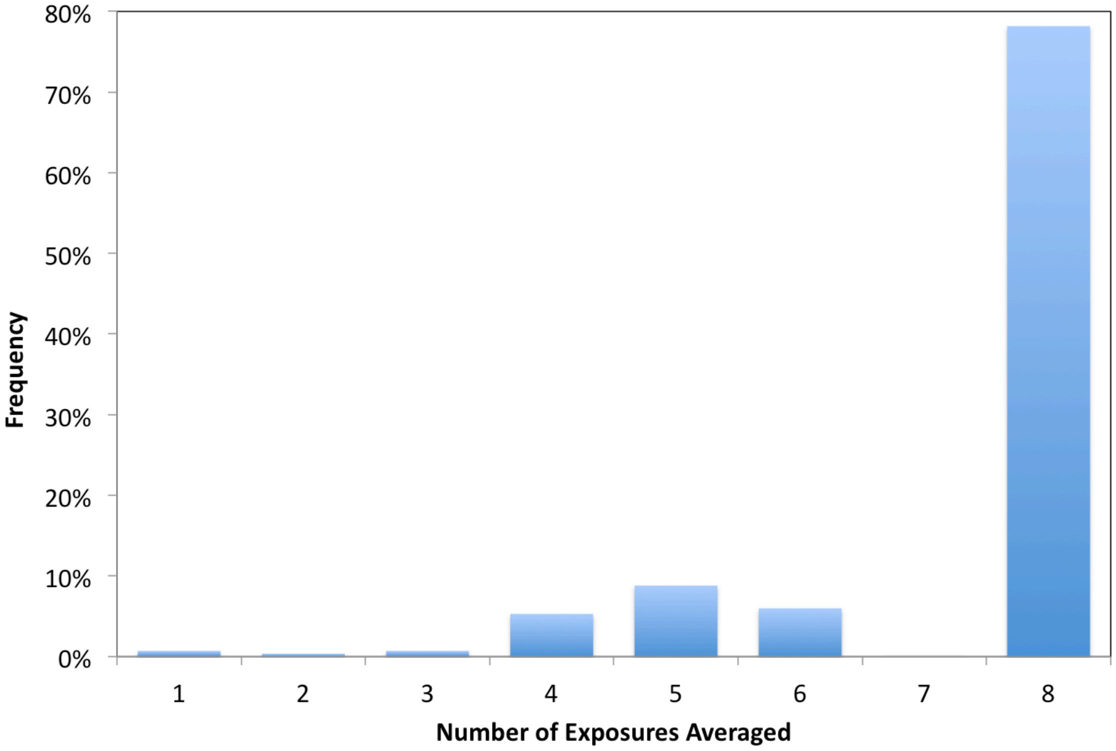


Figure 3.4 Frequency of the Number of Exposures Averaged. Most samples showed no radiation damage. For damaged samples, between 4 and 6 exposures were undamaged.

3.4.3.1 X Statistics

The similarity between the overall SAXS profiles of different exposures was measured using the χ statistic. The distribution of χ's for the 284 successful concentrations is shown in Figure 3.5. More than 95% of all concentrations showed an average χ value between 1.0 and 2.0, demonstrating that the vast majority of exposures within a concentration are similar, which is not surprising since the only difference between exposures is the total X-ray dose. The standard deviation of χ is shown in Figure 3.6. The standard deviation for most concentrations is less than 0.3 with 95% less than 0.8. For the 10 concentrations where radiation damage was detected for χ, the severity of the damage was relatively small, with 8 of 10 having slopes for the regression of less than 0.1 units per exposure, which is comparable to the majority of the standard deviations.

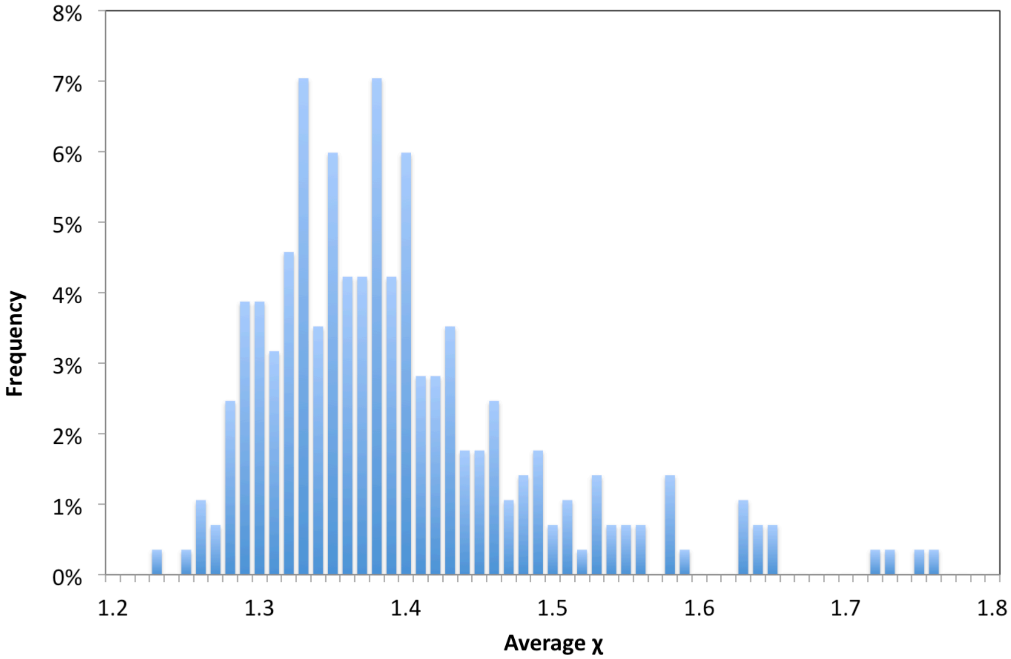


Figure 3.5 Distribution of Average χ Values. The majority of samples exhibit χ values between 1.2 and 1.6, showing that SAXS profiles of subsequent exposures are similar.

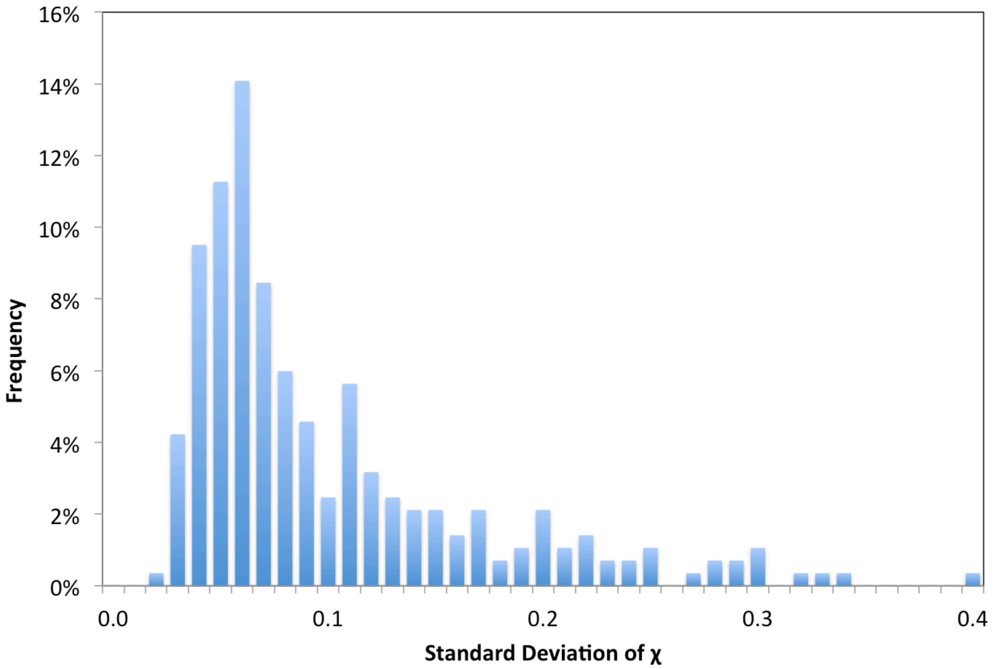


Figure 3.6 Distribution of the Standard Deviation of χ.

3.4.3.2 *P(r)* Distribution Statistics

The pair distribution function calculated by DATGNOM yields R_g and D_{max}. The distributions of average R_g and the standard deviation of R_g are shown in Figure 3.7 and Figure 3.8. The average R_g ranges from a minimum of 8 Å to a maximum of 130 Å with most particles between 10 Å and 50 Å. The standard deviation of R_g has a wide range with 75% of the population less than 5 Å. For the 17 concentrations experiencing damage for R_g, the severity of the damage ranged from -6.4 Å per exposure to +0.51 Å per exposure and was more heavily weighted towards negative values.

The average and standard deviation of D_{max} are shown in Figure 3.9 and Figure 3.10. The average D_{max} ranges from a minimum of 28 Å to a maximum of 405 Å. D_{max} has very large standard deviations with a significant population extending more than 30 Å. For the 22 concentrations showing radiation damage in D_{max}, the degree of severity was large ranging from -35.9 Å per exposure to +14.7 Å per exposure. However, while large, the degree of damage is within the range of standard deviations of D_{max}.

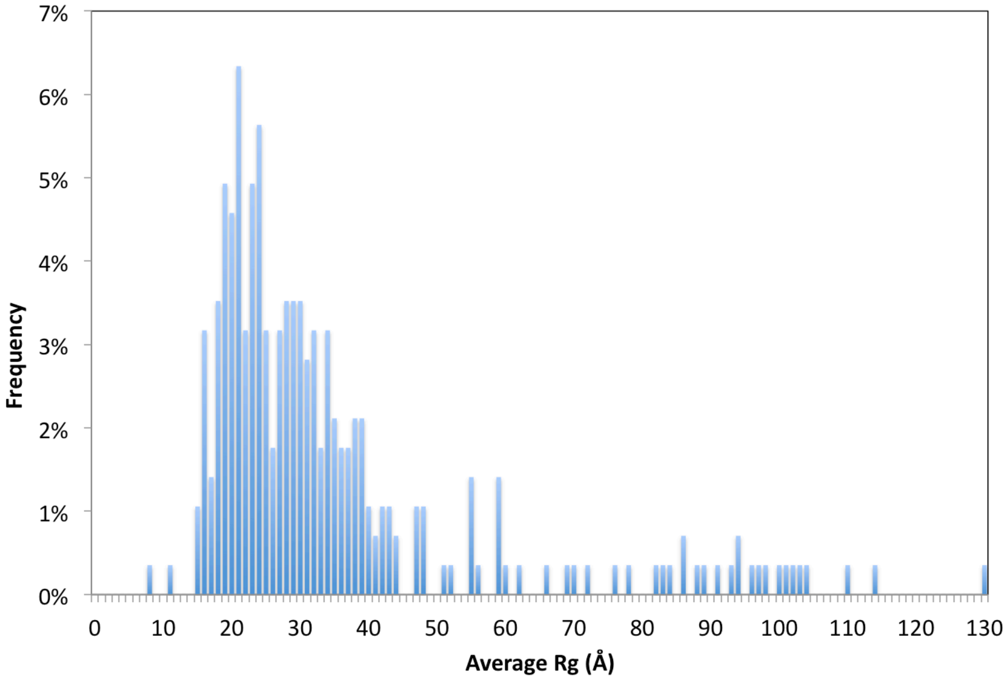


Figure 3.7 Distribution of Average R_g Calculated from *P(r)*. The R_g of most particles in the sample set is between 10 and 50 Å.

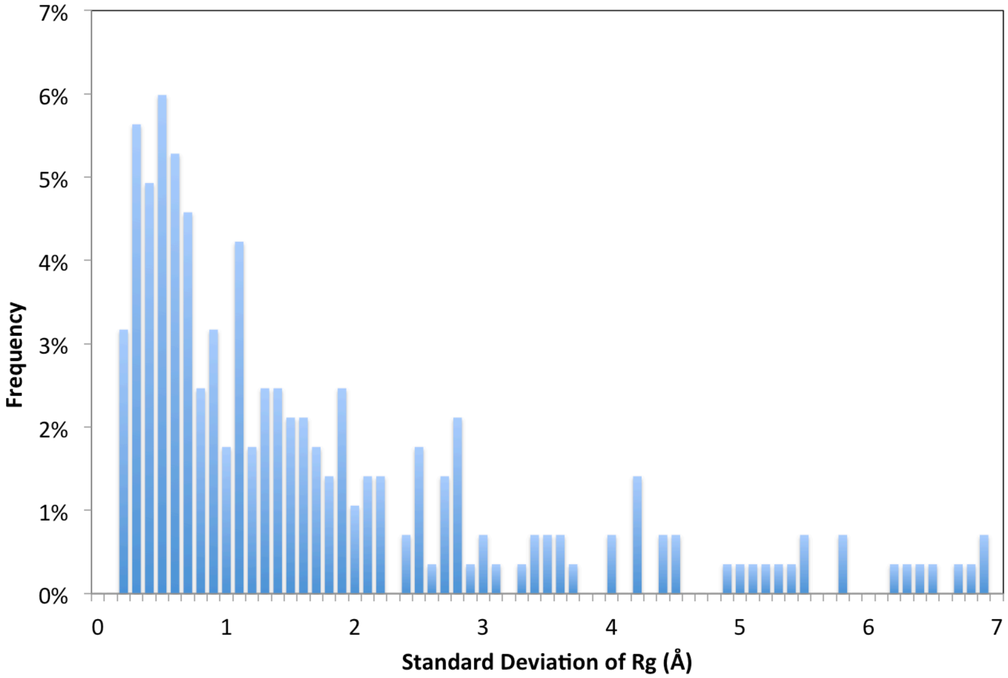


Figure 3.8 Distribution of Standard Deviation of R_g Calculated from *P(r)*. 75% of the standard deviations of R_g are below 5 Å.

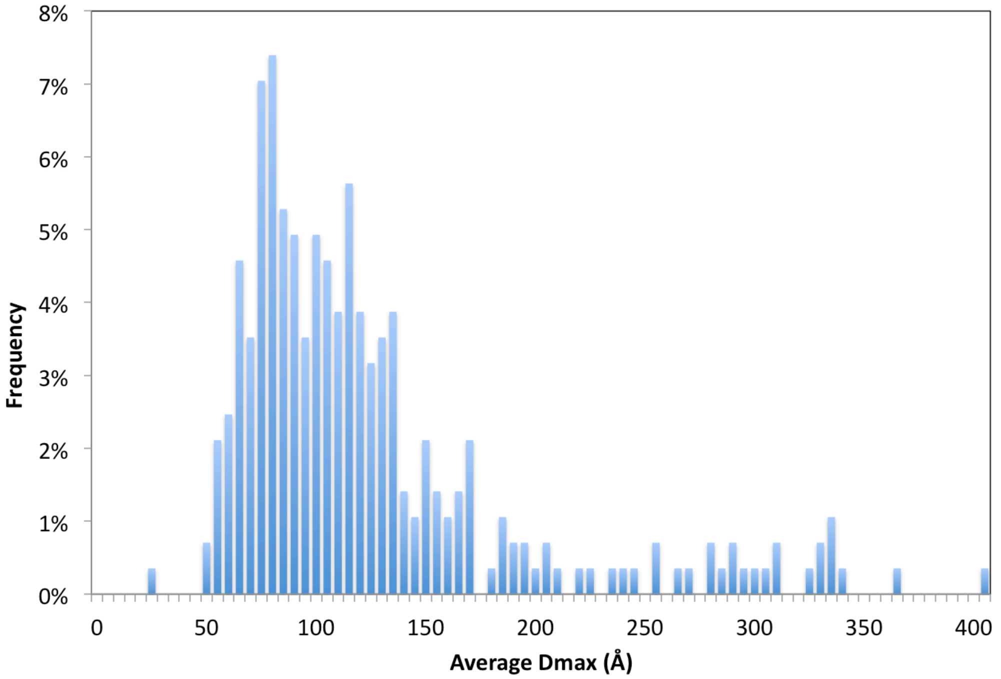


Figure 3.9 Distribution of Average D_{max}. The sample set shows a large range of particle sizes, with the majority less than 200 Å.

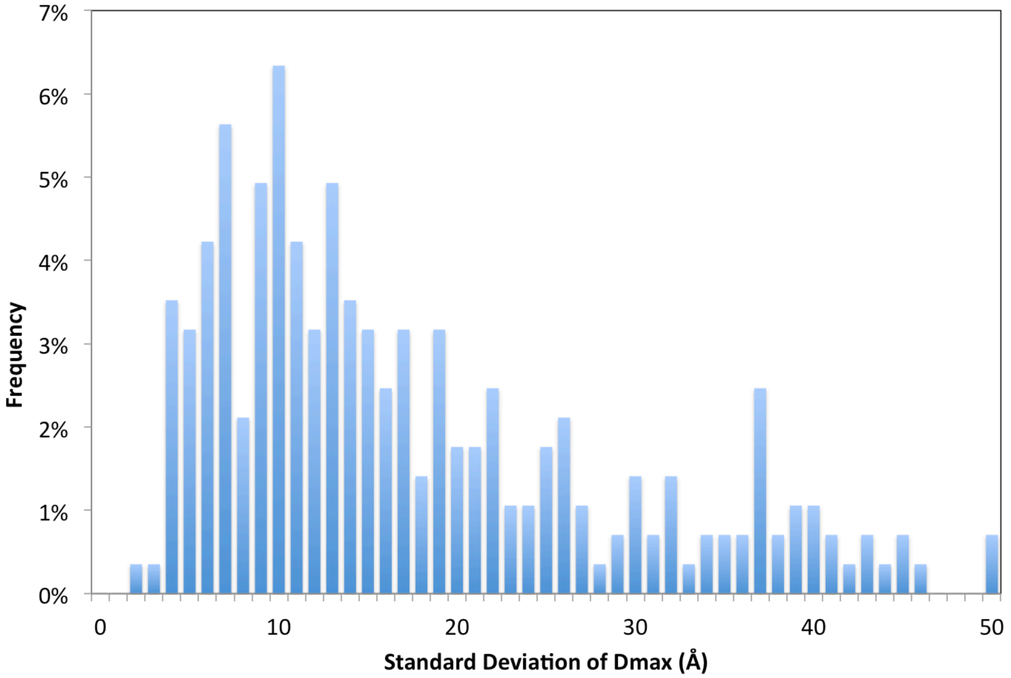


Figure 3.10 Distribution of Standard Deviation of D_{max}. D_{max} typically has very large standard deviations and varies widely between samples.

3.4.3.3 Guinier Approximation Statistics

The Guinier approximation was used to calculate R_g and *I(0)* in the interval defined by the R_g and D_{max} calculated using equation 3.21. The distributions of the average and standard deviation of R_g are shown in Figure 3.11 and Figure 3.12. The average R_g calculated using the Guinier approximation has a similar distribution to that calculated using the *P(r)* function, ranging from a minimum of 8 Å to a maximum of 133 Å. The distribution of the standard deviation of R_g calculated using the Guinier method shows a slightly narrower distribution than that calculated from *P(r)*, with 85% of the population less than 5 Å. For the 14 concentrations showing radiation damage in R_g, the severity of the damage ranged from -2.3 Å per exposure to +2.1 Å per exposure.

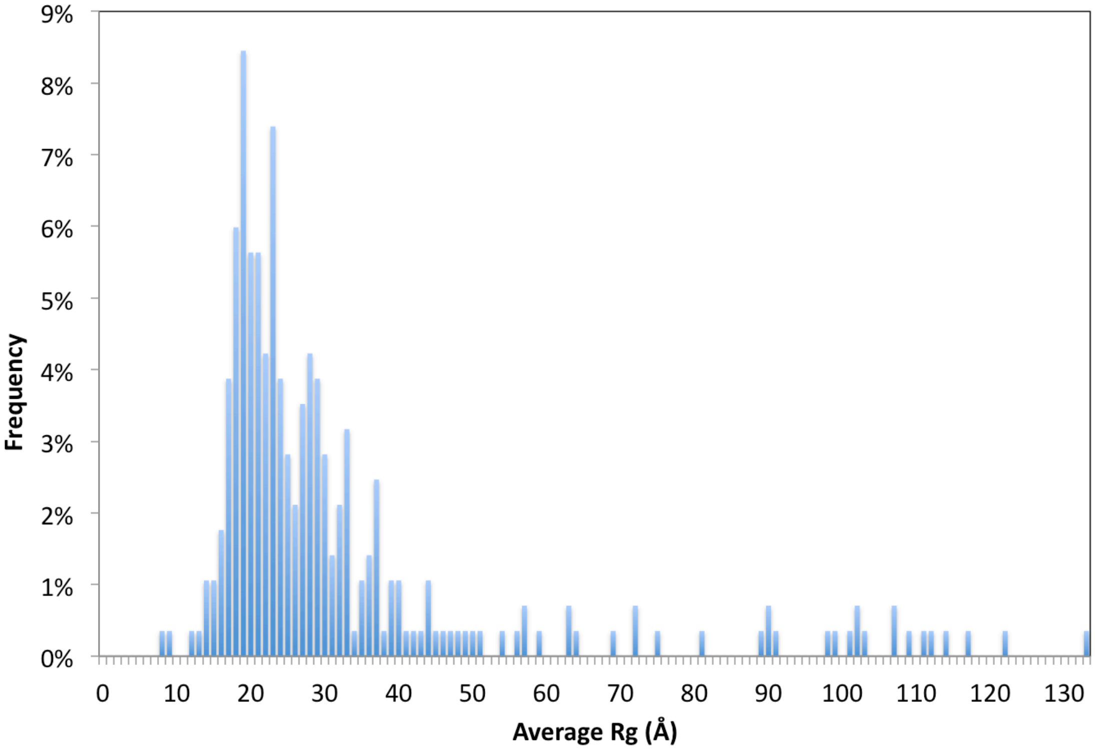


Figure 3.11 Distribution of Average R_g Calculated by the Guinier Method. The majority of R_g values is between 10 and 50 Å, similar to R_g calculated by *P(r)*.

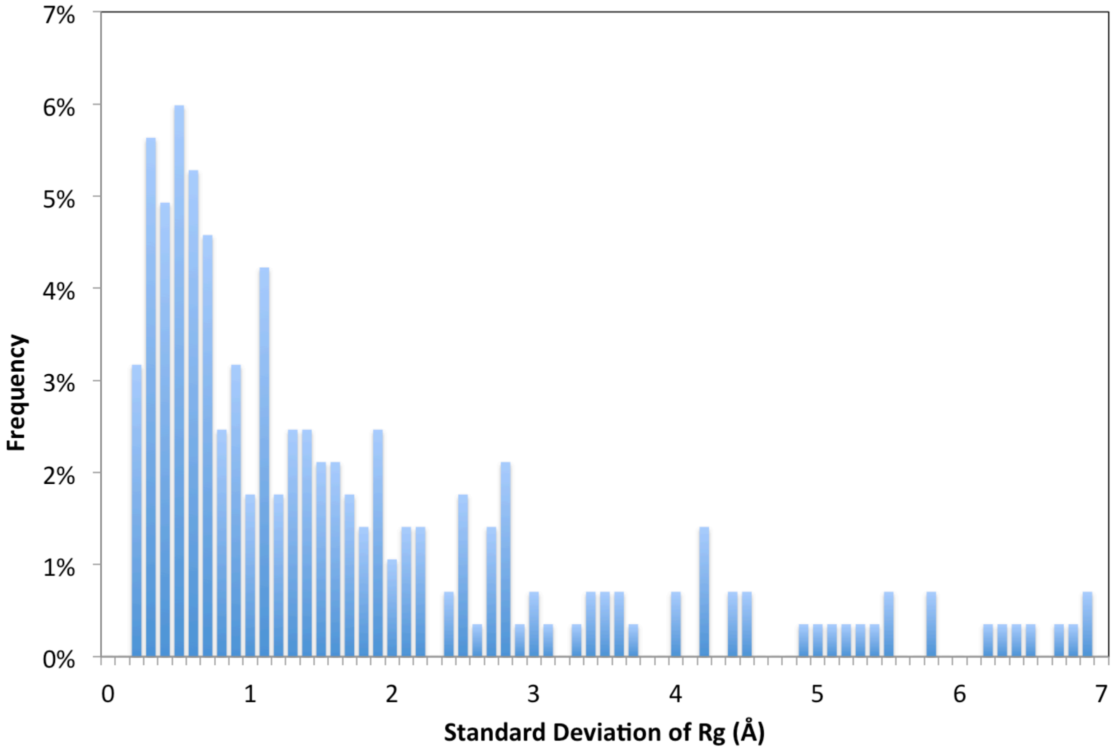


Figure 3.12 Distribution of Standard Deviation of R_g Calculated by the Guinier Method. 85% of the population shows standard deviations of less than 5Å.

To compare how similar the R_g values calculated from *P(r)* are to those calculated from the Guinier method, the difference between both values was calculated. The distribution of these differences is shown in Figure 3.13. More than half of the population showed very similar R_g values with a difference of less than 2 Å, 75% less than 5 Å, and only 10% greater than 10 Å. It is noteworthy that the distribution is not symmetrical about zero, with the peak of the distribution occurring at approximately +1 to +2 Å, showing that typically the R_g calculated from the *P(r)* function is slightly larger than that calculated using the Guinier approximation. One possible explanation for this may be that if a larger population of samples exhibits slightly repulsive interactions that this would result in an R_g that is underestimated by the Guinier approximation, since the Guinier region is more sensitive to interparticle interactions than *P(r)*.

The distribution of the average and standard deviation of *I(0)* is shown in Figure 3.14 and Figure 3.15. Greater than 90% of the population yielded *I(0)* values less than 4000 with more than 85% showing standard deviations less than 100. For the 14 concentrations showing that radiation damage affected *I(0)*, the severity of the damage ranged from -222 units per exposure to +32 units per exposure, with only one changing by more than 50 units per exposure.

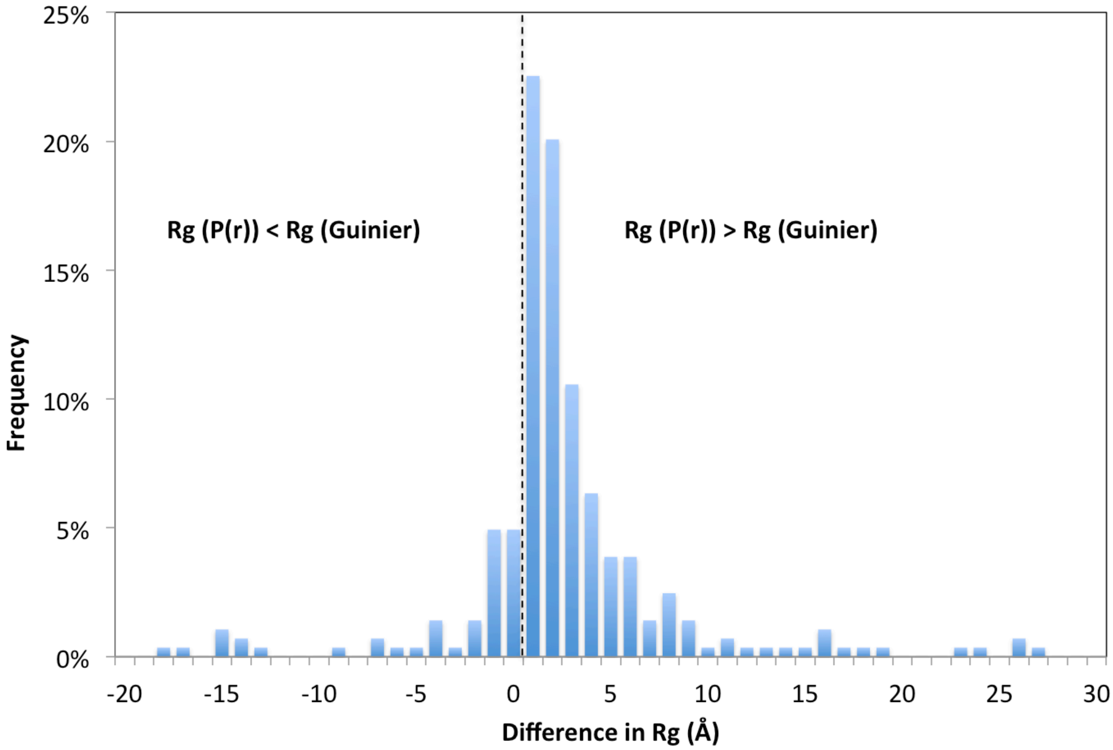


Figure 3.13 Distribution of the Difference in R_g Calculated Using Two Different Methods. The majority of samples show differences of less than 2 Å, 75% less than 5 Å, and 90% less than 10 Å.

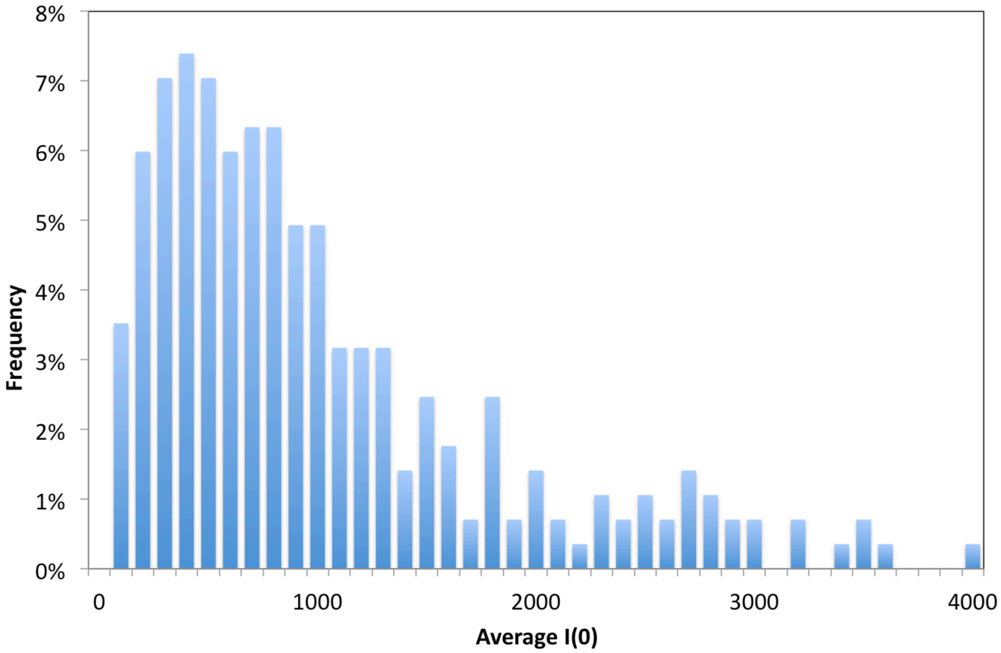


Figure 3.14 Distribution of Average *I(0)*. Values vary widely with 90% less than 4000.

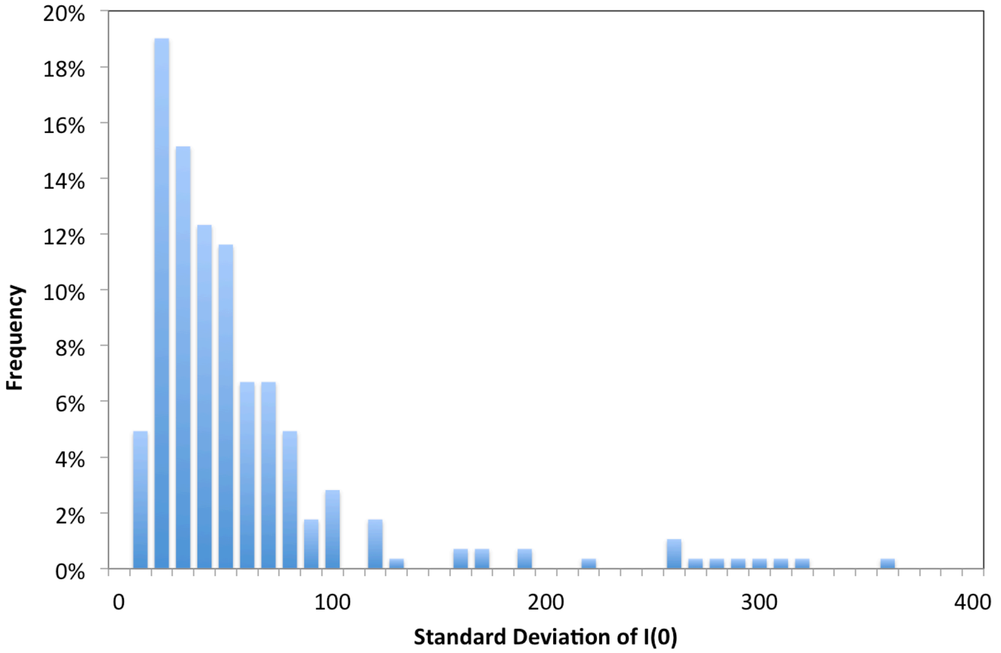


Figure 3.15 Distribution of the Standard Deviation of *I(0).* 85% of samples show standard deviations less than 100 units.

3.4.3.4 Kratky Ratio Statistics

In total, for 13 concentrations, SAXStats failed to find a maximum in the Kratky plot and these proteins were deemed unfolded. In each of these cases, the corresponding concentrations of the same protein were also either unfolded or had Kratky Ratios greater than 0.9. Of the remaining 271 concentrations, a very wide distribution of Kratky Ratios is seen (Figure 3.16 and Figure 3.17). No Kratky Ratios less than 0.13 were observed, which may signify a lower limit on the parameter, possibly due to either limitations of signal-to-noise at higher resolutions, issues with buffer subtraction, or to the breakdown of the Porod law due to contributions from internal structure. The distribution of Kratky Ratios appears to be slightly weighted towards more folded proteins, with 56% of the population between 0.13 and 0.50. Poor signal-to-noise will likely only cause ambiguity in determining whether or not a high Kratky Ratio suggests unfoldedness, and not in determining whether or not a low Kratky Ratio suggests globularity. Therefore, if signal-to-noise were improved, it may be that many of the Kratky Ratios greater than 0.5 would decrease to values less than 0.5, suggesting more globular particles. For the 13 concentrations that showed radiation damage effects in the Kratky Ratio, only two resulted in negative slopes, ranging from -0.018 units per exposure to +0.043 units per exposure. The heavily weighted positive distribution of slopes suggests that radiation damage only causes unfolding, which is not surprising considering that unfolding occurs due to radical formation that can break bonds holding together either secondary or tertiary structure elements, whereas it is highly unlikely that ionizing radiation will result in the formation of intramolecular contacts, resulting in compaction.

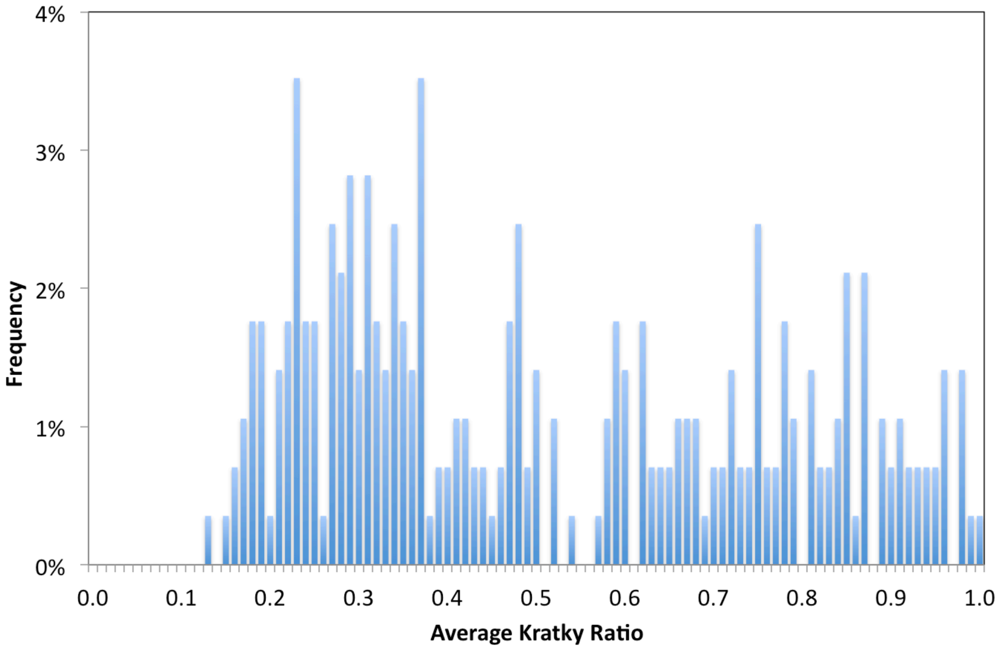


Figure 3.16 Distribution of the Average Kratky Ratio. Smaller values indicate a greater degree of foldedness. The distribution ranges from 0.13 to 1.0. Unfolded samples are not shown. The distribution is weighted to the left, with 56% of the samples less than 0.5.

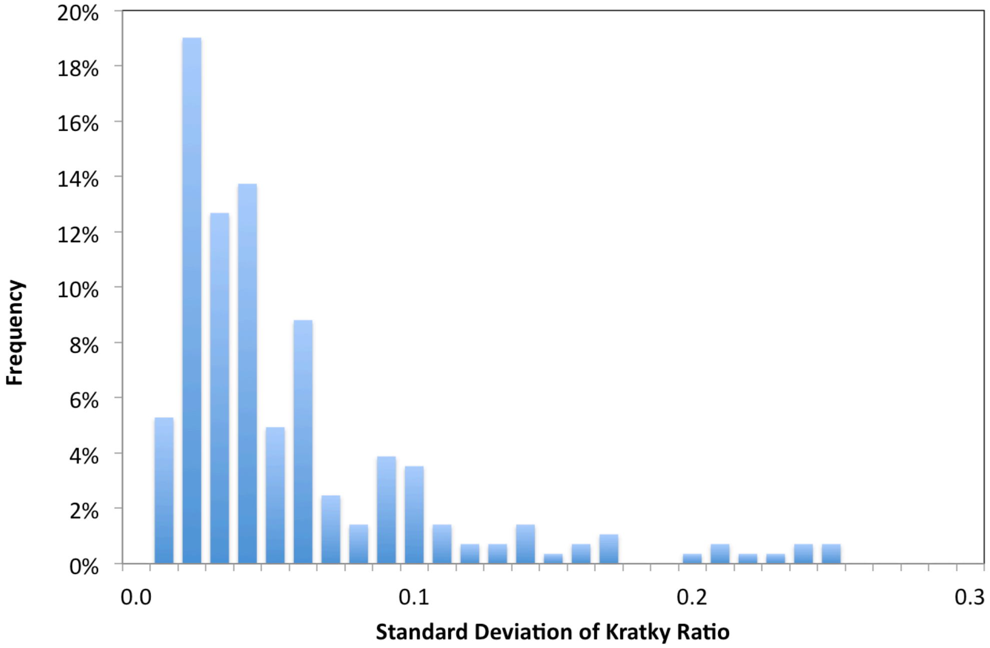


Figure 3.17 Distribution of the Standard Deviation of the Kratky Ratio. The majority of samples show standard deviations less than 0.1.

3.4.3.5 Comparing SAXS Parameters

Some SAXS parameters may be known more or less precisely than other parameters due to varying effects of the signal-to-noise ratio or particle size. The ability to directly compare different SAXS parameters would allow us to estimate which parameters are more precisely determined. Currently the above distributions for the various SAXS parameters cannot be compared directly to one another. One statistical measure than can be used to compare distributions of dissimilar parameters is called the coefficient of variation. The coefficient of variation is defined by the following equation :

$$c_v = \frac{\sigma}{\mu} \qquad (3.28.)$$

where c_v is the coefficient of variation, σ is the standard deviation and μ is the mean. The coefficient of variation allows us to compare the standard deviation of different SAXS parameters by normalizing each to the average value. The coefficient of variation for each SAXS parameter is shown in Figure 3.18 through Figure 3.23.

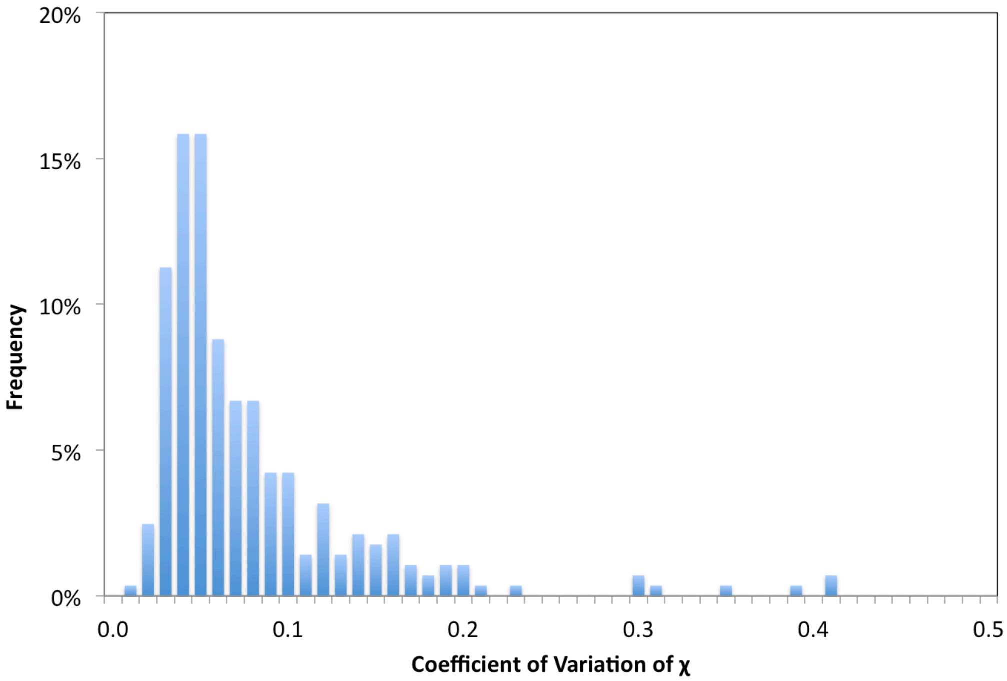


Figure 3.18 Distribution of the Coefficient of Variation of χ. χ is relatively well determined, with most of the population having c_v values less than 0.1.

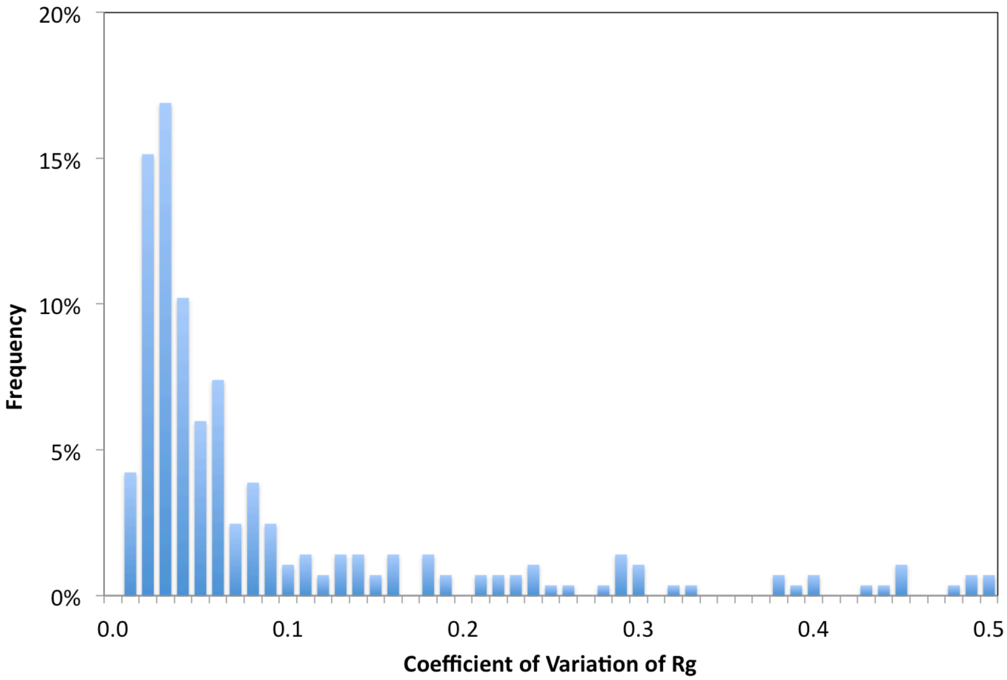


Figure 3.19 Distribution of the Coefficient of Variation of R_g Calculated Using *P(r)*. R_g is relatively well determined, with most of the population having c_v values less than 0.1

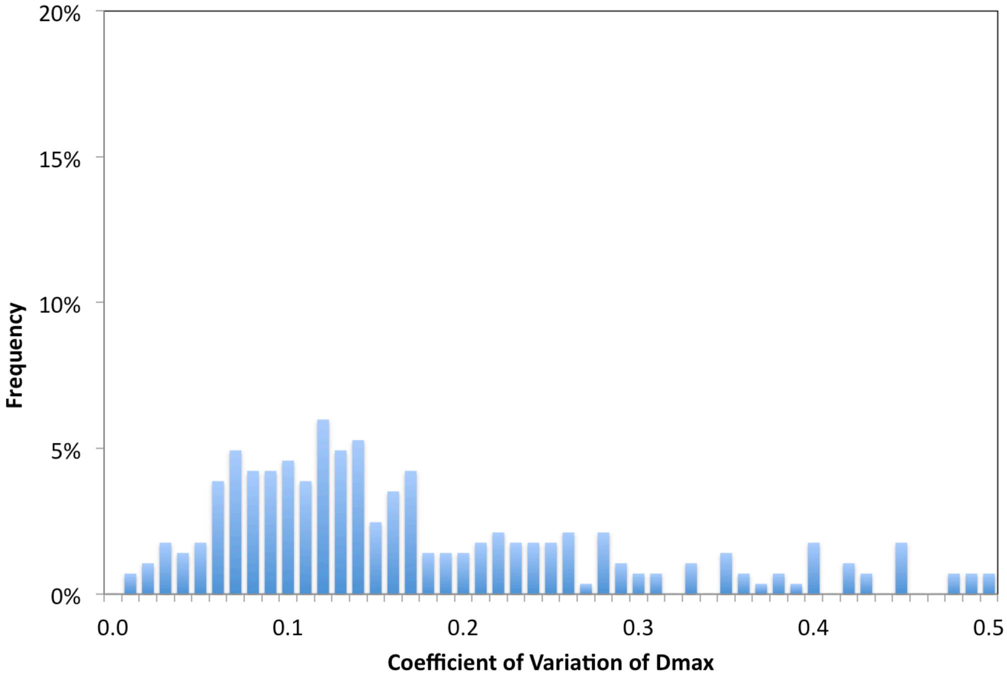


Figure 3.20 Distribution of the Coefficient of Variation of D_{max}. D_{max} is not relatively well determined, with most of the population having c_v values greater than 0.1.

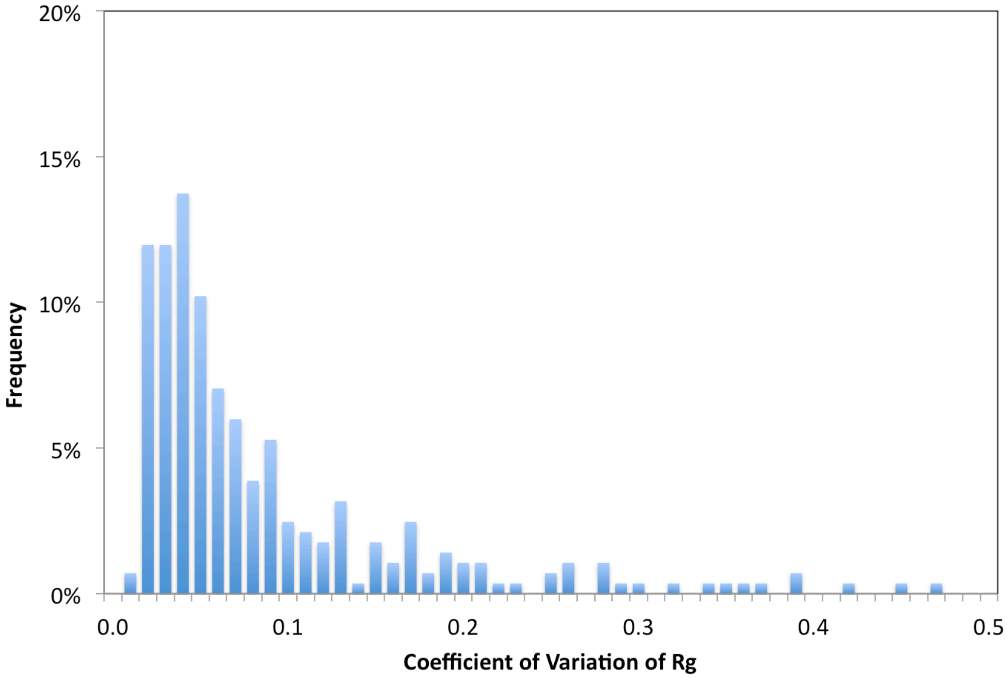


Figure 3.21 Distribution of the Coefficient of Variation of R_g Calculated by Guinier Method. R_g is relatively well determined, with most of the population having c_v values less than 0.1.

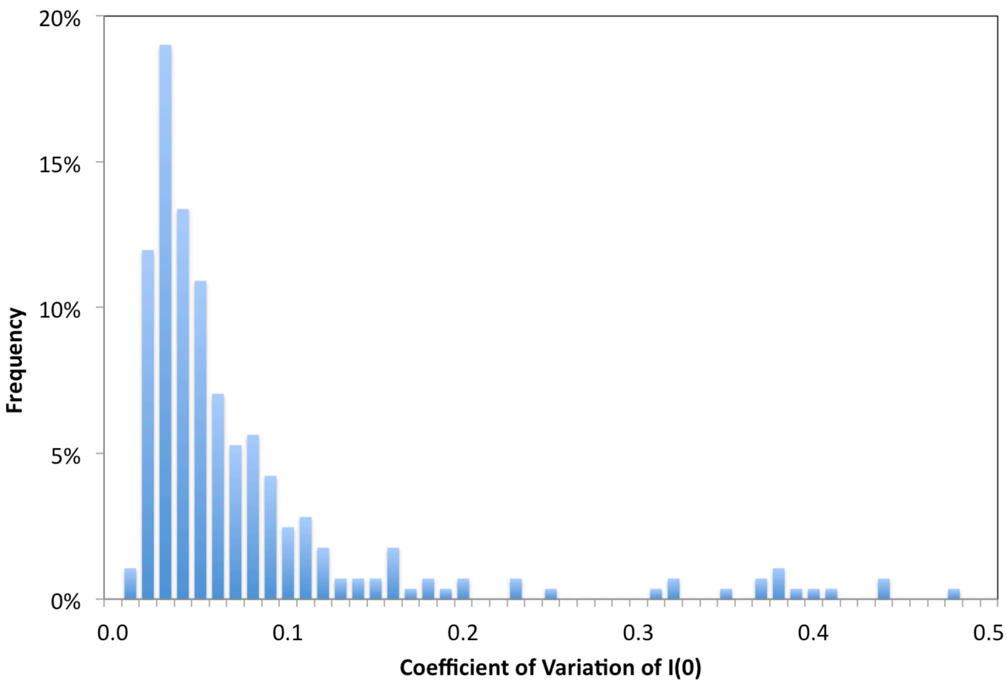


Figure 3.22 Distribution of the Coefficient of Variation of *I(0)*. *I(0)* is relatively well determined, with most of the population having c_v values less than 0.1.

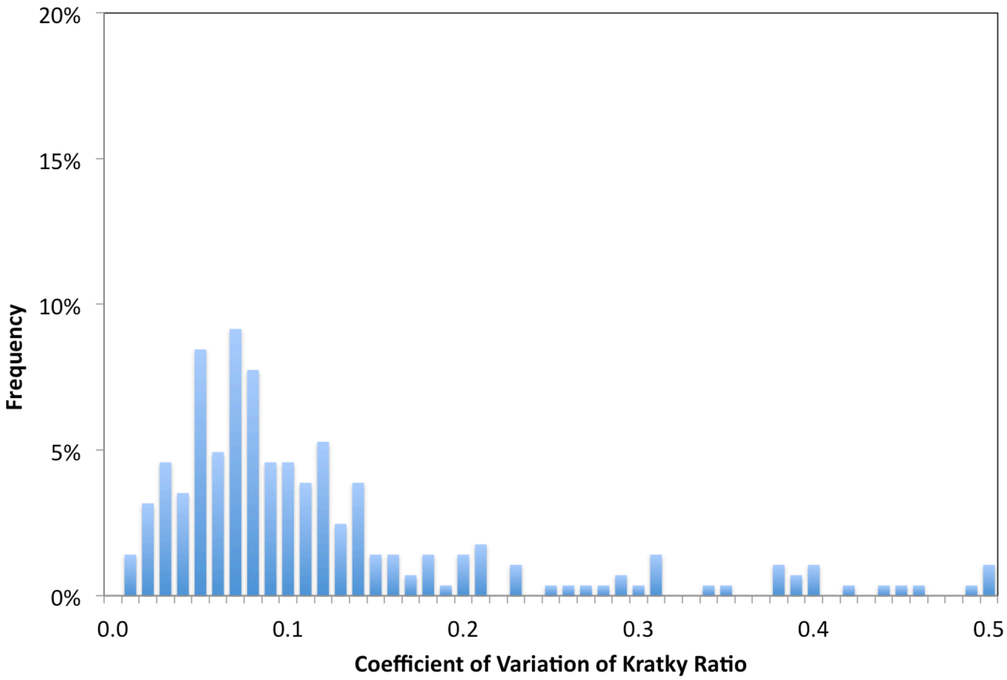


Figure 3.23 Distribution of the Coefficient of Variation of the Kratky Ratio. Kratky Ratio is not relatively well determined, with half of the population having c_v values greater than 0.1.

The distributions of the coefficient of variation are similar for χ, R_g (*P(r)*), R_g (Guinier), and *I(0)*. The majority of these distributions falls below 0.1 with peak values near 0.02, which suggests that these values are relatively well determined for most samples studied. However, the distributions for both D_{max} and the Kratky Ratio are much wider, suggesting greater dispersion in the values obtained from the multiple exposures and thus a lower confidence in the precise value of each parameter compared to the other parameters.

Since the coefficient of variation is dependent upon the standard deviation of the parameter, it is likely that samples with lower signal-to-noise will also result in higher coefficients of variation. The SAXS parameter most closely associated with signal-to-noise is *I(0)*. To determine whether or not a correlation exists between the coefficient of variation and *I(0)* and the degree of the correlation, the coefficient of variation versus the corresponding *I(0)* is plotted in Figure 3.24. The parameter used to assess the coefficient of variation is R_g calculated using the Guinier approximation, since it is one of the most commonly calculated parameters in SAXS analysis.

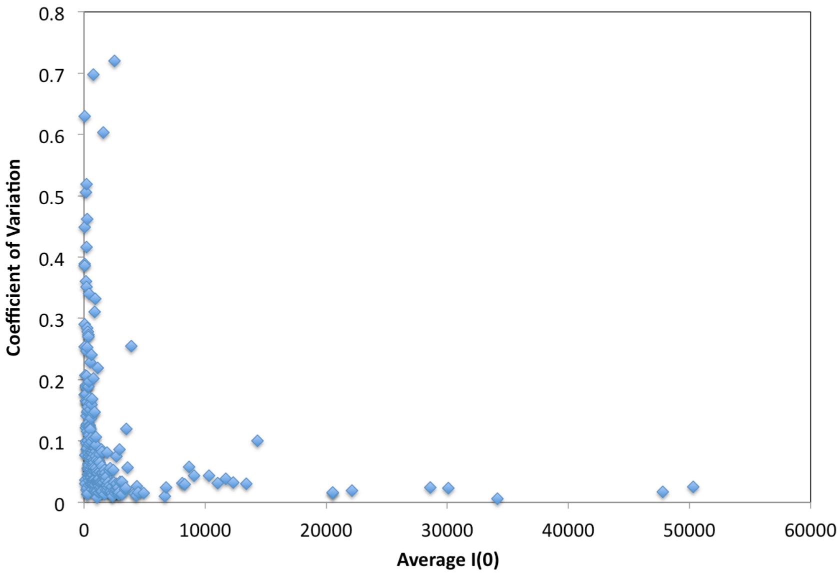


Figure 3.24 Coefficient of Variation of R_g as a Function of Signal-to-noise. A strong correlation exists between c_v and *I(0)*.

It is clear from Figure 3.24 that the coefficient of variation has a strong dependence on *I(0)*. 90% of concentrations with *I(0)* greater than 500 have a coefficient of variation of less than 0.1 and more than 90% of coefficients of variation greater than 0.1 have an *I(0)* less than 1000. There does appear to be a lower limit to the coefficient of variation, regardless of how high the signal-to-noise is. Since the vast majority of concentrations with high *I(0)*s have coefficients of variation that vary between 0 and 0.1, 0.1 be the lower limit of c_v.

Using this information about the relationship between the coefficient of variation and *I(0)* we can choose a cutoff value for c_v and determine what *I(0)* is required to have a desired likelihood of achieving it. Choosing a cutoff of 0.1 results in requiring an *I(0)* of at least 238 for an 80% likelihood of achieving a c_v less than 0.1, an *I(0)* of greater than 371 is required for a 85% likelihood, an *I(0)* of at least 491 is required for a 90% likelihood, and an *I(0)* of at least 1110 is required for a 95% likelihood. Knowing the minimum *I(0)* required to achieve a desired coefficient of variation is useful as it can be used to prepare a protein solution of sufficient concentration such that signal-to-noise is great enough. However, proteins of varying molecular weights will also impact the resulting *I(0)*, as will variations in instrumentation. To account for molecular weight and experimental set ups, we have used a solution of bovine serum albumin (BSA) as a standard for calibration, since it is well characterized and used as a model system for SAXS experiments. We performed the identical SAXS experiment using a 1 mg/mL solution of BSA. Knowing the molecular weight of BSA is 66,463 Da, and using equation 3.26, a plot of required concentration versus protein molecular weight is presented for achieving a coefficient of variation of less than 0.1 (Figure 3.25).

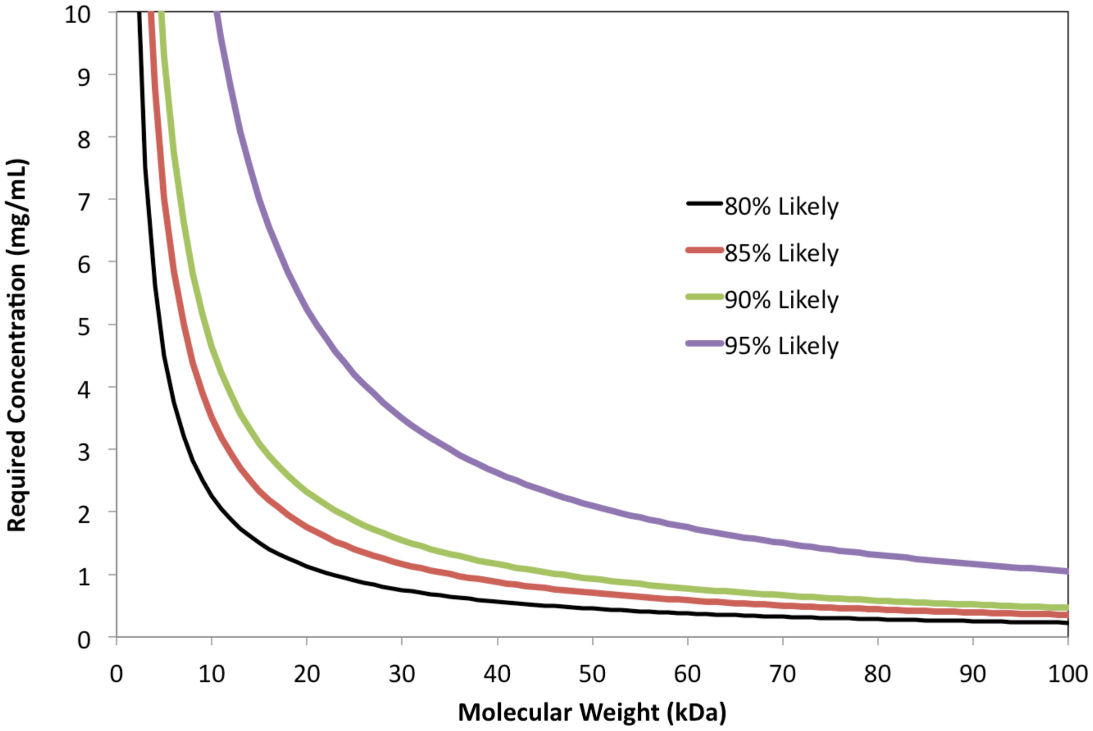


Figure 3.25 Required Protein Concentration for Achieving a c_v Less than 0.1. The curves represent the concentration required for an 80% likelihood (black), 85% likelihood (red), 90% likelihood (green) or 95% likelihood (purple) of achieving c_v less than 0.1 for the corresponding particle molecular weight.

3.4.4 Concentration Dependence Statistics

Now that the radiation damage portion of the SAXStats package has averaged all undamaged exposures to create one SAXS profile for each concentration, the three concentrations for each sample were then evaluated for concentration dependence. The concentration dependence script resulted in 82 samples proving successful for all statistics for all three concentrations. This is 6 samples fewer than that seen for the radiation damage script. These 6 samples were not detected as failures by the radiation damage script because they did not result in an R_g of zero. Upon further inspection it was discovered that each of these 6 samples had at least one concentration that suffered from air bubbles in the sample chamber, determined by characteristic 2D, anisotropic scattering seen in the raw images, which caused the scaling procedure to fail. For the remaining 12 samples the same statistics of success were seen as described for the radiation damage script. While a minimum of three concentrations is required to perform the linear regression t-test to detect concentration dependence, in cases where less than three concentrations were successful, the successful concentrations were averaged. In these cases, the user is cautioned that further experiments should be performed to ensure no concentration dependence, however the values for the average SAXS parameters are still reported.

The SAXS parameters that should remain the same independent of concentration are the R_g, D_{max}, and Porod molecular weight. Each of these parameters was tested for concentration dependence using the linear regression t-test. The total number of samples that showed concentration dependence for each parameter is shown in Figure 3.26. Since there are only three concentrations, and since interparticle interactions can cause distortions to the entire SAXS curve, two p values of less than 0.20 and 0.05 are reported. R_g calculated from the *P(r)* function showed the most samples suffering from concentration dependence, with 35 samples for p less than 0.20 and 12 samples for p less than 0.05, and D_{max} showed the least number of samples suffering from concentration dependence, with 20 samples for p less than 0.20 and 2 samples for p less than 0.05.

In total 55 of the 82 samples experienced concentration dependence in at least one SAXS parameter with p less than 0.20 and 25 samples with p less than 0.05 (Figure 3.27). For p less than 0.20, 20 samples saw concentration dependence in only one parameter, whereas 7 samples saw concentration in all four SAXS parameters. For p less than 0.05, 22 samples saw concentration dependence in only one parameter, while no samples saw concentration dependence for either 3 or 4 parameters. Only 27 of the 82 total samples did not experience any concentration dependence using either p value.

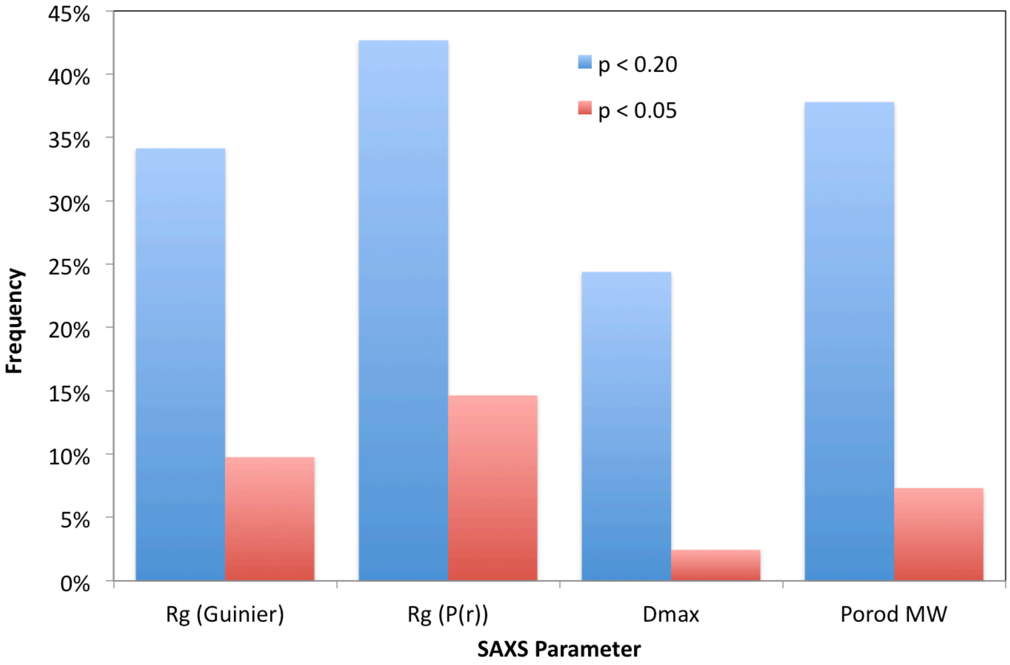


Figure 3.26 Frequency (Total = 82) of Concentration Dependence for Each Parameter. Concentration effects are seen in each parameter with varying frequency, with R_g calculated by *P(r)* demonstrating concentration dependence most frequently.

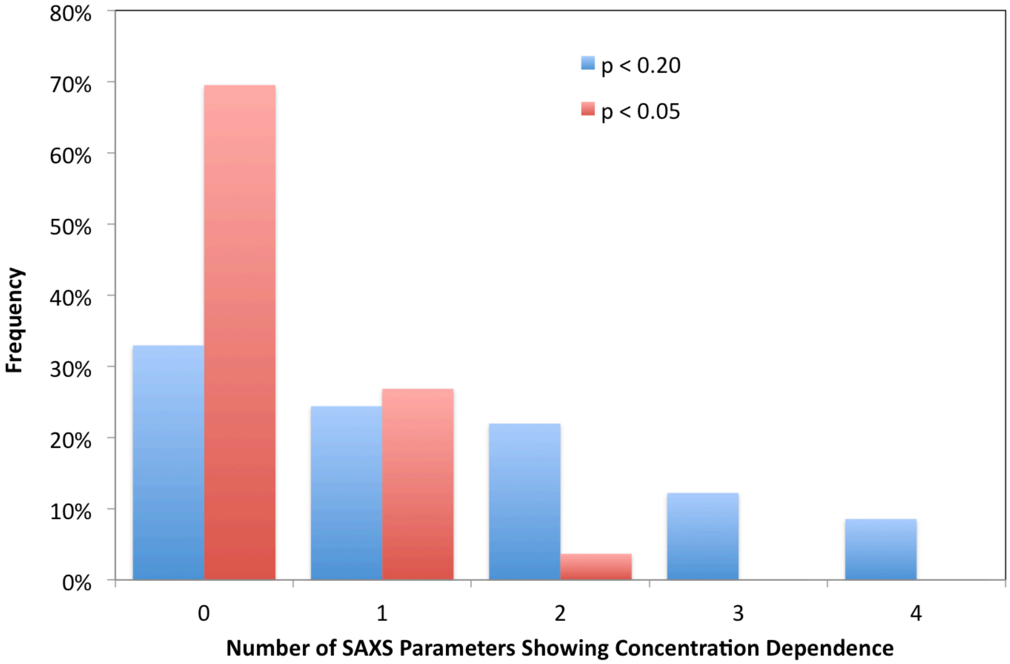


Figure 3.27 Frequency of Multiple SAXS Parameters Showing Concentration Dependence. Most samples did not show concentration dependence with a certainty of p<0.05, while significantly more samples demonstrated dependence with a certainty of p<0.20.

3.4.4.1 Interparticle Interaction Statistics

To test for interparticle interactions for each sample, curvature in the Guinier region for each concentration was assessed. Figure 3.28 shows the frequency that interparticle interactions occurred for both Guinier regions and for p<0.20 and 0.05. The Guinier region using the lower limit defined by D_{max}, with 49 concentrations experiencing interparticle interactions for p<0.20, proved to be less sensitive to curvature than the Guinier region without this lower limit, where 106 concentrations experienced interparticle interactions. Attractive interactions resulting in aggregation were more common than repulsive interactions. Figure 3.29 shows the distribution of the number of concentrations showing interparticle interactions for each sample for p<0.20. Using the most stringent requirements for sample quality, including a Guinier region of q < 1.3/R_g, no interparticle interactions for any concentration, and no concentration dependence with p<0.20, only 5 of the 82 samples would meet all requirements.

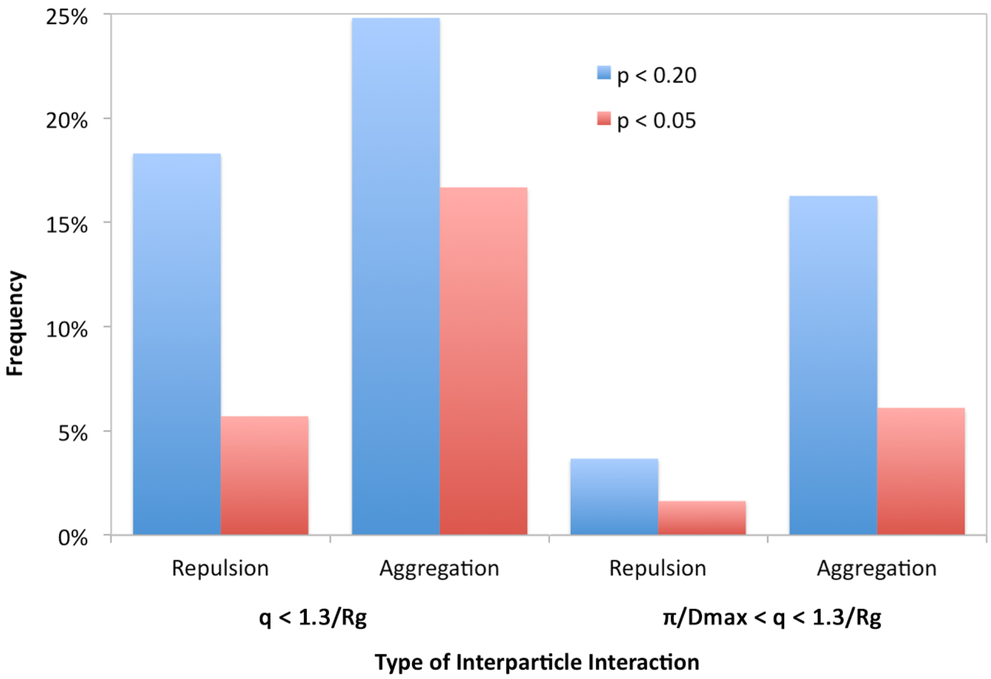


Figure 3.28 Frequency (Total = 246) of Interparticle Interactions. The Guinier region truncated by π/*D*$_{max}$ is less sensitive to curvature than the full Guinier region.

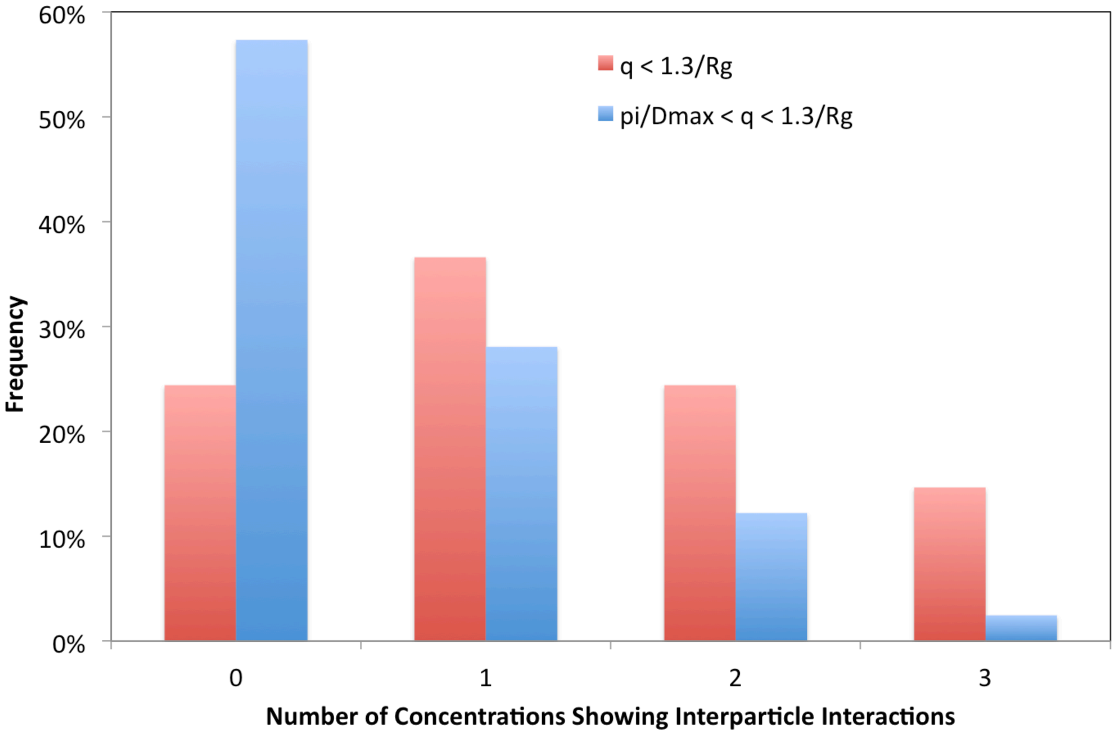


Figure 3.29 Concentrations Per Sample (Total = 82) Showing Interparticle Interactions. 75% of samples show interparticle interactions in at least one concentration using a Guinier region defined by q < 1.3/R_g, while more than 85% of samples have at least one concentration not showing interparticle interactions with the same Guinier region.

3.4.4.2 Severity of Concentration Dependence

While only 5 samples would meet all of the most stringent requirements for sample quality, it is important to know the degree of impact that concentration dependence has on the measurement of SAXS parameters. After estimating concentrations using BSA as a calibration standard, the slope of the linear regression for each parameter is used for determining the change in the parameter as a function of concentration. Figure 3.30 through Figure 3.33 show the impact that concentration has on the measurement of each parameter. For most of the samples, concentration changes R_g by less than 1 Å per mg/mL, however there are several samples that suffer from dependencies greater than 3 Å per mg/mL, which may distort conclusions about the particle depending on the question sought. For D_{max} greater changes are seen with a significant number of samples showing slopes greater than 5 Å per mg/mL, however most samples show slopes of less than 5 Å per mg/mL. Molecular weight shows very small changes as a function of concentration, with more than 70% of samples changing by less than 1 kDa per mg/mL.

From this analysis it is clear that while very few samples meet all of the most stringent requirements for sample quality, since many samples do not change the measured SAXS parameter by significant amounts, the stringent requirements may be relaxed and still result in drawing accurate conclusions, provided that the solution conditions are dilute enough. In particular, if oligomeric state is sought out using SAXS, which is one of the most common applications of the technique, it has been shown that under dilute enough conditions the effects of concentration dependence are unlikely to change the apparent oligomeric state with few exceptions. However, it should be noted that 30% of the 246 concentrations exceeded 5 mg/mL, which may cause significant distortion to SAXS parameters for many samples.

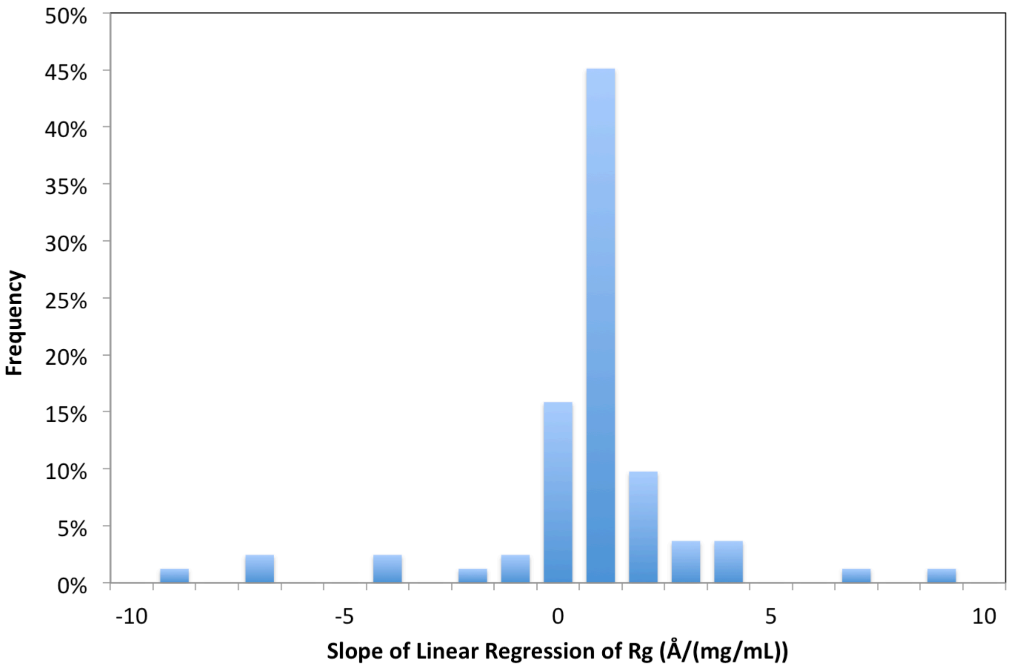


Figure 3.30 Distribution of the Slope of R_g Calculated using Guinier Method. More than half of all samples show less than a 1 Å/mg/mL trend in R_g.

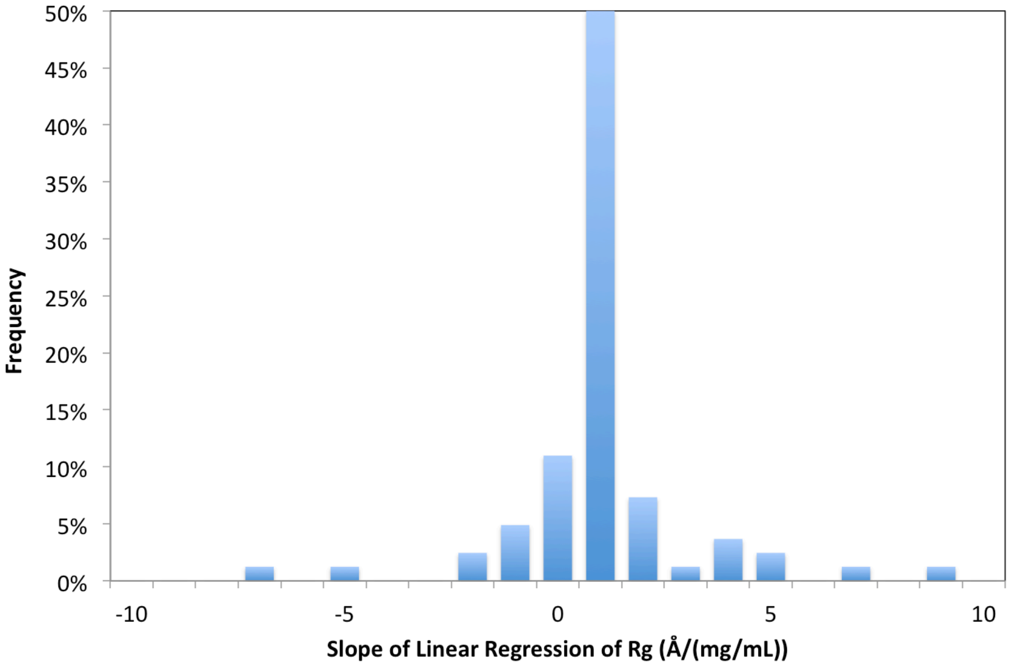


Figure 3.31 Distribution of the Slope of Rg Calculated using *P(r).* More than half of all samples show less than a 1 Å/mg/mL trend in *R_g*.

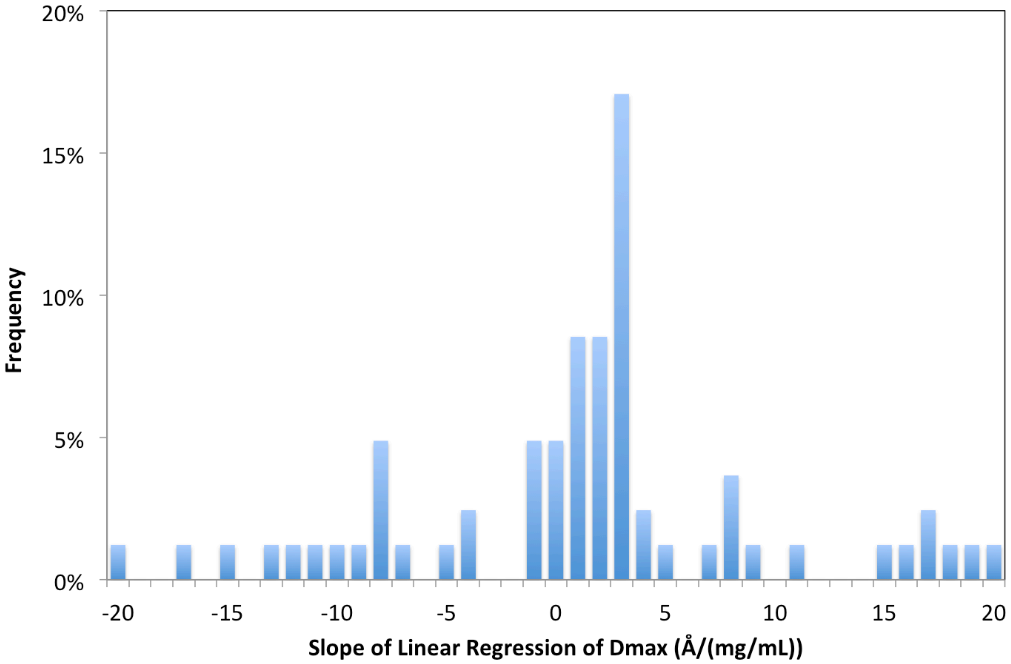


Figure 3.32 Distribution of the Slope of Dmax. A wide range of trends is observed with a significant number of samples showing trends greater 5 Å/mg/mL.

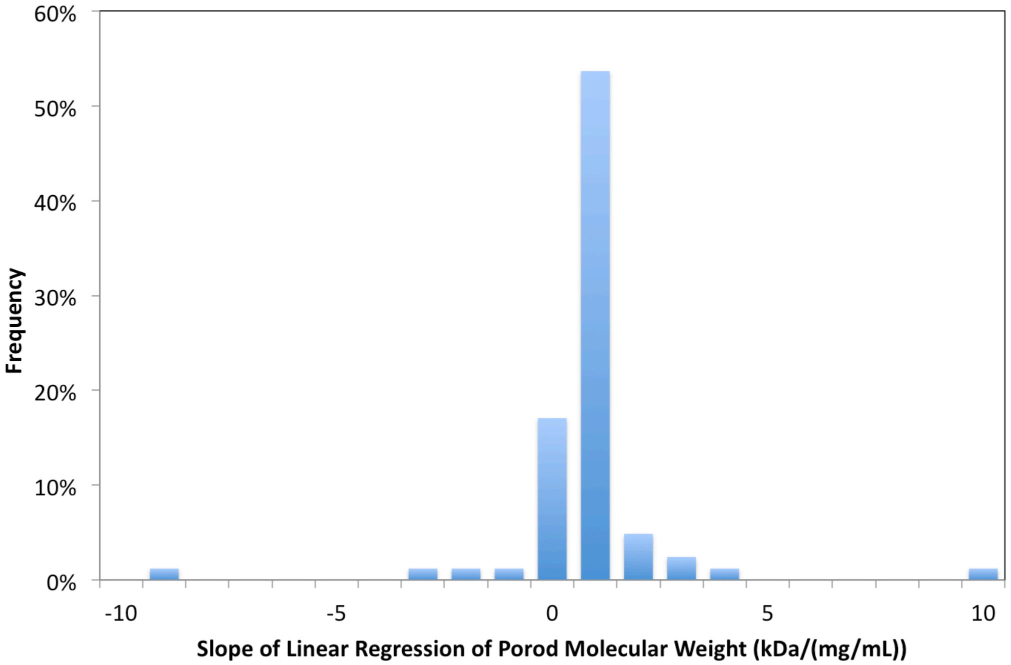


Figure 3.33 Distribution of the Slope of Porod Molecular Weight. Molecular weight changes very little in most samples, with the majority affected by less than 1 Å/mg/mL.

3.4.4.3 Second Virial Coefficients

For the 82 samples for which all statistics were determined for each of the three concentrations, second virial coefficients (B_{22}) were calculated using both the Guinier estimate of *I(0)* and the *P(r)* estimate of *I(0)*. The distributions of each are shown in Figure 3.34 and Figure 3.35. In addition to the currently studied samples, the distribution of B_{22} values of proteins used in the study determining the "crystallization slot" are also shown. The distribution of B_{22} calculated using the *P(r)* estimate of *I(0)* appears to be more disperse than that calculated using the Guinier approximation of *I(0)*. In the current sample set of NESG targets, 50% of B_{22} values calculated using the Guinier approximation are inside the boundaries of the crystallization slot and therefore may be more likely to crystallize than other samples.

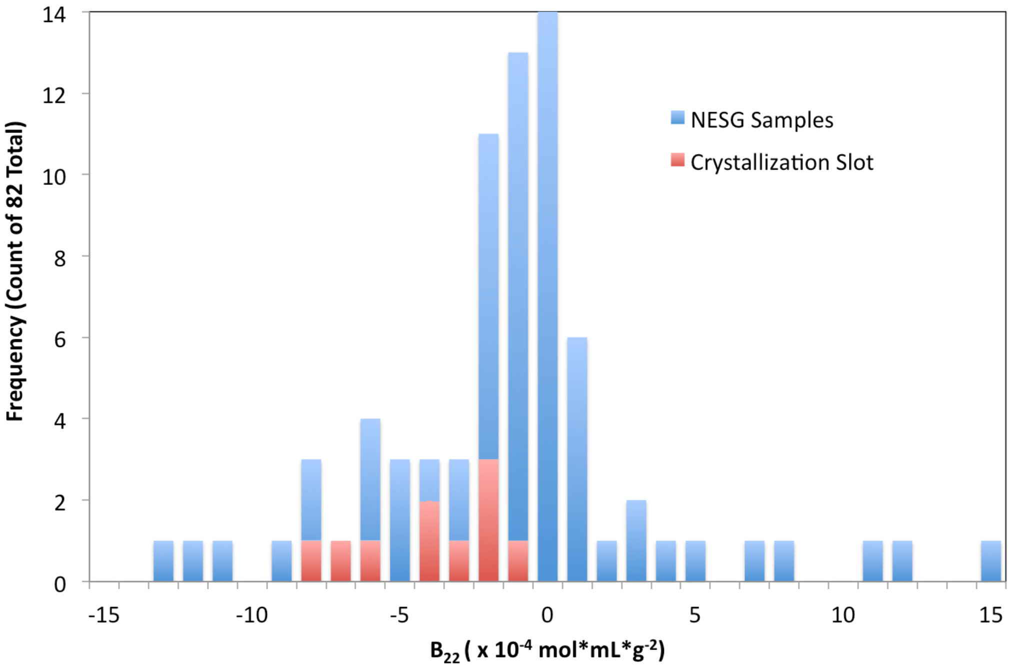


Figure 3.34 Distribution of B_{22} Calculated Using the Guinier Method. Half of the samples show B_{22} values that exist inside the so-called "crystallization slot".

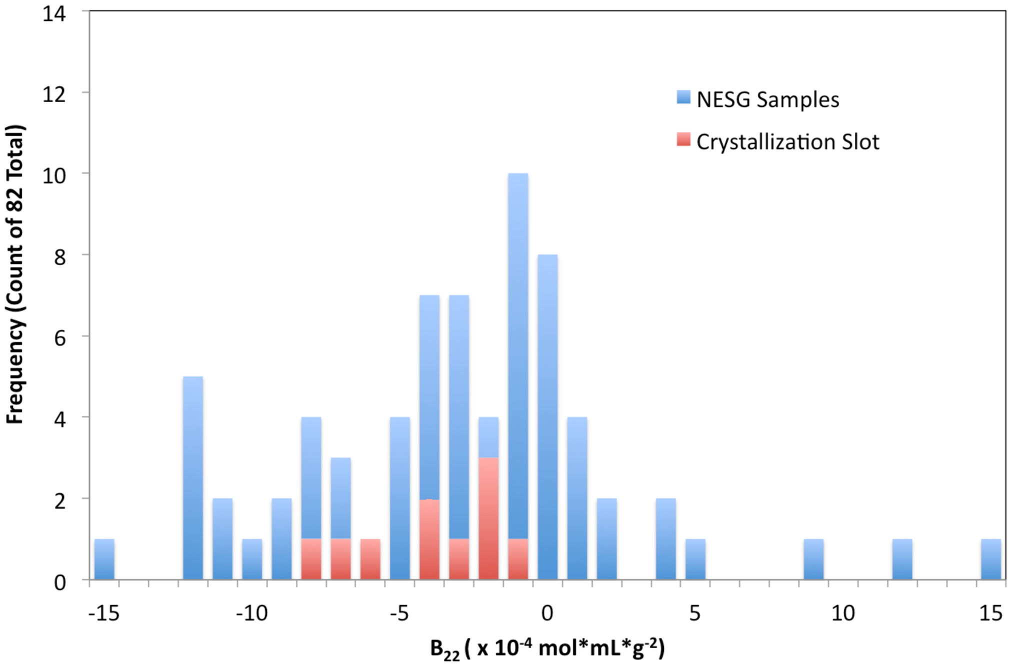


Figure 3.35 Distribution of B_{22} Calculated Using *P(r)*. The distribution is wider than observed for that calculated using the Guinier region.

3.4.5 Summary of Results

Using SAXStats, the amount of time used to process all eight exposures of three concentrations of 100 samples, a total of 2400 individual SAXS profiles, has been greatly reduced compared with that required for manual analysis. For an expert user, the time required to extract basic SAXS parameters from each of these scattering curves would require a minimum of approximately two to three minutes. Additional time required for plotting and analyzing parameters as a function of radiation and concentration would put an equivalent evaluation as that presented here at well over 100 hours. In comparison the computational time required for SAXStats to be run on the 100 samples studied using a single 2.53 GHz Intel processor was only 2 hours and 14 minutes, resulting in an average time of less than 90 seconds per sample. Such a large reduction in processing time makes high-throughput, real time SAXS analysis possible. In future iterations, rewriting SAXStats using more efficient programming techniques may further reduce the processing time required.

SAXStats performed statistical analyses on 284 SAXS profiles with varying degrees of signal-to-noise. By analyzing the coefficients of variation produced for all of these samples, we were able to present a chart of concentration that would be required to obtain a signal-to-noise ratio sufficient for producing highly precise SAXS parameters for proteins of varying molecular weight. This knowledge is extremely useful to SAXS scientists everywhere as it provides a guideline for preparing protein solutions prior to performing SAXS experiments by directing how high of a concentration they must obtain to determine SAXS parameters with sufficient precision. Additionally, in cases where protein characteristics prevent the production of sufficient protein concentration, possibly due to limitations presented by protein aggregation or precipitation, scientists can use this knowledge, coupled with protein standards for calibration for particular SAXS instrumentation setups, to estimate the minimum exposure time needed to obtain the required signal-to-noise for dilute protein solutions. As such, this information is a useful tool for virtually all scientists interested in performing SAXS experiments.

Currently SAXStats is being installed for use at SSRL BL 4-2 for all users to work in conjunction with their own automated image integration software. This set up provides a great advantage to scientists with the advent of near instantaneous sample quality feedback. Since gaining access to synchrotron resources is very difficult, and beam time applications are highly competitive, it is important that beam time shifts are as successful as possible. In many cases timely manual analysis of SAXS data cannot be performed before the beam time shift ends, and thus many samples are not identified as being of poor quality until the user arrives home and analyzes the data and realizes that further experiments must be performed under different conditions to promote a high quality, monodisperse solution. With near instantaneous feedback regarding sample quality, a user can be notified immediately when a sample requires solution conditions more conducive to monodispersity. With this information the user can then attempt to alter solution conditions such as salt concentration or protein concentration and repeat the SAXS experiment to obtain useable data. We have employed this procedure at BL 4-2 in a preliminary testing mode and it has worked well to alert us to problematic sample conditions requiring more dilute protein concentrations. Therefore, SAXStats has the potential to increase the success rate of SAXS experiments utilizing valuable X-ray resources, and subsequently the number of scientific results produced using these resources.

One of the greatest advantages of the SAXStats package is its use of the linear regression statistical analysis to identify radiation damage, concentration dependence, and interparticle interactions. Since SAXS data analysis does not have data quality measurements, such as an R_{free} as than in protein crystallography, historically much of

SAXS data analysis has been performed “by eye” in a highly subjective nature. In particular, linearity of the Guinier region can be difficult to assess and different users may disagree on the analysis. The statistical analysis performed by SAXStats provides an objective analysis on data quality. This is particularly useful considering the growing user community of SAXS resulting in the ability to accurately characterize sample quality for new users with little experience in the technique. Additionally, the ability to quantitatively evaluate data will give scientists the ability to compare data with other published works, placing SAXS data on equal footing independent of the experiment and user expertise.

3.5 References

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res *29*, 2884-2898.

Bonnete, F., and Vivares, D. (2002). Interest of the normalized second virial coefficient and interaction potentials for crystallizing large macromolecules. Acta crystallographica Section D, Biological crystallography *58*, 1571-1575.

Carter, P., Liu, J., and Rost, B. (2003). PEP: Predictions for Entire Proteomes. Nucleic Acids Res *31*, 410-413.

Davies, K.J., and Delsignore, M.E. (1987). Protein damage and degradation by oxygen radicals. III. Modification of secondary and tertiary structure. J Biol Chem *262*, 9908-9913.

Garrison, W.M. (1987). Reaction mechanisms in the radiolysis of peptides, polypeptides, and proteins. Chemical Reviews *87*, 381-398.

George, A., and Wilson, W.W. (1994). Predicting protein crystallization from a dilute solution property. Acta crystallographica Section D, Biological crystallography *50*, 361-365.

Glatter, O., and Kratky, O. (1982). Small angle x-ray scattering (London ; New York :, Academic Press).

Goh, C.S., Lan, N., Echols, N., Douglas, S.M., Milburn, D., Bertone, P., Xiao, R., Ma, L.C., Zheng, D., Wunderlich, Z.*, et al.* (2003). SPINE 2: a system for collaborative structural proteomics within a federated database framework. Nucleic Acids Res *31*, 2833-2838.

Goulden, C.H. (1956). Methods of Statistical Analysis, 2nd edn (New York, Wiley).

Grossmann, J.G. (2007). Biological solution scattering: recent achievements and future challenges. J Appl Crystallogr *40*, s217-s222.

Guinier, A., and Foumet, F. (1955). Small Angle Scattering of X-rays (New York, Wiley Interscience).

Jacques, D.A., and Trewhella, J. (2010). Small-angle scattering for structural biology--expanding the frontier while avoiding the pitfalls. Protein Sci *19*, 642-657.

Kenney, J.F., Keeping, E.S. (1962). Mathematics of Statistics, Vol 1, 3rd edn (Princeton, N.J., Van Nostrand).

Le Maire, M., Thauvette, L., de Foresta, B., Viel, A., Beauregard, G., and Potier, M. (1990). Effects of ionizing radiations on proteins. Evidence of non-random fragmentations and a caution in the use of the method for determination of molecular mass. Biochem J *267*, 431-439.

Li, Z., Li, D., Wu, Z., Wu, Z., and Liu, J. (2012). Optimization of a three slit collimation system for a SAXS camera with a divergent beam. Journal of X-Ray Science & Technology *20*, 331-338.

Liu, J., Hegyi, H., Acton, T.B., Montelione, G.T., and Rost, B. (2004). Automatic target selection for structural genomics on eukaryotes. Proteins *56*, 188-200.

Liu, J., and Rost, B. (2004). CHOP proteins into structural domain-like fragments. Proteins *55*, 678-688.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol *142*, 170-179.

Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D.T., Konarev, P.V., and Svergun, D.I. (2012). New developments in the ATSAS program package for small-angle scattering data analysis. J Appl Crystallogr *45*, 342-350.

Pollack, L. (2011). SAXS studies of ion-nucleic acid interactions. Annu Rev Biophys *40*, 225-242.

Porod, G. (1951). Die Röntgenkleinwinkelstreuung von dichtgepackten kolloiden Systemen. Colloid & Polymer Science *124*, 83-114.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys *40*, 191-285.

Rambo, R.P., and Tainer, J.A. (2011). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers *95*, 559-571.

Receveur-Brechot, V., and Durand, D. (2012). How random are intrinsically disordered proteins? A small angle scattering perspective. Curr Protein Pept Sci *13*, 55-75.

Semenyuk, A.V., and Svergun, D.I. (1991). GNOM - a program package for small-angle scattering data processing. J Appl Crystallogr *24*, 537-540.

Smolsky, I.L., Liu, P., Niebuhr, M., Ito, K., Weiss, T.M., and Tsuruta, H. (2007). Biological small-angle x-ray scattering facility at the Stanford synchrotron radiation laboratory. J Appl Crystallogr *40*, S453-S458.

Svergun, D.I., and Koch, M.H.J. (2003). Small-angle scattering studies of biological macromolecules in solution. Reports on Progress in Physics *66*, 1735.

Wignall, G.D., Lin, J.S., and Spooner, S. (1990). Reduction of parasitic scattering in small-angle X-ray scattering by a three-pinhole collimating system. J Appl Crystallogr *23*, 241-245.

Wunderlich, Z., Acton, T.B., Liu, J., Kornhaber, G., Everett, J., Carter, P., Lan, N., Echols, N., Gerstein, M., Rost, B*., et al.* (2004). The protein target list of the Northeast Structural Genomics Consortium. Proteins *56*, 181-187.

Xiao, R., Anderson, S., Aramini, J., Belote, R., Buchwald, W.A., Ciccosanti, C., Conover, K., Everett, J.K., Hamilton, K., Huang, Y.J*., et al.* (2010). The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. Journal of structural biology *172*, 21-33.

4 Structural Conservation of an Ancient tRNA Sensor in Eukaryotic Glutaminyl-tRNA Synthetase

4.1 Introduction

In chapter 2 it was demonstrated that with SAXS data of sufficient quality, as determined by manual analysis, parameters extracted from the scattering profile such as radius of gyration, maximum particle dimension, oligomeric state, and even low resolution molecular envelopes can be obtained with high fidelity in a high-throughput pipeline, substantiated by high-resolution structural data. SAXS data was discussed as being most useful when applied as a complement to other known structural or biochemical data to gain a more complete understanding of a biological system. Chapter 3 described and implemented a statistical methodology, SAXStats, by which SAXS data can be assessed. Using SAXStats yielded the ability to objectively evaluate the quality of the SAXS data and quickly obtain several parameters of interest in a high-throughput fashion. The remaining chapters have not used SAXStats in a high-throughput pipeline, but instead used it for data quality analysis and to obtain SAXS parameters for a specific biological system. To showcase SAXS as a complementary tool in the structural biologist's toolkit, SAXStats has been used in conjunction with several other structural, biochemical, and bioinformatics tools to generate a more complete understanding of the eukaryotic glutaminyl-tRNA synthetase, Gln4.

Gln4 is an aminoacyl-tRNA synthetase. Aminoacyl tRNA synthetases perform a critical function in conversion of the genetic code into amino acids by covalently attaching the correct amino acid to specific cognate tRNAs (Guo et al., 2010; Mirande, 2010). Since the ribosome has no mechanism of ensuring that the correct aminoacid has been attached to its cognate tRNA, tRNA synthetases act as the "codebook" for the genetic code, by associating an amino acid with its corresponding codon. These enzymes are divided into two structural classes, each arising from a common ancestor (Cusack et al., 1990; Eriani et al., 1990), and catalyze aminoacyl-tRNA formation by a two-step pathway: (i) an activated aminoacyl adenylate is first formed from ATP and the cognate amino acid; (ii) the amino acid is transferred to its cognate tRNA with release of AMP. Each synthetase nearly perfectly selects the correct tRNA among 20-22 different isoacceptor tRNA families (Marck and Grosjean, 2002) as well as the correct amino acid substrate; in some cases, this is achieved via the use of hydrolytic editing mechanisms to clear misactivated amino acid and/or misacylated tRNA (Cusack et al., 1990; Eriani et al., 1990). It is of particular interest that tRNA^{Gln} and tRNA^{Asn} are aminoacylated by distinct mechanisms in different kingdoms. For example, whereas Gln-tRNA^{Gln} is formed in the canonical manner in the eukaryotic cytoplasm, all archaea, many bacteria, and eukaryotic organelles possess an alternative two-step pathway. In this route, a nondiscriminating GluRS first misaminoacylates tRNA^{Gln}; next, the Glu-tRNA^{Gln} is converted to Gln-tRNA^{Gln} by a tRNA-dependent amidotransferase belonging to either the GatCAB family (bacteria and some archaea), or the GatDE family (archaea only) (Curnow et al., 1997; Ibba and Soll, 2004; Tumbula et al., 2000). Thus, glutaminyl-tRNA synthetase (GlnRS) is primarily a eukaryotic enzyme. Synthesis of cysteinyl-tRNA^{Cys} in methanogens and highly related archaea provides another example of a two-step pathway to cognate aminoacyl-tRNA, although the phylogenetic distribution of this pathway is much more limited (Sauerwald et al., 2005).

Eukaryotic tRNA synthetases are distinctly more complex than their prokaryotic homologs because they have progressively acquired and retained additional domains throughout evolution (Guo et al., 2010; Mirande, 2010). It is perplexing why tRNA synthetases, unlike other eukaryotic proteins, have been subject to massive progressive additions over the course of evolution (Guo et al., 2010). While some appended domains are shared among synthetase families and are similar to domains in other proteins implicated in either nucleic acid binding or protein-protein interactions (see (Mirande, 2010), at least eight domains are uniquely associated with a single synthetase family, and neither their structures nor their roles are generally understood (Guo et al., 2010). An exception is the C-terminal domain (CTD) of human CysRS, which is known to enhance anticodon discrimination at the expense of the aminoacylation rate, acting as a quality control step (Liu et al., 2007). This report focuses on the N-terminal domain (NTD) of GlnRS, which is itself unique because GlnRS likely originated in eukaryotes, evolving directly from a progenitor eukaryotic non-discriminating GluRS (Lamour et al., 1994; Nureki et al., 2010). Like other eukaryotic GlnRS species, *Saccharomyces cerevisiae* Gln4 contains both a highly conserved CTD with all of the known features of class I synthetases, as well as a less conserved appended NTD with no obvious sequence homology to any known protein domain.

The origin and function of the NTD in GlnRS are of particular interest. Most eukaryotic GlnRS proteins have an appended NTD, whereas the bacterial GlnRS proteins do not, although the bacterial proteins were almost certainly acquired by horizontal transfer from eukaryotes. *S. cerevisiae* GlnRS contains both a 595 amino acid CTD that contains the signature elements of a type I synthetase (Deniziak et al., 2007; Eriani et al., 1990; Ludmerer and Schimmel, 1987b; Rould et al., 1989), and suffices for both catalytic function and yeast viability (Ludmerer and Schimmel, 1987a; Ludmerer et al., 1993), and a 224 amino acid NTD that is uniquely associated with GlnRS in many eukaryotes (Guo et al., 2010). Although both *E. coli* and *D. radiodurans* GlnRS proteins share extensive identity with the conserved *S. cerevisiae* GlnRS CTD, *E. coli* GlnRS entirely lacks an NTD (Ludmerer and Schimmel, 1987b) and *D. radiodurans* GlnRS has an unrelated domain appended to the C-terminus of the conserved domain (Deniziak et al., 2007). Two observations imply that the *S. cerevisiae* NTD contributes to synthetase function: the NTD alone exhibits a non-specific RNA binding activity (Wang and Schimmel, 1999), and the addition of the NTD to *Ec*GlnRS results in a chimeric protein that can replace the native yeast gene (Whelihan and Schimmel, 1997). However, the precise role of the NTD in eukaryotic GlnRS function is unknown.

In this study we report a functional and structural analysis of the NTD of *S. cerevisiae* GlnRS, Gln4. Yeast mutants lacking the NTD exhibit growth defects, and Gln4 lacking the NTD has reduced complementarity for tRNA^{Gln} and glutamine. The 187 amino acid Gln4 NTD, crystallized and solved at 2.3 Å resolution, consists of two subdomains, each exhibiting an extraordinary structural resemblance to adjacent tRNA specificity-determining domains in the GatB subunit of the GatCAB amidotransferase, which forms Gln-tRNA^{Gln}. These subdomains are connected by an apparent hinge comprised of conserved residues. Mutation of these amino acids produces Gln4 variants with reduced affinity for tRNA^{Gln}, consistent with a hinge-closing mechanism proposed for GatB recognition of tRNA. Our results suggest a possible origin and function of the NTD that would link the phylogenetically diverse mechanisms of Gln-tRNA^{Gln} synthesis. Portions of this research have been published in (Grant et al., 2012). In this chapter, I performed crystallization experiments, crystal extraction, X-ray data collection, structure determination, SAXS data collection and analysis, and some bioinformatics analysis. The laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester performed genetic analysis of Gln4 mutants, protein expression and purification, and tRNA purification and EMSA binding assays while steady state analysis was performed by the laboratory of Dr. John Perona at the University of California at Santa Barbara.

4.2 Methods

4.2.1 Genetic analysis of *gln4* mutants

To construct a strain (MEM70) of genotype *gln4-Δ::kan*^{R} [*CEN URA3 GLN4*], a *CEN GLN4* plasmid was transformed into yeast strain BY4741, and then the *gln4-ΔKan* allele was introduced by transformation, using PCR primers HWI P239 and HWI P234 (Table 4.1) to amplify the fragment from the appropriate *GLN4*/ *gln4-ΔKan* heterozygous diploid (Open Biosystems ID 22424). To construct strains bearing an integrated copy of either *GLN4* or *gln4-Δ2-210*, we used an integrating cassette (Whipple et al., 2011) that carries *MET15* flanked by sequences homologous to *ADE2*, into which we inserted *GLN4* or the *gln4(211-809)* allele (constructed with a synthetic fragment made by Geneart). Plasmids were then digested with Stu I to release the integrating cassette and transformed into MEM70, and transformants were screened for Ade$^-$, and plated on FOA to select for removal of the *CEN URA3 GLN4* plasmid, generating the desired *gln4-Δ::kan*^{R} *ade2*$^-$::*GLN4*::*MET15* (MEM133) and *gln4-Δ::kan*^{R} *ade2*$^-$:: *gln4(211-809)*::*MET15* (MEM141) strains. To test for growth phenotypes, MEM133 and MEM141 were transformed with a [*CEN LEU2 GLN4*] or a control [*CEN LEU2*] vector, grown overnight in SD-Leu media (see (Sherman, 1986), diluted to OD_{600} of 1 and 2 μL of 10-fold serial dilutions were spotted onto plates containing either YPD or YP glycerol and incubated at the indicated temperatures for 1-7 days with similarly spotted control parent strains that were grown in YPD media. Oligonucleotides, yeast strains and plasmids used in these studies are reported in Table 4.1, Table 4.2, and Table 4.3. Genetic analysis of Gln4 mutants was performed by the laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester.

Table 4.1 Oligonucleotides for various Gln4 constructs

Oligonucleotide Name	Construct or Use	Sequence
QB832ADFP	Full length GLN4+ His6	AATTCCATCAACCTTAAA ATGGCTCACCATCACCATCACCAT ATGTCTTCTGTAGAAGAATTGACT
QB832ADRP	GLN4 to end	CTTCCAAACCACTCTTGGAAGTTGCGTCCTTCAA
QB1115ADFP	GLN4 187 – end	AATTCCATCAACCTTAAAATGATCAAGAAGAAGACCAAGAAT
QB1114ADFP	GLN4 211 – end	AATTCCATCAACCTTAAAATGTCTTCCGGTCCAAAGAGG
QB1034ADFP	GLN4 216 – end	AATTCCATCAACCTTAAAATGAGGACTATGTTCAATGAAGGTTTCC
QB1012ADFP	GLN4 1-187	AATTCCATCAACCTTAAAATGTCTTCTGTAGAAGAATTGACT
QB1012ADRP	GLN4 1-187	CTTCCAAACCACT GATTAAGTCTCTCTCATCCTT
HWI P239	gln4::Kan DNA	GAAGACATATATAAGAAACAAAAGGCAC
HWI P234	gln4::Kan DNA	GCCGTATATTGCTAAGGACACC
HWI P237	GLN4 -517	GGGTCCTGGTTCGAATCTGCAAATAAGCTTCCTAACGC
HWI P238	GLN4 +527	CTTGTTCGTGCTGTTTAAGAATCCCAGTGACAAGAATGACA
HWI P181	tRNA$^{Gln(CUG)}$	GGGTCCTGGTTCGAGAGCTTCTACTATAAACCTCACTC
HWI P182	tRNA$^{Gln(CUG)}$	CTTGTTCGTGCTGTTTACTATCAGGCTCTCAGAAGGC
HWI P235	tRNA$^{Gln(CUG)}$	TCTCCTGGATCCGGCTCGTATGTTGTGTGG
HWI P236	tRNA$^{Gln(CUG)}$	CCAGTGAATTCGAGCTCGGTAC
HWI P257	Purification of tRNA$^{Gln(CUG)}$	5`-Bio- TGGAGGTCCCACCCGGATTCGAACTGG 3`
HWI P237	JE1060AP and JE1062AP	GGGTCCTGGTTCGAATCTGCAAATAAGCTTCCTAACGC
HWI P238	JE1060Ap and JE1062AP	CTTGTTCGTGCTGTTTAAGAATCCCAGTGACAAGAATGACA
HWI P239	gln4-ΔKAN (-554)	GAAGACATATATAAGAAACAAAAGGCAC
HWI P234	gln4-ΔKAN (+586)	GCCGTATATTGCTAAGGACACC
HWI P40	Check gln4-ΔKAN (-790)	AGCAGCTGGAGCCACTAATGTTAG
P256	Check gln4-ΔKAN +811	CCAAATCCTAGCCCAAACTCTTCG
HWI P285	G-432 Xho1	CCTCCACTCGAG CCCTTCACGTTTCTGACAATAGTTCTG
HWI P288	G+288R Mlu1	CTCCAC ACGCGT CCATGTAGATTAACGTTATATTTTCCTTC

Table 4.2 Yeast strains used for Gln4 expression

Strain	Genotype	Source
BY4741	*MAT*a *his3Δ1 met15Δ0 leu2Δ0 ura3Δ0*	Open Biosystems
MEM 70	BY4741 *gln4-Δ::kan*^{R} [*CEN URA3 GLN4*]	This study
MEM 133	BY4741 *gln4-Δ::kan*^{R} *ade2*$^-$*::GLN4::MET15*	This study
MEM 141	BY4741 *gln4-Δ::kan*^{R} *ade2*$^-$*::gln4(211-809)::MET15*	This study
BCY123	*MAT*a, *pep4-3*::*HIS3*, *prb1*::*LEU2*, *bar1*::*HISG*, *lys2*::*GAL1/10-GAL4*, *can1*, *ade2*, *trp1*, *his3*, *ura3-52*, *leu2-3,112*	Mark Macbeth[23]
QB1012AD	BCY123 JE1012A [2 micron *URA3 GLN4*(1-187)-PT]	This study
EJG1117	*MATX leu2 trp1ura3 prb1-112 pep4-1 his3Δ-pGAL10-GAL4*	Erin O'Shea
EJG1473	EJG117 *sam1Δ::NatR sam2Δ::KanR*	This study
MA337	EJ1473 JE1012A [2 micron *URA3* GLN4(1-187)-PT]	This study

Table 4.3 Plasmids used in Gln4 study

Plasmid	Description	Source/ Reference
BG2483	2 micron, *URA3* vector with P$_{GAL1}$, LIC cloning sites, ORF is fused to PT tag containing a recognition site for protease 3C, HA epitope, His6, ZZ domain of protein A	Malkowski[24]
JE1012	*GLN*4(1-187) in BG2483	This study
JE1115	*GLN4*(187-809) in BG2483	This study
JE1033	*GLN4* (212-809) in BG2483	This study
JE1034	*GLN4* (216-809) in BG2483	This study
JE1135	*GLN4*-G$_{112}$A V$_{113}$A G$_{114}$A in BG2483	This study
JE1136	*GLN4*-G$_{112}$P in BG2483	This study
JE1137	*GLN4*-W$_{160}$A in BG2483	This study
JE1140	*GLN4*-W$_{160}$F in BG2483	This study
PYEX4T	2 micron, *URA3 leu2D*	Martzen[25]
JE1135	tRNA Glu-CUC (AB209-11) in PYEX4T	This study
JE1060	GLN4 +/- 500 in AVA579[URA3 CEN]	This study
AVA579	URA3 CEN plasmid derived from YCPlac33 by insertion of LIC cloning site	

4.2.2 Protein expression and purification

To express high levels of *GLN4* and its derivatives in yeast, ORFs were cloned under P_{GAL1} control into the previously described 2 μ *URA3* LIC vectors BG2483 or BG2663, in which ORFs are expressed with their C termini fused to a complex tag containing a 3C protease site, followed by an HA epitope, His_6, and the ZZ domain of protein A (Quartley et al., 2009), and expressed in yeast strain BCY123 (Macbeth et al., 2004). Gln4(1-187) was expressed in yeast strain EJG1473, which was grown in media containing selenomethionine and Ado-Methionine as described (Malkowski et al., 2007). Expressed proteins were purified by affinity purification on IgG sepharose, removal of GST-3C protease, concentration of samples and sizing on SuperdexHiLoad 1660 (GE Healthcare 17-1069, 10 x 300 mm bed dimension), as described (Quartley et al., 2009). Protein expression and purification was performed by the laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester.

4.2.3 tRNA purification and EMSA binding assay

To obtain native yeast $tRNA^{Gln(CUG)}$, we cloned the *tQ(CUG)M* gene into the *leu2-d URA3* vector pYEX4T (Martzen et al., 1999), transformed the plasmid into BY4741, grew transformants in SD-Ura media overnight, followed by overnight growth in SD-Leu-Ura media. We then prepared low molecular weight RNA, purified the $tRNA^{Gln}$ with biotinylated oligonucleotides oligo HWI P257 (Table 4.1), and performed HPLC analysis of modified nucleotides as described (Jackman et al., 2003). The ratio of modified to unmodified nucleotides was similar to that in strains with $tRNA^{Gln}$ on a lower copy plasmid.

tRNA binding was measured, as described (Wilkinson et al., 2007) in reaction mixtures containing Gln4 or its buffer, 2.4 nM 5'-[^{32}P]-labeled tRNA, in buffer containing 28 mM HEPES (pH 7.5), 80 mM NaCl, 5 mM $MgCl_2$, 0.5 mM DTT, 2.5 mM spermidine, 50 μg/ml BSA, 20μM EDTA, 200 μg/ml polyA, 4.6 mM Tris-Cl (pH 7.5), 1 mM β-mercaptoethanol, and 10 % glycerol. Reactions were incubated for 20 min on ice and loaded onto prerun 5% polyacrylamide gels containing 50 mM Tris Borate, pH 8.3, 1 mM EDTA, 5 mM $MgCl_2$ and 5 % glycerol, and run at 4°C in the same buffer without glycerol. tRNA purification and EMSA binding assays were performed by the laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester.

4.2.4 *In vitro* synthesis of tRNA transcripts

Duplex DNA templates for *in vitro* transcription of yeast $tRNA^{Gln}$ were synthesized from two single-stranded oligodeoxynucleotides containing a complementary overlap duplex region, as described (Sherlin et al., 2001). The two 3'-terminal deoxynucleotides on the noncoding strand incorporated 2'-O-methyl sugars (mU and mG in the sequences), to improve the fidelity of transcription termination by T7 RNA polymerase. Milligram quantities of each tRNA were transcribed with the Del(172-173) variant of T7 RNA polymerase, as described (Lyakhov et al., 1997; Sherlin et al., 2001), and purified by denaturing polyacrylamide gel electrophoresis. tRNA was stored at 200 μM in 10 mM Tris (pH 8.0), 1 mM EDTA (TE buffer). *In vitro* synthesis of tRNA transcipts was performed by the laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester and Dr. John Perona at the University of California at Santa Barbara.

4.2.5 Steady State Methods

$tRNA^{Gln}$ transcripts were ^{32}P-labeled at the 3'-terminal internucleotide linkage using the exchange reaction of tRNA nucleotidyltransferase (Bullock et al., 2003; Ibba et al., 1996; Uter and Perona, 2004), and purified again by gel electrophoresis. Steady state kinetics of tRNA aminoacylation reactions were performed in a buffer consisting of

50 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, and 10 mM β-mercaptoethanol. tRNA was first refolded by heating to 85°C in TE buffer for 3 minutes, followed by addition of $MgCl_2$ to 10 mM and slow-cooling to ambient temperature. Two μL aliquots from the reactions were added to 5 μL of a quenching solution containing 400 mM sodium acetate (pH 5.2) and 0.1% SDS, followed by addition of 3-5 mL of 0.01 – 0.1 mg/mL P1 nuclease (Fluka) to digest the tRNA to 5'-phosphorylated nucleosides. The digestion products were spotted on PEI-cellulose thin layer chromatography (TLC) plates and developed in a solution containing 100 mM ammonium acetate and 5% acetic acid. Raw data were quantified by phosphorimaging analysis, and corrected intensities were analyzed to obtain initial velocities. K_M and V_{max} were then obtained by Michaelis-Menten analysis. 5 mM ATP was used in all reactions; saturation was confirmed for both FL-GlnRS and the NTD variant. The glutamine concentrations used to determine $K_M^{(tRNA)}$ for FL-GlnRS and Gln4(187-809) were 10 mM and 60 mM, respectively; saturation was verified in each case. tRNA concentrations used were 20 nM – 3 μM for FL-GlnRS and 500 nM – 20 μM for Gln4(187-809). To determine K_M for glutamine, the tRNA concentrations used were 1 μM for FL-GlnRS and 15 μM for Gln4(187-809). Enzyme concentrations were maintained at least 20-fold below tRNA concentrations for all experiments to ensure multiple-turnover conditions. Steady state analysis was performed by the laboratory of Dr. John Perona at the University of California at Santa Barbara.

4.2.6 Crystallization and Structure Determination

Initial crystallization conditions were identified using a high-throughput microbatch-under-oil method (Luft et al., 2003). Crystals appeared after six-week's incubation at 22°C in conditions containing 0.2 µL protein solution (8.9 mg/mL protein in 100 mM NaCl, 5% (v/v) glycerol, 2 mM DTT, 0.025% (w/v) NaN_3, 20 mM HEPES buffer, pH 7.5) and 0.2 mL of precipitant solution (100 mM KCl, 100 mM Tris, pH 8 and 20%

(w/v) PEG 4000). Multiple attempts at crystallization optimization using vapor diffusion and batch under oil techniques at various temperatures (Luft et al., 2007) failed to produce larger crystals. Therefore, crystals were extracted directly from the 1536-well crystallization screening plate. To perform the extraction, a custom apparatus was designed. The crystallization plate was placed on an x-y stage with a scope and light source placed beneath. A syringe fitted with a 0.025 mm thin walled capillary with a diameter of 0.5 mm was then inserted into the well using fine adjustment gears attached to the syringe. After carefully placing the capillary directly over the crystal, the syringe plunger was raised to extract the crystal from the well (Figure 4.1). The contents of the capillary, including the protein crystal, were then expelled directly onto the loop used for crystal mounting and immediately flash frozen in liquid nitrogen.

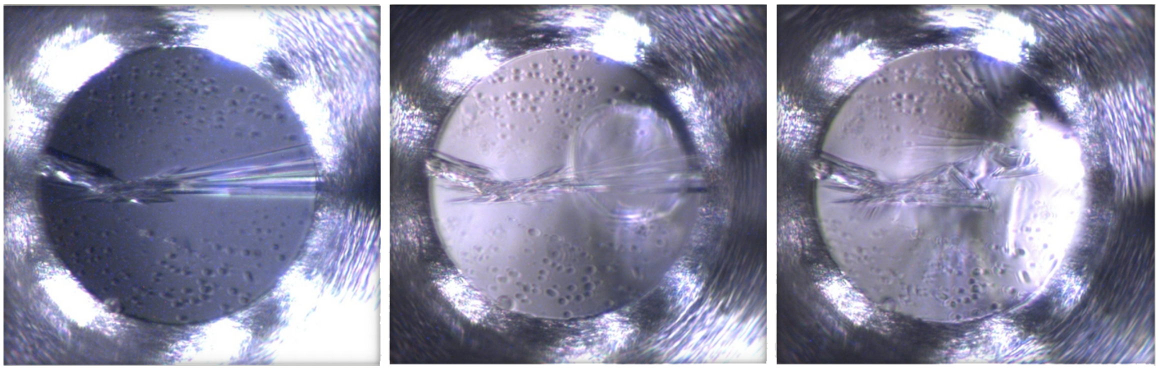


Figure 4.1 Crystal extraction from 1536-well crystallization screening plate. Three images show before (left), during (middle), and after (right) capillary insertion and crystal extraction. The NTD crystals can be seen spanning the horizontal width of the well. The capillary shown in the middle image can be seen as it impinges the crystal.

Remote MAD data collection was carried out at 100K on beamline 11-1 of the Stanford Synchrotron Radiation Lightsource (SSRL) (Soltis et al., 2008) with a MAR 325 CCD detector. To minimize radiation effects, the data collection protocol was designed with Best (Popov and Bourenkov, 2003) automated within the Web-Ice analysis package (Gonzalez et al., 2008). Integration, reduction and scaling took place with XDS (Kabsch, 2010). The structure was solved with Phenix (Adams et al., 2010). Using the remote wavelength data set the structure was refined through an iterative process using Phenix with manual model building with Coot (Emsley et al., 2010). Validation was carried out with Molprobity (Chen et al., 2010). The structure was deposited as PDB ID 3TL4. Experimental and refinement details are given in Table 4.5. Surface charge was calculated assuming vacuum electrostatics using PyMOL.

The sequences of several appended NTDs from GlnRS sequences of other organisms, listed in Table 4.6, were threaded to the Gln4(1-187) structure using SwissModel (Schwede et al., 2003). As a control the reversed sequence was also threaded. From the models a Z-score was calculated using Prosa2003 (Wiederstein and Sippl, 2007) with a 20-residue moving window. The typical combined, pairwise and surface Z-scores for native proteins are (-6 to -12), (-3 to -7.5), and (-3 to -8) respectively.

4.2.7 Small Angle X-ray Scattering

Small angle X-ray solution scattering data were collected on Beamline 4-2 of the SSRL (Smolsky et al., 2007). Data were collected from Gln4(1-187) at a wavelength of 1.3 Å for eight consecutive two-second exposures collected at four different concentrations ranging from 1.0 to 9.3 mg/mL. Data were collected from the flow-through buffer of the final purification column and subtracted from the total protein solution scattering. The data were integrated with SasTool (Smolsky et al., 2007). Analysis of eight consecutive time frames using SAXStats (Chapter 3) showed that radiation damage occurred in two to four exposures for three of the concentrations used. These exposures were removed from averaging. The SAXS data for different protein concentrations were investigated for aggregation and concentration dependence using SAXStats. No evidence of concentration dependence was seen when comparing all four concentrations and linearity in the Guinier region was conserved for each. CRYSOL (Svergun et al., 1995) was used to calculate the scattering profiles from crystal structures and fit them to the experimental scattering. DATGNOM was used to calculate the pair distribution function. Ten *ab initio* shape reconstructions were generated by DAMMIF (Franke and Svergun, 2009) and averaged with DAMAVER (Volkov and Svergun, 2003).

4.3 Results

4.3.1 Removal of the NTD impairs Gln4 function *in vivo* and *in vitro*

To determine if the NTD is important for the essential function of Gln4, we compared the growth of yeast strains expressing either full length *GLN4* or *gln4* lacking the NTD [*gln4*(211-809)] integrated into the chromosome under control of its own promoter, as the sole source of *Sc*GlnRS. Growth of the *gln4*(211-809) mutant is impaired at 14^{o}C and 19^{o}C, but not at 30^{o}C, on both YPD and YP glycerol media, and, as expected, this phenotype is complemented by full length *GLN4* on a single copy plasmid but not by an empty vector (Figure 4.2, A). In addition, the *gln4*(211-809) mutant is much more sensitive than wild type to L-methionine sulfoximine, a highly specific inhibitor of glutamine synthase (Manning et al., 1969), which results in reduced concentrations of intracellular glutamine (Figure 4.2, B). These observations demonstrate that the NTD plays an important role in the function of the native yeast enzyme *in vivo*.

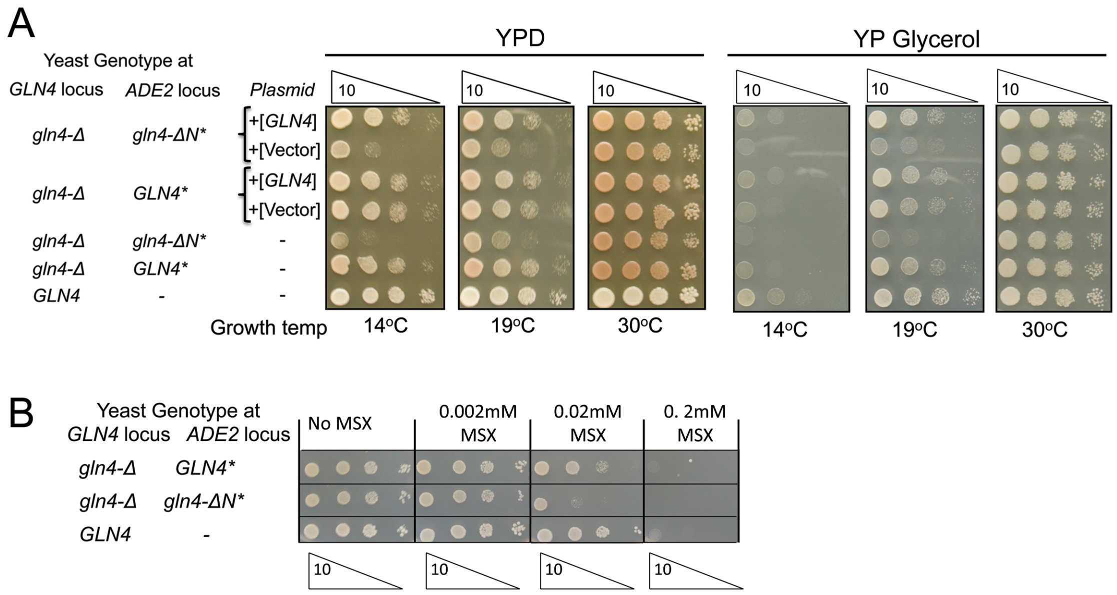


Figure 4.2 Deletion of the N-terminal domain of *GLN4* impairs function. A. Mutants bearing a *gln4* mutation in which amino acids 2-210 are deleted are defective in growth at low temperature on YP media containing glucose or glycerol as a carbon source. Serial dilutions of strains with either wild type *GLN4* or *gln4(*211-809) (marked *gln4-ΔN) integrated at the *ade2* locus in the *gln4-ΔKanR* mutant were grown as indicated. Indicated strains carry CEN plasmids either with or without *GLN4*. B. Mutants bearing a *gln4* mutation in which amino acids 2-210 are deleted are sensitive to the glutamine synthase inhibitor L-methionine sulfoximine (MSX).**

Steady-state kinetic parameters were measured to directly assess the effects of the NTD on $tRNA^{Gln}$ aminoacylation. Substantial differences between full length Gln4 and Gln4(187-809) were found. For the wild-type enzyme, similar K_M^{tRNA} (0.14 μM *versus* 0.19 μM) and k_{cat} (1.7 s^{-1} *versus* 1.4 s^{-1}) were measured for affinity-purified native $tRNA^{Gln}$ and an unmodified transcript, suggesting that post-transcriptional modifications do not have significant effects in this system. Using unmodified $tRNA^{Gln(CUG)}$ as substrate, we then found that Gln4(187-809) exhibits a 30-fold increase in K_M^{tRNA} (from 0.2 μM to 5.8 μM), and a 5.4-fold increase in K_M^{Gln} (from 1.7 mM to 9.3 mM) although the k_{cat} values are similar (1.4 sec^{-1} *versus* 1.7 sec^{-1}) (Table 4.4). We infer that the NTD influences the complementarity of both the tRNA and glutamine binding sites for their respective substrates, as also suggested by the sensitivity of the Gln4(211-809) mutant to L-methionine sulfoximide.

Table 4.4 Comparison of steady state kinetic parameters for Gln4 and Gln4 variants

	k_{cat} (s^{-1})	K_M^{tRNA} (μM)	k_{cat}/K_M^{tRNA} ($M^{-1}\cdot s^{-1}$)	K_M^{Gln} (mM)	k_{cat}/K_M^{Gln} ($M^{-1}\cdot s^{-1}$)
FL-Gln4	1.4 ± 0.2	0.19 ± 0.04	$7.6x10^6$	1.7 ± 0.2	$8.5x10^2$
Gln4 (187-809)	1.7 ± 0.3	5.85 ± 0.52	$2.9x10^5$	9.3 ± 0.3	$1.8x10^2$
PVG-GlnRS	2.8 ± 0.6	1.55 ± .51	$1.8x10^6$	N.A.	N.A.
FL-Gln4 + native tRNA	1.7 ± 0.1	0.144 ± 0.07	$1.2x10^7$	N.A.	N.A.

Since the kinetic analysis suggested a role for the NTD in $tRNA^{Gln}$ binding, we developed an EMSA assay to directly measure binding. We find that yeast Gln4 binds tightly and specifically to fully modified $tRNA^{Gln(CUG)}$ purified from *S. cerevisiae*, with ~25 nM Gln4 required for 50% binding (Figure 4.3, see Figure 4.10) while greater than 800 nM Gln4 is required to bind comparably to $tRNA^{Phe}$ (Figure 4.4). Remarkably, Gln4(187-809) binds only very weakly at 27 μM, 1000-fold above the apparent K_D of wild type Gln4 (Figure 4.3, A and B), and other Gln4 variants Gln4(211-809) and Gln4(216-809) do not detectably bind $tRNA^{Gln(CUG)}$ (Figure 4.3, A). Furthermore, there was no improvement in binding of Gln4(187-809) in the presence of other Gln4 substrates including glutamine, ATP, or the non-hydrolyzable ATP analog AMPPNP (Figure 4.5).

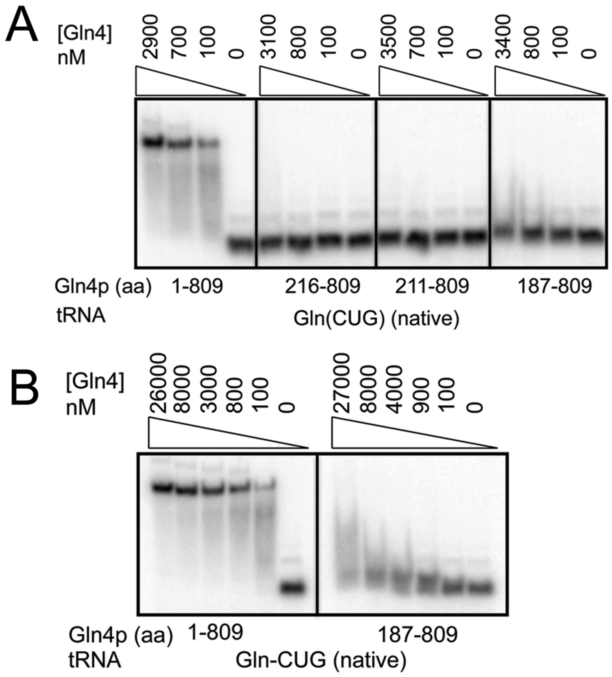


Figure 4.3 The N-terminal domain of Gln4 is required for specific binding to native tRNA^{Gln}(CUG). A. Gln4 variant proteins deleted for different amounts of the NTD exhibit reduced tRNA^{Gln}(CUG) binding. B. Gln4(187-809) protein exhibits detectable binding to tRNA^{Gln}(CUG) at high concentrations.

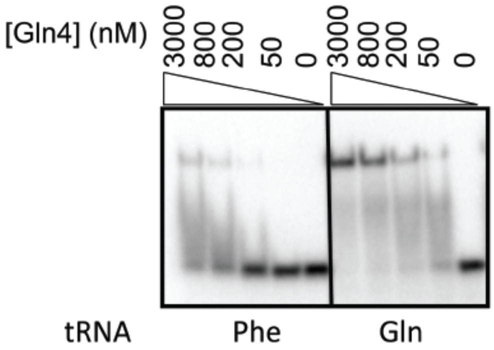


Figure 4.4 Gln4 binds more efficiently to native tRNA^{Gln}(CUG) than to native tRNA^{Phe}

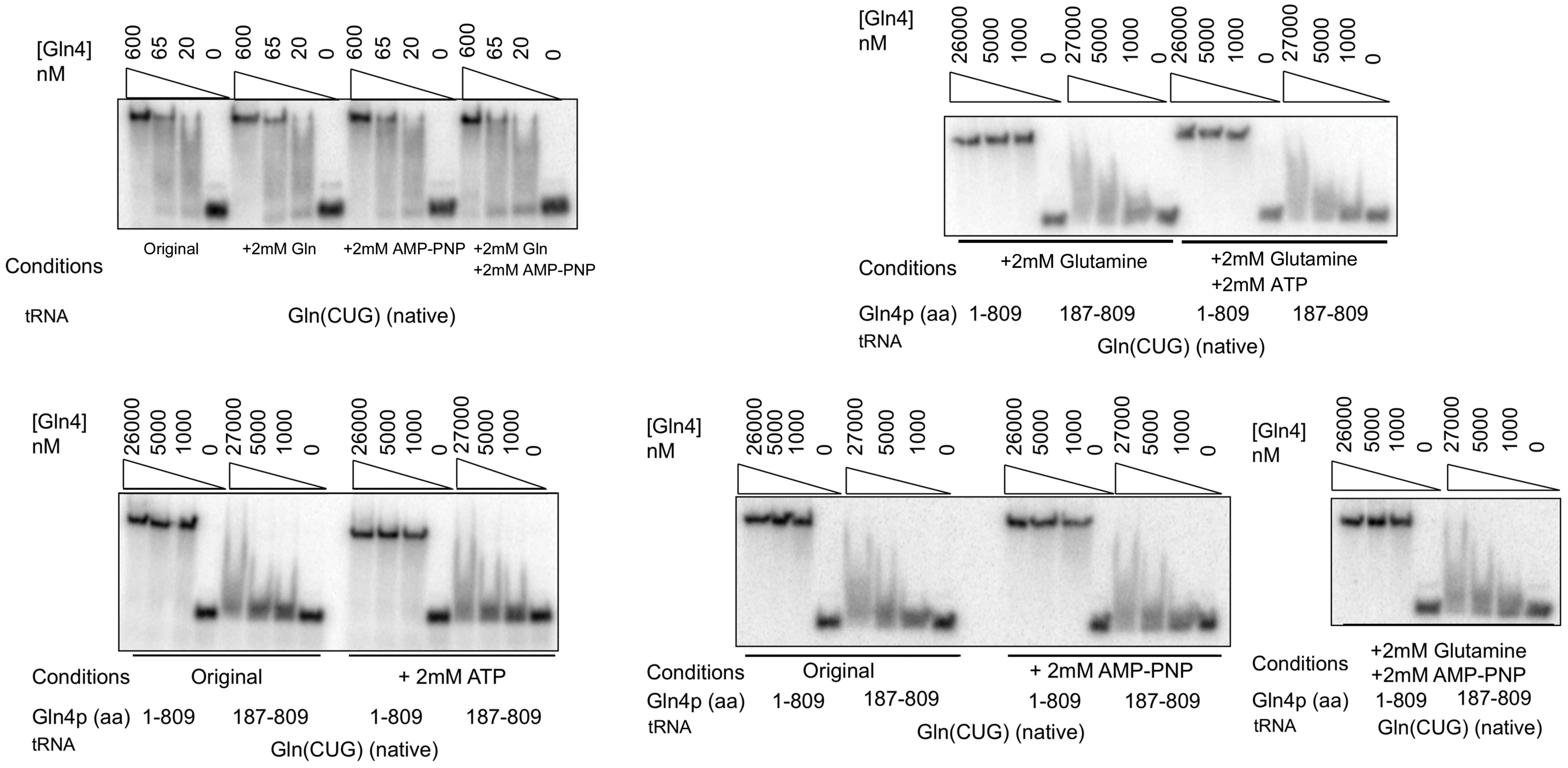


Figure 4.5 Effects of glutamine, ATP and AMP-PNP on binding of Gln4 and Gln4(187-809) to tRNA^{Gln}(CUG)

4.3.2 The Gln4 NTD is structurally similar to two subdomains in the amidotransferase that distinguish tRNA^{Gln} from tRNA^{Glu}.

To further discern the function of the NTD, we solved the structure of the isolated NTD, which behaves as a discrete unit to confer function when fused to the *E. coli* GlnRS (Whelihan and Schimmel, 1997). We purified three NTD variants ending at amino acids 187, which spans the region of extensive identity between the NTD of GlnRS from multiple species (see below), 215 and 228, which covers the entire region without extensive homology to *E. coli* GlnRS. We obtained crystals of Gln4(1-187) that diffracted to 2.3 Å, and solved the structure of a selenomethionine derivative purified from a yeast *sam1-Δ sam2-Δ* mutant (Malkowski et al., 2007) (Table 4.5).

Table 4.5 Data collection, phasing and refinement statistics for Gln4(1-187)

	Gln4 NTD		
Data collection			
Space group	P 1 2_1 1		
Cell dimensions			
a, *b*, *c* (Å)	40.79 34.61 74.25		
α, β, γ (°)	90.00 97.61 90.00		
	Peak	*Inflection*	*Remote*
Wavelength	0.97937	0.97904	0.91162
Resolution (Å)	29.2-2.30 (2.42-2.30)		
R_{merge}	0.098(0.485)	0.098(0.467)	0.100(0.490)
R_{pim}	0.062(0.309)	0.062(0.297)	0.063(0.312)
$I/\sigma I$	12.9(3.4)	12.7(3.6)	13.0(3.4)
Completeness (%)	100.0(100.0)	100.0(100.0)	99.9(99.8)
Redundancy	6.5(6.6)	6.5(6.6)	6.5(6.6)

Refinement	
Resolution (Å)	29.2-2.30
No. reflections	9385
$R_{work/}$ R_{free}	0.182/0.211
No. atoms	
Protein	1489
Ligand/ion	0
Water	82
B-factors	
Protein	35.10
Ligand/ion	
Water	42.36
R.m.s deviations	
Bond lengths (Å)	0.007
Bond angles (º)	1.035

*Highest resolution shell is shown in parenthesis.

Gln4(1-187) consists of two alpha helical domains, the first from residues 1-111 containing a seven-helix bundle, and the second from residues 119-187 containing a four-helix bundle, which are connected by a seven residue $G_{112}VG_{114}IGIT$ linker (Figure 4.6, A). One face of each domain is positively charged across the length of the domain, which might facilitate interactions with the negatively charged tRNA and provide the basis for the nonspecific RNA binding activity of this domain (Wang and Schimmel, 1999) (Figure 4.6, B).

Although the NTD lacks sequence homology to any available structure, a DALI search (Holm and Rosenstrom, 2010) of the NTD and the individual domains revealed substantial structural homology to the helical and tail domains of the GatB subunit of

GatCAB, the glutamyl-tRNA admidotransferase, from *Staphylococcus aureus* (PDB ID: 3IP4) (Nakamura et al., 2010) and *Thermotoga maritima* (PDB ID: 3AL0) (Ito and Yokoyama, 2010) (Figure 4.6, C and D, and Figure 4.7). The seven-helix bundle seen in the NTD yields an r.m.s. deviation of 3.75 Å using carbon alpha atoms in the alpha helices of *S. aureus* GatB and 4.01 Å when compared with *T. maritima*. However, a five-residue insertion between helix 4 and helix 5 appears to shift the orientation of the remaining three helices of *S. aureus* GatB. When aligning these three helices separately, an r.m.s. deviation of 1.89 Å is observed. The four-helix bundle of the C-terminal subdomain of the NTD has an r.m.s. deviation of only 1.64 Å compared with the *S. aureus* GatB tail domain, and 1.80 Å compared with the *T. maritima* GatB tail domain. Since the GatB helical and tail domains make specific and nonspecific contacts with tRNA^{Gln} (Ito and Yokoyama, 2010), we infer that the Gln4 NTD has similar biochemical function. Furthermore, it is likely that GlnRS NTDs from other eukaryotes adopt a similar structure, based on threading of these sequences to the Gln4(1-187) structure (Guex and Peitsch, 1997) (Table 4.6).

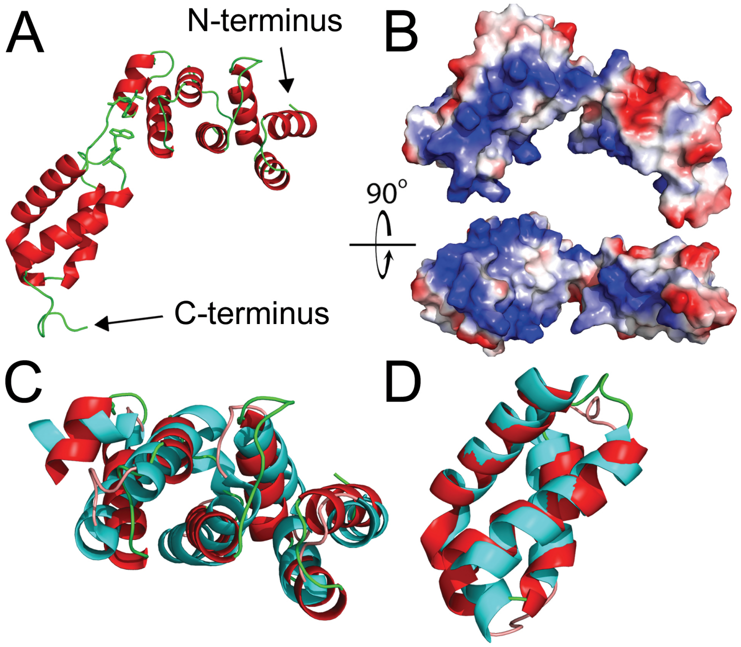


Figure 4.6 Structure of Gln4(1-187) with comparisons to domains in *S. aureus* GatB (PDB ID: 3IP4). A. Crystallographic structure of Gln4 residues 1-187 in cartoon representation. The proposed hinge region ($Gly_{112}Val_{113}Gly_{114}$) is highlighted together with the likely interacting residue Trp_{160}, and shown in stick representation. B. Surface electrostatic model of Gln4 residues 1-187, shown with two orientations rotated by 90° relative to each other, with positively charged residues colored blue. C, D. Structural alignment of helical and tail domains of Gln4 NTD and *S.aureus* GatB (PDB ID: 3IP4)(Nakamura et al., 2010). C. The crystal structure of Gln4(1-110) (red) is superposed to the helical domain of GatB(295-406) (cyan). D. The crystal structure of Gln4(119-178) (red) is superposed on the tail domains of GatB(414-475) (cyan).

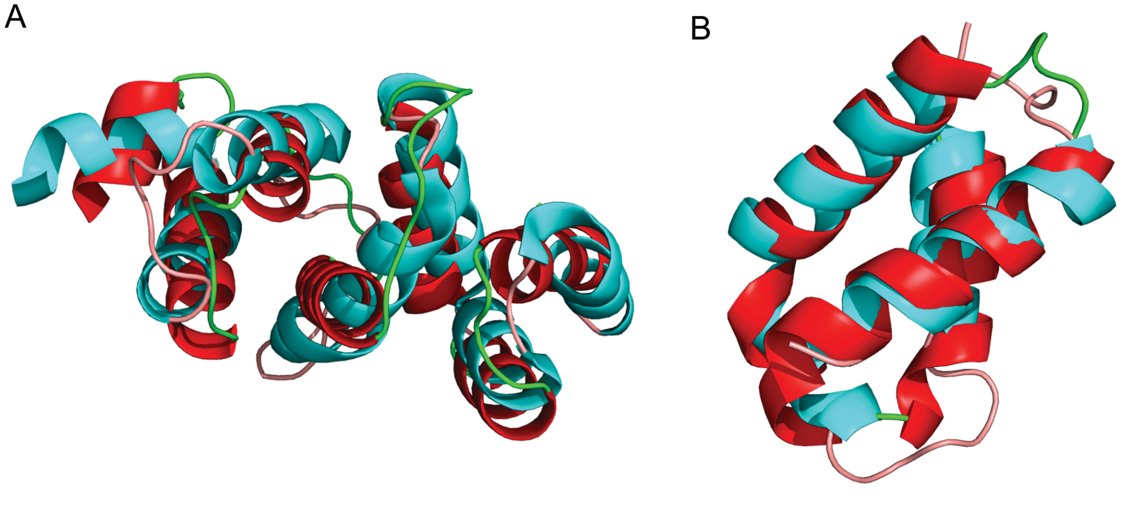


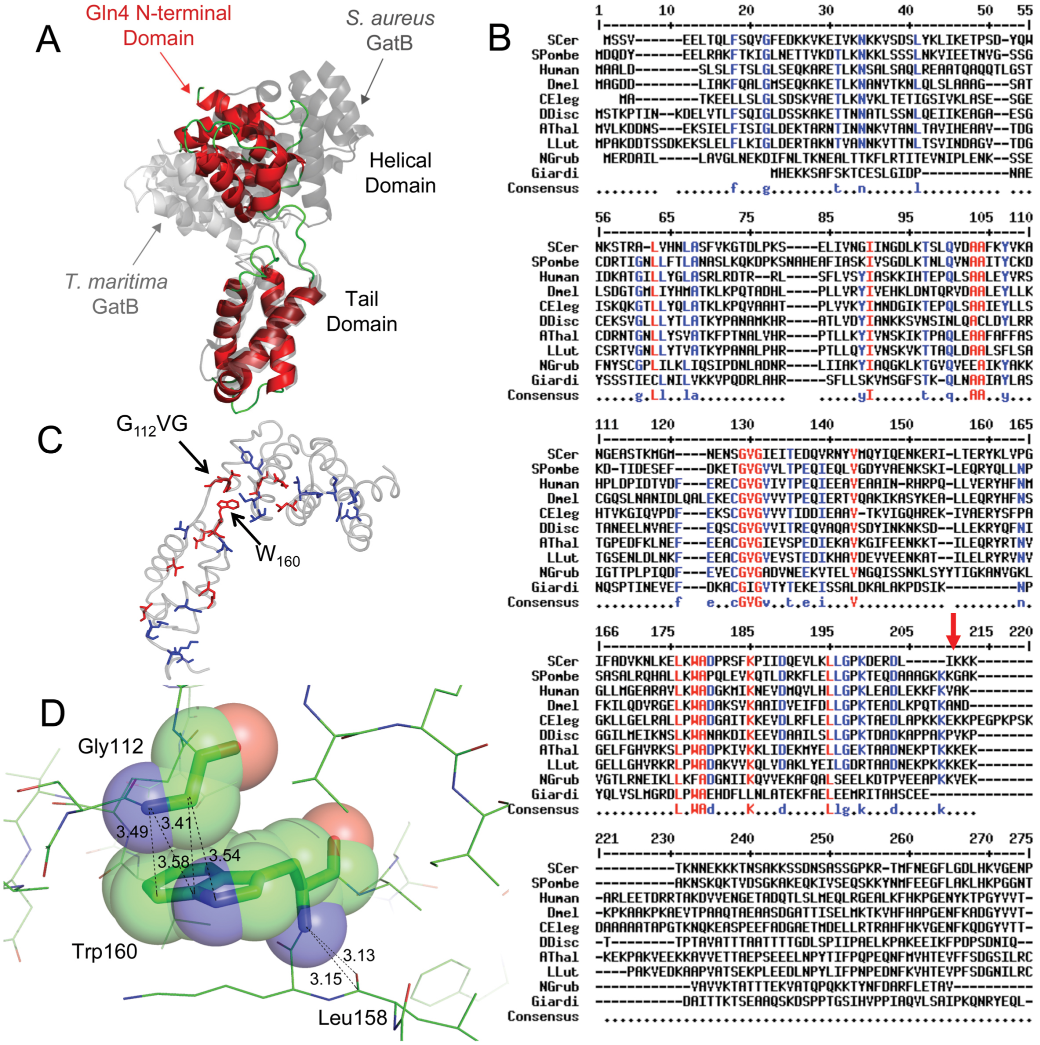
Figure 4.7 Structural alignment of helical and tail domains of Gln4 NTD and *Thermotoga maritima* GatB (PDB ID: 3AL0). A. The crystal structure of Gln4(2-110) (red) is superposed to the helical domain of *T. maritima* GatB(303-415) (cyan). B. The crystal structure of Gln4(119-178) (red) is superposed on the tail domains of GatB (422-481) (cyan).

Table 4.6 Comparison of sequences threaded to the Gln4(1-187) NTD structure.

Name	Species	Residues	Z-score		
			Combined	Pair	Surface
NTD	*S. cerevisiae*	186	-11.23	-7.98	-8.71
NTD reversed	*S. cerevisiae*	186	-0.30	-1.56	0.55
p13188	*S. cerevisiae*	186	-11.28	-7.99	-8.75
q9y7y8	*S. pombe*	190	-6.02	-0.79	-7.00
q9y105	*D. melanogaster*	188	-3.99	1.11	-5.62
q62431	*M. musculus*	183	-8.42	-6.44	-5.92
p47897	*H. sapiens*	185	-6.55	-3.69	-5.20
q3mhh4	*B. Taurus*	185	-6.62	-3.84	-5.22
p52780	*L. lupines*	188	-7.02	-1.73	-6.90
p14325	*D. discoideum*	185	-7.98	-4.06	-6.92
GatB	*T. maritima*	177	-10.43	-6.42	-8.91

4.3.3 The linker between the NTD subdomains is conserved and functionally important.

Three observations suggest that the linker that connects the two domains in Gln4 plays a crucial role in the tRNA binding function of this domain. First, the helical and tail domains of GatB are also connected by a linker, which appears to function as a flexible hinge that closes upon tRNA binding, based on differences in the orientation of the domains in the tRNA-bound (*T. maritima)* and tRNA-free (*S. aureus)* structures (Ito and Yokoyama, 2010; Nakamura et al., 2010). In this regard, we note that the domains in the Gln4 NTD are oriented at an angle between that of the *T. maritima* tRNA-bound GatB and the *S. aureus* tRNA-free GatB (Figure 4.8, A). Second, although the linker sequences in GlnRS differ from the sequences in GatB, the linker sequences in GlnRS are among the most highly conserved amino acids in the Gln4 NTD family (Figure 4.8, B). In a comparison of highly divergent eukaryotes, although neither the length nor the sequence of the N-terminal domain is highly conserved, three of the seven amino acids in the linker region $G_{112}V_{113}G_{114}$ are nearly 100% conserved (Figure 4.8, B and C). Furthermore, G_{112} appears to interact with W_{160}, one of the ten other highly conserved residues in the NTD; the α carbon of G112 is in van der Waals contact with C9 of W160 (Figure 4.6, A and Figure 4.8, C and D). Third, G_{114} is predicted to be a hinge residue, acting as a flexible connector of the two domains, based on an elastic network analysis with the program HingeProt (Emekli et al., 2008).



```
          1        10        20        30        40        50   55
          |--------+---------+---------+---------+---------+----|
     SCer  MSSV------EELTQLFSQVGFEDKKVKEIVKNKKVSDSLYKLIKETPSD-YQW
   SPombe MDQDY------EELRAKFTKIGLNETTVKDTLKNKKLSSSLNKVIEETNVG-SSG
    Human MAALD--------SLSLFTSLGLSEQKARETLKNSALSAQLREAATQAQQTLGST
     Dmel MAGDD--------LIAKFQALGMSEQKAKETLKNANVTKNLQLSLAAAG---SAT
    CEleg    MA--------TKEELLSLGLSDSKVAETLKNVKLTETIGSIVKLASE--SGE
    DDisc MSTKPTIN--KDELVTLFSQIGLDSSKAKETTNNATLSSNLQEIIKEAGA--ESG
    AThal MVLKDDNS---EKSIELFISIGLDEKTARNTINNNKVTANLTAVIHEAAV--TDG
     LLut MPAKDDTSSDKEKSLELFLKIGLDERTAKNTVANNKVTTNLTSVINDAGV--TDG
    NGrub  MERDAIL-----LAVGLNEKDIFNLTKNEALTTKFLRTITEVNIPLENK--SSE
   Giardi                       MHEKKSAFSKTCESLGIDP-----------NAE
Consensus ........  .......f...g.........t..n......l.......... ...

          56       65        75        85        95        105  110
          |--------+---------+---------+---------+---------+----|
     SCer NKSTRA-LVHNLASFVKGTDLPKS----ELIVNGIINGDLKTSLQVDAAFKYVKA
   SPombe CDRTIGNLLFTLANASLKQKDPKSNAHEAFIASKIVSGDLKTNLQVNAAITYCKD
    Human IDKATGILLYGLASRLRDTR--RL----SFLVSYIASKKIHTEPQLSAALEYVRS
     Dmel LSDGTGMLIYHMATKLKPQTADHL----PLLVRYIVEHKLDNTQRVDAALEYLLK
    CEleg ISKQKGTLLYQLATKLKPQVAAHT----PLVVKYIMNDGIKTEPQLSAAIEYLLS
    DDisc CEKSVGLLLYTLATKYPANAMKHR----ATLVDYIANKKSVNSINLQACLDYLRR
    AThal CDRNTGNLLYSVATKFPTNALVHR----PTLLKYIVNSKIKTPAQLEAAFAFFAS
     LLut CSRTVGNLLYTVATKYPANALPHR----PTLLQYIVNSKVKTTAQLDAALSFLSA
    NGrub FNYSCGPLILKLIQSIPDNLADNR----LIIAKYIAQGKLKTGVQVEEAIKYAKK
   Giardi YSSSTIECLNILVKKVPQDRLAHR-----SFLLSKVMSGFSTK-QLNAAIAYLAS
Consensus .....g.Ll..la...........    .....yI......t..q..AA..y...

          111      120       130       140       150       160  165
          |--------+---------+---------+---------+---------+----|
     SCer NGEASTKMGM----NENSGVGIEITEDQVRNYVMQYIQENKERI-LTERYKLVPG
   SPombe KD-TIDESEF----DKETGVGVVLTPEQIEQLVGDYVAENKSKI-LEQRYQLLNP
    Human HPLDPIDTVDF---ERECGVGVIVTPEQIEEAVEAAIN-RHRPQ-LLVERYHFNM
     Dmel CGQSLNANIDLQALEKECGVGVVVTPEQIERTVQAKIKASYKEA-LLEQRYHFNS
    CEleg HTVKGIQVPDF---EKSCGVGVVVTIDDIEAAV-TKVIGQHREK-IVAERYSFPA
    DDisc TANEELNVAEF---EQSCGVGVVITREQVAQAVSDYINKNKSD--LLEKRYQFNI
    AThal TGPEDFKLNEF---EEACGVGIEVSPEDIEKAVKGIFEENKKT--ILEQRYRTNV
     LLut TGSENLDLNKF---EEACGVGVEVSTEDIKHAVDEVVEENKAT--ILELRYRVNV
    NGrub IGTTPLPIQDF---EVECGVGADVNEEKVTELVNGQISSNKLSYYTIGKANVGKL
   Giardi NQSPTINEVEF---DKACGIGVTYTEKEISSALDKALAKPDSIK---------NP
Consensus ..........f   e..cGVGv..t.e.i...V............ ........n.

          166      175       185       195       205       215  220
          |--------+---------+---------+---------+---------+----|
     SCer IFADVKNLKELKWADPRSFKPIIDQEVLKLLGPKDERDL-----IKKK-------
   SPombe SASALRQHALLKWAPQLEVKQTLDRKFLELLGPKTEQDAAAGKKKGAK-------
    Human GLLMGEARAVLKWADGKMIKNEVDMQVLHLLGPKLEADLEKKFKVAK--------
     Dmel FKILQDVRGELKWADAKSVKAAIDVEIFDLLGPKTEADLKPQTKAND--------
    CEleg GKLLGELRALLPWADGAITKKEVDLRFLELLGPKTAEDLAPKKKEKKPEGPKPSK
    DDisc GGILMEIKNSLKWANAKDIKEEVDAAILSLLGPKTDADKAPPAKPVKP-------
    AThal GELFGHVRKSLPWADPKIVKKLIDEKMYELLGEKTAADNEKPTKKKEK-------
     LLut GELLGHVRKRLPWADAKVVKQLVDAKLYEILGDRTAADNEKPKKKKEK-------
    NGrub VGTLRNEIKLLKFADGNIIKQVVEKAFQALSEELKDTPVEEAPKKVEK-------
   Giardi YQLVSLMGRDLPWAEHDFLLNLATEKFAELEEMRITAHSCEE-------------
Consensus ..........L.WAd....K...d.....Llg.k...d.....k....

          221      230       240       250       260       270  275
          |--------+---------+---------+---------+---------+----|
     SCer ----------TKNNEKKKTNSAKKSSDNSASSGPKR-TMFNEGFLGDLHKVGENP
   SPombe ----------AKNSKQKTVDSGKAKEQKIVSEQSKKYNMFEEGFLAKLHKPGGNT
    Human -ARLEETDRRTAKDVVENGETADQTLSLMEQLRGEALKFHKPGENYKTPGYVVT-
     Dmel -KPKAAKPKAEVTPAAQTAEAASDGATTISELMKTKVHFHAPGENFKADGYVVT-
    CEleg DAAAAATAPGTKNQKEASPEEFADGAETMDELLRTRAHFHKVGENFKQDGYVTT-
    DDisc -T--------TPTAVATTTAATTTTGDLSPIIPAELKPAKEEIKFPDPSDNIQ--
    AThal -KEKPAKVEEKKAVVETTAEPSEEELNPYTIFPQPEQNFMVHTEVFFSDGSILRC
     LLut ----PAKVEDKAAPVATSEKPLEEDLNPYLIFPNPEDNFKVHTEVPFSDGNILRC
    NGrub ------------VAVVKTATTTEKVATQPQKKTYNFDARFLETAV----------
   Giardi -----------DAITTKTSEAAQSKDSPPTGSIHVPPIAQVLSAIPKQNRYEQL-
Consensus  ......................................................
```

Figure 4.8 The linker between the two domains in Gln4(1-187) likely behaves as a hinge, is highly conserved and is important for tRNA binding. A. Structure of Gln4(1-187) (red) superposed on *Tm*GatB (light gray) and *Sa*GatB (dark gray) by alignment of the tail domains. B. Conservation of GlnRS NTD sequences, red-≥90%; blue-≥ 70%, with arrow at $Gln4_{187}$., aligned using Multalin (Corpet, 1988). C. Conserved residues are highlighted on Gln4(1-187) according to the color code in B with the NTD backbone shown in light grey. D. Close contacts between W160 of the Gln4 NTD and other residues.

4.3.4 SAXS data reveals that the NTD structure is not distorted by crystal contacts

Given that the NTD crystal structure shows the position of the helical subdomain to be in between that of the tRNA-bound and tRNA-free GatB structures (Figure 4.8, A), it is possible that crystal contacts may have altered the relative orientation of the helical and tail domains of the NTD. To determine the native position of the helical domain in solution, we collected SAXS data on the NTD. Using the crystal structures of the NTD and both the tRNA-bound and tRNA-free GatB structures, the simulated scattering profiles were calculated and fit to the experimental SAXS data (Figure 4.9, A). While the differences are small, the fit of the Gln4 NTD structure was better than either the tRNA-bound or the tRNA-free GatB structures. Figure 4.9, B shows that the *ab initio* envelope reconstruction fits to the NTD structures better than to either the tRNA-bound or tRNA-free GatB structures. Our SAXS data, therefore, demonstrates that the crystal structure of the NTD accurately reflects the relative conformation of the helical and tail domains in solution and that it has not been altered by crystal contacts.

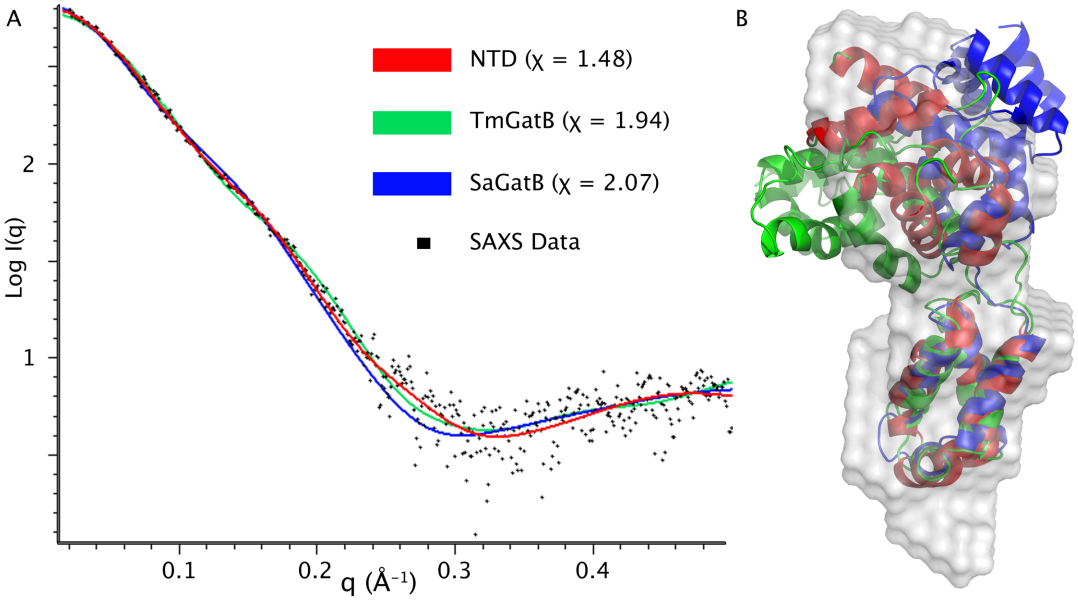


Figure 4.9 SAXS data shows that the NTD crystal structure is similar to that found in solution. A. Simulated scattering profiles calculated by CRYSOL for the Gln4 NTD (red), *Tm*GatB (green), and *Sa*GatB (blue) are shown overlaid on top of experimental SAXS data from the Gln4 NTD in solution. Goodness of fit values (χ) are given in parentheses. B. The *ab initio* envelope reconstructed from the experimental scattering profile of the Gln4 NTD is shown superimposed onto the crystal structures of the Gln4 NTD (red), *Tm*GatB (green), and *Sa*GatB (blue). The orientation of these structures is similar to that presented in Figure 4.8, A.

4.3.5 Point mutations in NTD linker impair tRNA binding integrity

Since the $G_{112}V_{113}G_{114}$ residues of the linker are highly conserved, and since hinges frequently mediate conformational changes upon ligand binding (Gerstein et al., 1994), we considered it likely that mutations in the linker region would impair function. Thus, we purified variant proteins in which $G_{112}V_{113}G_{114}$ was replaced with AAA and with PVG and in which W_{160} was replaced with F or A, and measured $tRNA^{Gln(CUG)}$ binding. Although the variant proteins all bind $tRNA^{Gln(CUG)}$, as measured by reduced mobility of the tRNA, all of the mutant proteins exhibit defects in binding (Figure 4.10, A and B).

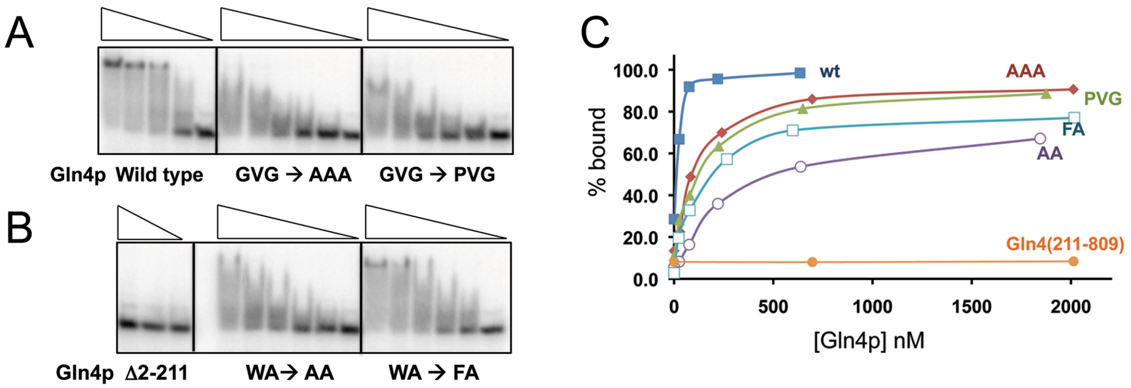


Figure 4.10 Mutations in conserved amino acids in the putative hinge of the NTD affect the interaction of Gln4 with native tRNA$^{Gln(CUG)}$. A., B. EMSA wild type and mutant Gln4 proteins (23 nM to 2,017 nM). C. Binding as a function of Gln4 protein concentration.

Three variants (Gln4-$A_{112}A_{113}A_{114}$, Gln4-$G_{112}P$, Gln4-$W_{160}A$) fail to form stable complexes with tRNA$^{Gln(CUG)}$, as judged by lack of comigration of the complexed tRNA with that formed by wild type Gln4, and all four variant proteins exhibit an apparently reduced affinity for tRNA$^{Gln(CUG)}$, requiring 4 to 12 times more protein than the wild type to bind comparable amounts of tRNA (Figure 4.10, C). Moreover, the Gln4-$G_{112}P$ variant exhibits a 10-fold increase in the K_M^{tRNA} (from 0.19 μM to 1.6 μM) as well as a slight increase in k_{cat} (1.4 s^{-1} versus 2.8 s^{-1}) (Table 4.4). Thus, we conclude that the linker region is important for binding, and speculate that it acts as a hinge facilitating closure between the helical and tail domains upon tRNA binding.

4.4 Discussion

The observations that the NTD of *S. cerevisiae* GlnRS bears a substantial structural resemblance to two domains of the bacterial GatB amidotransferase that distinguish tRNA^{Gln} from tRNA^{Glu}, and that the NTD also participates in tRNA^{Gln} binding, imply that there is a connection between the indirect pathways for formation of Gln-tRNA^{Gln} in bacteria and archaea, and the direct pathway that evolved in eukaryotes. Since it is thought that tRNA^{Gln} was present in the last universal common ancestor, it has been puzzling that aminoacylation of this tRNA is achieved by different routes in each of the three kingdoms. Sheppard and Soll proposed that both GatCAB and GatDE were present prior to the split between archaea and bacteria (Sheppard and Soll, 2008), while the specific GlnRS evolved in eukaryotes. We propose that the tRNA^{Gln} recognition domain from an amidotransferase was most likely conscripted as an NTD to a progenitor nondiscriminating GluRS, and thus played an integral part in the development of the eukaryotic GlnRS family. In particular, evolution of GlnRS from an early nondiscriminating GluRS required selectivity determinants in favor of tRNA^{Gln} to evolve, while negative determinants against tRNA^{Glu} would also appear. The proximity of the NTD to the tRNA-synthetase core domain suggests that eukaryotes may have exploited the NTD domain to provide subtle structural discrimination between tRNA^{Gln} and tRNA^{Glu} prior to the appearance of discriminatory residues in the core synthetase globular domain.

In support of this, we find evidence that the NTD of GlnRS likely existed in the common eukaryotic ancestor, based on comparative genomic reconstruction of the Gln4 family (Fritz-Laylin et al., 2010). Thus, GlnRS proteins from highly diverse, free living eukaryotes, spanning lineages from the ancient JEH and POD clades through more recent clades (including Plantae, Amoebozoans and Opisthokonts) share a recognizably homologous, but diverse, NTD of 210 to 259 amino acids (Figure 4.8, B and Figure 4.11). Curiously, we also find that the appended domain is absent in some eukaryotes, including parasitic protozoa such as*Trypanosoma brucei* and *Leishmania major*, as well as the *Eurotiomycetidae*, *Trichocomaceae* fungi. There also appears to be a correlation between the presence of the appended domain and the use of U_{73} as the discriminator base (Figure 4.12). Thus, although an appended domain is not required to construct a specific GlnRS, such a domain was likely a part of the specific GlnRS in the eukaryotic common ancestor and may have played a crucial role in the development of a specific GlnRS.

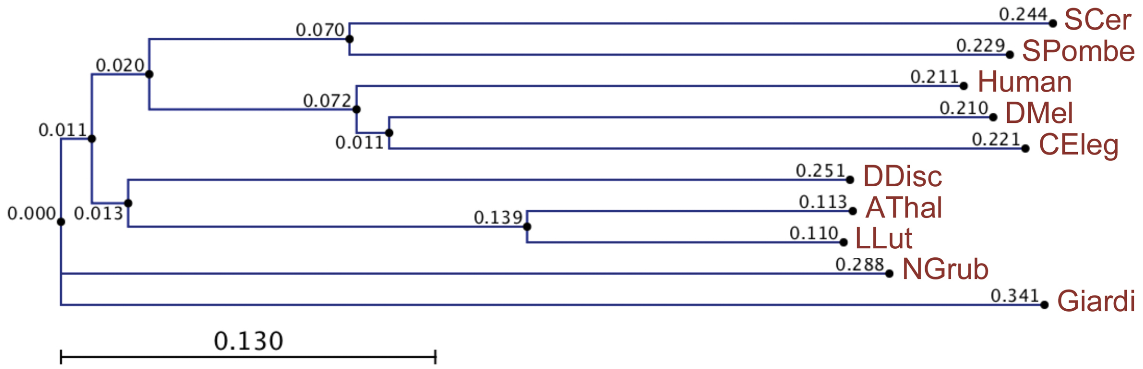


Figure 4.11 Phylogenetic tree indicating the relationships between the organisms from which GlnRS NTDs were compared in Figure 4.7B. The scale bar indicates the number of amino acid substitutions per site.

Our findings also point to a parallel between the appended domains in eukaryotic GlnRS proteins and in GlnRS in the bacterium *D. radiodurans* (Deniziak et al., 2007), even though the eukaryotic domains are located on the N terminus, upstream of the conserved core, while the appended domain of the *D. radiodurans* GlnRS is on the C terminus, downstream of the conserved core. Although the Gln4 NTD and the *D. radiodurans* GlnRS CTD have no significant sequence similarity (Deniziak et al., 2007), and are at opposite termini, it is likely that the *D. radiodurans* GlnRS CTD, like the Gln4 NTD, is structurally related to GatB, because the CTD has weak sequence homology with regions of GatB, and cross-reacts with GatB antibody (Deniziak et al., 2007).

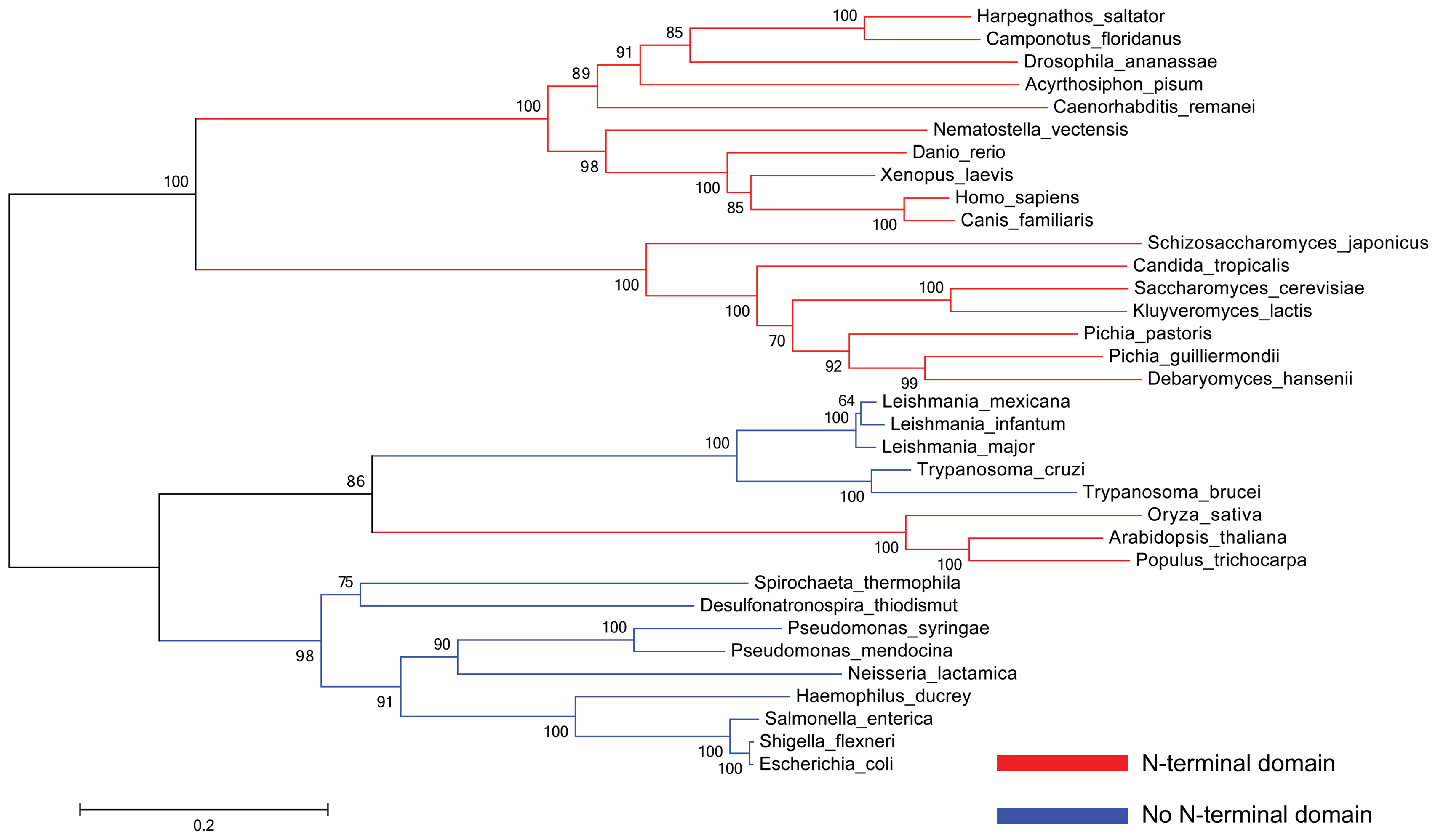


Figure 4.12 Phylogenetic distribution of GlnRSs with and without N-terminal appended domains. Eukaryotic and bacterial GlnRS sequences were aligned using ClustalW and a phylogenetic tree was constructed using MEGA5, with 200 bootstraps carried out to test statistical relevance. In addition to bacteria, a set of Euglenozoa protists lacks the appended domain.

4.5 Summary

In this chapter the structure of the Gln4 NTD was presented and found to be structurally homologous to the B subunit of GatCAB from bacterial species. Since GatCAB is the enzyme involved in the indirect route of Gln-tRNA^{gln} formation, this suggests a phylogenetic link between the ancient and modern pathways of glutaminyl-tRNA aminoacylation. Biochemical studies also found that specific residues in the linker region of the NTD likely act as a hinge between the helical and tail domains. Using SAXS data coupled with data quality analysis performed by SAXStats it was shown that the solution structure of the NTD is more similar to the crystal structure of the NTD than to either the open or closed conformations of GatB, suggesting that crystal contacts have not distorted the conformation of the NTD. This suggests that while the NTD is very similar to the GatB structure, differences in the conformational orientation of the helical and tail domains are real, suggesting that the open conformation of the NTD of Gln4 is slightly different than the open conformation of GatB. This study shows the complementary nature of SAXS analysis using SAXStats to objectively characterize data quality as part of several methods the structural biologist can use to gain a more complete understanding of a biological system.

4.6 References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W., *et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica *66*, 213-221.

Bullock, T.L., Uter, N., Nissan, T.A., and Perona, J.J. (2003). Amino acid discrimination by a class I aminoacyl-tRNA synthetase specified by negative determinants. J Mol Biol *328*, 395-408.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D *66*, 12-21.

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res *16*, 10881-10890.

Curnow, A.W., Hong, K., Yuan, R., Kim, S., Martins, O., Winkler, W., Henkin, T.M., and Soll, D. (1997). Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. Proc Natl Acad Sci U S A *94*, 11819-11826.

Cusack, S., Berthet-Colominas, C., Hartlein, M., Nassar, N., and Leberman, R. (1990). A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature *347*, 249-255.

Deniziak, M., Sauter, C., Becker, H.D., Paulus, C.A., Giege, R., and Kern, D. (2007). Deinococcus glutaminyl-tRNA synthetase is a chimer between proteins from an ancient and the modern pathways of aminoacyl-tRNA formation. Nucleic Acids Res *35*, 1421-1431.

Emekli, U., Schneidman-Duhovny, D., Wolfson, H.J., Nussinov, R., and Haliloglu, T. (2008). HingeProt: automated prediction of hinges in protein structures. Proteins *70*, 1219-1227.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta crystallographica Section D, Biological crystallography *66*, 486-501.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature *347*, 203-206.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr *42*, 342-346.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J.*, et al.* (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell *140*, 631-642.

Gerstein, M., Lesk, A.M., and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. Biochemistry *33*, 6739-6749.

Gonzalez, A., Moorhead, P., McPhillips, S.E., Song, J., Sharp, K., Taylor, J.R., Adams, P.D., Sauter, N.K., and Soltis, S.M. (2008). Web-Ice: integrated data collection and analysis for macromolecular crystallography. J Appl Crystallogr *41*, 176-184.

Grant, T.D., Snell, E.H., Luft, J.R., Quartley, E., Corretore, S., Wolfley, J.R., Snell, M.E., Hadd, A., Perona, J.J., Phizicky, E.M.*, et al.* (2012). Structural conservation of an ancient tRNA sensor in eukaryotic glutaminyl-tRNA synthetase. Nucleic Acids Res *40*, 3723-3731.

Guex, N., and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis *18*, 2714-2723.

Guo, M., Yang, X.L., and Schimmel, P. (2010). New functions of aminoacyl-tRNA synthetases beyond translation. Nat Rev Mol Cell Biol *11*, 668-674.

Holm, L., and Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-549.

Ibba, M., Hong, K.W., Sherman, J.M., Sever, S., and Soll, D. (1996). Interactions between tRNA identity nucleotides and their recognition sites in glutaminyl-tRNA synthetase determine the cognate amino acid affinity of the enzyme. Proc Natl Acad Sci U S A *93*, 6953-6958.

Ibba, M., and Soll, D. (2004). Aminoacyl-tRNAs: setting the limits of the genetic code. Genes Dev *18*, 731-738.

Ito, T., and Yokoyama, S. (2010). Two enzymes bound to one transfer RNA assume alternative conformations for consecutive reactions. Nature *467*, 612-616.

Jackman, J.E., Montange, R.K., Malik, H.S., and Phizicky, E.M. (2003). Identification of the yeast gene encoding the tRNA m1G methyltransferase responsible for modification at position 9. Rna *9*, 574-585.

Kabsch, W. (2010). Xds. Acta crystallographica *66*, 125-132.

Lamour, V., Quevillon, S., Diriong, S., N'Guyen, V.C., Lipinski, M., and Mirande, M. (1994). Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci U S A *91*, 8670-8674.

Liu, C., Gamper, H., Shtivelband, S., Hauenstein, S., Perona, J.J., and Hou, Y.M. (2007). Kinetic quality control of anticodon recognition by a eukaryotic aminoacyl-tRNA synthetase. J Mol Biol *367*, 1063-1078.

Ludmerer, S.W., and Schimmel, P. (1987a). Construction and analysis of deletions in the amino-terminal extension of glutamine tRNA synthetase of Saccharomyces cerevisiae. J Biol Chem *262*, 10807-10813.

Ludmerer, S.W., and Schimmel, P. (1987b). Gene for yeast glutamine tRNA synthetase encodes a large amino-terminal extension and provides a strong confirmation of the signature sequence for a group of the aminoacyl-tRNA synthetases. J Biol Chem *262*, 10801-10806.

Ludmerer, S.W., Wright, D.J., and Schimmel, P. (1993). Purification of glutamine tRNA synthetase from Saccharomyces cerevisiae. A monomeric aminoacyl-tRNA synthetase with a large and dispensable NH2-terminal domain. J Biol Chem *268*, 5519-5523.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol *142*, 170-179.

Luft, J.R., Wolfley, J.R., Said, M.I., Nagel, R.M., Lauricella, A.M., Smith, J.L., Thayer, M.H., Veatch, C.K., Snell, E.H., Malkowski, M.G.*, et al.* (2007). Efficient optimization of crystallization conditions by manipulation of drop volume ratio and temperature. Protein Sci *16*, 715-722.

Lyakhov, D.L., He, B., Zhang, X., Studier, F.W., Dunn, J.J., and McAllister, W.T. (1997). Mutant bacteriophage T7 RNA polymerases with altered termination properties. J Mol Biol *269*, 28-40.

Macbeth, M.R., Lingam, A.T., and Bass, B.L. (2004). Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. Rna *10*, 1563-1571.

Malkowski, M.G., Quartley, E., Friedman, A.E., Babulski, J., Kon, Y., Wolfley, J., Said, M., Luft, J.R., Phizicky, E.M., DeTitta, G.T.*, et al.* (2007). Blocking S-adenosylmethionine synthesis in yeast allows selenomethionine incorporation and multiwavelength anomalous dispersion phasing. Proc Natl Acad Sci U S A *104*, 6678-6683.

Manning, J.M., Moore, S., Rowe, W.B., and Meister, A. (1969). Identification of L-methionine S-sulfoximine as the diastereoisomer of L-methionine SR-sulfoximine that inhibits glutamine synthetase. Biochemistry *8*, 2681-2685.

Marck, C., and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. Rna *8*, 1189-1232.

Martzen, M.R., McCraith, S.M., Spinelli, S.L., Torres, F.M., Fields, S., Grayhack, E.J., and Phizicky, E.M. (1999). A biochemical genomics approach for identifying genes by the activity of their products. Science *286*, 1153-1155.

Mirande, M. (2010). Processivity of translation in the eukaryote cell: role of aminoacyl-tRNA synthetases. FEBS Lett *584*, 443-447.

Nakamura, A., Sheppard, K., Yamane, J., Yao, M., Soll, D., and Tanaka, I. (2010). Two distinct regions in Staphylococcus aureus GatCAB guarantee accurate tRNA recognition. Nucleic Acids Res *38*, 672-682.

Nureki, O., O'Donoghue, P., Watanabe, N., Ohmori, A., Oshikane, H., Araiso, Y., Sheppard, K., Soll, D., and Ishitani, R. (2010). Structure of an archaeal non-discriminating glutamyl-tRNA synthetase: a missing link in the evolution of Gln-tRNAGln formation. Nucleic Acids Res *38*, 7286-7297.

Popov, A.N., and Bourenkov, G.P. (2003). Choice of data-collection parameters based on statistic modelling. Acta crystallographica *59*, 1145-1153.

Quartley, E., Alexandrov, A., Mikucki, M., Buckner, F.S., Hol, W.G., DeTitta, G.T., Phizicky, E.M., and Grayhack, E.J. (2009). Heterologous expression of L. major proteins in S. cerevisiae: a test of solubility, purity, and gene recoding. J Struct Funct Genomics *10*, 233-247.

Rould, M.A., Perona, J.J., Soll, D., and Steitz, T.A. (1989). Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science *246*, 1135-1142.

Sauerwald, A., Zhu, W., Major, T.A., Roy, H., Palioura, S., Jahn, D., Whitman, W.B., Yates, J.R., 3rd, Ibba, M., and Soll, D. (2005). RNA-dependent cysteine biosynthesis in archaea. Science *307*, 1969-1972.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res *31*, 3381-3385.

Sheppard, K., and Soll, D. (2008). On the evolution of the tRNA-dependent amidotransferases, GatCAB and GatDE. J Mol Biol *377*, 831-844.

Sherlin, L.D., Bullock, T.L., Nissan, T.A., Perona, J.J., Lariviere, F.J., Uhlenbeck, O.C., and Scaringe, S.A. (2001). Chemical and enzymatic synthesis of tRNAs for high-throughput crystallization. Rna *7*, 1671-1678.

Sherman, F., Fink,G., and Hicks,J.B. (1986). In Methods in Yeast Genetics (New York, Cold Spring Harbor Laboratory Press), pp. 145-149.

Smolsky, I.L., Liu, P., Niebuhr, M., Ito, K., Weiss, T.M., and Tsuruta, H. (2007). Biological small-angle x-ray scattering facility at the Stanford synchrotron radiation laboratory. J Appl Crystallogr *40*, S453-S458.

Soltis, S.M., Cohen, A.E., Deacon, A., Eriksson, T., Gonzalez, A., McPhillips, S., Chui, H., Dunten, P., Hollenbeck, M., Mathews, I.*, et al.* (2008). New paradigm for macromolecular crystallography experiments at SSRL: automated crystal screening and remote data collection. Acta Crystallogr D *64*, 1210-1221.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL - a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. J Appl Crystallogr *28*, 768-773.

Tumbula, D.L., Becker, H.D., Chang, W.Z., and Soll, D. (2000). Domain-specific recruitment of amide amino acids for protein synthesis. Nature *407*, 106-110.

Uter, N.T., and Perona, J.J. (2004). Long-range intramolecular signaling in a tRNA synthetase complex revealed by pre-steady-state kinetics. Proc Natl Acad Sci U S A *101*, 14396-14401.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.

Wang, C.C., and Schimmel, P. (1999). Species barrier to RNA recognition overcome with nonspecific RNA binding domains. J Biol Chem *274*, 16508-16512.

Whelihan, E.F., and Schimmel, P. (1997). Rescuing an essential enzyme-RNA complex with a non-essential appended domain. EMBO J *16*, 2968-2974.

Whipple, J.M., Lane, E.A., Chernyakov, I., D'Silva, S., and Phizicky, E.M. (2011). The yeast rapid tRNA decay pathway primarily monitors the structural integrity of the acceptor and T-stems of mature tRNA. Genes Dev *25*, 1173-1184.

Wiederstein, M., and Sippl, M.J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res *35*, W407-410.

Wilkinson, M.L., Crary, S.M., Jackman, J.E., Grayhack, E.J., and Phizicky, E.M. (2007). The 2'-O-methyltransferase responsible for modification of yeast tRNA at position 4. Rna *13*, 404-413.

5 The Structure of Yeast Glutaminyl-tRNA Synthetase and Modeling Its Interaction with tRNA

5.1 Introduction

Aminoacyl-tRNA synthetases are required in all three domains of life for covalently attaching amino acids to their cognate tRNA molecule for use in protein synthesis (Ibba and Soll, 2000). While in most cases one synthetase exists for each amino acid, an exception occurs for glutamine and asparagine (Curnow et al., 1997). In eukaryotes and some bacteria, the traditional pathway of aminoacylation exists for glutamine, in which glutaminyl-tRNA synthetase (GlnRS) binds to tRNA^{gln}, glutamine and ATP and first forms a glutaminyl adenylate molecule that is then covalently attached to the 3'-end of tRNA^{gln} with the release of AMP. A different pathway exists in most bacteria and all archaea, where a non-discriminating glutamyl-tRNA synthetase (GluRS) attaches glutamic acid to both tRNA^{glu} and tRNA^{gln}. The misacylated glu-tRNA^{gln} is then converted to gln-tRNA^{gln} by the GatCAB amidotransferase enzyme in bacteria and some archaea, or by the GatDE amidotransferases in other archaea. Since this indirect pathway for aminoacylation exists in most prokaryotes, GlnRS is primarily a eukaryotic enzyme, and its presence in a small number of bacteria is believed to have occurred through a horizontal gene transfer event (Lamour et al., 1994). Nonetheless, there are significant differences between the prokaryotic and eukaryotic GlnRS enzymes.

Saccharomyces cerevisiae glutaminyl-tRNA synthetase (*Sc*GlnRS) is an 809-residue protein that contains a 215-residue domain appended to its N-terminus. This domain is nearly ubiquitous among eukaryotic GlnRS species, but absent in prokaryotic homologs (Ludmerer et al., 1993). Eukaryotic tRNA synthetases have often been shown to contain additional domains appended to their N-terminal or C-terminal ends, compared to their prokaryotic homologs (Guo et al., 2010). Some of these domains are known to be involved in various roles, including nucleic acid binding, protein-protein interactions, and hydrolytic editing mechanisms (Cusack et al., 1990; Eriani et al., 1990); however, the function of many of these domains remains uncertain. In Chapter 4 we described the structure of the N-terminal domain (NTD) of *Sc*GlnRS, revealing that it has an extraordinary structural resemblance to the region of the B subunit of the GatCAB amidotransferase (Ito and Yokoyama, 2010) that binds to tRNA^{gln} (Grant et al., 2012). Although deletion of the NTD distinctly affects catalytic activity, growth in yeast, and tRNA binding (Grant et al., 2012), the manner in which tRNA binding occurs is still unknown.

Structural data for two prokaryotic GlnRS species exists (Deniziak et al., 2007; Rould et al., 1989), yet no structure has been reported for any full-length eukaryotic GlnRS. Here we present the first crystallographic structure of the CTD of *Sc*GlnRS from crystals of full-length GlnRS where the NTD is disordered. Based on this structure, the structure of the isolated NTD (Grant et al., 2012) and small angle X-ray scattering (SAXS) data of the full-length enzyme evaluated by SAXStats, we present a model of the full-length enzyme in solution. We extend this to model the full-length enzyme bound to tRNA^{gln} from the crystallographic structures and homology with known transamidosome and GlnRS-tRNA complex structures (Ito and Yokoyama, 2010; Rath et al., 1998) yielding new insights into the structural rearrangements occurring in eukaryotic GlnRS-tRNA^{gln} complex formation. SAXS has proven to be a valuable tool to complement the multiple biochemical, structural, and bioinformatics approaches used to understand this biological system.

In this chapter the laboratories of Drs. Eric Phizicky and Elizabeth Grayhack at the University of Rochester performed protein expression and purification. The laboratory of Dr. Edward Snell performed full-length Gln4 crystallization and Dr. Snell determined the structure. I performed bioinformatics sequence analysis, collected and analyzed SAXS data, created all models of full-length Gln4, performed molecular dynamics simulations, and performed limited crystallization, structure solution, and refinement of Gln4.

5.2 Experimental Procedures

5.2.1 Protein Expression and Purification

To obtain highly purified *Sc*GlnRS protein and its derivatives, ORFs were cloned into the previously described LIC vectors BG2483 or BG2663 under P_{GAL1} control (Malkowski et al., 2007) as described (Grant et al., 2012), and expressed in yeast strain BCY123 (Macbeth et al., 2004). Large-scale growth, affinity purification on IgG Sepharose, removal of GST-3C protease, concentration of samples and sizing on Superdex HiLoad 16/60 (GE Healthcare 17-1069, 10 x 300 mm bed dimension) were performed as previously described (Quartley et al., 2009).

To ensure that full-length polypeptide was purified for crystallography, an N-terminal Met-Ala-His_6 tag was added at the N-terminus of *GLN4*, during PCR amplification of the *GLN4* gene using QB832ADFP and QB832ADRP primers (Table 5.1). The C-terminal domain of *GLN4* beginning at amino acid 216 was amplified with oligonucleotides QB1034ADFP and QB832ADRP, while full-length *GLN4* was amplified with QB1012ADFP and QB832ADRP.

Table 5.1 Oligonucleotides for Gln4 protein constructs

Primer	Nucleotide Sequence
QB1012ADFP	AATTCCATCAACCTTAAAATGTCTTCTGTAGAAGAATTGACT
QB832ADFP	AATTCCATCAACCTTAAAATGGCTCACCATCACCATCACCATATGTCTTCTGTAGAAGAATTGACT
QB832ADRP	CTTCCAAACCACTCTTGGAAGTTGCGTCCTTCAA
QB1034ADFP	AATTCCATCAACCTTAAAATGAGGACTATGTTCAATGAAGGTTTCC

For crystallography of the full-length His_6-Gln4 protein, the His_6-Gln4-pt was purified from strain QB832AD (21.6 liters at 6.5OD/liter) on 24 ml IgG Sepharose and eluted with 64 ml 3C cleavage buffer (Quartley et al., 2009). After elution from IgG Sepharose, and removal of GST-3C protease but prior to sizing, the sample containing the full-length *Sc*GlnRS protein with the N-terminal His_6 tag (QB832AD) was diluted with an equal volume of buffer T (25 mM HEPES, pH 7.5, 850 mM NaCl, 10 % glycerol, 2 mM β-mercaptoethanol (BME), 2 mM PMSF), mixed with 3ml prewashed Talon resin (washed in Wash buffer, 20mM Hepes pH 7.5, 0.5M NaCl, 5% Glycerol, 2 mM BME), incubated for 1 hour at 4°C, followed by centrifugation at 2K for 2 min and removal of supernatant, after which the Talon resin was washed once for 10 min with 40 ml Wash buffer containing 0.5 M NaCl (20mM Hepes pH 7.5, 0.5M NaCl, 5% Glycerol, 2 mM BME), centrifuged at 2K for 2 min, followed by two more washes of the resin with Wash buffer containing 1 M NaCl , then a wash with Wash buffer containing 0.5 M NaCl, followed by a wash with Wash buffer containing 0.5 M NaCl and 10 mM imidazole. The protein was eluted from the Talon resin with 4 sequential washes of the resin with Wash buffer containing 0.5 M NaCl containing 250 mM imidazole, each of which was mixed for 10 min prior to the low speed spin. Three fractions (elutions 1 and 2 as well as the 10 mM imidazole wash) were combined and diluted with 75 mL of NoSalt buffer to bring the NaCl to 0.2 M, followed by concentration to 5 mL using an Amicon (Millipore UFC901024), and loaded onto a Superdex 200 sizing column as described (Quartley et al., 2009). The laboratories of Drs. Eric Phizicky and Elizabeth Grayhack performed protein expression and purification.

5.2.2 Crystallization.

Initial crystallization conditions were identified using a high-throughput microbatch-under-oil method with a 1536 condition, incomplete-factorial based screen

(Luft et al., 2003). Conditions that produced crystals appearing suitable for optimization were prioritized according to their ease of cryoprotection (Kempkes et al., 2008) and optimized using a drop volume ratio versus temperature technique (DVR/T) (Luft et al., 2007). Crystals of the full-length ScGlnRS were prepared for diffraction by combining 3.5 µL of protein solution (13.9 mg/mL protein in 200mM NaCl, 5% (v/v) glycerol, 2mM DTT, 0.025% (w/v) NaN3, 20mM HEPES buffer, pH 7.5) with 2.0µL of precipitant solution (50mM NH4Br, 50mM KC2H3O2, 100mM HEPES, pH 7.5 and 20%(w/v) PEG 20K), incubated at 14°C. The optimized 16:9 protein to precipitant ratio and 14°C temperature were determined from DVR/T. Crystals appeared after four weeks.

5.2.3 Single crystal data collection and structure solution.

Crystals of the full-length protein were harvested and cryoprotected, then shipped to Stanford Synchrotron Radiation Lightsource (SSRL). Single crystal X-ray data was collected remotely on beamline 11-1 (Soltis et al., 2008). An initial excitation scan revealed the presence of zinc. Data were collected to 2.15 Å, integrated with XDS (Kabsch, 1998, b) and reduced with Scala (Evans, 2006), part of the CCP4 package (Collaborative Computational Project, 1994). Initial molecular replacement (MR) with Phenix (Adams et al., 2010) using a ~40% sequence homology *E. coli* Gln-tRNA synthetase, (PDB ID 1GTS (Perona et al., 1993)) failed, but a combined MR/SAD approach with the zinc signal was successful. An iterative process of Phenix refinement and manual model building through COOT (Emsley et al., 2010) was employed with validation using Molprobity (Chen et al., 2010). The coordinates were deposited as PDB ID 4H3S. Data collection, processing and refinement statistics are given in Table 5.2. The N-terminal region, residues 1-214, was unresolved in the electron density map. Crystals were dissolved by washing in cocktail solution at 14°C and centrifuging at 8,000 RPM, and repeated three times after discarding supernatant. 6x SDS solution was mixed with protein buffer and vortexed at 23°C. SDS-PAGE gels indicated that the NTD was present in the crystals. A similar procedure was followed to determine the structure of the His-tag purified protein. Western blot analysis again confirmed that the N-terminal residues were present in the crystals. PyMOL (Schrodinger, 2010) was used to generate symmetry mates, analyze solvent channels and crystallographic packing and to produce images.

5.2.4 Bioinformatics

The DISOPRED2 prediction of protein disorder server was used to predict disordered residues (Ward et al., 2004). Primary sequences were taken from the UniProt database (Magrane and Consortium, 2011). Structures of GlnRS from *E. coli* (PDB ID: 1NYL), *M. thermoautotrophicus* (PDB ID: 3AII), and *D. radiodurans* (PDB ID: 2HZ7) were taken from the Protein Data Bank (Bernstein et al., 1977; Deniziak et al., 2007; Nureki et al., 2010; Sherlin and Perona, 2003). The Basic Local Alignment Search Tool (BLAST) was used to calculate expectation values (E values) for comparing sequences (Altschul et al., 1990; Karlin and Altschul, 1990). tRNA sequences were taken from the tRNAdb 2009 online database (Juhling et al., 2009). Sequence data and structural data were combined for use in the PROMALS3D multiple sequence and structure alignment server (Pei et al., 2008) and the resulting alignment was used to identify structural motifs discussed in the text. ClustalW was used to calculate sequence similarity scores for homologous regions of adjacent proteins in Figure 5.16 (Goujon et al., 2010; Larkin et al., 2007).

5.2.5 Small Angle X-ray Scattering

Small Angle X-ray Solution Scattering data were collected on Beamline 4-2 of the SSRL (Smolsky et al., 2007) at a wavelength of 1.3 Å for eight consecutive two-second exposures. Solutions were prepared at five different concentrations ranging from 1.0 to 7.7 mg/mL for full-length Gln4 and 1.9 to 8.4 mg/mL for the CTD. Data were collected from the flow-through buffer of the final purification column and subtracted from the total protein solution scattering. The data were integrated with SasTool (Smolsky et al., 2007) and then examined SAXStats. No radiation damage was seen to be present in either sample. Concentration dependence resulting from repulsion appeared to be present in R_g estimates for the CTD, however the impact was less than 0.08 Å/mg/mL, and thus the highest concentration yielding the greatest signal to noise was used for further analysis. Lysozyme was used as a protein standard to estimate the molecular weight from I(0) extrapolated from the scattering curve. Porod-Debye analysis and Porod volume estimations were performed using a pre-release version of the software SCÅTTER (Rambo and Tainer, 2011). Ten *ab initio* shape reconstructions were generated by DAMMIF (Franke and Svergun, 2009) and averaged with DAMAVER (Volkov and Svergun, 2003). Ensemble modeling was carried out with the Ensemble Optimization Method (Bernado et al., 2007) using default parameters. Fifty identical runs were performed using 20 conformers in the ensemble, and rigid body modeling was performed similarly using only one conformer in the ensemble. OLIGOMER was used to assess possible mixtures of models (Konarev et al., 2003).

5.2.6 Molecular Dynamics

The initial model used for molecular dynamics simulation was generated as described in the text. All structural alignments were performed using the “fit” function of PyMOL using carbon alpha atoms. Molecular dynamics simulations were performed in GROMACS with the AMBER99SB force field (Hess et al., 2008; Hornak et al., 2006). The initial model was solvated using a cubic SPC/E water model (Berendsen et al., 1987) and neutralized with ions prior to minimization via steepest descents. Distance restraints were added to keep the zinc ion in place. The model was then equilibrated under an isothermal-isochoric ensemble for 100 picoseconds at 300K followed by equilibration under an isothermal-isobaric ensemble for 100 picoseconds. Simulations were then performed at the Center for Computational Resources (Green and Miller, 2003) on 512 processors. Total simulation time was 70 ns. DynDom (Hayward and Berendsen, 1998) was used to compare models. RMS fluctuation was calculated as the deviation from the starting structure after fitting to the starting structure.

5.3 Overall Structure

We solved the crystal structure of *Sc*GlnRS to 2.15 Å resolution (PDB ID: 4H3S) by single-wavelength anomalous dispersion (SAD) with a zinc anomalous signal identified from an initial excitation scan, coupled with molecular replacement using the crystal structure of the *E. coli* GlnRS as a template (Figure 5.1, Table 5.2). Of the 809 amino acids in the primary sequence of *Sc*GlnRS, ~30% were not resolved in the electron density. The unresolved regions are residues 1-187, which correspond to the NTD, residues 188-214, which encode a nonconserved 26-residue region that is predicted to be unstructured and links the NTD and CTD (Figure 5.2), and residues 672-678, which encodes a disordered loop within the CTD. The presence of the NTD in purified protein and in the crystals was confirmed by expressing *Sc*GlnRS with a His-6 tag on its N-terminus, followed by purification using Ni^{2+} affinity chromatography, crystallization, and antibody detection with anti-His antibody on dissolved crystals (Figure 5.3). This protein yielded structural data similar to that from the protein lacking the His-6 tag, with electron density still absent for residues 1-214.

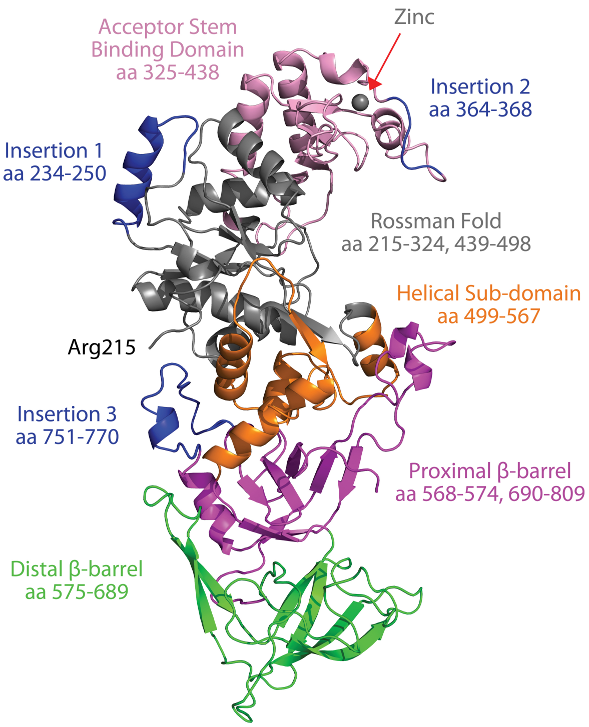


Figure 5.1 X-ray Crystal Structure of *Sc*GlnRS. Cartoon representation of the *Sc*GlnRS structure is shown color-coded according to domains and insertions relative to *E. coli* GlnRS. Residues 1-214 are missing in the electron density. Domains are labeled with corresponding amino acid numbers.

Table 5.2 Data collection and refinement statistics for full-length Gln4

	*Sc*GlnRS
Data collection	
Beamline	SSRL BL 11-1
Wavelength (Å)	1.169
Space group	P 3_1 2 1
Cell dimensions	
a, *b*, *c* (Å)	176.611, 176.611, 72.1884
α, β, γ (°)	90, 90, 120
Resolution (Å) *	52.49 – 2.15 (2.23 – 2.15)
R_{sym} or R_{merge} *	0.068 (0.348)
Completeness (%) *	99.86 (99.84)
I/σ*I* *	23.26 (2.98)
Unique reflections *	70276 (6963)
Redundancy *	11.2 (4.5)
Wilson B-factor (Å^2)	33.55

Refinement	
Resolution (Å)	52.49 – 2.15
$R_{\text{work/}}$ R_{free} *	0.1633/0.1826 (0.2232/0.2514)
No. atoms	10537
Protein	5043
Ligand/ion	75
Water	449
B-factors (Å^2)	
Protein	40.40
Ligand/ion	34.47
Water	44.90
R.m.s deviations	
Bond lengths (Å)	0.005
Bond angles (°)	0.90
Ramachandran favored (%)	98.0
Ramachandran outliers (%)	0.17
Clashscore	6.55

*Highest resolution shell is shown in parenthesis.

Since the NTD of *Sc*GlnRS is known to be structured (Grant et al., 2012), and is present in the full-length protein in the crystal, but no electron density is observed for this region, we conclude that the NTD is crystallographically disordered. It seems likely that the NTD is located in large solvent channels approximately 145-175 Å wide observed along the z-axis (Figure 5.4). Consistent with this, we note that the channels can comfortably accommodate the structure of the NTD (Grant et al., 2012), which has dimensions of ~25 x 25 x 75 Å and a macromolecular volume of ~25,000 Å^3. Moreover, the first residue resolved in the electron density, Arg215, is located adjacent to the solvent channels, and the preceding residues, 188-214, are predicted to be unstructured (Figure 5.2). It is likely that this region is the cause of the crystallographic disorder of the NTD.

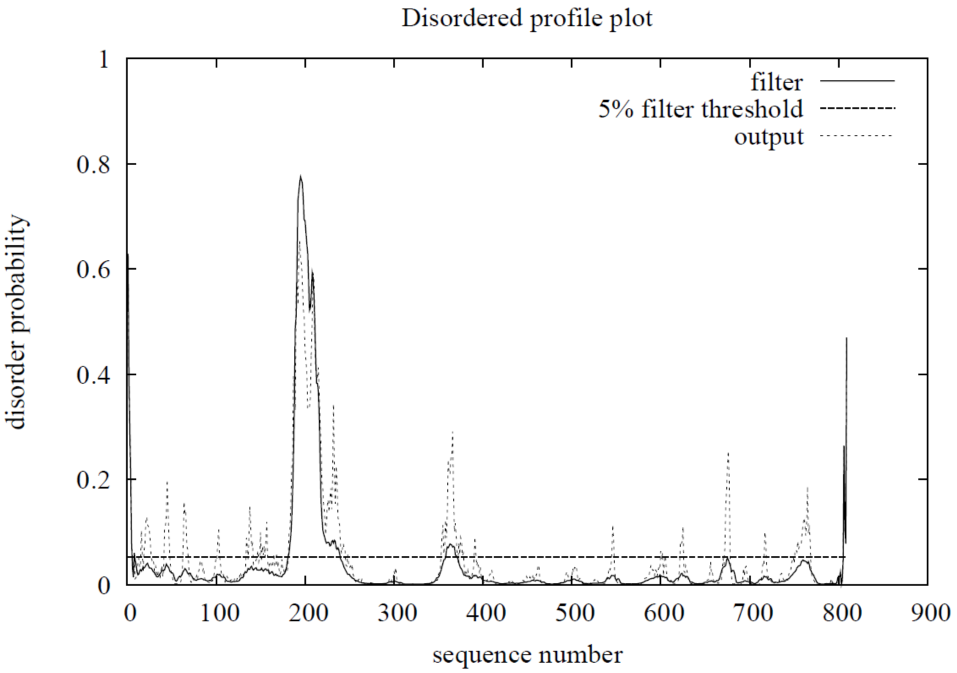


Figure 5.2 Disorder Prediction Analysis of the Primary Sequence of *Sc*GlnRS. The probability of disorder is shown on the y-axis and the residue number is shown on the x-axis. The linker connecting the N-terminal and C-terminal domains extends from residue 188 to 214. Disorder probability was calculated using DISOPRED2.

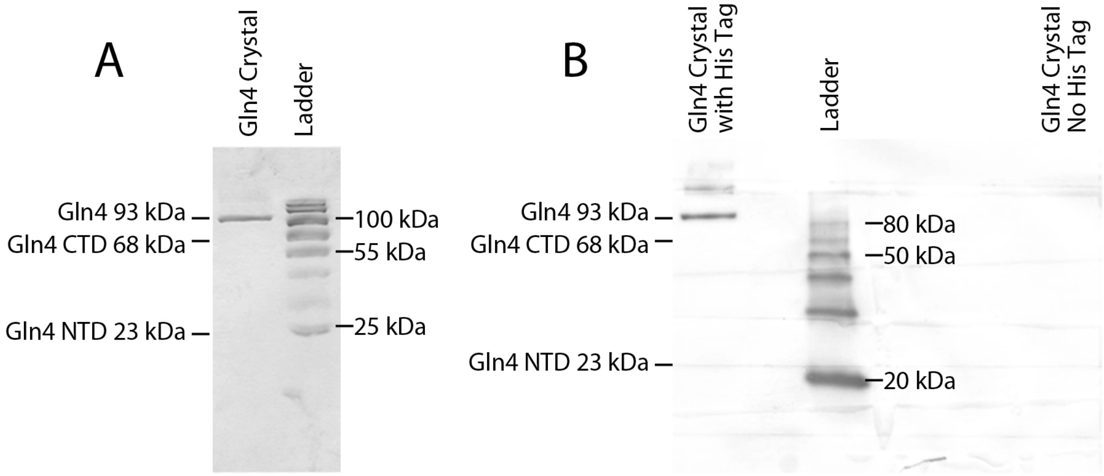


Figure 5.3 Dissolved *Sc*GlnRS crystals show only full-length *Sc*GlnRS protein. A. SDS PAGE gel showing dissolved Gln4 protein crystals is shown in the left lane, and the molecular weight ladder is shown in the right lane. Labels for the full-length protein, and both the NTD and CTD fragments are given. The presence of full-length Gln4 and absence of NTD and CTD fragments indicates that only the full-length protein is present in the crystal. B. Western blot using an anti-His antibody for crystals containing both His-tagged (left-most lane) and non-His-tagged (right-most lane) Gln4 protein. The molecular weight ladder is shown in the middle lane.

Using the naming convention previously described for *E. coli* GlnRS (Perona et al., 1993), the overall structure of *Sc*GlnRS consists of a folded NTD (residues 1-187); a positively charged 26-residue region linking the NTD and CTD (residues 188-214); a catalytic domain consisting of a Rossman fold domain (residues 215-324,439-498), an acceptor stem binding domain (325-438) and a helical sub-domain (499-567); and an anticodon binding domain consisting of proximal (568-574, 690-809) and distal (575-689) beta barrels (Figure 5.1). The structure of the CTD of *Sc*GlnRS (residues 215-809) is highly similar to *E. coli* GlnRS (PDB ID: 1NYL; RMSD=1.86Å, (Sherlin and Perona, 2003)) but has multiple insertions relative to the bacterial homolog. The *Sc*GlnRS catalytic domain (residues 215-567) is highly conserved in sequence and structure, showing an E value of 1 x 10^{-101} and a RMSD of only 1.03 Å. The *Sc*GlnRS anticodon binding domain (residues 568-809) shares moderate sequence homology but high structural conservation showing an E value of 5 x 10^{-32} and a RMSD of 1.13 Å.

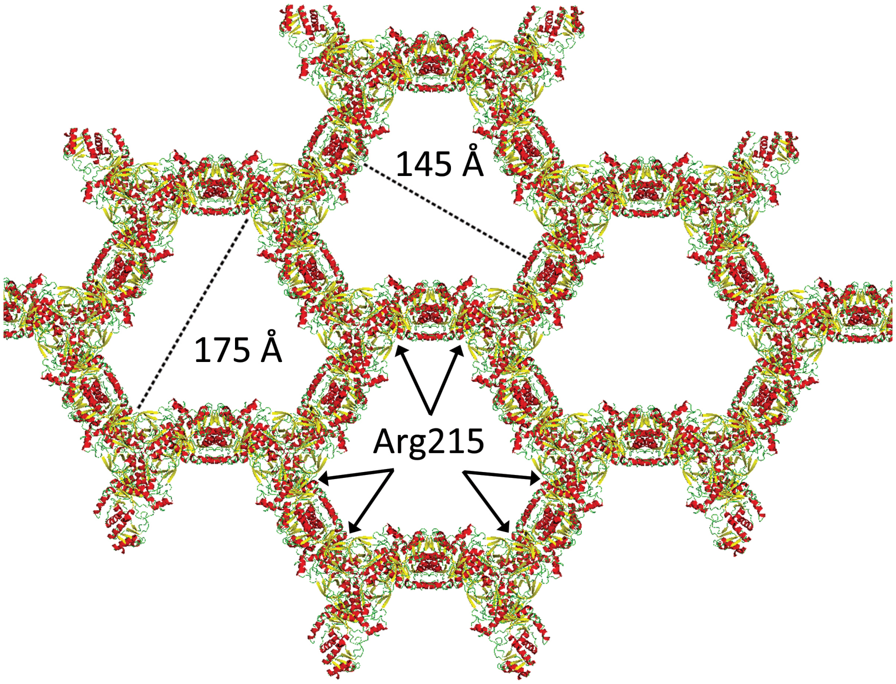


Figure 5.4 Crystallographic Packing Arrangement of *Sc*GlnRS. Shown oriented with the z-axis perpendicular to the page. Minimum and maximum diameters of solvent channels are labeled with dashed lines. Arg215 is labeled.

Multiple sequence analysis among eukaryotic GlnRS species revealed three insertions relative to *E. coli* GlnRS (Figure 5.1, Figure 5.5). Insertion 1, located at residues 234-250 of *Sc*GlnRS, is present in all eukaryotes, including those species lacking an NTD, although its length and sequence are not conserved among different kingdoms. Comparison with the *E. coli* GlnRS structure reveals that this insertion is part of the Rossmann fold domain in the CTD. Insertion 2 is located at residues 364-368 and is situated in a zinc-binding motif consisting of Cys346-X-Cys348-X_{24}-Cys372-X-His374 (Figure 5.6). The zinc-binding motif and Insertion 2, which are always found together, are present in fungi GlnRS species but is absent in other eukaryotes. Insertion 3 is a loop located at residues 751-770 and is also only conserved among fungi and absent in other eukaryotes.

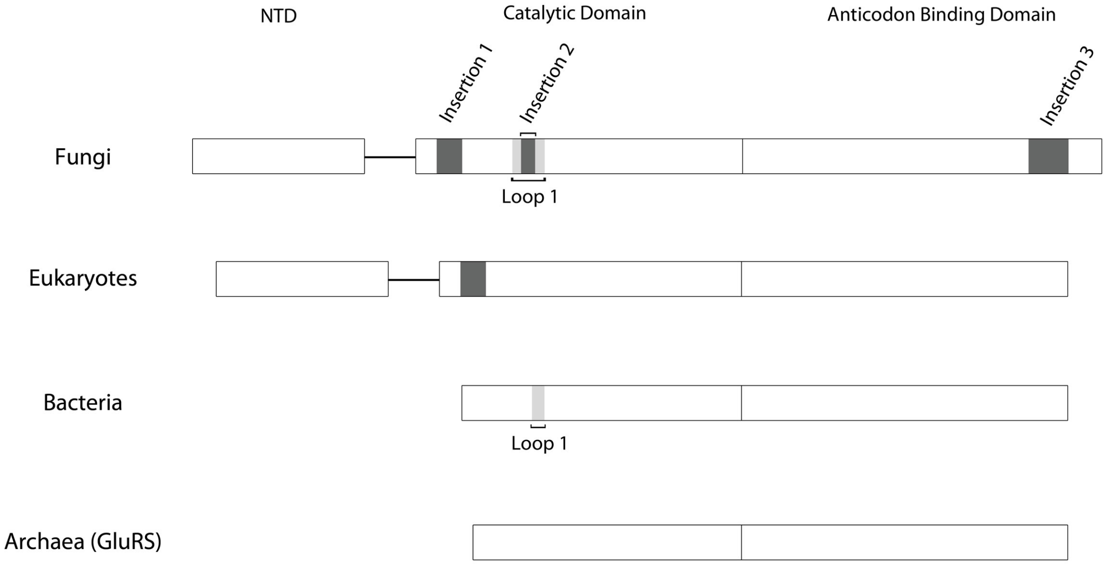


Figure 5.5 Domain Architecture of GlnRS. Domains for GlnRS are shown for Fungi, Eukaryotes (other than Fungi), and Bacteria and Archaeal GluRS. NTD = N-terminal Domain. Major domains are labeled NTD, Catalytic Domain and Anticodon Binding Domain. Insertion 1, Insertion 2, Insertion 3, and Loop 1 are labeled.

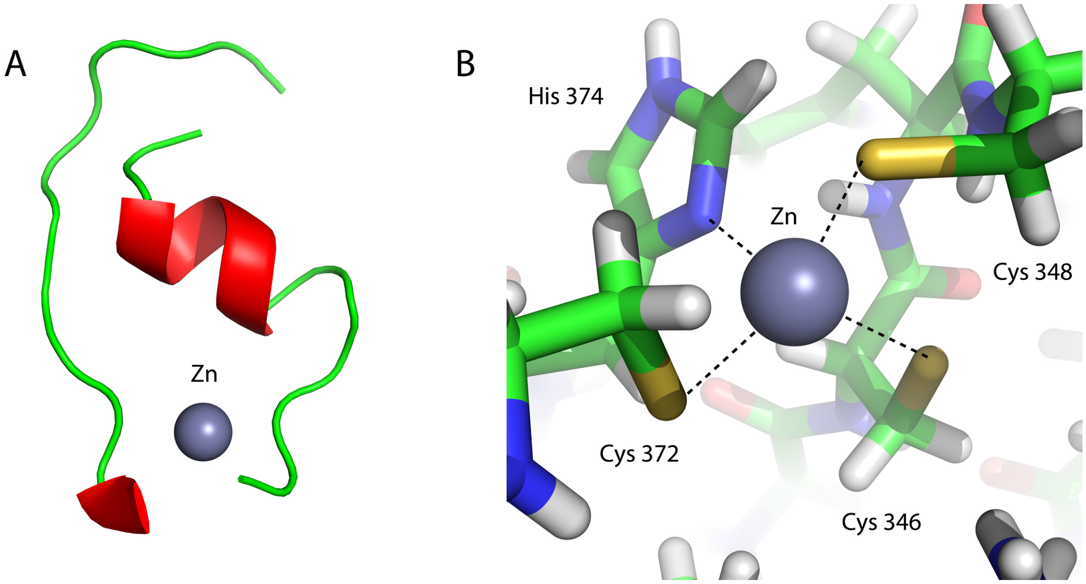


Figure 5.6 Zinc Finger Motif in *Sc*GlnRS. A. Location of zinc ion shown as gray sphere. B. Zinc binding motif residues shown coordinating zinc ion.

GlnRS selects both the correct amino acid substrate, glutamine, and the cognate tRNA^{gln} molecule. To gain insight into the tRNA^{gln} discrimination mechanism in *Sc*GlnRS, we performed further sequence analysis incorporating the archaeal non-discriminating GluRS, an evolutionary precursor of GlnRS that has the ability to recognize both tRNA^{gln} and tRNA^{glu} (Nureki et al., 2010). This revealed that Insertion 2 is part of a larger nine-residue loop, Loop 1, which is absent in archaeal GluRS (Figure 5.5). In *E. coli* GlnRS, Loop 1 is only four residues long and has been suggested to provide the ability to discriminate against the G1:C72 base pair in tRNA^{glu} and instead only recognizes tRNA^{gln} by disrupting the weak U1:A72 base pair (Nureki et al., 2010). Our results suggest that Loop 1 may play a different role in fungi and other eukaryotes than it does in bacteria. Although Loop 1 is present in *Sc*GlnRS, it differs in amino acid sequence compared to *E. coli* GlnRS and, with Insertion 2, is five residues longer. Additionally, while in *E. coli* the first base pair of tRNA^{gln} is U1:A72, in fungi and other eukaryotes the first base pair of tRNA^{gln} is G1:C72. It is therefore unlikely that the extended Loop 1 seen in *Sc*GlnRS functions similarly to *E. coli* GlnRS Loop 1 by disrupting the first base pair, since if the extended Loop 1 in *Sc*GlnRS were to disrupt the strong G1:C72 base pair in tRNA^{gln} it would also be likely to disrupt the weak U1:A72 base pair in tRNA^{glu} and fail to discriminate between them. While our observations are based on length, it is unknown whether sequence-specific contacts in Loop 1 contribute tRNA discrimination ability to *Sc*GlnRS. However, in all eukaryotes other than fungi, Loop 1 is entirely absent and eukaryotic GlnRS more closely resembles archaeal GluRS suggesting that Loop 1 is not likely to be a critical tRNA discriminatory motif in eukaryotes with the possible exception of fungi.

5.4 Motion of NTD in Solution is Limited in the Absence of tRNA

We collected small angle X-ray scattering (SAXS) data on the full-length *Sc*GlnRS and the CTD, (Figure 5.7), because SAXS is a solution based technique that is sensitive to the overall shape and size of a molecule and can provide insight into protein dynamics (Grant et al., 2011; Putnam et al., 2007). Analysis of the pair distance distribution function yields a maximum particle dimension of ~147 Å and a radius of gyration (R_g) of 43.51 Å for the full length protein, in close agreement with the Guinier estimate of 43.32 Å, and shows a large shoulder compared to the CTD, likely corresponding to the NTD appended domain (Figure 5.7, D). A Porod-Debye analysis gave a Porod exponent of 3.4, suggesting mild flexibility (Rambo and Tainer, 2011), compared to an exponent of 4.0 for the CTD, reflecting rigidity (Figure 5.7, C). The molecular weight estimated from the Porod volume was 92.6 kDa, similar to the molecular weight estimated from the forward scattering, I(0), of 89.1 kDa and the expected molecular weight of 93.1 kDa, demonstrating that the volume occupied by the NTD is approximately limited to the size of the domain and does not occupy a larger region of conformational space.

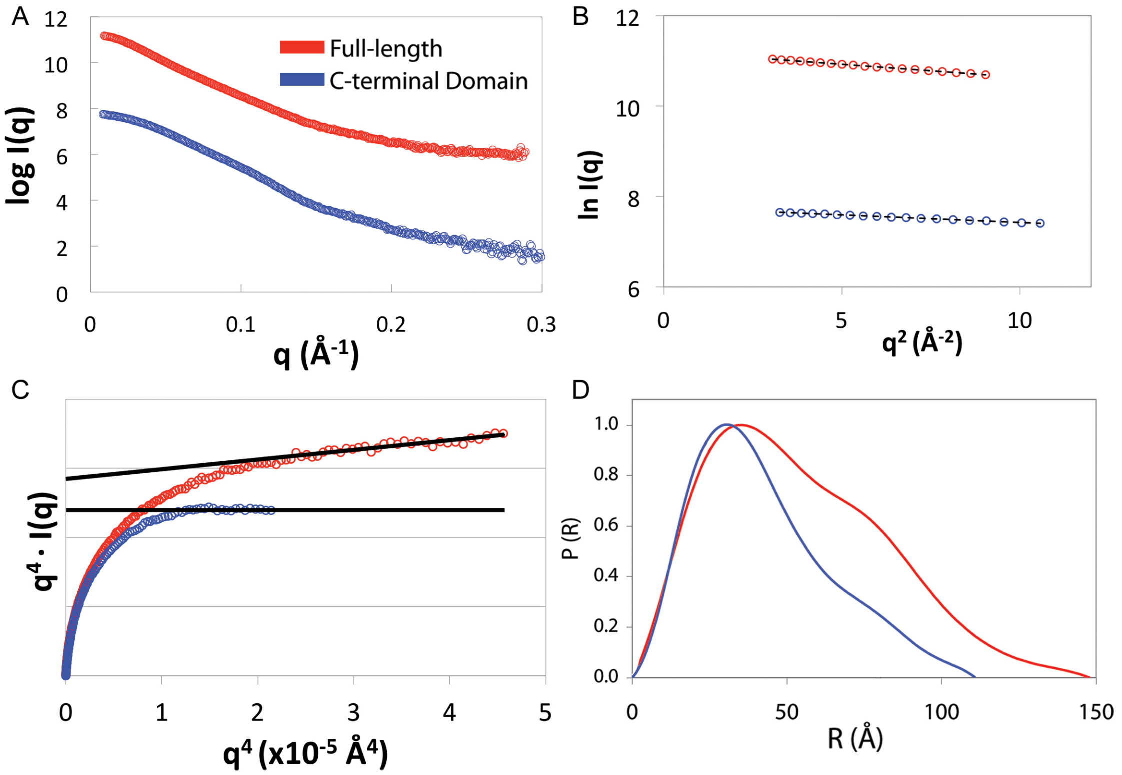


Figure 5.7 SAXS Data of Full-length *Sc*GlnRS. For clarity, all plots have been scaled. A. The log of the scattering intensity plotted as a function of momentum transfer of full-length *Sc*GlnRS (red) and the CTD (blue). B. Guinier plots of full-length *Sc*GlnRS (red) and the CTD (blue). The dashed line shows linearity in Guinier region. C. Porod-Debye plots of full-length *Sc*GlnRS (red) and the CTD (blue). The linear fits are based on the corresponding Porod region. Full-length *Sc*GlnRS has a Porod exponent of 3.4 and the CTD has a Porod exponent of 4.0. D. Pair distribution functions of full-length *Sc*GlnRS (red) and the CTD (blue).

Ten *ab initio* envelope reconstructions of full-length *Sc*GlnRS were created and averaged, exhibiting an average Normalized Spatial Discrepancy (NSD) of 0.633 ± 0.022, reflecting uniqueness and low flexibility (Bernado et al., 2007). A rigid body model of the CTD and the NTD of *Sc*GlnRS superposed onto the *ab initio* envelope is shown in Figure 5.8. The rigid body model fits well to the experimental scattering data with a χ^2 = 1.82. *Ab initio* reconstructions and rigid body modeling do not take into account the dynamic information present in SAXS data but represent an average of all conformations present in solution. Ensemble modeling can overcome this by representing a protein structure as an ensemble of multiple conformations. Using the Ensemble Optimization

Method (Bernado et al., 2007) with multiple conformations of the NTD did improve the overall fit to the scattering data (χ^2 = 1.75); however, the improvement was only marginal. The R_g distribution of conformers present in the ensemble was quite broad compared to the pool of 10,000 random conformations (Figure 5.9). Normally this would suggest considerable flexibility. However, as a control we simulated SAXS data for the rigid body model, which is completely inflexible, and performed the ensemble modeling. Interestingly, we found the R_g distribution to also be quite broad and very similar to that seen for the experimental SAXS data (Figure 5.9). This is counterintuitive since a perfectly rigid system, which is represented by the simulated scattering profile of the rigid body model, should not show a broad distribution but a sharp, narrow peak. This may be caused by the somewhat elongated, more rod-like shape of the full-length protein and reflect a limitation of ensemble modeling. Since the R_g distributions are similar for both the simulated data from the rigid body model and the experimental scattering profile, we conclude that the large degree of flexibility suggested by the broad R_g distribution from ensemble modeling is misleading. However, the limited flexibility from the Porod-Debye analysis, the low NSD for multiple *ab initio* envelope reconstructions, and the marginal improvement from an ensemble model compared to the rigid body model demonstrate that the mobility of the NTD in solution is limited.

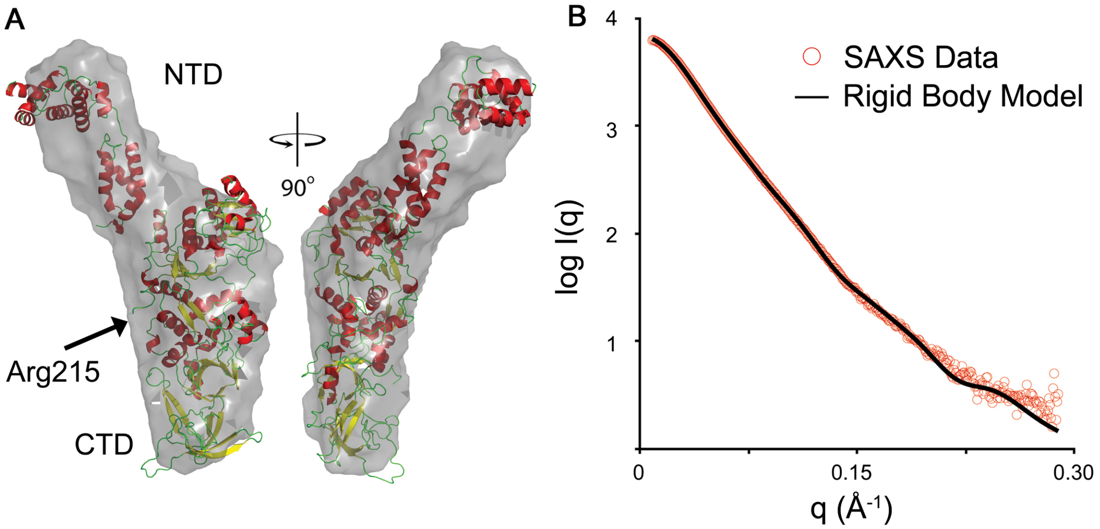


Figure 5.8 SAXS Rigid Body Model of Full-length *Sc*GlnRS. A. Two orientations rotated 90° relative to each other are shown in cartoon representation and colored according to secondary structure and superposed onto the *ab initio* envelope shown in gray. B. The calculated scattering of the rigid body model (solid black line) fit to the experimental SAXS data (red circles).

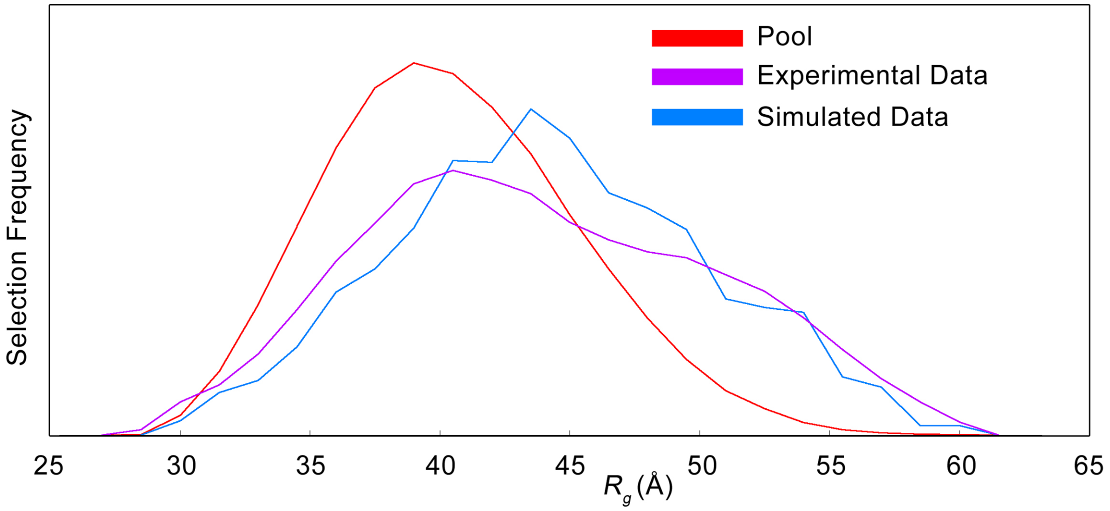


Figure 5.9 R_g Distribution of Ensembles Calculated by the Ensemble Optimization Method. The R_g distribution of the pool of 10,000 random conformers (red) is shown with the distribution of the ensemble calculated using experimental SAXS data (purple) and using the simulated scattering profile of the rigid body model (blue).

5.5 A Model of Full-length *Sc*GlnRS Bound to tRNA^{gln} Suggests a Substantial Conformational Reorientation of the NTD

We used the structure of *E. coli* GlnRS bound to tRNA (Rath et al., 1998) and the structure of the transamidosome from *Thermus thermophilus* (Ito and Yokoyama, 2010) to obtain a model of eukaryotic glutaminyl-tRNA synthetase bound to tRNA^{gln}. First, the *E. coli* GlnRS co-crystal structure solved with tRNA^{gln} (PDB ID: 1QTQ) was superposed onto the CTD residues 215-809 of *Sc*GlnRS. The structure of the tRNA molecule was then extracted, providing an initial model of *E. coli* tRNA^{gln} complexed with the CTD of *Sc*GlnRS. The nucleotide sequence of *E. coli* tRNA^{gln} was computationally mutated to match that of *S. cerevisiae* tRNA^{gln} and a geometric minimization performed using ModeRNA (Rother et al., 2011a; Rother et al., 2011b). To correctly orient the NTD of *Sc*GlnRS in the full-length complex with tRNA, we utilized known structural homology of the NTD with subunit B of the GatCAB amidotransferase from *T. thermophilus* (PDB ID: 3AL0), which was solved in complex with tRNA. The tail subdomain (residues 119-187) of the NTD of *Sc*GlnRS was superposed to the tail subdomain of GatB due to the high level of structural homology between these two regions (Grant et al., 2012). MODELLER (Sali et al., 1995) was used for *de novo* modeling of the predicted flexible linker (residues 188-214) (Figure 5.10).

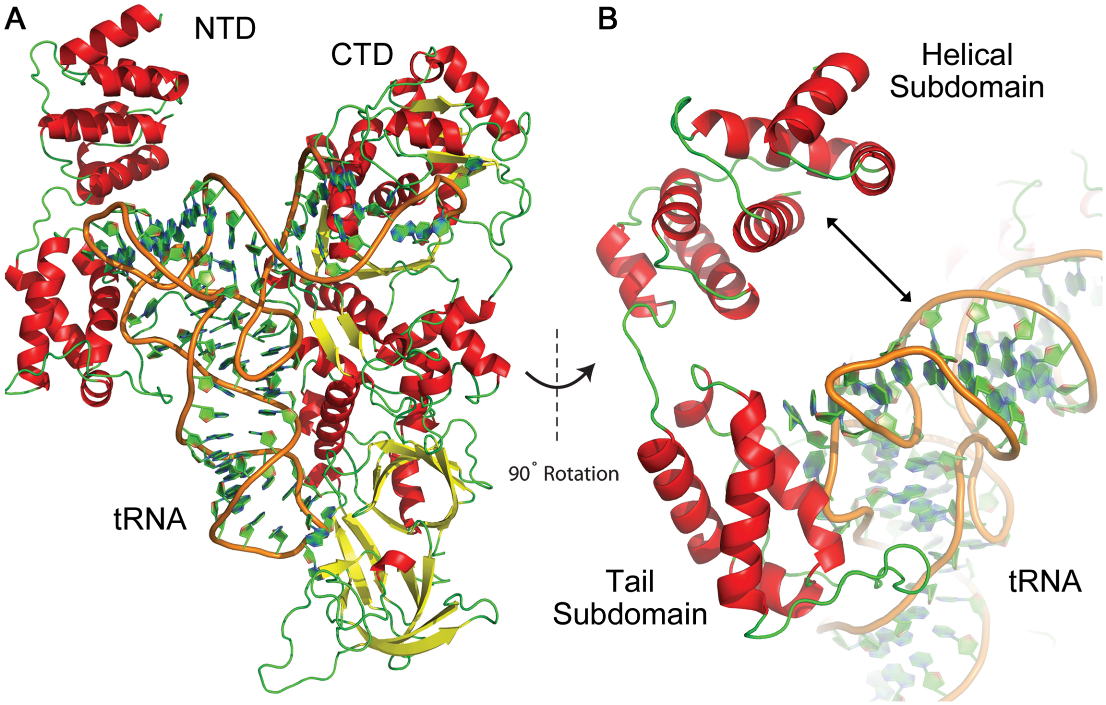


Figure 5.10 Homology Model of Full-length *Sc*GlnRS Bound to tRNA^{gln}. A. Full-length *Sc*GlnRS shown bound to tRNA^{gln}. B. Enlarged and rotated model showing gap between NTD helical subdomain and tRNA molecule.

Our model of the full-length *Sc*GlnRS bound to tRNA^{gln} shows a significant change in the NTD position when compared to the tRNA^{gln}-free, SAXS-derived conformation (Figure 5.11). The model shows a ~160° rotation and a ~40 Å translation of the NTD with respect to the solution conformation according to an analysis by DynDom (Hayward and Berendsen, 1998). Fitting the simulated scattering of the protein portion of the protein-tRNA complex to the experimental SAXS data resulted in a poor fit, yielding a χ^2 = 12.25 compared to 1.82 for the rigid body model (Figure 5.11). The limited flexibility of the NTD, coupled with the poor fit of the simulated scattering of the protein portion of the model bound to tRNA^{gln}, suggests that without tRNA bound, this conformation does not exist in solution. Analysis with OLIGOMER (Konarev et al., 2003), to see if a mixture of the rigid body model and the homology model exists simultaneously in solution, supported this result showing that only the rigid body model exists in solution, while the homology model does not. Taken together, these observations suggest that CTD binding of tRNA^{gln} induces substantial conformational reorientation of the NTD required for interactions with tRNA^{gln}.

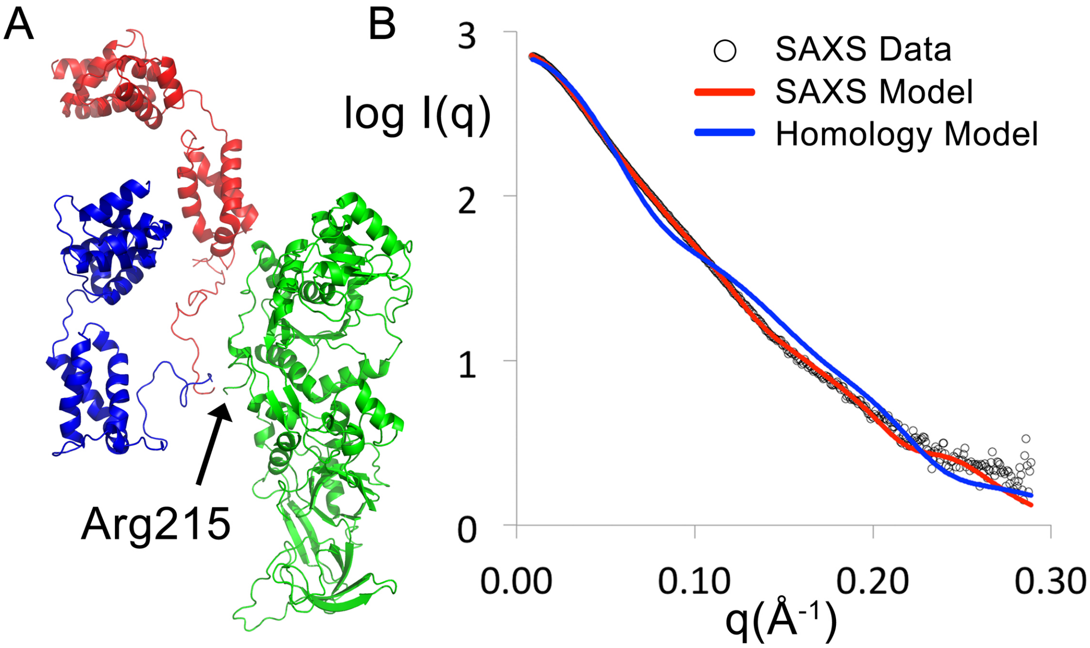


Figure 5.11 Comparison of NTD Position Before and After tRNA Binding. A. *Sc*GlnRS is shown in cartoon representation. The NTD position prior to tRNA binding determined by SAXS rigid body modeling is shown in red. The NTD position upon binding tRNA predicted by homology modeling is shown in blue. The CTD is shown in green. B. The calculated scattering of the rigid body model (red line) and the homology model (blue line) are shown fitted to the experimental SAXS data (circles).

5.6 Conformational Change in NTD Subdomains is Predicted Upon Interaction with tRNA^{gln}

In the compiled model, the NTD appears to be in an "open" conformation resulting in its helical subdomain (residues 1-110) being too distant from the tRNA molecule to form stable contacts (Figure 5.10, B). This conformation more closely resembles that observed in the *Staphylococcus aureus* GatCAB structure solved without tRNA bound (Grant et al., 2012) than the conformation seen in tRNA-bound *T. thermophilus* GatCAB. Thus, we considered that the molecule might undergo a conformational change upon tRNA binding. To probe the dynamics of the full-length glutaminyl-tRNA synthetase in complex with tRNA^{gln}, we carried out 70 nanoseconds of molecular dynamics simulation. The model stabilized within approximately 10 ns of simulation time and remained relatively unchanged for the remaining 60 ns (Figure 5.12), showing that the simulation time captured the relevant dynamics. The resulting molecular dynamics trajectory (Figure 5.12, Figure 5.13, Figure 5.14) shows that the helical subdomain of the NTD rotates about a conserved hinge by 22.9°, forming electrostatic interactions between the positively charged side chains of lysine residues 19, 20, and 26, which have been previously implicated in tRNA binding integrity (Wang et al., 2000), and the negatively charged phosphate backbone of tRNA. This is in agreement with the tRNA-bound conformation seen in the *T. thermophilus* GatB structure. This large conformational change in the NTD could provide *Sc*GlnRS the heightened affinity for tRNA^{gln} since it has been shown that mutations in the hinge residues significantly reduce tRNA binding (Grant et al., 2012). In addition our modeling predicts that residues K29, K63, G64, T65, and D66 of the helical subdomain of the NTD make several contacts with the CTD, including contacts with residues P238 and M241 of Insertion 1. Since Insertion 1 was shown above to be exclusive to eukaryotes we speculate that this insertion may provide a means of communication between the NTD and CTD.

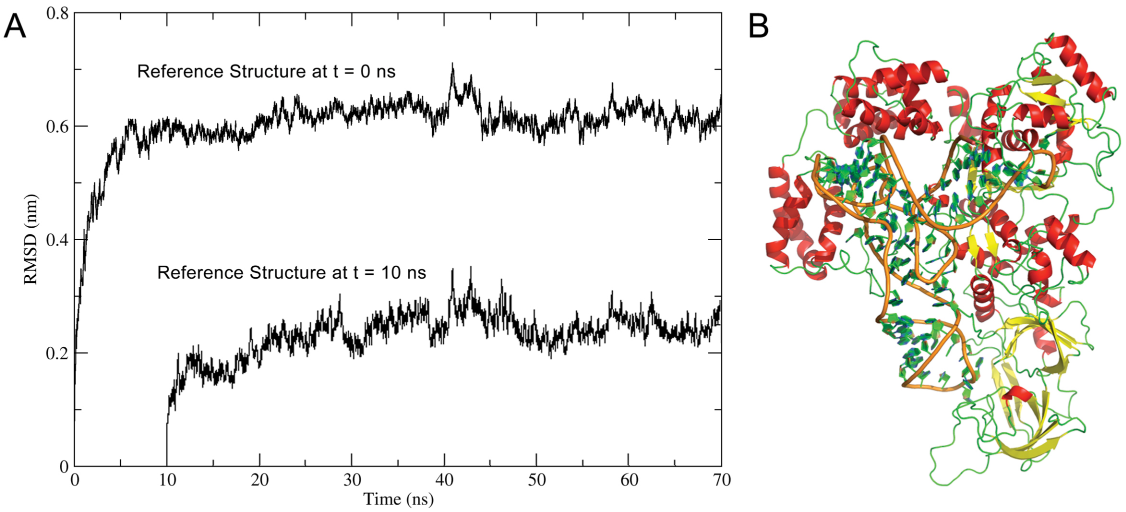


Figure 5.12 Molecular Dynamics Simulation of *Sc*GlnRS bound to tRNA^{gln}. A. Plot of backbone RMSD of molecular dynamics trajectory as a function of time fit to the structure at time t=0 ns and t=10 ns. B. *Sc*GlnRS bound to tRNA^{gln} after 70 nanoseconds of molecular dynamics simulation colored according to secondary structure.

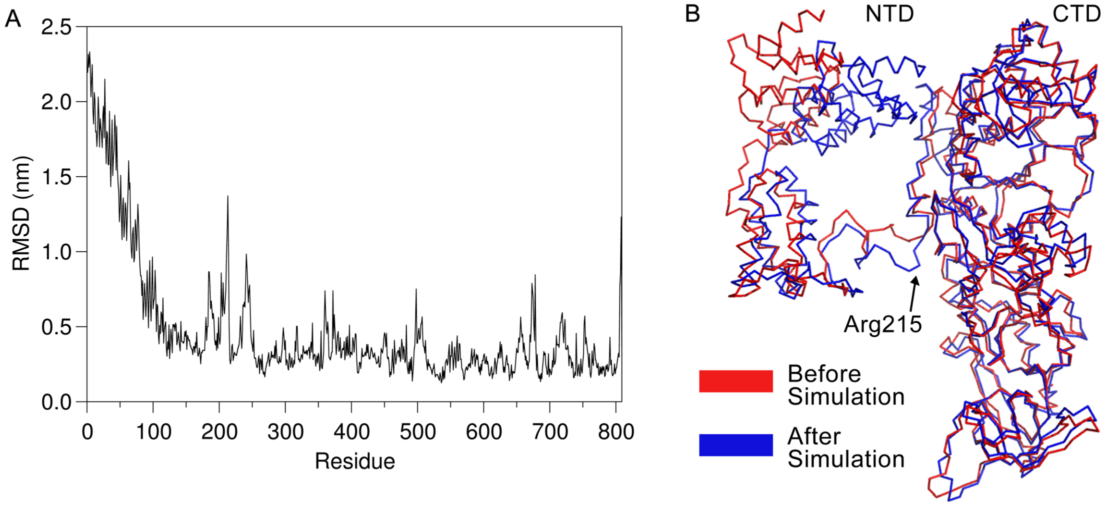


Figure 5.13 NTD undergoes conformational change after binding to tRNA. A. Plot of backbone RMS fluctuation as a function of residue. The RMS has been calculated as the deviation from the starting structure. B. Structure of *Sc*GlnRS before (red) and after (blue) molecular dynamics simulation. The tRNA molecule has been removed for clarity.

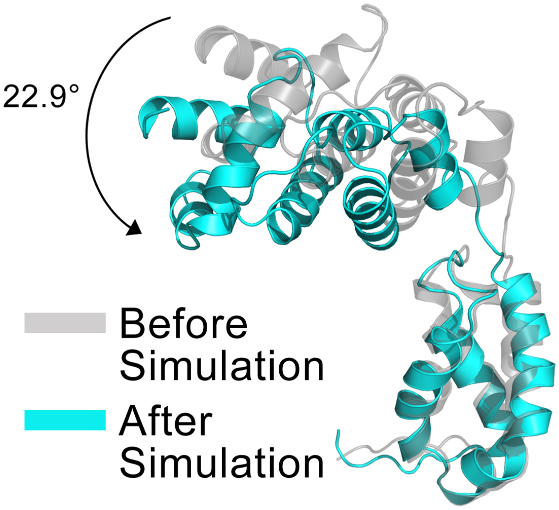
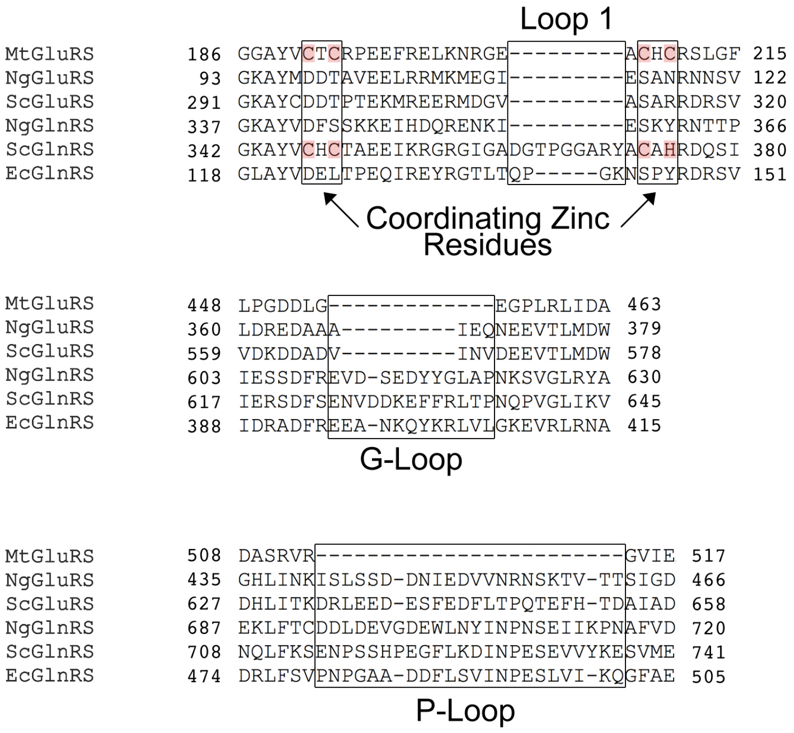


Figure 5.14 Comparison of NTD Structure Before and After Molecular Dynamics Simulation. The position of the NTD before the simulation is shown in gray and after the simulation in cyan. The solid black arrow shows the degree and direction of angular motion calculated by DynDom.

5.7 GlnRS Phylogenetic Analysis

Having produced a model of the structure and binding mechanism of a eukaryotic glutaminyl-tRNA synthetase-tRNA^{gln} complex, we performed a phylogenetic analysis. Two routes of gln-tRNA^{gln} formation exist in organisms. The indirect route is believed to be the more ancient method requiring a non-discriminating GluRS to misacylate tRNA^{gln} with glutamate, after which an amidotransferase converts the glu-tRNA^{gln} to gln-tRNA^{gln} (Curnow et al., 1997). The more recent route is the direct method, commonly seen among other aminoacyl-tRNA synthesis mechanisms, which utilizes a dedicated GlnRS to attach glutamine to the cognate tRNA molecule. GlnRS first arose in eukaryotes from a progenitor non-discriminating GluRS and was subsequently transferred to bacteria (Lamour et al., 1994). An important feature of eukaryotic GlnRS that is absent in prokaryotes is the presence of the appended NTD. The presence of appended domains in eukaryotic homologs of proteins is of special importance for aminoacyl-tRNA synthetases, since it has been shown that this particular class of enzymes experiences domain addition more often than most enzyme classes (Guo et al., 2010). If the NTD was present before the transfer to bacteria, then the domain must have been lost at some point in the evolutionary process.

To assess the likely sequence of evolutionary events leading to the eukaryotic development and prokaryotic retention of the GlnRS enzyme, we performed a phylogenetic and structural motif analysis using the amino acid sequences and available high-resolution crystal structures of various GluRS and GlnRS enzymes throughout the three domains of life. Using structures of *Mt*GluRS, *Ec*GlnRS, and *Sc*GlnRS, we were able to determine the specific amino acids forming structural motifs and translate this information to sequences where structure is not available (Figure 5.15). Of particular note is the publication of the *Naegleria gruberi* genome (Fritz-Laylin et al., 2010), which is believed to be the most ancient eukaryotic genome sequenced to date and therefore provides insight into the phylogenetic history of early eukaryotic organisms following the split from archaea.



```
MtGluRS     186 GGAYVCTCRPEEFRELKNRGE---------ACHCRSLGF 215
NgGluRS      93 GKAYMDDTAVEELRRMKMEGI---------ESANRNNSV 122
ScGluRS     291 GKAYCDDTPTEKMREERMDGV---------ASARRDRSV 320
NgGlnRS     337 GKAYVDFSSKKEIHDQRENKI---------ESKYRNTTP 366
ScGlnRS     342 GKAYVCHCTAEEIKRGRGIGADGTPGGARYACAHRDQSI 380
EcGlnRS     118 GLAYVDELTPEQIREYRGTLTQP-----GKNSPYRDRSV 151

MtGluRS     448 LPGDDLG-------------EGPLRLIDA 463
NgGluRS     360 LDREDAAA---------IEQNEEVTLMDW 379
ScGluRS     559 VDKDDADV---------INVDEEVTLMDW 578
NgGlnRS     603 IESSDFREVD-SEDYYGLAPNKSVGLRYA 630
ScGlnRS     617 IERSDFSENVDDKEFFRLTPNQPVGLIKV 645
EcGlnRS     388 IDRADFREEA-NKQYKRLVLGKEVRLRNA 415

MtGluRS     508 DASRVR------------------------GVIE 517
NgGluRS     435 GHLINKISLSSD-DNIEDVVNRNSKTV-TTSIGD 466
ScGluRS     627 DHLITKDRLEED-ESFEDFLTPQTEFH-TDAIAD 658
NgGlnRS     687 EKLFTCDDLDEVGDEWLNYINPNSEIIKPNAFVD 720
ScGlnRS     708 NQLFKSENPSSHPEGFLKDINPESEVVYKESVME 741
EcGlnRS     474 DRLFSVPNPGAA-DDFLSVINPESLVI-KQGFAE 505
```

Figure 5.15 Conserved Structural Motifs of GlxRS Family of Enzymes. Sections of residues are shown to demonstrate the conservation of select motifs identified through PROMALS3D sequence and structure alignment. Structures of *Mt*GluRS, *Sc*GlnRS and

***Ec*GlnRS were used to aid sequence alignment and identify motifs. Motifs are boxed and labeled as described in the text. Zinc coordinating residues of the zinc-binding motif are highlighted in red.**

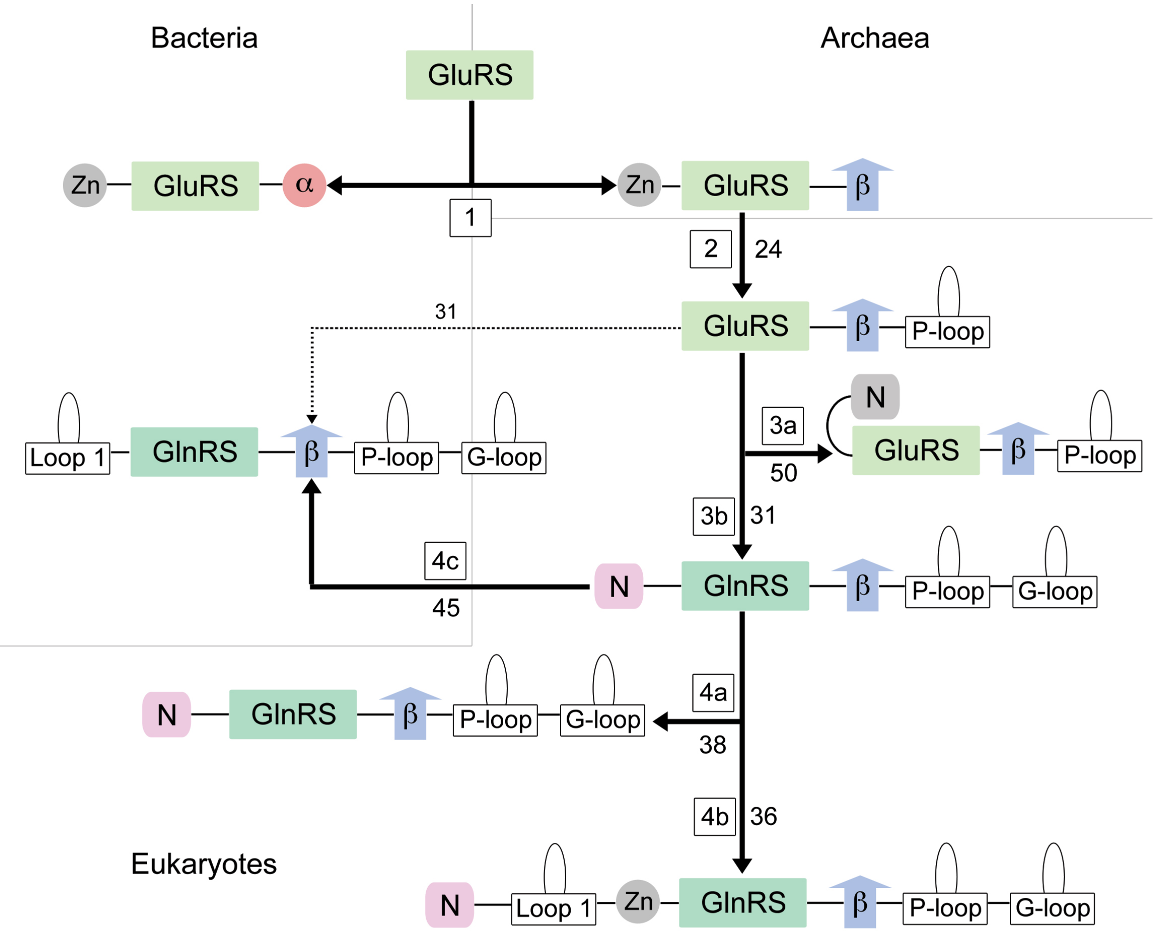


Figure 5.16 GlxRS Evolutionary Pathway. Boxed numbers refer to steps in the pathway. Unboxed numbers are ClustalW scores for adjacent protein sequences. Further details are given in the text. Steps: 1. The last universal common ancestor containing only the GluRS catalytic domain splits and adds an alpha helical anticodon binding domain in bacteria and a beta barrel anticoding binding domain in archaea. 2. The archaeal GluRS evolves in early eukaryotes by adding the P-loop in the anticodon binding domain. 3a. The early eukaryotic GluRS evolves into the modern eukaryotic GluRS by adding an NTD. 3b. The early eukaryotic GluRS evolves into the first dedicated GlnRS in early eukaryotes by altering the catalytic domain, and adding the G-loop and the NTD. 4a. The early eukaryotic GlnRS experiences little evolutionary change in modern eukaryotes. 4b. Only in fungi, Loop 1 is added with the zinc-binding motif. 4c. The early eukaryotic GlnRS evolves into the bacterial GlnRS by losing the appended NTD and gaining Loop 1. The dashed line represents the less likely route of GlnRS horizontal gene transfer to prokaryotes.

The catalytic domain of GluRS was present in the last universal common ancestor and subsequently added an α-helical anti-codon binding domain in bacteria whereas a β-barrel anti-codon binding domain was added in archaea (Siatecka et al.,

1998) (Figure 5.16, step 1). The zinc-binding motif is present in both the bacterial (*E. coli)* and archaeal (*M. thermoautotrophicus)* GluRS sequences (Figure 5.15), and the presence of zinc has additionally been confirmed in the archaeal GluRS crystal structure (Nureki et al., 2010).

As archaea evolved into early eukaryotes (represented by *Naegleria gruberi*), the non-discriminating GluRS lost the zinc finger motif and added a distinct loop to the proximal beta barrel (P-loop, step 2), providing a communication link between the anticodon binding domain and the catalytic core of the enzyme (Uter and Perona, 2004) possibly providing a means of substrate discrimination to GluRS. The sequence of *Sc*GluRS is highly similar to *Ng*GluRS (ClustalW score of 50), suggesting little has changed in modern eukaryotic GluRSs (step 3a), except for an additional N-terminal appended domain in *Sc*GluRS that has neither sequence nor structural homology to the appended NTD of *Sc*GlnRS (PDB ID: 2HRA, (Simader et al., 2006)). The acquisition of an additional loop (G-loop) in the anticodon binding region that recognizes discriminatory base G36 of tRNA^{gln}, which is C36 in tRNA^{glu}, a sequence modified P-loop, and the addition of a large NTD gave the enzyme a heightened affinity for tRNA^{gln} and the ability to discriminate against tRNA^{glu} (Grant et al., 2012; Nureki et al., 2010; Saha et al., 2009). Coupled with various modifications in the catalytic core affecting amino acid substrate recognition (Bullock et al., 2008; Nureki et al., 2010), these changes resulted in the first dedicated GlnRS enzyme (step 3b). The subsequent reemergence of the zinc binding motif, and the insertion of the unpairing loop discussed earlier, Loop 1, in the acceptor stem binding domain gave rise to the modern eukaryotic GlnRS (step 4b).

This leaves two possible avenues for the horizontal gene transfer resulting in prokaryotic GlnRS. The first possible pathway is that an early eukaryotic GluRS was transferred to bacteria and evolved in a convergent fashion to form GlnRS (dotted line in Figure 7). The second pathway (step 4c) is that the early eukaryotic GlnRS was transferred to bacteria and lost the appended NTD and acquired Loop 1 whose length and sequence is dissimilar to modern eukaryotes (Figure 5.15). The presence of the G-loop discriminatory motif and the significantly higher ClustalW score (45 vs. 31) suggest that it is more likely that GlnRS was first formed in eukaryotes and then subsequently transferred to prokaryotes, losing the appended NTD.

5.8 Conclusion

The structure of the CTD of *Sc*GlnRS presented here was shown to be highly similar to the *E. coli* GlnRS structure. However, multiple insertions relative to *E. coli* GlnRS revealed insights into one structural motif common to all eukaryotes and two motifs specific to fungi. Since most eukaryotic GlnRS species lack the bacterial unpairing loop that is proposed to play a role in tRNA discrimination, then another mechanism must be employed to discriminate between $tRNA^{glu}$ and $tRNA^{gln}$ in eukaryotic GlnRS species. Therefore, in eukaryotes there may be a compensating mechanism to discriminate between G1:C72 of $tRNA^{gln}$ from U1:A72 of $tRNA^{glu}$, or, as in the case of non-discriminating archaeal GluRS (Nureki et al., 2010), the first base pair does not play a significant role in tRNA discrimination.

Our structure-based models of the first full-length eukaryotic GlnRS with and without $tRNA^{gln}$ bound suggest that CTD binding to tRNA results in a large conformational reorientation of the NTD allowing for interactions between the NTD and the tRNA. Given the distinct increases in K_M and K_D for $tRNA^{gln}$ following deletion of the NTD (Grant et al., 2012; Ludmerer et al., 1993), the solution model of the full-length *Sc*GlnRS presented here suggests that the NTD plays a direct role in tRNA binding. Our molecular dynamics simulation revealed that the helical and tail subdomains of the NTD undergo a hinge motion after binding to tRNA, allowing for tighter binding between the NTD and tRNA. Our structural results and modeling suggest the intriguing possibility that the NTD communicates with the CTD through Insertion 1, which is found in all eukaryotes. The absence of such an interaction may explain the loss of the NTD in bacterial GlnRS evolution. In addition, since the NTD and the active site of *Sc*GlnRS are too distant to interact directly, and since deletion of the NTD also increases K_M for glutamine and ATP, it seems plausible that the effects on glutamine and ATP are due to the concerted conformational changes in *Sc*GlnRS that occur upon tRNA binding as was observed in *E. coli* GlnRS (Grant et al., 2012; Ludmerer et al., 1993; Rath et al., 1998; Sherlin and Perona, 2003).

We have shown that, in addition to its usefulness in high-throughput applications of SAXS, SAXStats is also useful as a tool to objectively evaluate SAXS data quality for specific biological systems. In chapter 2 we described how SAXS data is most effective when used as a complementary tool with other structural data. In the case of eukaryotic glutaminyl-tRNA synthetase we found that the crystallographic data failed to yield information about the full-length Gln4, lacking electron density for residues 1-214. This suggested that the NTD was disordered in the crystal and may be mobile in solution. Using SAXS we were able to place the NTD relative to the CTD and showed that the mobility of the appended domain is limited in solution. Additionally, using structural homology with *Ec*GlnRS and GatB coupled with molecular dynamics we proposed a model of the full-length *Sc*GlnRS bound to tRNA. Using multiple structural, biochemical and bioinformatics tools we have developed a far more complete understanding of eukaryotic GlnRS than could be achieved with any method alone.

5.9 References

Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.W., Kapral, G.J., Grosse-Kunstleve, R.W.*, et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta crystallographica *66*, 213-221.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Berendsen, H.J.C., Grigera, J.R., and Straatsma, T.P. (1987). The missing term in effective pair potentials. The Journal of Physical Chemistry *91*, 6269-6271.

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc *129*, 5656-5664.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem *80*, 319-324.

Bullock, T.L., Rodriguez-Hernandez, A., Corigliano, E.M., and Perona, J.J. (2008). A rationally engineered misacylating aminoacyl-tRNA synthetase. Proc Natl Acad Sci U S A *105*, 7428-7433.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D *66*, 12-21.

Collaborative Computational Project, N. (1994). The CCP4 suite: programs for protein crystallography. Acta Cryst D Biological Crystallography *50*, 760-763.

Curnow, A.W., Hong, K., Yuan, R., Kim, S., Martins, O., Winkler, W., Henkin, T.M., and Soll, D. (1997). Glu-tRNAGln amidotransferase: a novel heterotrimeric enzyme required for correct decoding of glutamine codons during translation. Proc Natl Acad Sci U S A *94*, 11819-11826.

Cusack, S., Berthet-Colominas, C., Hartlein, M., Nassar, N., and Leberman, R. (1990). A second class of synthetase structure revealed by X-ray analysis of Escherichia coli seryl-tRNA synthetase at 2.5 A. Nature *347*, 249-255.

Deniziak, M., Sauter, C., Becker, H.D., Paulus, C.A., Giege, R., and Kern, D. (2007). Deinococcus glutaminyl-tRNA synthetase is a chimer between proteins from an ancient and the modern pathways of aminoacyl-tRNA formation. Nucleic Acids Res *35*, 1421-1431.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of Coot. Acta crystallographica Section D, Biological crystallography *66*, 486-501.

Eriani, G., Delarue, M., Poch, O., Gangloff, J., and Moras, D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature *347*, 203-206.

Evans, P. (2006). Scaling and assessment of data quality. Acta crystallographica *62*, 72-82.

Franke, D., and Svergun, D.I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Crystallogr *42*, 342-346.

Fritz-Laylin, L.K., Prochnik, S.E., Ginger, M.L., Dacks, J.B., Carpenter, M.L., Field, M.C., Kuo, A., Paredez, A., Chapman, J., Pham, J.*, et al.* (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell *140*, 631-642.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res *38*, W695-699.

Grant, T.D., Luft, J.R., Wolfley, J.R., Tsuruta, H., Martel, A., Montelione, G.T., and Snell, E.H. (2011). Small angle X-ray scattering as a complementary tool for high-throughput structural studies. Biopolymers *95*, 517-530.

Grant, T.D., Snell, E.H., Luft, J.R., Quartley, E., Corretore, S., Wolfley, J.R., Snell, M.E., Hadd, A., Perona, J.J., Phizicky, E.M.*, et al.* (2012). Structural conservation of an ancient tRNA sensor in eukaryotic glutaminyl-tRNA synthetase. Nucleic Acids Res *40*, 3723-3731.

Green, M.L., and Miller, R. (2003). Grid Computing in Buffalo, New York. Annals of the European Academy of Sciences, 191-218.

Guo, M., Yang, X.L., and Schimmel, P. (2010). New functions of aminoacyl-tRNA synthetases beyond translation. Nat Rev Mol Cell Biol *11*, 668-674.

Hayward, S., and Berendsen, H.J. (1998). Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins *30*, 144-154.

Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. Journal of Chemical Theory and Computation *4*, 435-447.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics *65*, 712-725.

Ibba, M., and Soll, D. (2000). Aminoacyl-tRNA synthesis. Annu Rev Biochem *69*, 617-650.

Ito, T., and Yokoyama, S. (2010). Two enzymes bound to one transfer RNA assume alternative conformations for consecutive reactions. Nature *467*, 612-616.

Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F., and Putz, J. (2009). tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res *37*, D159-162.

Kabsch, W. (1998). Automatic indexing of rotation diffraction patterns. J Appl Cryst *21*, 67-72.

Kabsch, W. (2010a). Integration, scaling, space-group assignment and post-refinement. Acta crystallographica *66*, 133-144.

Kabsch, W. (2010b). Xds. Acta crystallographica *66*, 125-132.

Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A *87*, 2264-2268.

Kempkes, R., Stofko, E., Lam, K., and Snell, E.H. (2008). Glycerol concentrations required for the successful vitrification of cocktail conditions in a high-throughput crystallization screen. Acta Cryst D *64*, 287-301.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. J Appl Crystallogr *36*, 1277-1282.

Lamour, V., Quevillon, S., Diriong, S., N'Guyen, V.C., Lipinski, M., and Mirande, M. (1994). Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci U S A *91*, 8670-8674.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Ludmerer, S.W., Wright, D.J., and Schimmel, P. (1993). Purification of glutamine tRNA synthetase from Saccharomyces cerevisiae. A monomeric aminoacyl-tRNA synthetase with a large and dispensable NH2-terminal domain. J Biol Chem *268*, 5519-5523.

Luft, J.R., Collins, R.J., Fehrman, N.A., Lauricella, A.M., Veatch, C.K., and DeTitta, G.T. (2003). A deliberate approach to screening for initial crystallization conditions of biological macromolecules. J Struct Biol *142*, 170-179.

Luft, J.R., Wolfley, J.R., Said, M.I., Nagel, R.M., Lauricella, A.M., Smith, J.L., Thayer, M.H., Veatch, C.K., Snell, E.H., Malkowski, M.G.*, et al.* (2007). Efficient optimization of crystallization conditions by manipulation of drop volume ratio and temperature. Protein Science *16*, 715-722.

Macbeth, M.R., Lingam, A.T., and Bass, B.L. (2004). Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. Rna *10*, 1563-1571.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) *2011*, bar009.

Malkowski, M.G., Quartley, E., Friedman, A.E., Babulski, J., Kon, Y., Wolfley, J., Said, M., Luft, J.R., Phizicky, E.M., DeTitta, G.T.*, et al.* (2007). Blocking S-adenosylmethionine synthesis in yeast allows selenomethionine incorporation and multiwavelength anomalous dispersion phasing. Proc Natl Acad Sci U S A *104*, 6678-6683.

Nureki, O., O'Donoghue, P., Watanabe, N., Ohmori, A., Oshikane, H., Araiso, Y., Sheppard, K., Soll, D., and Ishitani, R. (2010). Structure of an archaeal non-discriminating glutamyl-tRNA synthetase: a missing link in the evolution of Gln-tRNAGln formation. Nucleic Acids Res *38*, 7286-7297.

Pei, J., Tang, M., and Grishin, N.V. (2008). PROMALS3D web server for accurate multiple protein sequence and structure alignments. Nucleic Acids Res *36*, W30-34.

Perona, J.J., Rould, M.A., and Steitz, T.A. (1993). Structural basis for transfer RNA aminoacylation by Escherichia coli glutaminyl-tRNA synthetase. Biochemistry *32*, 8758-8771.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys *40*, 191-285.

Quartley, E., Alexandrov, A., Mikucki, M., Buckner, F.S., Hol, W.G., DeTitta, G.T., Phizicky, E.M., and Grayhack, E.J. (2009). Heterologous expression of L. major proteins in S. cerevisiae: a test of solubility, purity, and gene recoding. J Struct Funct Genomics *10*, 233-247.

Rambo, R.P., and Tainer, J.A. (2011). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers *95*, 559-571.

Rath, V.L., Silvian, L.F., Beijer, B., Sproat, B.S., and Steitz, T.A. (1998). How glutaminyl-tRNA synthetase selects glutamine. Structure *6*, 439-449.

Rother, M., Milanowska, K., Puton, T., Jeleniewicz, J., Rother, K., and Bujnicki, J.M. (2011a). ModeRNA server: an online tool for modeling RNA 3D structures. Bioinformatics *27*, 2441-2442.

Rother, M., Rother, K., Puton, T., and Bujnicki, J.M. (2011b). ModeRNA: a tool for comparative modeling of RNA 3D structure. Nucleic Acids Res *39*, 4007-4022.

Rould, M.A., Perona, J.J., Soll, D., and Steitz, T.A. (1989). Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science *246*, 1135-1142.

Saha, R., Dasgupta, S., Basu, G., and Roy, S. (2009). A chimaeric glutamyl:glutaminyl-tRNA synthetase: implications for evolution. Biochem J *417*, 449-455.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. Proteins *23*, 318-326.

Schrodinger, LLC (2010). The PyMOL Molecular Graphics System, Version 1.3r1.

Sherlin, L.D., and Perona, J.J. (2003). tRNA-dependent active site assembly in a class I aminoacyl-tRNA synthetase. Structure *11*, 591-603.

Siatecka, M., Rozek, M., Barciszewski, J., and Mirande, M. (1998). Modular evolution of the Glx-tRNA synthetase family--rooting of the evolutionary tree between the bacteria and archaea/eukarya branches. Eur J Biochem *256*, 80-87.

Simader, H., Hothorn, M., and Suck, D. (2006). Structures of the interacting domains from yeast glutamyl-tRNA synthetase and tRNA-aminoacylation and nuclear-export cofactor Arc1p reveal a novel function for an old fold. Acta crystallographica Section D, Biological crystallography *62*, 1510-1519.

Smolsky, I.L., Liu, P., Niebuhr, M., Ito, K., Weiss, T.M., and Tsuruta, H. (2007). Biological small-angle x-ray scattering facility at the Stanford synchrotron radiation laboratory. J Appl Crystallogr *40*, S453-S458.

Soltis, S.M., Cohen, A.E., Deacon, A., Eriksson, T., Gonzalez, A., McPhillips, S., Chui, H., Dunten, P., Hollenbeck, M., Mathews, I*., et al.* (2008). New paradigm for macromolecular crystallography experiments at SSRL: automated crystal screening and remote data collection. Acta Crystallogr D *64*, 1210-1221.

Uter, N.T., and Perona, J.J. (2004). Long-range intramolecular signaling in a tRNA synthetase complex revealed by pre-steady-state kinetics. Proc Natl Acad Sci U S A *101*, 14396-14401.

Volkov, V.V., and Svergun, D.I. (2003). Uniqueness of ab initio shape determination in small-angle scattering. J Appl Crystallogr *36*, 860-864.

Wang, C.C., Morales, A.J., and Schimmel, P. (2000). Functional redundancy in the nonspecific RNA binding domain of a class I tRNA synthetase. J Biol Chem *275*, 17180-17186.

Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F., and Jones, D.T. (2004). The DISOPRED server for the prediction of protein disorder. Bioinformatics *20*, 2138-2139.

6 Summary and Future Work

6.1 Thesis Summary

In this thesis, the utility of SAXS has been discussed as a complementary tool used to fully understand biological systems. Chapter 2 demonstrated that, with SAXS data of sufficient quality, structural information extracted from the data, such as radius of gyration, maximum particle dimension, and even low-resolution molecular envelopes, agreed well with high-resolution structures for each of the 28 targets studied. Moreover, it was shown that, in several cases, additional information was provided by SAXS, such as oligomeric state and visualization of regions of structure unresolved by X-ray crystallography, which provided a more complete understanding of the biological system. In chapter 2, data quality was evaluated through a combination of automated and manual analysis, where numerical measurements were determined by automated software programs while data quality indicators such as linearity in the Guinier region and trends as a function of radiation damage and concentration were evaluated manually. In chapter 3, however, a fully automated, statistical approach to data evaluation, called SAXStats, was created and employed on a set of 100 proteins. This statistical approach removed the manual, subjective aspect of data quality evaluation while greatly improving the time required for SAXS data analysis in a high-throughput application.

Prior to this work, the only objective measures of SAXS data regarded the determination of size parameters such as radius of gyration, maximum particle dimension, or Porod volume. However, objective data quality measurements were virtually absent. This thesis provides the community with SAXStats, a statistical method for objectively quantifying the quality of SAXS data as determined by radiation damage, concentration dependence and linearity in the Guinier region. In addition, it provides the community with a method for quickly determining the second virial coefficient, a quantity describing the nature and degree of interparticle interactions in solution. This thesis also presents a guideline with which to prepare protein solutions for standard SAXS experiments, which may lead to increased success rates of the technique.

While chapters 2 and 3 reported on data collected as part of a high-throughput structural pipeline, chapters 4 and 5 instead focused on using SAXS as a complementary tool to advance the scientific understanding of a specific biological system, that of glutaminylation of tRNA in eukaryotes. SAXStats was used to objectively evaluate SAXS data for the full-length yeast glutaminyl-tRNA synthetase, Gln4, as well as for N-terminal (NTD) and C-terminal domain (CTD) fragments. In this study several other experimental and modeling techniques were used such as X-ray crystallography, bioinformatics, and molecular biology tools.

Prior to this study, little was known about the structural and biological nature of the N-terminal domain of Gln4. In 1985 the genomic sequence of yeast GlnRS revealed that it contains an appended domain compared to its prokaryotic homologs, and ever since the question has remained as to why this domain exists (Ludmerer and Schimmel, 1985). Studies throughout the 1980s and 1990s suggested that the NTD is not essential for yeast viability (Ludmerer and Schimmel, 1987; Ludmerer et al., 1993; Whelihan and Schimmel, 1997). It was further shown that the NTD fused to the *E. coli* GlnRS could substitute for the native yeast gene *in vivo* (Wang and Schimmel, 1999; Whelihan and Schimmel, 1997). In 2000 it was shown that two specific lysine-rich patches in the appended N-terminal extension cooperate to enhance tRNA binding (Wang et al., 2000). The precise purpose of this domain, however, has remained elusive.

This thesis demonstrates major advances in understanding the origin and function of this domain. It was shown that while the NTD is not essential for yeast viability in ideal conditions, deletion of the domain severely impairs growth at 14°C and

19°C, demonstrating that the NTD is important to enzyme function. Further studies showed that while the NTD does not play a role in catalysis, it functions to significantly improve binding of $tRNA^{gln}$. The structure of the NTD was determined and shown to be structurally homologous to the GatCAB amidotransferase enzymes that play a role in the indirect pathway of glutaminylation found in prokaryotes, yielding insights into its origin. The structure of the C-terminal domain of Gln4, coupled with SAXS data of the enzyme, yielded the orientation of the two fragments in the full-length enzyme in solution. Bioinformatics and molecular dynamics simulations enabled the production of a model of the full-length enzyme bound to tRNA. This study demonstrates the utility of SAXS as it is most commonly practiced, as one of many tools in the structural biologist's toolkit to fully understand a biological system.

6.2 Combining SAXS Data with Information from the NESG Database to Supplement the High-throughput Structural Pipeline

The NESG consortium has created the Structural Proteomics in the NorthEast (SPINE) database (Bertone et al., 2001) for collecting and organizing all data related to protein targets for quick access by SPINE users. This database tracks target, construct, expression, purification, biophysical characterization, X-ray and NMR structure data. Included in this database are any results from mass spectrometry, gel filtration, circular dichroism, static and dynamic light scattering, HSQC spectra, residual dipolar couplings and other X-ray and NMR data. Chapters 2 and 3 discussed the SAXS data associated with some 128 different protein targets that are part of this database. SAXS data have also been collected on more than 500 additional targets that are yet to be analyzed. We will work with the NESG to populate this database with SAXS data, thereby supplementing the known properties of targets with parameters such as R_g, D_{max}, oligomeric state and assembly in solution, globularity, overall shape and size, molecular envelopes, flexibility in solution, and visualization of regions that may be unresolved in high resolution X-ray structures. This data can be very useful to the NESG scientists working on specific projects. For example, for many of these targets, Nuclear Overhauser effect spectroscopy, or NOESY, data has been collected for use in NMR structure determination. To correctly interpret NOESY data it is necessary to know the oligomeric state of the protein. Typical methods used by the NESG to determine molecular weight include SDS-PAGE, gel filtration, and NMR-determined rotational correlation time measurements. However, SDS-PAGE dissociates proteins providing only subunit molecular weights, and gel filtration can sometimes yield ambiguous results, and rotational correlation times become inaccurate for proteins approaching 25 kDa or more, which is particularly a problem for oligomers (Chevalier, 2010; Luginbühl and Wüthrich, 2002). SAXS data can yield accurate information regarding oligomeric state in solution and can also be used on proteins of almost unlimited size. Additionally, this SAXS data can be combined with NMR data to enhance the accuracy and success of structure determination (Gabel et al., 2006; Grishaev et al., 2008) or to validate existing structures (Schwieters et al., 2010).

6.3 Improvements to SAXStats

Chapter 3 presented a novel method, called SAXStats, for evaluating SAXS data using a statistical approach that removes the subjective nature inherent in manual SAXS analysis. In the current implementation of SAXStats, the statistical analysis uses linear regression that inherently assumes a linear trend in either radiation damage or concentration dependence. It may be the case, however, that radiation damage or concentration dependence follows a nonlinear trend that may cause the current statistical analysis to report less accurate results than if a nonlinear regression were employed. For example, when considering concentration dependence, as concentration increases interparticle interactions become more likely due to the fact that the average distance between particles, *r*, decreases. If the interparticle interactions are dominated by electrostatic interactions, then the changes from once concentration to the next may follow a $1/r^2$ dependence. Future implementations can utilize nonlinear regression procedures that can be used to decide thresholds for rejection of damaged exposures or more accurate statistical analysis of concentration dependence.

For the 82 studied samples that were provided by the NESG consortium, it was shown that only 5 samples met the most stringent sample quality requirements determined by SAXStats. One of the reasons for this may be that each of the sample volumes has been recovered from crystallization screening trials and each sample has experienced at least two freeze-thaw cycles, and is therefore a "left-over" from crystallization screening. Additionally, for many of these samples several months may have passed from protein preparation to data collection that may disturb the solution conditions, such that a degree of aggregation or interparticle interactions result. Protein solutions prepared specifically for SAXS experiments that will not experience two freeze-thaw cycles or the extended length of time will likely result in higher quality SAXS data. Additionally, these protein solutions could be compared to the previously studied solutions to determine if these factors have adversely affected sample quality. It may be seen from these experiments that solution conditions properly prepared are more amenable to SAXS than the present analysis suggests, and once again highlights the requirement that accurate sample characterization be performed prior to experiment.

6.4 Probing the Acceptability of SAXS Data Quality

SAXS has been used in a wide variety of biological applications including determining oligomeric state and assembly (Mertens and Svergun, 2010), assessing conformational changes upon ligand binding (Lipfert and Doniach, 2007), validation of structures in solution (Grishaev et al., 2008), and several others. While only 5 of the 82 samples studied in Chapter 3 met the most stringent data quality requirements regarding concentration dependence and linearity in the Guinier region, it may be that, depending on the specific biological question being asked, the SAXS results of other samples are still useful. In section 3.4.4.2 the severity of concentration dependence was discussed. For many samples, while concentration dependence may be present, the impact of this dependence may be small enough so as to not distort the biological conclusions. The impact of less-than-ideal quality SAXS data, however, depends entirely upon the biological question. If the oligomeric state of a protein under certain conditions is being sought out, then, depending on the molecular weight of the protein, some concentration dependence can be tolerated since more than 70% of proteins experienced less than 1 kDa change in molecular weight for every mg/mL increase in concentration, which is sufficient for determining oligomeric state for most proteins at reasonable concentrations. However, if, for example, two conformations of a protein exist that differ by only a few angstroms in R_g, then highly stringent data quality requirements may need to be met in order to determine which conformation exists in solution, since concentration dependencies of several angstroms per mg/mL were seen.

In an attempt to determine the degree of data quality required to answer a variety of typical questions sought out by SAXS experiments, we will apply the statistical methods employed in SAXStats to SAXS data for targets whose high-resolution structures are known, many of which were discussed in Chapter 2. By working with the NESG scientists familiar with these specific projects, we will determine whether or not the SAXS data for these cases is of sufficient quality to address the biological puzzle. For example, in chapter 2, SAXS data demonstrated that samples 2, 3, 4, 5, 6, 9, 11, 16 and 17 exhibit oligomeric states in solution that are alternatives to those seen in the crystallographic structures. We can combine this information with other NESG targets and determine whether or not the oligomeric state can be determined unambiguously by SAXS, possibly corroborating this information with other biophysical characterization methods such as SDS-PAGE, gel filtration, light scattering or NMR relaxation measurements. From this we can then determine which oligomeric states determined by SAXS were accurate even with poor quality data as evaluated by SAXStats. By performing this analysis for a significant number of samples, we may be able to provide limits on the acceptability of data quality to determine oligomeric state. A similar analysis for other NESG targets where different questions are being asked, such as determining precise subunit assembly, determining which conformational state exists in solution, assessing degrees of foldedness for various cofactors or solution conditions, etc., may allow us to postulate a general model for SAXS data quality requirements. Doing so would provide scientists with a guide for preparing solutions for SAXS experiments to maximize the likelihood of success, success rated not by absolute measurements of data quality, but rather by whether or not the biological question has been answered.

6.5 Probing Long Time-scale Protein Dynamics Using SAXS Envelopes

SAXS data is often collected to understand protein dynamics in solution. In many cases simple analyses are performed that require no modeling, such as Kratky plots or Porod-Debye plots, which can yield information about the degree of flexibility and conformational motion present in solution. More sophisticated modeling software (Bernado et al., 2007; Pelikan et al., 2009) can lead to understanding protein dynamics in three dimensions. Another popular method for uncovering protein dynamics is molecular dynamics simulations. Atomistic simulations can help describe conformational or catalytic processes on relatively short timescales of picoseconds up to hundreds of nanoseconds using vast computing resources (Klepeis et al., 2009). Coarse-grained modeling can significantly reduce the computational time and resources required to perform molecular dynamics simulations at long time scales, but at the expense of accuracy (Flores et al., 2012). However, oftentimes, important biological processes occur in the several microseconds to millisecond timescales that are currently unattainable for most of these methods. In an effort to understand biological processes at these timescales, researchers at the University at Leeds have developed algorithms utilizing Finite Element Analysis (FEA) coupled with known flexibility characteristics of proteins to approximate large scale molecular motions that occur over significant lengths of time (Oliver et al., 2012). This algorithm, called Fluctuating Finite Element Analysis (FFEA), also known as "Jelly Modeling", has the added advantage that it does not require a high-resolution structure to work. Low-resolution finite element meshes generated from high-resolution structures were shown to provide similar results of large-scale dynamic motions using FFEA as those using standard all-atom molecular dynamics simulations (Oliver et al., 2012). They have shown that these simulations can also be performed using low-resolution reconstructions from cryoelectron microscopy data.

SAXS data from the high-throughput structural pipeline could be used to supply these scientists with low-resolution molecular envelopes fitting a variety of shapes and sizes to further develop this algorithm in cases where high-resolution structures are unavailable. These low-resolution envelopes could therefore be used to gain insight into protein dynamics for globular proteins with no high-resolution structural information, whereas current SAXS algorithms used to model dynamics are usually limited to proteins with large degrees of intrinsic disorder, or those where large portions of the structure are known (Bernado et al., 2007; Pelikan et al., 2009). In addition to FFEA using these envelopes as starting models, we are also working to develop methods to corroborate the results of the FFEA analysis using the SAXS data, such that the dynamic model fits the SAXS data better than the static envelope. Jelly Modeling therefore provides the SAXS community with a valuable resource for gaining additional dynamic information from *ab initio* molecular envelopes than currently exists.

6.6 Expanding the Structural Knowledge of Gln4

Chapters 4 and 5 discussed the biochemical, structural, and bioinformatics characterization of the yeast glutaminyl-tRNA synthetase (Gln4). In Chapter 4 it was shown that the N-terminal domain (NTD) of Gln4 plays a vital role in binding to tRNA^{gln}. In Chapter 5, homology with bacterial GlnRS and GatCAB enzymes, coupled with molecular dynamics simulation, was used to create a model for the full-length Gln4 complexed with tRNA^{gln}. This is the first model of any eukaryotic GlnRS-tRNA complex. However, to date there is no experimental structural evidence for this model. SAXS experiments performed on the full-length enzyme-tRNA complex would be able to confirm the overall shape and size of this model in solution. Additionally, the co-crystal structure of the complex would yield high-resolution structural information that can uncover specific contacts between the tRNA and the enzyme. This information would be able to prove or disprove the speculation that the NTD interacts with the eukaryotic-specific Insertion 1, which would yield insights into the evolutionary process resulting in the genesis of the NTD in eukaryotes. The high-resolution structure would also uncover details about the reorganization of specific residues in the active site that allow binding of ATP and glutamine necessary for catalysis.

The X-ray diffraction data for full-length Gln4 presented in Chapter 5 showed that the NTD was crystallographically disordered. However, the SAXS data for Gln4 in solution showed that the mobility of the NTD was very limited. To gain further insights into the position and motion of the NTD in solution, the rigid body model determined by SAXS will be subjected to molecular dynamics simulations to probe the degree of motion of the NTD with higher resolution than SAXS can provide. Additionally, the position of the NTD in solution was also shown to be very different from the position of the NTD when bound to tRNA, which suggests that C-terminal domain (CTD) binding to tRNA causes conformational reorientation of the NTD. This suggestion will be further investigated using molecular dynamics simulations, which will start with either the SAXS rigid body model of Gln4 or the same model after equilibration with molecular dynamics. The tRNA will be superimposed to the CTD based on structural homology with *E. coli* GlnRS and subjected to molecular dynamics simulation. The ~40 Å translation and ~160° rotation of the NTD may be able to be captured in a reasonable amount of simulation time using sufficient computational resources yielding insights into the conformational reorientation of the NTD when bound to tRNA. However, given the exceedingly large conformational motion involved, coarse-grained molecular dynamics approaches may need to be employed instead in order to capture the movement in a reasonable amount of time.

6.7 References

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc *129*, 5656-5664.

Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T., and Gerstein, M. (2001). SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res *29*, 2884-2898.

Chevalier, F. (2010). Highlights on the capacities of "Gel-based" proteomics. Proteome Sci *8*, 23.

Flores, S.C., Bernauer, J., Shin, S., Zhou, R., and Huang, X. (2012). Multiscale modeling of macromolecular biosystems. Brief Bioinform *13*, 395-405.

Gabel, F., Simon, B., and Sattler, M. (2006). A target function for quaternary structural refinement from small angle scattering and NMR orientational restraints. Eur Biophys J *35*, 313-327.

Grishaev, A., Tugarinov, V., Kay, L.E., Trewhella, J., and Bax, A. (2008). Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. J Biomol NMR *40*, 95-106.

Klepeis, J.L., Lindorff-Larsen, K., Dror, R.O., and Shaw, D.E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol *19*, 120-127.

Lipfert, J., and Doniach, S. (2007). Small-angle X-ray scattering from RNA, proteins, and protein complexes. Annu Rev Biophys Biomol Struct *36*, 307-327.

Ludmerer, S.W., and Schimmel, P. (1985). Cloning of GLN4: an essential gene that encodes glutaminyl-tRNA synthetase in Saccharomyces cerevisiae. J Bacteriol *163*, 763-768.

Ludmerer, S.W., and Schimmel, P. (1987). Construction and analysis of deletions in the amino-terminal extension of glutamine tRNA synthetase of Saccharomyces cerevisiae. J Biol Chem *262*, 10807-10813.

Ludmerer, S.W., Wright, D.J., and Schimmel, P. (1993). Purification of glutamine tRNA synthetase from Saccharomyces cerevisiae. A monomeric aminoacyl-tRNA synthetase with a large and dispensable NH2-terminal domain. J Biol Chem *268*, 5519-5523.

Luginbühl, P., and Wüthrich, K. (2002). Semi-classical nuclear spin relaxation theory revisited for use with biological macromolecules. Progress in Nuclear Magnetic Resonance Spectroscopy *40*, 199-247.

Mertens, H.D., and Svergun, D.I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J Struct Biol *172*, 128-141.

Oliver, R., Read, D.J., Harlen, O.G., and Harris, S.A. (2012). A Stochastic Finite Element Model for the Dynamics of Globular Macromolecules. ArXiv e-prints.

Pelikan, M., Hura, G.L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys *28*, 174-189.

Schwieters, C.D., Suh, J.Y., Grishaev, A., Ghirlando, R., Takayama, Y., and Clore, G.M. (2010). Solution structure of the 128 kDa enzyme I dimer from Escherichia coli and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. J Am Chem Soc *132*, 13026-13045.

Wang, C.C., Morales, A.J., and Schimmel, P. (2000). Functional redundancy in the nonspecific RNA binding domain of a class I tRNA synthetase. J Biol Chem *275*, 17180-17186.

Wang, C.C., and Schimmel, P. (1999). Species barrier to RNA recognition overcome with nonspecific RNA binding domains. J Biol Chem *274*, 16508-16512.

Whelihan, E.F., and Schimmel, P. (1997). Rescuing an essential enzyme-RNA complex with a non-essential appended domain. EMBO J *16*, 2968-2974.

Appendix A. Converting t-statistic to p-value.

Table A.1 Converting t-statistic to p-value. Top row represents p-values assuming a two-tailed test. Left column refers to number of degrees of freedom, DF.

DF	0.200	0.100	0.050	0.020	0.010	0.005	0.002	0.001
1	3.078	6.314	12.70	31.82	63.65	127.3	318.3	636.6
2	1.886	2.920	4.303	6.965	9.925	14.08	22.32	31.59
3	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.897	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.625	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.584	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.090	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	1.309	1.695	2.040	2.453	2.744	3.022	3.375	3.633
32	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	1.306	1.688	2.028	2.434	2.719	2.991	3.333	3.582
37	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566
39	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
42	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
44	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
46	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
48	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
70	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
80	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	1.291	1.662	1.987	2.369	2.632	2.878	3.183	3.402
100	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.391
120	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
150	1.287	1.655	1.976	2.351	2.609	2.849	3.145	3.357
200	1.286	1.652	1.972	2.345	2.601	2.839	3.131	3.340
300	1.284	1.650	1.968	2.339	2.592	2.828	3.118	3.323
500	1.283	1.648	1.965	2.334	2.586	2.820	3.107	3.310

Appendix B. Script for detecting radiation damage in SAXS Profiles

```
#!/bin/bash

while getopts hf:b:e:i:l: opt; do
  case $opt in
    h)
    echo
    echo " ------------------------------------------------------------------------------ "
      echo
      echo " This script checks radiation damage in SAXS data profiles."
      echo " In addition to calculating chi between exposures, Rg and "
      echo " I(0), it calculates a linear regression and generates a p-value for "
      echo " the likelihood of a trend being present, suggesting damage."
      echo " If damage is present (p < 0.05), it will only average those "
      echo " exposures inside the standard deviation of all exposures, "
      echo " otherwise it will average all exposures into one file."
      echo
      echo " Statistics are written to <sample>_rdam_stats.txt"
      echo " and a log file is written."
      echo
      echo " AutoRg and datcmp (ATSAS) must be in \$PATH."
      echo
      echo " If gnuplot exists in \$PATH, a plot will be printed to the terminal"
      echo " showing the trend in I(0) for each exposure. "
      echo
      echo " Usage:  rdam.sh [options] "
    echo
    echo " where [options] are:"
    echo
    echo " -h:  Print this help menu and exit"
    echo " -f:  File name of the first exposure of a sample (required)"
    echo " -b:  First point of Guinier region (optional, defaults to pi/dmax from datGNOM)"
    echo " -e:  Last point of Guinier region (optional, defaults to 1.3/Rg from datGNOM)"
    echo " -i:  First point in data to take (optional)"
    echo " -l:  Last point in data to take (required if -i is given)"
      echo
      echo " The file name must observe the SSRL BL 4-2 standard naming protocol, i.e."
      echo " <samplename>_<tube-ID>_<sample-number>_0_<exposurenumber>.dat  "
      echo
      echo " For example:  Protein-1_02B_S025_0_01.dat "
      echo
      echo " The extension of the file must be .dat and the data must already "
      echo " be buffer subtracted (use the \"-s yes\" option in the SAStool parameters file)."
    echo
    echo " ----------------------------------------------------------------------------- "
    echo
    exit 0
    ;;
    f)
      firstexp=$OPTARG
      ;;
    b)
      begp=$OPTARG
      ;;
    e)
      endp=$OPTARG
      ;;
    i)
      firstp=$OPTARG
      ;;
    l)
      lastp=$OPTARG
      ;;
   \?)
      echo "Invalid option: -$OPTARG" >&2
      exit 1
      ;;
    :)
      echo "Option -$OPTARG requires an argument." #>&2
      exit 1
      ;;
  esac
done

if [ -z "$firstexp" ]
then
```

```
      echo " To get a quick help, type rdam.sh -h"
      echo
      echo " Enter file name of first exposure of sample: "
      read firstexp
fi

for i in $firstexp
do

      numexp=0
      l=1
      j=1
      k=1

      sample=`echo ${i%.*} | awk 'BEGIN {FS="_"} {for (i=1; i<(NF-3); i++) printf "%s_", $i;print ""}'`
      conc=`echo ${i%.*} | awk 'BEGIN {FS="_"} {print $(NF-3)}'`
      snum=`echo ${i%.*} | awk 'BEGIN {FS="_"} {print $(NF-2)}'`
      preexp=`echo ${i%.*} | awk 'BEGIN {FS="_"} {print $(NF-1)}'`
      exp=`echo ${i%.*} | awk 'BEGIN {FS="_"} {print $(NF)}'`
    concl=`echo $conc | awk '{print substr($1,3,1)}'`


      for i in ${sample}${conc}_${snum}_${preexp}_*[0-9][0-9].dat
      do
        let numexp++
      done

  #create log file using if statement piped into "tee" command
  if [ 1 ]
  then

      echo
      echo " ---------------------------   ${sample}${conc}_${snum}  -------------------------------"
      echo

      for_avg=""
      chi_avgs=""
      rgs=""
      I0s=""
      chis_to_1=""


      if [ "${firstp}" ]
      then
        j=1
        while [ $j -le $numexp ]
        do
                 k=`echo $j | awk '{printf "%02d",$1; print ""}'`
                 awk '{if (NR>='"${firstp}"' && NR<='"${lastp}"') print}'
${sample}${conc}_${snum}_${preexp}_${k}.dat > copy-${sample}${conc}_${snum}_${preexp}_${k}.dat
                 let j++
        done
        sample=copy-${sample}
      fi



      while [ $l -eq 1 ]
      do
        j=1
        chis=""

        while [ $j -le $numexp ]
        do
                 k=`echo $j | awk '{printf "%02d",$1; print ""}'`

                 #calculate and extract chi and Rg values
            chi[j]=`datcmp ${sample}${conc}_${snum}_${preexp}_0${l}.dat
${sample}${conc}_${snum}_${preexp}_${k}.dat | awk '{printf "%-2.2f", $1}'`

                 chis="$chis ${chi[j]}"

            if [ $j -eq 1 ]
                 then
                m=$j
            fi

            #new way of setting Guinier region based on Dmax and Rg estimates from datGNOM
```

```
                 gnom[j]=`datgnom ${sample}${conc}_${snum}_${preexp}_${k}.dat`
                 rg_gnom[j]=`echo ${gnom[j]} | awk '{printf "%.2f", $NF}'`
                 dmax[j]=`echo ${gnom[j]} | awk '{printf "%.2f", $3}'`

                 echo "${rg_gnom[j]} ${dmax[j]}"

                 #if begp and endp are not set, then set begp > endp, so that the following if statement will
set it automatically
                 if [ -z "${begp}" ]
                 then
                         begp=1
                 fi

                 if [ -z "${endp}" ]
                 then
                         endp=0
                 fi

                 if [ $endp -le $begp ]
                 then
                         qmin=`echo ${dmax[j]} | awk '{print 3.14159/$1}'`
                         qmax=`echo ${rg_gnom[j]} | awk '{print 1.3/$1}'`
                         begp=`awk '{if ($1 < "'$qmin'") i=NR} END {print i}'
${sample}${conc}_${snum}_${preexp}_01.dat`
                         endp=`awk '{if ($1 < "'$qmax'") i=NR} END {print i}'
${sample}${conc}_${snum}_${preexp}_01.dat`
            fi

            echo "${qmin} - ${qmax}"
            echo "${begp} - ${endp}"

        let j++

        done
        let l++
    done
      j=1

      while [ $j -le $numexp ]
          do
            k=`echo $j | awk '{printf "%02d",$1; print ""}'`
            chis_to_1="$chis_to_1 ${j}:${chi[j]}"

                                  # linreg.awk: An awk script to compute linear regression
                                  # Input columns x and y, outputs a=slope and b=intercept
                                  # Usage: awk -f linreg.awk file
                                  #
                                  rg_fit=`awk 'NR>='"${begp}"' && NR<='"${endp}"' {print
$1**2,log(sqrt($2*$2)) }' ${sample}${conc}_${snum}_${preexp}_${k}.dat | awk '

                                          {
                                           delta = $2 - avg;
                                           avg += delta / NR;
                                           mean2 += delta * ($2 - avg);
                                           x[NR] = $1; y[NR] = $2;
                                           sx += x[NR]; sy += y[NR];
                                           sxx += x[NR]*x[NR];
                                           sxy += x[NR]*y[NR];
                                           syy += y[NR]*y[NR];
                                          }

                                          END{
                                           sd = sqrt (mean2/NR);
                                           det = NR*sxx - sx*sx;
                                           a = (NR*sxy - sx*sy)/det;
                                           b = (-sx*sxy+sxx*sy)/det;
                                           se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
                                           sb = sqrt(NR*se*se/det);
                                           t = a/sb;
                                           #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
                                           print a, b, t, avg, sd;
                                          }'`

                                  rg[j]=`echo $rg_fit | awk '{print sqrt(-3*$1)}'`
                                  I0[j]=`echo $rg_fit | awk '{print exp($2)}'`

                    if [ "${rg[j]}" != "nan" ] && [ "${I0[j]}" != "nan" ]
                                  then
                    rgs="$rgs ${j}:${rg[j]}"
```

```
                rgs_gnom="$rgs_gnom ${j}:${rg_gnom[j]}"
                dmaxs="$dmaxs ${j}:${dmax[j]}"
                              I0s="$I0s ${j}:${I0[j]}"
                fi

        let j++
     done

qmin=`awk '{if (NR=="'${begp}'") print $1}' ${firstexp}`
qmax=`awk '{if (NR=="'${endp}'") print $1}' ${firstexp}`

rename=`echo ${sample#copy-}`

    n=1
    until [ "${rg[n]}" != "nan" ]
    do
        let n++
    done

     chis=`echo $chis_to_1 | awk 'BEGIN{FS=":"; RS=" "; ORS=" ";} {if (NR>1) print}' `

     #compute linear regression for chi, Rg, and I(0) with increasing radiation exposure

     # linreg.awk: An awk script to compute linear regression
     # Input columns x and y, outputs a=slope and b=intercept
     # Usage: awk -f linreg.awk file
     #
     chi_fit=`echo $chis | awk '
             BEGIN{
              FS=":"; RS=" ";
             }

             {
              delta = $2 - avg;
              avg += delta / NR;
              mean2 += delta * ($2 - avg);
              x[NR] = $1; y[NR] = $2;
              sx += x[NR]; sy += y[NR];
              sxx += x[NR]*x[NR];
              sxy += x[NR]*y[NR];
              syy += y[NR]*y[NR];
             }

             END{
              sd = sqrt (mean2/NR);
              det = NR*sxx - sx*sx;
              a = (NR*sxy - sx*sy)/det;
              b = (-sx*sxy+sxx*sy)/det;
          se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
          sb = sqrt(NR*se*se/det);
          t = a/sb;
          #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
          print a, b, t, avg, sd;
             }' `

     chi_trend=`echo $chi_fit | awk '{printf "%5.3f", $1}'`
     chi_yint=`echo $chi_fit | awk '{printf "%5.3f", $2}'`
     chi_t=`echo $chi_fit | awk '{printf "%5.3f", $3}'`
     chi_mean=`echo $chi_fit | awk '{printf "%5.3f", $4}'`
     chi_sd=`echo $chi_fit | awk '{printf "%5.3f", $5}'`

     # linreg.awk: An awk script to compute linear regression
     # Input columns x and y, outputs a=slope and b=intercept
     # Usage: awk -f linreg.awk file
     #
     rg_fit=`echo $rgs | awk '
             BEGIN{
              FS=":"; RS=" ";
             }

             {
              delta = $2 - avg;
              avg += delta / NR;
              mean2 += delta * ($2 - avg);
              x[NR] = $1; y[NR] = $2;
              sx += x[NR]; sy += y[NR];
              sxx += x[NR]*x[NR];
              sxy += x[NR]*y[NR];
              syy += y[NR]*y[NR];
             }

             END{
              sd = sqrt (mean2/NR);
              det = NR*sxx - sx*sx;
              a = (NR*sxy - sx*sy)/det;
              b = (-sx*sxy+sxx*sy)/det;
          se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
          sb = sqrt(NR*se*se/det);
          t = a/sb;
          #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
          print a, b, t, avg, sd;
             }' `

     rg_trend=`echo $rg_fit | awk '{printf "%5.3f", $1}'`
     rg_yint=`echo $rg_fit | awk '{printf "%5.3f", $2}'`
     rg_t=`echo $rg_fit | awk '{printf "%5.3f", $3}'`
     rg_mean=`echo $rg_fit | awk '{printf "%5.3f", $4}'`
     rg_sd=`echo $rg_fit | awk '{printf "%5.3f", $5}'`

     # linreg.awk: An awk script to compute linear regression
     # Input columns x and y, outputs a=slope and b=intercept
     # Usage: awk -f linreg.awk file
     #
     rg_gnom_fit=`echo $rgs_gnom | awk '
             BEGIN{
              FS=":"; RS=" ";
             }

             {
              delta = $2 - avg;
              avg += delta / NR;
              mean2 += delta * ($2 - avg);
              x[NR] = $1; y[NR] = $2;
              sx += x[NR]; sy += y[NR];
              sxx += x[NR]*x[NR];
              sxy += x[NR]*y[NR];
              syy += y[NR]*y[NR];
             }

             END{
              sd = sqrt (mean2/NR);
              det = NR*sxx - sx*sx;
              a = (NR*sxy - sx*sy)/det;
              b = (-sx*sxy+sxx*sy)/det;
          se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
          sb = sqrt(NR*se*se/det);
          t = a/sb;
          #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
          print a, b, t, avg, sd;
             }' `

     rg_gnom_trend=`echo $rg_gnom_fit | awk '{printf "%5.3f", $1}'`
     rg_gnom_yint=`echo $rg_gnom_fit | awk '{printf "%5.3f", $2}'`
     rg_gnom_t=`echo $rg_gnom_fit | awk '{printf "%5.3f", $3}'`
     rg_gnom_mean=`echo $rg_gnom_fit | awk '{printf "%5.3f", $4}'`
     rg_gnom_sd=`echo $rg_gnom_fit | awk '{printf "%5.3f", $5}'`

     # linreg.awk: An awk script to compute linear regression
     # Input columns x and y, outputs a=slope and b=intercept
     # Usage: awk -f linreg.awk file
     #
     dmax_fit=`echo $dmaxs | awk '
             BEGIN{
              FS=":"; RS=" ";
             }

             {
              delta = $2 - avg;
              avg += delta / NR;
              mean2 += delta * ($2 - avg);
              x[NR] = $1; y[NR] = $2;
              sx += x[NR]; sy += y[NR];
              sxx += x[NR]*x[NR];
              sxy += x[NR]*y[NR];
              syy += y[NR]*y[NR];
             }

             END{
```

```
                  sd = sqrt (mean2/NR);
                  det = NR*sxx - sx*sx;
                  a = (NR*sxy - sx*sy)/det;
                  b = (-sx*sxy+sxx*sy)/det;
             se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
             sb = sqrt(NR*se*se/det);
             t = a/sb;
             #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
             print a, b, t, avg, sd;
                 }' `

        dmax_trend=`echo $dmax_fit | awk '{printf "%5.3f", $1}'`
        dmax_yint=`echo $dmax_fit | awk '{printf "%5.3f", $2}'`
        dmax_t=`echo $dmax_fit | awk '{printf "%5.3f", $3}'`
        dmax_mean=`echo $dmax_fit | awk '{printf "%5.3f", $4}'`
        dmax_sd=`echo $dmax_fit | awk '{printf "%5.3f", $5}'`

        # linreg.awk: An awk script to compute linear regression
        # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
        # Usage: awk -f linreg.awk file
        #
        I0_fit=`echo $I0s | awk '
                 BEGIN{
                  FS=":"; RS=" ";
                 }

                 {
                  delta = $2 - avg;
                  avg += delta / NR;
                  mean2 += delta * ($2 - avg);
                  x[NR] = $1; y[NR] = $2;
                  sx += x[NR]; sy += y[NR];
                  sxx += x[NR]*x[NR];
                  sxy += x[NR]*y[NR];
                  syy += y[NR]*y[NR];
                 }

                 END{
                  sd = sqrt (mean2/NR);
                  det = NR*sxx - sx*sx;
                  a = (NR*sxy - sx*sy)/det;
                  b = (-sx*sxy+sxx*sy)/det;
             se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
             sb = sqrt(NR*se*se/det);
             t = a/sb;
             #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
             print a, b, t, avg, sd;
                 }' `

        I0_trend=`echo $I0_fit | awk '{printf "%5.3f", $1}'`
        I0_yint=`echo $I0_fit | awk '{printf "%5.3f", $2}'`
        I0_t=`echo $I0_fit | awk '{printf "%5.3f", $3}'`
        I0_mean=`echo $I0_fit | awk '{printf "%5.3f", $4}'`
        I0_sd=`echo $I0_fit | awk '{printf "%5.3f", $5}'`

      #make simple plot in terminal for Rg and I(0) vs exposure number
      echo $rgs | awk 'BEGIN {FS=":"; RS=" ";OFS=" "} {print $1,$2}' > rgs_plot_${rename}${conc}.dat
      echo $I0s | awk 'BEGIN {FS=":"; RS=" ";OFS=" "} {print $1,$2}' > I0s_plot_${rename}${conc}.dat
      echo $I0s | awk 'BEGIN {FS=":"; RS=" ";OFS=" "} {print $1,$2}' > I0s_plot.dat

      if [ `command -v gnuplot` ]
      then
        gnuplot << endoffile
         set terminal dumb
         set xrange [0:${numexp}]
         set tics out
         set yrange [*:*]
         set ylabel "I(0)"
         plot "I0s_plot.dat" u 1:2 title "I(0)" pt 1
endoffile
      fi

#*************************************************
#     SECTION FOR CHECKING CHANGES IN FOLDEDNESS    *
#*************************************************

      echo " Calculating changes in the ratio of the peak and trough values of the Kratky plot... "

    j=1
    ratios=""

    while [ $j -le $numexp ]
    do

    k=`echo $j | awk '{printf "%02d",$1; print ""}'`
      m=10
    o=$((m - 5))
      data_trends=""

    avg[o]=0

    #create a file containing the GNOM fit to the data rather than the raw data points
    awk 'BEGIN {i=0} {if ($2=="J" && $3=="EXP") i=1} {if ($1=="Distance") i=0} {if (i==1) print $0}'
${sample}${conc}_${snum}_${preexp}_${k}.out | awk 'NR>2 {if ($5!="") print $1,$5}' >
${sample}${conc}_${snum}_${preexp}_${k}.fit

      num_pts=`awk 'END {print NR}' ${sample}${conc}_${snum}_${preexp}_${k}.fit`


    #get maximum values of kratky plot by using moving averages. Once the moving average decreases, mark
previous average as maximum.
      while [ $m -le $num_pts ]
        do
                 n=$((m + 10))  # block of points to determine slope from
                 o=$((m - 5))   # compare current block of points to previous block of points

                 awk 'BEGIN {print "'$m'"RS"'$n'"} {print}' ${sample}${conc}_${snum}_${preexp}_${k}.fit >
tmp.dat

            #create kratky values from intensity values
                 awk '{if (NR==1) m=$1; if (NR==2) n=$1}  {if (NR>=m && NR<=n) print $1*1 FS $2*$1*$1}'
tmp.dat > range.dat

            #average y values of kratky plot in block of points
            avg[m]=`awk '{sum+=$2} END {print sum/NR}' range.dat`

            p=`echo "${avg[m]} ${avg[o]}" | awk '{if ($1<$2) print "1"}'`

            #set xmax and ymax if average of current block is less than average of previous block
            if [ "${p}" = "1" ]
            then
                xmax=${m}           #since point m is in the middle of the previous block, set xmax to be m
                ymax=${avg[o]}      #set ymax to be the average of the previous block, since thats higher than
the average of the current block
                break
            fi

                 let m+=5  # increment block of points by 5
        done

        echo > avgs.dat
        echo > avgs-sorted.dat     #clear file

        while [ $m -le $num_pts ]
        do
            n=$((m + 10))  # block of points to determine slope from
           o=$((m - 5))   # compare current block of points to previous block of points

                 awk 'BEGIN {print "'$m'"RS"'$n'"} {print}' ${sample}${conc}_${snum}_${preexp}_${k}.fit >
tmp.dat

            #create kratky values from intensity values
                 awk '{if (NR==1) m=$1; if (NR==2) n=$1}  {if (NR>=m && NR<=n) print $1*1 FS $2*$1*$1}'
tmp.dat > range.dat

            avg[m]=`awk '{sum+=$2} END {print sum/NR}' range.dat`

            echo "${m} ${avg[m]}" >> avgs.dat

                 let m+=5  # increment block of points by 5
        done

        xymin=`sort -k 2 avgs.dat | awk 'NR==2 {print $1, $2}'`
        xmin=`echo ${xymin} | awk '{print $1}'`
        ymin=`echo ${xymin} | awk '{print $2}'`

        echo "ymin: $ymin ymax: $ymax"
        if [ $xmax ]
        then
                 ratio[j]=`echo $ymin $ymax | awk '{print $1/$2}'`
                 ratios="${ratios} ${j}:${ratio[j]}"
                 echo " ${sample}${conc}_${snum}_${preexp}_${k}.fit:  Xmax = $xmax  Xmin = $xmin  Ymin/Ymax =
${ratio[j]}"
        else
                 ratio[j]="unfolded"
                 ratios="${ratios} ${j}:${ratio[j]}"
                 echo " ${sample}${conc}_${snum}_${preexp}_${k}.fit has no peak and appears to be unfolded."
        fi

        let j++

      done

    unf=`echo ${ratio[@]} | awk '{if ($0 ~ "unfolded") print "unfolded"}'`

    if [ "${unf}" != "unfolded" ]
    then
        # linreg.awk: An awk script to compute linear regression
        # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
        # Usage: awk -f linreg.awk file
        #
        ratio_fit=`echo $ratios | awk '
            BEGIN{
             FS=":"; RS=" ";
            }

            {
             delta = $2 - avg;
             avg += delta / NR;
             mean2 += delta * ($2 - avg);
             x[NR] = $1; y[NR] = $2;
             sx += x[NR]; sy += y[NR];
             sxx += x[NR]*x[NR];
             sxy += x[NR]*y[NR];
             syy += y[NR]*y[NR];
            }

            END{
             sd = sqrt (mean2/NR);
             det = NR*sxx - sx*sx;
             a = (NR*sxy - sx*sy)/det;
             b = (-sx*sxy+sxx*sy)/det;
             se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
             sb = sqrt(NR*se*se/det);
             t = a/sb;
             print a, b, t, avg, sd;
            }' `

        ratio_trend=`echo $ratio_fit | awk '{printf "%5.3f", $1}'`
        ratio_yint=`echo $ratio_fit | awk '{printf "%5.3f", $2}'`
        ratio_t=`echo $ratio_fit | awk '{printf "%5.3f", $3}'`
        ratio_mean=`echo $ratio_fit | awk '{printf "%5.3f", $4}'`
        ratio_sd=`echo $ratio_fit | awk '{printf "%5.3f", $5}'`
    else
        ratio_trend="unfolded"
        ratio_yint="unfolded"
        ratio_t="unfolded"
        ratio_mean="unfolded"
        ratio_sd="unfolded"
    fi

#********************************************************

      chi_p=rg_p=rg_gnom_p=dmax_p=I0_p=ratio_p=""

      chi_t=`echo $chi_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
      rg_t=`echo $rg_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
      rg_gnom_t=`echo $rg_gnom_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
      dmax_t=`echo $dmax_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
      I0_t=`echo $I0_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`

      ndf=`expr $numexp - 2`  #number of degrees of freedom = (number of exposures - 2); two-tailed test

      if [ $ndf -gt 40 ]
      then
        ndf=40
      fi

      #calculate p-values from t-statistic.
      chi_p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$chi_t"' < $2) print "0.3";
         else if ('"$chi_t"'>$9) print "0.001";
         else if ('"$chi_t"'>$8) print "0.002";
         else if ('"$chi_t"'>$7) print "0.005";
         else if ('"$chi_t"'>$6) print "0.01";
         else if ('"$chi_t"'>$5) print "0.02";
         else if ('"$chi_t"'>$4) print "0.05";
         else if ('"$chi_t"'>$3) print "0.1";
         else if ('"$chi_t"'>$2) print "0.2";
        }}' <<TDIST1
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
1     3.078      6.314    12.70    31.82    63.65    127.3    318.3    636.6
2     1.886      2.920    4.303    6.965    9.925    14.08    22.32    31.59
3     1.638      2.353    3.182    4.541    5.841    7.453    10.21    12.92
4     1.533      2.132    2.776    3.747    4.604    5.598    7.173    8.610
5     1.476      2.015    2.571    3.365    4.032    4.773    5.893    6.869
6     1.440      1.943    2.447    3.143    3.707    4.317    5.208    5.959
7     1.415      1.895    2.365    2.998    3.499    4.029    4.785    5.408
8     1.397      1.860    2.306    2.897    3.355    3.833    4.501    5.041
9     1.383      1.833    2.262    2.821    3.250    3.690    4.297    4.781
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
17    1.333      1.740    2.110    2.567    2.898    3.222    3.646    3.965
18    1.330      1.734    2.101    2.552    2.878    3.197    3.610    3.922
19    1.328      1.729    2.093    2.539    2.861    3.174    3.579    3.883
20    1.325      1.725    2.086    2.528    2.845    3.153    3.552    3.850
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST1
`

      rg_p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$rg_t"' < $2) print "0.3";
         else if ('"$rg_t"'>$9) print "0.001";
         else if ('"$rg_t"'>$8) print "0.002";
```

```
          else if ('"$rg_t"'>$7) print "0.005";
          else if ('"$rg_t"'>$6) print "0.01";
          else if ('"$rg_t"'>$5) print "0.02";
          else if ('"$rg_t"'>$4) print "0.05";
          else if ('"$rg_t"'>$3) print "0.1";
          else if ('"$rg_t"'>$2) print "0.2";
        }}' <<TDIST2
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
1     3.078      6.314    12.70    31.82    63.65    127.3    318.3    636.6
2     1.886      2.920    4.303    6.965    9.925    14.08    22.32    31.59
3     1.638      2.353    3.182    4.541    5.841    7.453    10.21    12.92
4     1.533      2.132    2.776    3.747    4.604    5.598    7.173    8.610
5     1.476      2.015    2.571    3.365    4.032    4.773    5.893    6.869
6     1.440      1.943    2.447    3.143    3.707    4.317    5.208    5.959
7     1.415      1.895    2.365    2.998    3.499    4.029    4.785    5.408
8     1.397      1.860    2.306    2.897    3.355    3.833    4.501    5.041
9     1.383      1.833    2.262    2.821    3.250    3.690    4.297    4.781
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
17    1.333      1.740    2.110    2.567    2.898    3.222    3.646    3.965
18    1.330      1.734    2.101    2.552    2.878    3.197    3.610    3.922
19    1.328      1.729    2.093    2.539    2.861    3.174    3.579    3.883
20    1.325      1.725    2.086    2.528    2.845    3.153    3.552    3.850
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST2
`

      rg_gnom_p=`awk '
        {if (NR==('"$ndf"'+1)) {
          if ('"$rg_gnom_t"' < $2) print "0.3";
          else if ('"$rg_gnom_t"'>$9) print "0.001";
          else if ('"$rg_gnom_t"'>$8) print "0.002";
          else if ('"$rg_gnom_t"'>$7) print "0.005";
          else if ('"$rg_gnom_t"'>$6) print "0.01";
          else if ('"$rg_gnom_t"'>$5) print "0.02";
          else if ('"$rg_gnom_t"'>$4) print "0.05";
          else if ('"$rg_gnom_t"'>$3) print "0.1";
          else if ('"$rg_gnom_t"'>$2) print "0.2";
        }}' <<TDIST2
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
```

```
1      3.078      6.314     12.70     31.82     63.65     127.3     318.3     636.6
2      1.886      2.920     4.303     6.965     9.925     14.08     22.32     31.59
3      1.638      2.353     3.182     4.541     5.841     7.453     10.21     12.92
4      1.533      2.132     2.776     3.747     4.604     5.598     7.173     8.610
5      1.476      2.015     2.571     3.365     4.032     4.773     5.893     6.869
6      1.440      1.943     2.447     3.143     3.707     4.317     5.208     5.959
7      1.415      1.895     2.365     2.998     3.499     4.029     4.785     5.408
8      1.397      1.860     2.306     2.897     3.355     3.833     4.501     5.041
9      1.383      1.833     2.262     2.821     3.250     3.690     4.297     4.781
10     1.372      1.812     2.228     2.764     3.169     3.581     4.144     4.587
11     1.363      1.796     2.201     2.718     3.106     3.497     4.025     4.437
12     1.356      1.782     2.179     2.681     3.055     3.428     3.930     4.318
13     1.350      1.771     2.160     2.650     3.012     3.372     3.852     4.221
14     1.345      1.761     2.145     2.625     2.977     3.326     3.787     4.140
15     1.341      1.753     2.131     2.602     2.947     3.286     3.733     4.073
16     1.337      1.746     2.120     2.584     2.921     3.252     3.686     4.015
17     1.333      1.740     2.110     2.567     2.898     3.222     3.646     3.965
18     1.330      1.734     2.101     2.552     2.878     3.197     3.610     3.922
19     1.328      1.729     2.093     2.539     2.861     3.174     3.579     3.883
20     1.325      1.725     2.086     2.528     2.845     3.153     3.552     3.850
21     1.323      1.721     2.080     2.518     2.831     3.135     3.527     3.819
22     1.321      1.717     2.074     2.508     2.819     3.119     3.505     3.792
23     1.319      1.714     2.069     2.500     2.807     3.104     3.485     3.768
24     1.318      1.711     2.064     2.492     2.797     3.090     3.467     3.745
25     1.316      1.708     2.060     2.485     2.787     3.078     3.450     3.725
26     1.315      1.706     2.056     2.479     2.779     3.067     3.435     3.707
27     1.314      1.703     2.052     2.473     2.771     3.057     3.421     3.690
28     1.313      1.701     2.048     2.467     2.763     3.047     3.408     3.674
29     1.311      1.699     2.045     2.462     2.756     3.038     3.396     3.659
30     1.310      1.697     2.042     2.457     2.750     3.030     3.385     3.646
31     1.309      1.695     2.040     2.453     2.744     3.022     3.375     3.633
32     1.309      1.694     2.037     2.449     2.738     3.015     3.365     3.622
33     1.308      1.692     2.035     2.445     2.733     3.008     3.356     3.611
34     1.307      1.691     2.032     2.441     2.728     3.002     3.348     3.601
35     1.306      1.690     2.030     2.438     2.724     2.996     3.340     3.591
36     1.306      1.688     2.028     2.434     2.719     2.991     3.333     3.582
37     1.305      1.687     2.026     2.431     2.715     2.985     3.326     3.574
38     1.304      1.686     2.024     2.429     2.712     2.980     3.319     3.566
39     1.304      1.685     2.023     2.426     2.708     2.976     3.313     3.558
40     1.303      1.684     2.021     2.423     2.704     2.971     3.307     3.551
42     1.302      1.682     2.018     2.418     2.698     2.963     3.296     3.538
44     1.301      1.680     2.015     2.414     2.692     2.956     3.286     3.526
46     1.300      1.679     2.013     2.410     2.687     2.949     3.277     3.515
48     1.299      1.677     2.011     2.407     2.682     2.943     3.269     3.505
50     1.299      1.676     2.009     2.403     2.678     2.937     3.261     3.496
60     1.296      1.671     2.000     2.390     2.660     2.915     3.232     3.460
70     1.294      1.667     1.994     2.381     2.648     2.899     3.211     3.435
80     1.292      1.664     1.990     2.374     2.639     2.887     3.195     3.416
90     1.291      1.662     1.987     2.369     2.632     2.878     3.183     3.402
100    1.290      1.660     1.984     2.364     2.626     2.871     3.174     3.391
120    1.289      1.658     1.980     2.358     2.617     2.860     3.160     3.373
150    1.287      1.655     1.976     2.351     2.609     2.849     3.145     3.357
200    1.286      1.652     1.972     2.345     2.601     2.839     3.131     3.340
300    1.284      1.650     1.968     2.339     2.592     2.828     3.118     3.323
500    1.283      1.648     1.965     2.334     2.586     2.820     3.107     3.310
TDIST2
`
      dmax_p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$dmax_t"' < $2) print "0.3";
         else if ('"$dmax_t"'>$9) print "0.001";
         else if ('"$dmax_t"'>$8) print "0.002";
         else if ('"$dmax_t"'>$7) print "0.005";
         else if ('"$dmax_t"'>$6) print "0.01";
         else if ('"$dmax_t"'>$5) print "0.02";
         else if ('"$dmax_t"'>$4) print "0.05";
         else if ('"$dmax_t"'>$3) print "0.1";
         else if ('"$dmax_t"'>$2) print "0.2";
        }}' <<TDIST2
DF     0.200      0.100     0.050     0.020     0.010     0.005     0.002     0.001
1      3.078      6.314     12.70     31.82     63.65     127.3     318.3     636.6
2      1.886      2.920     4.303     6.965     9.925     14.08     22.32     31.59
3      1.638      2.353     3.182     4.541     5.841     7.453     10.21     12.92
4      1.533      2.132     2.776     3.747     4.604     5.598     7.173     8.610
5      1.476      2.015     2.571     3.365     4.032     4.773     5.893     6.869
6      1.440      1.943     2.447     3.143     3.707     4.317     5.208     5.959
7      1.415      1.895     2.365     2.998     3.499     4.029     4.785     5.408
8      1.397      1.860     2.306     2.897     3.355     3.833     4.501     5.041
9      1.383      1.833     2.262     2.821     3.250     3.690     4.297     4.781
```

```
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
17    1.333      1.740    2.110    2.567    2.898    3.222    3.646    3.965
18    1.330      1.734    2.101    2.552    2.878    3.197    3.610    3.922
19    1.328      1.729    2.093    2.539    2.861    3.174    3.579    3.883
20    1.325      1.725    2.086    2.528    2.845    3.153    3.552    3.850
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST2
`


      I0_p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$I0_t"' < $2) print "0.3";
         else if ('"$I0_t"'>$9) print "0.001";
         else if ('"$I0_t"'>$8) print "0.002";
         else if ('"$I0_t"'>$7) print "0.005";
         else if ('"$I0_t"'>$6) print "0.01";
         else if ('"$I0_t"'>$5) print "0.02";
         else if ('"$I0_t"'>$4) print "0.05";
         else if ('"$I0_t"'>$3) print "0.1";
         else if ('"$I0_t"'>$2) print "0.2";
        }}' <<TDIST3
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
1     3.078      6.314    12.70    31.82    63.65    127.3    318.3    636.6
2     1.886      2.920    4.303    6.965    9.925    14.08    22.32    31.59
3     1.638      2.353    3.182    4.541    5.841    7.453    10.21    12.92
4     1.533      2.132    2.776    3.747    4.604    5.598    7.173    8.610
5     1.476      2.015    2.571    3.365    4.032    4.773    5.893    6.869
6     1.440      1.943    2.447    3.143    3.707    4.317    5.208    5.959
7     1.415      1.895    2.365    2.998    3.499    4.029    4.785    5.408
8     1.397      1.860    2.306    2.897    3.355    3.833    4.501    5.041
9     1.383      1.833    2.262    2.821    3.250    3.690    4.297    4.781
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
```

```
17    1.333     1.740   2.110   2.567   2.898   3.222   3.646   3.965
18    1.330     1.734   2.101   2.552   2.878   3.197   3.610   3.922
19    1.328     1.729   2.093   2.539   2.861   3.174   3.579   3.883
20    1.325     1.725   2.086   2.528   2.845   3.153   3.552   3.850
21    1.323     1.721   2.080   2.518   2.831   3.135   3.527   3.819
22    1.321     1.717   2.074   2.508   2.819   3.119   3.505   3.792
23    1.319     1.714   2.069   2.500   2.807   3.104   3.485   3.768
24    1.318     1.711   2.064   2.492   2.797   3.090   3.467   3.745
25    1.316     1.708   2.060   2.485   2.787   3.078   3.450   3.725
26    1.315     1.706   2.056   2.479   2.779   3.067   3.435   3.707
27    1.314     1.703   2.052   2.473   2.771   3.057   3.421   3.690
28    1.313     1.701   2.048   2.467   2.763   3.047   3.408   3.674
29    1.311     1.699   2.045   2.462   2.756   3.038   3.396   3.659
30    1.310     1.697   2.042   2.457   2.750   3.030   3.385   3.646
31    1.309     1.695   2.040   2.453   2.744   3.022   3.375   3.633
32    1.309     1.694   2.037   2.449   2.738   3.015   3.365   3.622
33    1.308     1.692   2.035   2.445   2.733   3.008   3.356   3.611
34    1.307     1.691   2.032   2.441   2.728   3.002   3.348   3.601
35    1.306     1.690   2.030   2.438   2.724   2.996   3.340   3.591
36    1.306     1.688   2.028   2.434   2.719   2.991   3.333   3.582
37    1.305     1.687   2.026   2.431   2.715   2.985   3.326   3.574
38    1.304     1.686   2.024   2.429   2.712   2.980   3.319   3.566
39    1.304     1.685   2.023   2.426   2.708   2.976   3.313   3.558
40    1.303     1.684   2.021   2.423   2.704   2.971   3.307   3.551
42    1.302     1.682   2.018   2.418   2.698   2.963   3.296   3.538
44    1.301     1.680   2.015   2.414   2.692   2.956   3.286   3.526
46    1.300     1.679   2.013   2.410   2.687   2.949   3.277   3.515
48    1.299     1.677   2.011   2.407   2.682   2.943   3.269   3.505
50    1.299     1.676   2.009   2.403   2.678   2.937   3.261   3.496
60    1.296     1.671   2.000   2.390   2.660   2.915   3.232   3.460
70    1.294     1.667   1.994   2.381   2.648   2.899   3.211   3.435
80    1.292     1.664   1.990   2.374   2.639   2.887   3.195   3.416
90    1.291     1.662   1.987   2.369   2.632   2.878   3.183   3.402
100   1.290     1.660   1.984   2.364   2.626   2.871   3.174   3.391
120   1.289     1.658   1.980   2.358   2.617   2.860   3.160   3.373
150   1.287     1.655   1.976   2.351   2.609   2.849   3.145   3.357
200   1.286     1.652   1.972   2.345   2.601   2.839   3.131   3.340
300   1.284     1.650   1.968   2.339   2.592   2.828   3.118   3.323
500   1.283     1.648   1.965   2.334   2.586   2.820   3.107   3.310
TDIST3
`

    if [ "${unf}" != "unfolded" ]
    then
        ratio_t=`echo $ratio_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`

      ratio_p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$ratio_t"' < $2) print "0.3";
         else if ('"$ratio_t"'>$9) print "0.001";
         else if ('"$ratio_t"'>$8) print "0.002";
         else if ('"$ratio_t"'>$7) print "0.005";
         else if ('"$ratio_t"'>$6) print "0.01";
         else if ('"$ratio_t"'>$5) print "0.02";
         else if ('"$ratio_t"'>$4) print "0.05";
         else if ('"$ratio_t"'>$3) print "0.1";
         else if ('"$ratio_t"'>$2) print "0.2";
        }}' <<TDIST3
DF    0.200     0.100   0.050   0.020   0.010   0.005   0.002   0.001
1     3.078     6.314   12.70   31.82   63.65   127.3   318.3   636.6
2     1.886     2.920   4.303   6.965   9.925   14.08   22.32   31.59
3     1.638     2.353   3.182   4.541   5.841   7.453   10.21   12.92
4     1.533     2.132   2.776   3.747   4.604   5.598   7.173   8.610
5     1.476     2.015   2.571   3.365   4.032   4.773   5.893   6.869
6     1.440     1.943   2.447   3.143   3.707   4.317   5.208   5.959
7     1.415     1.895   2.365   2.998   3.499   4.029   4.785   5.408
8     1.397     1.860   2.306   2.897   3.355   3.833   4.501   5.041
9     1.383     1.833   2.262   2.821   3.250   3.690   4.297   4.781
10    1.372     1.812   2.228   2.764   3.169   3.581   4.144   4.587
11    1.363     1.796   2.201   2.718   3.106   3.497   4.025   4.437
12    1.356     1.782   2.179   2.681   3.055   3.428   3.930   4.318
13    1.350     1.771   2.160   2.650   3.012   3.372   3.852   4.221
14    1.345     1.761   2.145   2.625   2.977   3.326   3.787   4.140
15    1.341     1.753   2.131   2.602   2.947   3.286   3.733   4.073
16    1.337     1.746   2.120   2.584   2.921   3.252   3.686   4.015
17    1.333     1.740   2.110   2.567   2.898   3.222   3.646   3.965
18    1.330     1.734   2.101   2.552   2.878   3.197   3.610   3.922
19    1.328     1.729   2.093   2.539   2.861   3.174   3.579   3.883
20    1.325     1.725   2.086   2.528   2.845   3.153   3.552   3.850
```

```
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST3
`
    else
        ratio_p="unfolded"
    fi

      sel_exp=1

      l=$sel_exp
      j=1

      if [ "$chi_p" = "0.3" ]
      then
        chi_ps=">0.2"
        chi_pe=`echo $chi_p | awk '{print int($1*10000)}'`
      else
        chi_ps=$chi_p
        chi_pe=`echo $chi_p | awk '{print int($1*10000)}'`
      fi

      if [ "$rg_p" = "0.3" ]
      then
        rg_ps=">0.2"
        rg_pe=`echo $rg_p | awk '{print int($1*10000)}'`
      else
        rg_ps=$rg_p
        rg_pe=`echo $rg_p | awk '{print int($1*10000)}'`
      fi

      if [ "$rg_gnom_p" = "0.3" ]
      then
        rg_gnom_ps=">0.2"
        rg_gnom_pe=`echo $rg_gnom_p | awk '{print int($1*10000)}'`
      else
        rg_gnom_ps=$rg_gnom_p
        rg_gnom_pe=`echo $rg_gnom_p | awk '{print int($1*10000)}'`
      fi

      if [ "$dmax_p" = "0.3" ]
      then
        dmax_ps=">0.2"
        dmax_pe=`echo $dmax_p | awk '{print int($1*10000)}'`
      else
        dmax_ps=$dmax_p
```

```
         dmax_pe=`echo $dmax_p | awk '{print int($1*10000)}'`
      fi

      if [ "$I0_p" = "0.3" ]
      then
         I0_ps=">0.2"
         I0_pe=`echo $I0_p | awk '{print int($1*10000)}'`
      else
         I0_ps=$I0_p
         I0_pe=`echo $I0_p | awk '{print int($1*10000)}'`
      fi

    if [ "$ratio_p" = "0.3" ]
      then
         ratio_ps=">0.2"
         ratio_pe=`echo $ratio_p | awk '{print int($1*10000)}'`
      elif [ "${unf}" != "unfolded" ]
      then
      ratio_ps=$ratio_p
         ratio_pe=`echo $ratio_p | awk '{print int($1*10000)}'`
    else
        ratio_ps="unfolded"
         ratio_pe="unfolded"
      fi

        chi_test_p=`echo $chi_p | awk '{if ($1>0.05) print "good"}'`
        rg_test_p=`echo $rg_p | awk '{if ($1>0.05) print "good"}'`
        rg_gnom_test_p=`echo $rg_gnom_p | awk '{if ($1>0.05) print "good"}'`
        dmax_test_p=`echo $dmax_p | awk '{if ($1>0.05) print "good"}'`
        I0_test_p=`echo $I0_p | awk '{if ($1>0.05) print "good"}'`
        ratio_test_p=`echo $ratio_p | awk '{if ($1>0.05) print "good"}'`

         echo
         echo " exposure    chi(to1)    Rg    Rg(GNOM)    Dmax       I(0)    Kratky   "
         echo " ---------  --------  -------  --------  --------  --------  --------  "

         for_avg=""
         for_avg_num=""
        n=1
        sdmultiplier="2"

         while [ $j -le $numexp ]
         do

                 i=$((j - 1))
                 k=`echo $j | awk '{printf "%02d",$1; print ""}'`

                 #chi=`datcmp ${sample}${conc}_${snum}_${preexp}_0${l}.dat
${sample}${conc}_${snum}_${preexp}_${k}.dat | awk '{printf "%0.2f", $1}'`

                 #select only those exposures yielding statistics within standard deviation of
                 #extrapolated zero radiation (y-int), i.e. do not include radiation damaged data
                 #in averaging.

                 chi_test=rg_test=rg_gnom_test=dmax_test=I0_test=ratio_test=""
            if [ "$chi_test_p" != "good" ]
            then
                echo chi
                chi_test=`echo ${chi[j]} $chi_yint $chi_sd $chi_p $sdmultiplier | awk '{if ($1 < $2+$5*$3)
print "good"; else print "bad"}'`
                 elif [ "$rg_test_p" != "good" ]
            then
                rg_test=`echo ${rg[j]} $rg_yint $rg_sd ${rg[n]} $rg_p $sdmultiplier | awk '{if ($1 < $2+$6*$3
&& $1 > $2-$6*$3) print "good"; else print "bad"}'`
                 elif [ "$rg_gnom_test_p" != "good" ]
            then
                rg_gnom_test=`echo ${rg_gnom[j]} $rg_gnom_yint $rg_gnom_sd ${rg_gnom[n]} $rg_gnom_p
$sdmultiplier | awk '{if ($1 < $2+$6*$3 && $1 > $2-$6*$3) print "good"; else print "bad"}'`
                 elif [ "$dmax_test_p" != "good" ]
            then
                dmax_test=`echo ${dmax[j]} $dmax_yint $dmax_sd ${dmax[n]} $dmax_p $sdmultiplier | awk '{if ($1
< $2+$6*$3 && $1 > $2-$6*$3) print "good"; else print "bad"}'`
                 elif [ "$I0_test_p" != "good" ]
            then
                I0_test=`echo ${I0[j]} $I0_yint $I0_sd ${I0[n]} $I0_p $sdmultiplier | awk '{if ($1 < $2+$6*$3
&& $1 > $2-$6*$3) print "good"; else print "bad"}'`
            elif [ "$ratio_test_p" != "good" ]
            then
                if [ "${unf}" != "unfolded" ]
                then
                    ratio_test=`echo ${ratio[j]} $ratio_yint $ratio_sd ${ratio[n]} $ratio_p $sdmultiplier | awk
'{if ($1 <= $2+$6*$3 && $1 >= $2-$6*$3) print "good"; else print "bad"}'`
                fi
            fi

                 if  [ "$chi_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "               ^^^^                                    "
                 elif  [ "$rg_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                        ^^^^^                           "
                 elif  [ "$rg_gnom_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                                  ^^^^^                 "
                 elif  [ "$dmax_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                                            ^^^^^        "
                 elif  [ "$I0_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                                                     ^^^^"
            elif  [ "$ratio_test" = "bad" ]
                 then
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}  <<<"
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                                                              ^^^^"
                 else
                         echo -e "${k} ${chi[j]} ${rg[j]} ${rg_gnom[j]} ${dmax[j]} ${I0[j]} ${ratio[j]}     "
| awk '{printf "%10s %8s %9s %9s %9s %9s %9s\n", $1, $2, $3, $4, $5, $6, $7}'
                         echo "                                                               "
                         #mark good exposures for averaging
                         for_avg="$for_avg ${sample}${conc}_${snum}_${preexp}_${k}.dat"
                         for_avg_num="${for_avg_num} ${j},"
                 fi
                 let j++

        done

      concl=`echo $conc | awk '{print substr($1,3,1)}'`

      numgood=`echo $for_avg | awk '{print NF}'`

      if [ $numgood -eq 0 ]
    then
        echo " ERROR IN DATA FILES.  NO GOOD EXPOSURES."
        cp ${sample}${conc}_${snum}_${preexp}_01.dat ${rename}${concl}.dat
        echo " Exposure 1 will be written to ${rename}${concl}.dat"
        echo
    elif [ $numgood -eq 1 ]
    then
        echo " Too few exposures usable.  No averaging done."
        cp $for_avg ${rename}${concl}.dat
        echo " Exposure $for_avg_num will be written to ${rename}${concl}.dat"
        echo
      else
        #average good exposures
        dataver $for_avg -o ${rename}${concl}.dat
        echo " Exposures $for_avg_num will be averaged"
        echo
        echo " ${rename}${concl}.dat written"
        echo
      fi

      #create tab delimited log file for radiation damage check for each sample in directory
      if [ 1 ]
      then
      echo " ${rename}${conc}"
    echo " Exposures used in averaging:  ${for_avg_num}"
    echo
    echo " Guinier Range:  ${qmin} < q < ${qmax}    (points [${begp},${endp}])"
      echo
    echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n",
"","Chi","Rg","Rg(GNOM)","Dmax","I(0)","Kratky Ratio"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n", "Exposure
1","'${chi[n]}'","'${rg[n]}'","'${rg_gnom[n]}'","'${dmax[n]}'","'${I0[n]}'","'${ratio[n]}'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n",
"Average","'$chi_mean'","'$rg_mean'","'$rg_gnom_mean'","'$dmax_mean'","'$I0_mean'","'$ratio_mean'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n", "Std
Dev","'$chi_sd'","'$rg_sd'","'$rg_gnom_sd'","'$dmax_sd'","'$I0_sd'","'$ratio_sd'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n",
"Slope","'$chi_trend'","'$rg_trend'","'$rg_gnom_trend'","'$dmax_trend'","'$I0_trend'","'$ratio_trend'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n", "Y-
intercept","'$chi_yint'","'$rg_yint'","'$rg_gnom_yint'","'$dmax_yint'","'$I0_yint'","'$ratio_yint'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n", "T-
statistic","'$chi_t'","'$rg_t'","'$rg_gnom_t'","'$dmax_t'","'$I0_t'","'$ratio_t'"}'
      echo 1 | awk '{printf "%12s %10s %10s %10s %10s %10s %10s\n", "P-
value","'$chi_ps'","'$rg_ps'","'$rg_gnom_ps'","'$dmax_ps'","'$I0_ps'","'$ratio_ps'"}'
      echo
      fi | tee ${rename}${concl}_rdam_stats.txt

  fi | tee ${rename}${concl}.log
done
```

Appendix C. Script for detecting concentration dependence in SAXS Profiles

```
#!/bin/bash

while getopts hf:n:c:m:M:C:I: opt; do
  case $opt in
    h)
    echo
    echo " ------------------------------------------------------------------------------ "
      echo
      echo " This script checks concentration dependence in SAXS data profiles."
      echo " In addition to calculating Rg and I(0) and detecting aggregation,"
      echo " it calculates a linear regression and generates a p-value for the "
      echo " likelihood of a trend being present, suggesting interparticle interactions."
      echo " The second virial coefficient will be calculated."
      echo
      echo " Statistics are written to <sample>_conc_stats.txt"
      echo
      echo " AutoRg, datcmp, and datop (from ATSAS) must be in \$PATH."
      echo
      echo " If gnuplot exists in \$PATH, a plot will be printed to the terminal"
      echo " showing the trend in Rg and 1/S(0) vs. concentration. "
      echo
      echo " Usage:  conc.sh [options] "
      echo
      echo " where [options] are:"
    echo
    echo " -h:  Print this help menu and exit"
    echo " -f:  File name of the first concentration of a series (required)"
    echo " -n:  Number of concentrations to use (optional, defaults to # of conc with file name)"
    echo " -c:  The concentration (in mg/ml) of the first solution in the series (optional, but required to 
calculate A2)"
    echo " -m:  Molecular weight in kDa (optional, otherwise calculated from Porod volume)"
    echo " -M:  Molecular weight in kDa of protein standard (optional, required to calculate A2)"
    echo " -C:  Concentration (in mg/ml) of protein standard (optional, required to calculate A2)"
    echo " -I:  I(0) of protein standard (optional, required to calculate A2)"
      echo
      echo " The file name must observe the following naming protocol:"
    echo
      echo " <samplename>_<concentration>.dat  "
      echo
      echo " where <concentration> is a capital letter from B through H."
    echo
    echo " For example:  conc.sh -f Protein-1_B.dat -n 5 -m 50 -M 66.1 -C 1 -I 703"
    echo
    echo " will estimate the trend for Protein-1_B.dat, Protein-1_C.dat, ..., Protein-1_F.dat"
    echo " and calculate the A2 value assuming a molecular weight of 50 kDa for Protein-1, and"
    echo " a protein standard with a mol. wt. of 66.1 kDa at a concentration of 1 mg/ml and an I(0)=703."
    echo
    echo " * If -c is given, it will be given priority if the protein standard is also given."
      echo
      echo " The extension of the file name must be .dat"
    echo
    echo " ----------------------------------------------------------------------------- "
    echo
    exit 0
    ;;
    f)
      firstconc=$OPTARG
      ;;
    n)
      numconc=$OPTARG
      ;;
    c)
      concvalue=$OPTARG
      ;;
    m)
      mw_given=$OPTARG
      ;;
    M)
      mw_standard=$OPTARG
      ;;
    C)
      conc_standard=$OPTARG
      ;;
    I)
      I0_standard=$OPTARG
      ;;
```

```
   \?)
      echo "Invalid option: -$OPTARG" >&2
      exit 1
      ;;
    :)
      echo "Option -$OPTARG requires an argument." #>&2
      exit 1
      ;;
  esac
done

if [ "$firstconc" == "" ]
then
      echo
      echo " To get a quick help, type conc.sh -h"
      echo
      echo " Enter file name of first concentration of sample: "
      read firstconc
fi

sample=`echo ${firstconc%.*} | awk 'BEGIN {FS="_"} {for (i=1; i<NF; i++) printf "%s_", $i;print ""}'`
conc=`echo ${firstconc%.*} | awk 'BEGIN {FS="_"} {print $NF}'`

#get number of concentrations for default use
numconcfiles=0

for i in ${sample}[B-H].dat
do
    let numconcfiles++
done

if [ -z $numconc ]
then
    numconc=${numconcfiles}
fi

inttest=`echo $numconc | awk '{if (int($1) == $1) print "int"}'`

if [[ $inttest != "int" ]]
then
      echo " Error:  Number of concentrations must be an integer value!"
      exit
elif [[ $numconc -lt 3 ]]
then
    echo
    echo " You must have at least three concentrations to calculate a trend!"
    echo " Rg's will be calculated and plotted, but no trend or A2 value will be determined."
    calctrend="no"
fi

if [ -z $mw_given ]
then
    echo
    echo " No molecular weight given. A2 will be calculated using Porod estimation. "
else
    inttest=`echo $mw_given | awk '{if (int($1) == $1) print "int"}'`
    if [[ $inttest != "int" ]]
    then
        echo " Error:  Molecular weight must be an integer value in kDa!"
        exit 0
    fi
fi

if [[ $mw_standard ]] && [[ $conc_standard ]] && [[ $I0_standard ]]
then
    calcconc="yes"
    echo " Concentrations will be estimated from standard."
    calcA2="yes"
elif [[ $concvalue ]] && [[ $numconc -ge 3 ]]
then
    calcA2="yes"
    echo " Concentration given."
fi



for i in $firstconc
do
      sample=`echo ${i%.*} | awk 'BEGIN {FS="_"} {for (i=1; i<NF; i++) printf "%s_", $i;print ""}'`
      conc=`echo ${i%.*} | awk 'BEGIN {FS="_"} {print $NF}'`
      agg_alert="no"

      #increment concentration by one letter each
      files=`echo "${conc} ${numconc} ${sample}" | awk 'BEGIN {split("BCDEFGH",A,""); ORS=" "} {if($1=="B")
i=1; else if($1=="C") i=2; else if($1=="D") i=3; else if($1=="E") i=4; else if($1=="F") i=5; else if($1=="G")
i=6; else if($1=="H") i=7} {for(j=i;j<=7;j++) print $3 A[j]".dat"}'`

      l=1
    numconcfiles=0
      while [ $l -le $numconc ]
      do
        file[l]=`echo $files | awk '{print $'"$l"'}'`
        if [ ! -f ${file[l]} ]
        then
            unset file[l]
            break
        fi
        let l++
      done

    numconcfiles=${#file[*]}

    numconc=$numconcfiles

    if [[ $numconc -lt 3 ]]
    then
        echo
        echo " You must have at least three concentrations to calculate a trend!"
        echo " Rg's will be calculated and plotted, but no trend or A2 value will be determined."
    fi

      echo
      echo " ---------------------------   ${sample}${conc}  ------------------------------"
      echo

      # Determine the relative concentrations straight from the data instead.  Divide the data
      # of the first concentration by the second, and take the median difference in points 100-200
      # to estimate the relative concentrations.  Determine trend in Rgs using these concentrations.

      l=1
      while [ $l -le ${numconc} ]
      do
        med[l]=`datop DIV ${file[1]} ${file[l]} | awk '{if (NR>75 && NR<125) print $2}' | sort -g |\
                 awk '{x[NR]=$1} END {num=int((NR+1)/2); if (NR%2==1) print x[num]; else print (x[num] +
x[num+1])/2}'`
        medi[l]=`echo ${med[l]} | awk '{print 1/$1}'`
        isneg[l]=`echo ${med[l]} | awk '{if ($1<0) print "yes"}'`
        if [ "${isneg[l]}" == "yes" ]
        then
            cat ${file[l]} > ${sample}${l}.dat
            echo " ERROR:  Negative Scale Factor"
            echo " No trends or A2 will be calculated"
            calctrend="no"
            calcA2="no"
            isneg="yes"
        else
            datop DIV ${file[l]} ${medi[l]} > ${sample}${l}.dat
        fi
        let l++
      done

      j=1
      for k in ${sample}[1-${numconc}].dat
      do
        gnom[j]=`datgnom $k`
        rg_gnom[j]=`echo ${gnom[j]} | awk '{printf "%.2f", $NF}'`
        dmax[j]=`echo ${gnom[j]} | awk '{printf "%.2f", $3}'`
        I0_gnom[j]=`tail -1 ${k%.*}.out | awk '{printf "%.2f", $(NF-2)}'`

        porod[j]=`datporod ${k%.*}.out | awk '{printf "%.0f", $1*1}'`
        mw_porod[j]=`echo ${porod[j]} | awk '{print $1/1.66/1000}'`

        echo "Concentration $j: Rg[GNOM]: ${rg_gnom[j]} I(0)[GNOM]: ${I0_gnom[j]} Dmax: ${dmax[j]} MW[Porod]:
${mw_porod[j]}"

        qmin[j]=`echo ${dmax[j]} | awk '{printf "%.5f", 3.14159/$1}'`
        qmax[j]=`echo ${rg_gnom[j]} | awk '{printf "%.5f", 1.3/$1}'`
        begp[j]=`awk '{q='"${qmin[j]}"'; if ($1 < q) i=NR} END {print i}' $firstconc`
        endp[j]=`awk '{q='"${qmax[j]}"'; if ($1 < q) i=NR} END {print i}' $firstconc`

        if [ "${begp[j]}" == "" -o "${endp[j]}" == "" ]
        then
           begp[j]=2
           endp[j]=6
        fi

        bediff=`echo ${begp[j]} ${endp[j]} | awk '{print $2-$1}'`

        if [ ${bediff} -lt 6 ]
        then
            if [ ${endp[j]} -gt 6 ]
            then
                begp[j]=`echo ${endp[j]} | awk '{print $1-5}'`
            else
               begp[j]=2
               endp[j]=6
            fi
        fi

            if [ -n "${begp[j]}" ]
            then
                # linreg.awk: An awk script to compute linear regression
                # Input columns x and y, outputs a=slope and b=intercept
                # Usage: awk -f linreg.awk file
                #
                rg_fit=`awk 'NR>='"${begp[j]}"' && NR<='"${endp[j]}"' {print $1**2,log(sqrt($2*$2)) }' $k | awk
'

                {
                 delta = $2 - avg;
                 avg += delta / NR;
                 mean2 += delta * ($2 - avg);
                 x[NR] = $1; y[NR] = $2;
                 sx += x[NR]; sy += y[NR];
                 sxx += x[NR]*x[NR];
                 sxy += x[NR]*y[NR];
                 syy += y[NR]*y[NR];
                }

                END{
                 sd = sqrt (mean2/NR);
                 det = NR*sxx - sx*sx;
                 a = (NR*sxy - sx*sy)/det;
                 b = (-sx*sxy+sxx*sy)/det;
                 se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
                 sb = sqrt(NR*se*se/det);
                 t = a/sb;
                 #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
                 print a, b, t, avg, sd;
                }'`
                rg[j]=`echo $rg_fit | awk '{print sqrt(-3*$1)}'`
                I0[j]=`echo $rg_fit | awk '{print exp($2)}'`
            fi

    #*******************************************************
    #test for aggregation by looking for significant trend
    #in slope of every three points in guinier region
    #use first point in data file (not pi/dmax, i.e. $begp) to 1.3/Rg (already calculated as $endp)
    #do it again using $begp
    #*******************************************************

      m=1
    #create file with points in guinier plot, i.e. ln(I) vs q^2
    awk '{if (NR>=1 && NR<='"${endp[j]}"') print $1*$1,log(sqrt($2*$2)*1)}' ${k} > guinier.dat
    num_pts=`awk 'END {print NR}' guinier.dat`
    max_pt=`echo $num_pts | awk '{print $1-2}'`
    if [ ${max_pt} -lt 3 ]
    then
        max_pt=3
    fi

    while [ $m -le $max_pt ]
      do
        n=$((m + 2))  # block of points to determine slope from

        awk '{if (NR>='"$m"' && NR<='"$n"') print}' guinier.dat > tmp.dat
```

```
         # linreg.awk: An awk script to compute linear regression
         # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
         # Usage: awk -f linreg.awk file
         #

         slope[m]=`awk '
         { x[NR] = $1; y[NR] = $2;
          sx += x[NR]; sy += y[NR];
          sxx += x[NR]*x[NR];
          sxy += x[NR]*y[NR];
         }

         END{
          det = NR*sxx - sx*sx;
          a = (NR*sxy - sx*sy)/det;
          print a;
         }' tmp.dat`

        slopes="${slopes} ${m}:${slope[m]}"
         let m+=1  # increment block of points

      done

        # linreg.awk: An awk script to compute linear regression
        # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
        # Usage: awk -f linreg.awk file
        #
        slopes_fit=`echo $slopes | awk '
            BEGIN{
             FS=":"; RS=" ";
            }

            {
             delta = $2 - avg;
             avg += delta / NR;
             mean2 += delta * ($2 - avg);
             x[NR] = $1; y[NR] = $2;
             sx += x[NR]; sy += y[NR];
             sxx += x[NR]*x[NR];
             sxy += x[NR]*y[NR];
             syy += y[NR]*y[NR];
            }

            END{
             sd = sqrt (mean2/NR);
             det = NR*sxx - sx*sx;
             a = (NR*sxy - sx*sy)/det;
             b = (-sx*sxy+sxx*sy)/det;
             se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
             sb = sqrt(NR*se*se/det);
             t = a/sb;
             print a, b, t, avg, sd;
            }' `


        slopes_trend=`echo $slopes_fit | awk '{printf "%5.3f", $1}'`
        slopes_yint=`echo $slopes_fit | awk '{printf "%5.3f", $2}'`
        slopes_t=`echo $slopes_fit | awk '{printf "%5.3f", $3}'`
        slopes_mean=`echo $slopes_fit | awk '{printf "%5.3f", $4}'`
        slopes_sd=`echo $slopes_fit | awk '{printf "%5.3f", $5}'`

      slopes_t=`echo $slopes_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`

    if [ ${max_pt} -gt 2 ]
    then
      ndf=`expr $max_pt - 2`  #number of degrees of freedom = (number of points in region - 2); two-tailed
test
      if [ $ndf -gt 40 ]
    then
        ndf=40
    fi

      #calculate p-values from t-statistic

      slopes_p=`awk '
        {if ($1==('"$ndf"')) {
         if ('"$slopes_t"' <= $2) print ">0.2";
         else if ('"$slopes_t"'>$9) print "0.001";
         else if ('"$slopes_t"'>$8) print "0.002";
```

```
          else if ('"$slopes_t"'>$7) print "0.005";
          else if ('"$slopes_t"'>$6) print "0.01";
          else if ('"$slopes_t"'>$5) print "0.02";
          else if ('"$slopes_t"'>$4) print "0.05";
          else if ('"$slopes_t"'>$3) print "0.1";
          else if ('"$slopes_t"'>$2) print "0.2";
         }
        }' <<TDIST1
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
1     3.078      6.314    12.70    31.82    63.65    127.3    318.3    636.6
2     1.886      2.920    4.303    6.965    9.925    14.08    22.32    31.59
3     1.638      2.353    3.182    4.541    5.841    7.453    10.21    12.92
4     1.533      2.132    2.776    3.747    4.604    5.598    7.173    8.610
5     1.476      2.015    2.571    3.365    4.032    4.773    5.893    6.869
6     1.440      1.943    2.447    3.143    3.707    4.317    5.208    5.959
7     1.415      1.895    2.365    2.998    3.499    4.029    4.785    5.408
8     1.397      1.860    2.306    2.897    3.355    3.833    4.501    5.041
9     1.383      1.833    2.262    2.821    3.250    3.690    4.297    4.781
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
17    1.333      1.740    2.110    2.567    2.898    3.222    3.646    3.965
18    1.330      1.734    2.101    2.552    2.878    3.197    3.610    3.922
19    1.328      1.729    2.093    2.539    2.861    3.174    3.579    3.883
20    1.325      1.725    2.086    2.528    2.845    3.153    3.552    3.850
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST1
`
fi
    slopesp[j]="${slopes_p}"
    if [ "${slopes_p}" != ">0.2" ]
    then
        if [ "`echo 1 | awk '{if ('"${slopes_trend}"' > 0) print "positive"}'`" = "positive" ]
        then
            aggs[j]="*** Aggregation likely ***"
        elif [ "`echo 1 | awk '{if ('"${slopes_trend}"' < 0) print "negative"}'`" = "negative" ]
        then
            aggs[j]="*** Repulsion likely ***"
        fi
    fi
    echo
    slopes=""

    #*************************************
m=1
      awk '{if (NR>='"${begp[j]}"' && NR<='"${endp[j]}"') print $1*$1,log($2*1)}' ${k} > guinier2.dat
    num_pts2=`awk 'END {print NR}' guinier2.dat`
    max_pt2=`echo $num_pts2 | awk '{print $1-2}'`

if [ ${max_pt2} -lt 3 ]
    then
        max_pt2=3
    fi

    while [ $m -le $max_pt2 ]
      do
         n=$((m + 2))  # block of points to determine slope from

         awk '{if (NR>='"$m"' && NR<='"$n"') print}' guinier2.dat > tmp.dat

         # linreg.awk: An awk script to compute linear regression
         # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
         # Usage: awk -f linreg.awk file
         #

         slope2[m]=`awk '
         { x[NR] = $1; y[NR] = $2;
          sx += x[NR]; sy += y[NR];
          sxx += x[NR]*x[NR];
          sxy += x[NR]*y[NR];
         }

         END{
          det = NR*sxx - sx*sx;
          a = (NR*sxy - sx*sy)/det;
          print a;
         }' tmp.dat`

        slopes2="${slopes2} ${m}:${slope2[m]}"
         let m+=1  # increment block of points

      done

        # linreg.awk: An awk script to compute linear regression
        # Input columns x and y, outputs a=slope and b=intercept and t=t-statistic
        # Usage: awk -f linreg.awk file
        #
        slopes2_fit=`echo $slopes2 | awk '
            BEGIN{
             FS=":"; RS=" ";
            }

            {
             delta = $2 - avg;
             avg += delta / NR;
             mean2 += delta * ($2 - avg);
             x[NR] = $1; y[NR] = $2;
             sx += x[NR]; sy += y[NR];
             sxx += x[NR]*x[NR];
             sxy += x[NR]*y[NR];
             syy += y[NR]*y[NR];
            }

            END{
             sd = sqrt (mean2/NR);
             det = NR*sxx - sx*sx;
             a = (NR*sxy - sx*sy)/det;
             b = (-sx*sxy+sxx*sy)/det;
             se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
             sb = sqrt(NR*se*se/det);
             t = a/sb;
             print a, b, t, avg, sd;
            }' `


        slopes2_trend=`echo $slopes2_fit | awk '{printf "%5.3f", $1}'`
        slopes2_yint=`echo $slopes2_fit | awk '{printf "%5.3f", $2}'`
        slopes2_t=`echo $slopes2_fit | awk '{printf "%5.3f", $3}'`
        slopes2_mean=`echo $slopes2_fit | awk '{printf "%5.3f", $4}'`
        slopes2_sd=`echo $slopes2_fit | awk '{printf "%5.3f", $5}'`
      slopes2_t=`echo $slopes2_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`

    if [ ${max_pt2} -gt 2 ]
    then
      ndf=`expr $max_pt2 - 2`  #number of degrees of freedom = (number of points in region - 2); two-tailed
test
      if [ $ndf -gt 40 ]
    then
        ndf=40
    fi

      #calculate p-values from t-statistic

      slopes2_p=`awk '
        {if ($1==('"$ndf"')) {
         if ('"$slopes2_t"' <= $2) print ">0.2";
         else if ('"$slopes2_t"'>$9) print "0.001";
         else if ('"$slopes2_t"'>$8) print "0.002";
         else if ('"$slopes2_t"'>$7) print "0.005";
         else if ('"$slopes2_t"'>$6) print "0.01";
         else if ('"$slopes2_t"'>$5) print "0.02";
         else if ('"$slopes2_t"'>$4) print "0.05";
         else if ('"$slopes2_t"'>$3) print "0.1";
         else if ('"$slopes2_t"'>$2) print "0.2";
        }
        }' <<TDIST1
DF    0.200      0.100    0.050   0.020    0.010   0.005    0.002   0.001
1     3.078      6.314    12.70   31.82    63.65   127.3    318.3   636.6
2     1.886      2.920    4.303   6.965    9.925   14.08    22.32   31.59
3     1.638      2.353    3.182   4.541    5.841   7.453    10.21   12.92
4     1.533      2.132    2.776   3.747    4.604   5.598    7.173   8.610
5     1.476      2.015    2.571   3.365    4.032   4.773    5.893   6.869
6     1.440      1.943    2.447   3.143    3.707   4.317    5.208   5.959
7     1.415      1.895    2.365   2.998    3.499   4.029    4.785   5.408
8     1.397      1.860    2.306   2.897    3.355   3.833    4.501   5.041
9     1.383      1.833    2.262   2.821    3.250   3.690    4.297   4.781
10    1.372      1.812    2.228   2.764    3.169   3.581    4.144   4.587
11    1.363      1.796    2.201   2.718    3.106   3.497    4.025   4.437
12    1.356      1.782    2.179   2.681    3.055   3.428    3.930   4.318
13    1.350      1.771    2.160   2.650    3.012   3.372    3.852   4.221
14    1.345      1.761    2.145   2.625    2.977   3.326    3.787   4.140
15    1.341      1.753    2.131   2.602    2.947   3.286    3.733   4.073
16    1.337      1.746    2.120   2.584    2.921   3.252    3.686   4.015
17    1.333      1.740    2.110   2.567    2.898   3.222    3.646   3.965
18    1.330      1.734    2.101   2.552    2.878   3.197    3.610   3.922
19    1.328      1.729    2.093   2.539    2.861   3.174    3.579   3.883
20    1.325      1.725    2.086   2.528    2.845   3.153    3.552   3.850
21    1.323      1.721    2.080   2.518    2.831   3.135    3.527   3.819
22    1.321      1.717    2.074   2.508    2.819   3.119    3.505   3.792
23    1.319      1.714    2.069   2.500    2.807   3.104    3.485   3.768
24    1.318      1.711    2.064   2.492    2.797   3.090    3.467   3.745
25    1.316      1.708    2.060   2.485    2.787   3.078    3.450   3.725
26    1.315      1.706    2.056   2.479    2.779   3.067    3.435   3.707
27    1.314      1.703    2.052   2.473    2.771   3.057    3.421   3.690
28    1.313      1.701    2.048   2.467    2.763   3.047    3.408   3.674
29    1.311      1.699    2.045   2.462    2.756   3.038    3.396   3.659
30    1.310      1.697    2.042   2.457    2.750   3.030    3.385   3.646
31    1.309      1.695    2.040   2.453    2.744   3.022    3.375   3.633
32    1.309      1.694    2.037   2.449    2.738   3.015    3.365   3.622
33    1.308      1.692    2.035   2.445    2.733   3.008    3.356   3.611
34    1.307      1.691    2.032   2.441    2.728   3.002    3.348   3.601
35    1.306      1.690    2.030   2.438    2.724   2.996    3.340   3.591
36    1.306      1.688    2.028   2.434    2.719   2.991    3.333   3.582
37    1.305      1.687    2.026   2.431    2.715   2.985    3.326   3.574
38    1.304      1.686    2.024   2.429    2.712   2.980    3.319   3.566
39    1.304      1.685    2.023   2.426    2.708   2.976    3.313   3.558
40    1.303      1.684    2.021   2.423    2.704   2.971    3.307   3.551
42    1.302      1.682    2.018   2.418    2.698   2.963    3.296   3.538
44    1.301      1.680    2.015   2.414    2.692   2.956    3.286   3.526
46    1.300      1.679    2.013   2.410    2.687   2.949    3.277   3.515
48    1.299      1.677    2.011   2.407    2.682   2.943    3.269   3.505
50    1.299      1.676    2.009   2.403    2.678   2.937    3.261   3.496
60    1.296      1.671    2.000   2.390    2.660   2.915    3.232   3.460
70    1.294      1.667    1.994   2.381    2.648   2.899    3.211   3.435
80    1.292      1.664    1.990   2.374    2.639   2.887    3.195   3.416
90    1.291      1.662    1.987   2.369    2.632   2.878    3.183   3.402
100   1.290      1.660    1.984   2.364    2.626   2.871    3.174   3.391
120   1.289      1.658    1.980   2.358    2.617   2.860    3.160   3.373
```

```
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST1
`
fi

    slopesp2[j]="${slopes2_p}"

    if [ "${slopes2_p}" != ">0.2" ]
    then
        if [ "`echo 1 | awk '{if ('"${slopes2_trend}"' > 0) print "positive"}'`" = "positive" ]
        then
            aggs2[j]="*** Aggregation likely ***"
        elif [ "`echo 1 | awk '{if ('"${slopes2_trend}"' < 0) print "negative"}'`" = "negative" ]
        then
            aggs2[j]="*** Repulsion likely ***"
        fi
    fi
    echo
    slopes2=""
#************************************************************

    let j++
      done

      l=1
      while [ $l -le ${numconc} ]
      do
         coord[l]=`echo ${med[l]} | awk '{printf "%-0.1f", 1/$1}'`
         let l++
      done


      l=1
      rgs=""
      I0s=""
      while [ $l -le ${numconc} ]
      do
        if [ "${rg[l]}" == "0" -o "${rg[l]}" == "" -o "${rg[l]}" = "nan" -o "${rg_gnom[l]}" == "0" -o
"${rg_gnom[l]}" == "0.00" ]
        then
            badconc[l]="yes"
        fi
        let l++
    done

    problem=`echo ${numconc} ${badconc[*]} | awk '{if ($0~"yes" || $1<3) print "yes"}'`

    l=1
    while [ $l -le ${numconc} ]
      do
        if [ "${rg[l]}" != "nan" ] && [ "${rg[l]}" != "" ] && [ "${rg[l]}" != "0" ] && [ "${I0[l]}" != "nan" ]
&& [ "${I0[l]}" != "" ]
        then
            if [ "${problem}" != "yes" ]
                then
                    rgs="$rgs ${coord[l]}:${rg[l]}"
                    I0s="$I0s ${coord[l]}:${I0[l]}"
                    rgs_gnom="$rgs_gnom ${coord[l]}:${rg_gnom[l]}"
                    dmaxs="$dmaxs ${coord[l]}:${dmax[l]}"
                    I0s_gnom="$I0s_gnom ${coord[l]}:${I0_gnom[l]}"
                    mw_porods="$mw_porods ${coord[l]}:${mw_porod[l]}"
                else
                    rgs="$rgs ${rg[l]}"
                    I0s="$I0s ${I0[l]}"
                    rgs_gnom="$rgs_gnom ${rg_gnom[l]}"
                    dmaxs="$dmaxs ${dmax[l]}"
                    I0s_gnom="$I0s_gnom ${I0_gnom[l]}"
                    mw_porods="$mw_porods ${mw_porod[l]}"
            fi
        fi
        let l++
      done

if [ "${problem}" != "yes" ]
then
```

```
rg_fit=`echo $rgs | awk '
  BEGIN{
   FS=":"; RS=" ";
  }

  {
   delta = $2 - avg;
   avg += delta / NR;
   mean2 += delta * ($2 - avg);
   x[NR] = $1; y[NR] = $2;
   sx += x[NR]; sy += y[NR];
   sxx += x[NR]*x[NR];
   sxy += x[NR]*y[NR];
   syy += y[NR]*y[NR];
  }

  END{
   sd = sqrt (mean2/NR);
   det = NR*sxx - sx*sx;
   a = (NR*sxy - sx*sy)/det;
   b = (-sx*sxy+sxx*sy)/det;
   se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
   sb = sqrt(NR*se*se/det);
   t = a/sb;
   #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
   print a, b, t, avg, sd;
  }' `

  rg_trend=`echo $rg_fit | awk '{printf "%5.3f", $1}'`
  rg_yint=`echo $rg_fit | awk '{printf "%5.3f", $2}'`
  rg_t=`echo $rg_fit | awk '{printf "%5.3f", $3}'`
  rg_mean=`echo $rg_fit | awk '{printf "%5.3f", $4}'`
  rg_sd=`echo $rg_fit | awk '{printf "%5.3f", $5}'`

  rg_gnom_fit=`echo $rgs_gnom | awk '
  BEGIN{
   FS=":"; RS=" ";
  }

  {
   delta = $2 - avg;
   avg += delta / NR;
   mean2 += delta * ($2 - avg);
   x[NR] = $1; y[NR] = $2;
   sx += x[NR]; sy += y[NR];
   sxx += x[NR]*x[NR];
   sxy += x[NR]*y[NR];
   syy += y[NR]*y[NR];
  }

  END{
   sd = sqrt (mean2/NR);
   det = NR*sxx - sx*sx;
   a = (NR*sxy - sx*sy)/det;
   b = (-sx*sxy+sxx*sy)/det;
   se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
   sb = sqrt(NR*se*se/det);
   t = a/sb;
   #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
   print a, b, t, avg, sd;
  }' `

  rg_gnom_trend=`echo $rg_gnom_fit | awk '{printf "%5.3f", $1}'`
  rg_gnom_yint=`echo $rg_gnom_fit | awk '{printf "%5.3f", $2}'`
  rg_gnom_t=`echo $rg_gnom_fit | awk '{printf "%5.3f", $3}'`
  rg_gnom_mean=`echo $rg_gnom_fit | awk '{printf "%5.3f", $4}'`
  rg_gnom_sd=`echo $rg_gnom_fit | awk '{printf "%5.3f", $5}'`

dmax_fit=`echo $dmaxs | awk '
  BEGIN{
   FS=":"; RS=" ";
  }

  {
   delta = $2 - avg;
   avg += delta / NR;
   mean2 += delta * ($2 - avg);
   x[NR] = $1; y[NR] = $2;
   sx += x[NR]; sy += y[NR];
```

```
       sxx += x[NR]*x[NR];
       sxy += x[NR]*y[NR];
       syy += y[NR]*y[NR];
      }

      END{
       sd = sqrt (mean2/NR);
       det = NR*sxx - sx*sx;
       a = (NR*sxy - sx*sy)/det;
       b = (-sx*sxy+sxx*sy)/det;
       se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
       sb = sqrt(NR*se*se/det);
       t = a/sb;
       #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
       print a, b, t, avg, sd;
      }' `

      dmax_trend=`echo $dmax_fit | awk '{printf "%5.3f", $1}'`
      dmax_yint=`echo $dmax_fit | awk '{printf "%5.3f", $2}'`
      dmax_t=`echo $dmax_fit | awk '{printf "%5.3f", $3}'`
      dmax_mean=`echo $dmax_fit | awk '{printf "%5.3f", $4}'`
      dmax_sd=`echo $dmax_fit | awk '{printf "%5.3f", $5}'`

    mw_porod_fit=`echo $mw_porods | awk '
      BEGIN{
       FS=":"; RS=" ";
      }

      {
       delta = $2 - avg;
       avg += delta / NR;
       mean2 += delta * ($2 - avg);
       x[NR] = $1; y[NR] = $2;
       sx += x[NR]; sy += y[NR];
       sxx += x[NR]*x[NR];
       sxy += x[NR]*y[NR];
       syy += y[NR]*y[NR];
      }

      END{
       sd = sqrt (mean2/NR);
       det = NR*sxx - sx*sx;
       a = (NR*sxy - sx*sy)/det;
       b = (-sx*sxy+sxx*sy)/det;
       se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
       sb = sqrt(NR*se*se/det);
       t = a/sb;
       #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
       print a, b, t, avg, sd;
      }' `

      mw_porod_trend=`echo $mw_porod_fit | awk '{printf "%5.3f", $1}'`
      mw_porod_yint=`echo $mw_porod_fit | awk '{printf "%5.3f", $2}'`
      mw_porod_t=`echo $mw_porod_fit | awk '{printf "%5.3f", $3}'`
      mw_porod_mean=`echo $mw_porod_fit | awk '{printf "%5.3f", $4}'`
      mw_porod_sd=`echo $mw_porod_fit | awk '{printf "%5.3f", $5}'`
else

    rg_fit=`echo $rgs | awk '
      BEGIN{
       RS=" ";
      }

      {
       delta = $1 - avg;
       avg += delta / NR;
       mean2 += delta * ($1 - avg);
      }

      END{
       sd = sqrt (mean2/NR);
       print avg, sd;
      }' `

      rg_mean=`echo $rg_fit | awk '{printf "%5.3f", $1}'`
      rg_sd=`echo $rg_fit | awk '{printf "%5.3f", $2}'`

      rg_gnom_fit=`echo $rgs_gnom | awk '
      BEGIN{
       RS=" ";
      }

      {
       delta = $1 - avg;
       avg += delta / NR;
       mean2 += delta * ($1 - avg);
      }

      END{
       sd = sqrt (mean2/NR);
       print avg, sd;
      }' `

      rg_gnom_mean=`echo $rg_gnom_fit | awk '{printf "%5.3f", $1}'`
      rg_gnom_sd=`echo $rg_gnom_fit | awk '{printf "%5.3f", $2}'`

    dmax_fit=`echo $dmaxs | awk '
      BEGIN{
       RS=" ";
      }

      {
       delta = $1 - avg;
       avg += delta / NR;
       mean2 += delta * ($1 - avg);
      }

      END{
       sd = sqrt (mean2/NR);
       print avg, sd;
      }' `

      dmax_mean=`echo $dmax_fit | awk '{printf "%5.3f", $1}'`
      dmax_sd=`echo $dmax_fit | awk '{printf "%5.3f", $2}'`

    mw_porod_fit=`echo $mw_porods | awk '
      BEGIN{
       RS=" ";
      }

      {
       delta = $1 - avg;
       avg += delta / NR;
       mean2 += delta * ($1 - avg);
      }

      END{
       sd = sqrt (mean2/NR);
       print avg, sd;
      }' `

      mw_porod_mean=`echo $mw_porod_fit | awk '{printf "%5.3f", $1}'`
      mw_porod_sd=`echo $mw_porod_fit | awk '{printf "%5.3f", $2}'`
fi

mwporod=$mw_porod_mean

if [ -z "${mw_given}" ]
then
      mw=$mwporod
else
      mw=$mw_given
fi

    #calculate second virial coefficient
      #curves have been scaled together and written to ${sample}_1.dat, ${sample}_2.dat, etc. and I(c,0)
      #has been calculated from scaled curves.
      #calculate S(c,0) by dividing I(c,0) by I(0,0) taken from y-intercept of I(c,0) linear regression
      #here I calculate the inverse to achieve 1/S(c,0), i.e. I(0,0)/I(c,0).  Assume I(0,0), i.e. the
      #form factor, is the lowest concentration given.


      j=1
      while [ $j -le ${numconc} ]
      do
        S0inv[j]=`echo "${I0_gnom[1]} ${I0[j]}" | awk '{print $1/$2}'`
        S0inv_gnom[j]=`echo "${I0_gnom[1]} ${I0_gnom[j]}" | awk '{print $1/$2}'`
        let j++
      done

      #now determine slope of equation:  1/S(c,0) = 1 + [2 * mol.wt.(Da) * A2] * concentration
      # A2 = slope/(2*mw)

      #S0s="${coord[1]}:${S0inv[1]} ${coord[2]}:${S0inv[2]} ${coord[3]}:${S0inv[3]}"


      l=1
      S0s=""
    S0s_gnom=""

         if [ "$calcconc" = "yes" ]
        then
            if [ "$isneg" == "yes" ]
            then
              while [ $l -le ${numconc} ]
                do
                    conc[l]=`echo ${I0_gnom[l]} ${mw_porod[l]} $mw_standard $conc_standard $I0_standard | awk
'{print ($1/$2)*($3*$4/$5)}'`
                    let l++
                done
            else
                concvalue=`echo ${I0_gnom[1]} ${mw_porod[l]} $mw_standard $conc_standard $I0_standard | awk
'{print ($1/$2)*($3*$4/$5)}'`
              while [ $l -le ${numconc} ]
                do
                    conc[l]=`echo ${concvalue} ${coord[l]} | awk '{print $1*$2}'`
                    let l++
                done
            fi
        else
            if [ "$isneg" == "yes" ]
            then
                echo " ERROR: Negative Scale Factors.  Cannot calculate concentrations."
            else
              while [ $l -le ${numconc} ]
                do
                    conc[l]=`echo ${concvalue} ${coord[l]} | awk '{print $1*$2}'`
                    let l++
                done
            fi
        fi

    if [ "${conc[1]}" != "0" ] && [ "$problem" != "yes" ]
    then
        rg_trend=`echo $rg_trend ${conc[1]}| awk '{printf "%.3f\n", $1/$2}'`
        rg_gnom_trend=`echo $rg_gnom_trend ${conc[1]}| awk '{printf "%.3f\n", $1/$2}'`
        dmax_trend=`echo $dmax_trend ${conc[1]}| awk '{printf "%.3f\n", $1/$2}'`
        mw_porod_trend=`echo $rg_trend ${conc[1]}| awk '{printf "%.3f\n", $1/$2}'`
    fi

    l=1

if [ "$calcA2" = "yes" ]
then
      while [ $l -le ${numconc} ]
      do
        if [ "${S0inv[l]}" != "nan" ] && [ "${S0inv[l]}" != "" ]
        then
            S0s="$S0s ${conc[l]}:${S0inv[l]}"
        fi
        S0s_gnom="$S0s_gnom ${conc[l]}:${S0inv_gnom[l]}"
         let l++
      done

    #add (0,1) to the S(c,0) plot, since the equation assumes the y-intercept is 1
    S0s="0:1 $S0s"

      S0_fit=`echo $S0s | awk '
      BEGIN{
       FS=":"; RS=" ";
      }

      {
       delta = $2 - avg;
       avg += delta / NR;
       mean2 += delta * ($2 - avg);
       x[NR] = $1; y[NR] = $2;
       sx += x[NR]; sy += y[NR];
       sxx += x[NR]*x[NR];
       sxy += x[NR]*y[NR];
       syy += y[NR]*y[NR];
      }

      END{
       sd = sqrt (mean2/NR);
       det = NR*sxx - sx*sx;
       a = (NR*sxy - sx*sy)/det;
       b = (-sx*sxy+sxx*sy)/det;
       se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
       sb = sqrt(NR*se*se/det);
       t = a/sb;
       #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
       print a, b, t, avg, sd;
      }' `

      S0_trend=`echo $S0_fit | awk '{printf "%5.3f", $1}'`
      S0_yint=`echo $S0_fit | awk '{printf "%5.3f", $2}'`
      S0_t=`echo $S0_fit | awk '{printf "%5.3f", $3}'`
      S0_mean=`echo $S0_fit | awk '{printf "%5.3f", $4}'`
      S0_sd=`echo $S0_fit | awk '{printf "%5.3f", $5}'`

    S0_gnom_fit=`echo $S0s_gnom | awk '
      BEGIN{
       FS=":"; RS=" ";
      }

      {
       delta = $2 - avg;
       avg += delta / NR;
       mean2 += delta * ($2 - avg);
       x[NR] = $1; y[NR] = $2;
       sx += x[NR]; sy += y[NR];
       sxx += x[NR]*x[NR];
       sxy += x[NR]*y[NR];
       syy += y[NR]*y[NR];
      }

      END{
       sd = sqrt (mean2/NR);
       det = NR*sxx - sx*sx;
       a = (NR*sxy - sx*sy)/det;
       b = (-sx*sxy+sxx*sy)/det;
       se = sqrt((1/(NR*(NR-2)))*(NR*syy - sy*sy - a*a*det));
       sb = sqrt(NR*se*se/det);
       t = a/sb;
       #for(i=1;i<=NR;i++) print x[i],a*x[i]+b;
       print a, b, t, avg, sd;
      }' `

      S0_gnom_trend=`echo $S0_gnom_fit | awk '{printf "%5.3f", $1}'`
      S0_gnom_yint=`echo $S0_gnom_fit | awk '{printf "%5.3f", $2}'`
      S0_gnom_t=`echo $S0_gnom_fit | awk '{printf "%5.3f", $3}'`
      S0_gnom_mean=`echo $S0_gnom_fit | awk '{printf "%5.3f", $4}'`
      S0_gnom_sd=`echo $S0_gnom_fit | awk '{printf "%5.3f", $5}'`
fi

      #make simple plot in terminal for Rg and S(0) vs concentration
#     echo $rgs | awk -v FS=':' -v RS=' ' -v OFS=' ' -v ORS='\n' '{print $1 OFS $2 OFS $4}' > rgs_plot.dat
      echo $S0s | awk -v FS=':' -v RS=' ' -v OFS=' ' -v ORS='\n' '{print $1 OFS $2 OFS $4}' > S0s_plot.dat
      echo $S0s_gnom | awk -v FS=':' -v RS=' ' -v OFS=' ' -v ORS='\n' '{print $1 OFS $2 OFS $4}' >
S0s_gnom_plot.dat


#     if [ `command -v gnuplot` -a "${calcA2}" == "yes" ]
#     then
#gnuplot << endoffile
# set terminal dumb
# set xrange [0:*]
# set tics out
# set y2label "Rg"
# set y2tics border
# set ylabel "1/S(0)"
# set y2range [*:*]
# plot "S0s_plot.dat" title "1/S(0)" pt 1, "S0s_gnom_plot.dat" axes x1y2 title "S0s_gnom" pt 2
#endoffile
#fi
```

```
if [ "${calctrend}" != "no" ]
then
    if [ -n $mw ]
    then
    A2=""
      A2=`echo $S0_trend $mw | awk '{print $1/(2*$2)}'`
    A2_gnom=`echo $S0_gnom_trend $mw | awk '{print $1/(2*$2)}'`
      fi

      # calculate approximate p-value from t-statistic assuming one degree of freedom
      # and a two-tailed test.  p-values binned into groups of 0.05.

      rg_t=`echo $rg_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
    rg_gnom_t=`echo $rg_gnom_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
    dmax_t=`echo $dmax_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`
    mw_porod_t=`echo $mw_porod_t | awk '{if ($1 < 0) print (-1)*$1; else print $1}'`

      ndf=`expr $numconc - 2`  #number of degrees of freedom = (number of exposures - 2); two-tailed test

      #calculate p-values from t-statistic

#*******

t_to_p () {

      p=`awk '
        {if (NR==('"$ndf"'+1)) {
         if ('"$1"' < $2) print ">0.2";
         else if ('"$1"'>$9) print "0.001";
         else if ('"$1"'>$8) print "0.002";
         else if ('"$1"'>$7) print "0.005";
         else if ('"$1"'>$6) print "0.01";
         else if ('"$1"'>$5) print "0.02";
         else if ('"$1"'>$4) print "0.05";
         else if ('"$1"'>$3) print "0.1";
         else if ('"$1"'>$2) print "0.2";
        }}' <<TDIST1
DF    0.200      0.100    0.050    0.020    0.010    0.005    0.002    0.001
1     3.078      6.314    12.70    31.82    63.65    127.3    318.3    636.6
2     1.886      2.920    4.303    6.965    9.925    14.08    22.32    31.59
3     1.638      2.353    3.182    4.541    5.841    7.453    10.21    12.92
4     1.533      2.132    2.776    3.747    4.604    5.598    7.173    8.610
5     1.476      2.015    2.571    3.365    4.032    4.773    5.893    6.869
6     1.440      1.943    2.447    3.143    3.707    4.317    5.208    5.959
7     1.415      1.895    2.365    2.998    3.499    4.029    4.785    5.408
8     1.397      1.860    2.306    2.897    3.355    3.833    4.501    5.041
9     1.383      1.833    2.262    2.821    3.250    3.690    4.297    4.781
10    1.372      1.812    2.228    2.764    3.169    3.581    4.144    4.587
11    1.363      1.796    2.201    2.718    3.106    3.497    4.025    4.437
12    1.356      1.782    2.179    2.681    3.055    3.428    3.930    4.318
13    1.350      1.771    2.160    2.650    3.012    3.372    3.852    4.221
14    1.345      1.761    2.145    2.625    2.977    3.326    3.787    4.140
15    1.341      1.753    2.131    2.602    2.947    3.286    3.733    4.073
16    1.337      1.746    2.120    2.584    2.921    3.252    3.686    4.015
17    1.333      1.740    2.110    2.567    2.898    3.222    3.646    3.965
18    1.330      1.734    2.101    2.552    2.878    3.197    3.610    3.922
19    1.328      1.729    2.093    2.539    2.861    3.174    3.579    3.883
20    1.325      1.725    2.086    2.528    2.845    3.153    3.552    3.850
21    1.323      1.721    2.080    2.518    2.831    3.135    3.527    3.819
22    1.321      1.717    2.074    2.508    2.819    3.119    3.505    3.792
23    1.319      1.714    2.069    2.500    2.807    3.104    3.485    3.768
24    1.318      1.711    2.064    2.492    2.797    3.090    3.467    3.745
25    1.316      1.708    2.060    2.485    2.787    3.078    3.450    3.725
26    1.315      1.706    2.056    2.479    2.779    3.067    3.435    3.707
27    1.314      1.703    2.052    2.473    2.771    3.057    3.421    3.690
28    1.313      1.701    2.048    2.467    2.763    3.047    3.408    3.674
29    1.311      1.699    2.045    2.462    2.756    3.038    3.396    3.659
30    1.310      1.697    2.042    2.457    2.750    3.030    3.385    3.646
31    1.309      1.695    2.040    2.453    2.744    3.022    3.375    3.633
32    1.309      1.694    2.037    2.449    2.738    3.015    3.365    3.622
33    1.308      1.692    2.035    2.445    2.733    3.008    3.356    3.611
34    1.307      1.691    2.032    2.441    2.728    3.002    3.348    3.601
35    1.306      1.690    2.030    2.438    2.724    2.996    3.340    3.591
36    1.306      1.688    2.028    2.434    2.719    2.991    3.333    3.582
37    1.305      1.687    2.026    2.431    2.715    2.985    3.326    3.574
38    1.304      1.686    2.024    2.429    2.712    2.980    3.319    3.566
39    1.304      1.685    2.023    2.426    2.708    2.976    3.313    3.558
40    1.303      1.684    2.021    2.423    2.704    2.971    3.307    3.551
42    1.302      1.682    2.018    2.418    2.698    2.963    3.296    3.538
44    1.301      1.680    2.015    2.414    2.692    2.956    3.286    3.526
46    1.300      1.679    2.013    2.410    2.687    2.949    3.277    3.515
48    1.299      1.677    2.011    2.407    2.682    2.943    3.269    3.505
50    1.299      1.676    2.009    2.403    2.678    2.937    3.261    3.496
60    1.296      1.671    2.000    2.390    2.660    2.915    3.232    3.460
70    1.294      1.667    1.994    2.381    2.648    2.899    3.211    3.435
80    1.292      1.664    1.990    2.374    2.639    2.887    3.195    3.416
90    1.291      1.662    1.987    2.369    2.632    2.878    3.183    3.402
100   1.290      1.660    1.984    2.364    2.626    2.871    3.174    3.391
120   1.289      1.658    1.980    2.358    2.617    2.860    3.160    3.373
150   1.287      1.655    1.976    2.351    2.609    2.849    3.145    3.357
200   1.286      1.652    1.972    2.345    2.601    2.839    3.131    3.340
300   1.284      1.650    1.968    2.339    2.592    2.828    3.118    3.323
500   1.283      1.648    1.965    2.334    2.586    2.820    3.107    3.310
TDIST1
`

}

t_to_p $rg_t
rg_p=$p

t_to_p $rg_gnom_t
rg_gnom_p=$p

t_to_p $dmax_t
dmax_p=$p

t_to_p $mw_porod_t
mw_porod_p=$p


fi



#*********

    i=1

    #create tab delimited log file for concentration dependence check for each sample in directory
      if [ 1 ]
      then
    echo -e " ---------------------${sample}${conc}-------------------------"
      echo
      while [ $i -le $numconc ]
    do
        currentconc=`jot -c $i ${conc} | awk '{if (NR=="'${i}'") print}'`
        if [ "${rg[i]}" == "0" ]
        then
            echo -e " *** FAILED:  ${sample}${currentconc}.dat "
        fi
        echo -e " Concentration:  ${sample}${currentconc}.dat (${conc[i]} mg/ml)"
        echo -e " Interparticle Interaction P-value (q < 1.3/Rg):\t${slopesp[i]}  \t${aggs[i]}"
        echo -e " Interparticle Interaction P-value (pi/dmax < q < 1.3/Rg):\t${slopesp2[i]}  \t${aggs2[i]}"
        echo -e " Guinier Region: ${qmin[i]} < q < ${qmax[i]}  Points: [${begp[i]}:${endp[i]}]"
        echo -e " Rg (Guinier): ${rg[i]}"
        echo -e " Rg (GNOM):    ${rg_gnom[i]}"
        echo -e " I(0) (GNOM):  ${I0_gnom[i]}"
        echo -e " Dmax:         ${dmax[i]}"
        echo -e " Porod MW:     ${mw_porod[i]}"
        echo -e " 1/S(0):       ${S0inv[i]}"
        echo -e " 1/S(0)-GNOM:  ${S0inv_gnom[i]}"
        echo -e " -------------------------------------------------------"
        let i++
      done
    echo
    echo -e " # of Concentrations: $numconc"
      if [ -z "${mw_given}" ]
      then
        echo " Average Porod Molecular Weight:  $mwporod kDa"
      else
        echo " Molecular Weight Given:  $mw_given kDa"
        echo " Porod Molecular Weight:  $mwporod kDa"
      fi
    echo
    echo -e "        \t\t  Rg   \t GNOM  \t Dmax  \t Porod MW "
      echo -e "        \t\t------ \t------ \t------ \t----------"
      echo -e " Average:\t\t$rg_mean\t$rg_gnom_mean\t$dmax_mean\t $mw_porod_mean "
      echo -e " Std Dev:\t\t$rg_sd\t$rg_gnom_sd\t$dmax_sd\t $mw_porod_sd "
    if [ "$problem" != "yes" ]
    then
    if [ "${conc[1]}" != "0" ]
    then
    echo -e " Slope (ang/mg/mL):\t$rg_trend\t$rg_gnom_trend\t$dmax_trend\t $mw_porod_trend "
    else
    echo -e " Slope:\t\t\t$rg_trend\t$rg_gnom_trend\t$dmax_trend\t $mw_porod_trend "
    fi
      echo -e " Y-Int:\t\t\t$rg_yint\t$rg_gnom_yint\t$dmax_yint\t $mw_porod_yint "
      echo -e " T-stat:\t\t$rg_t\t$rg_gnom_t\t$dmax_t\t $mw_porod_t "
      echo -e " P-Value:\t\t$rg_p\t$rg_gnom_p\t$dmax_p\t $mw_porod_p "
    echo
    if [ "${conc[1]}" != "0" ]
    then
    echo -e " A2 = $A2 mol.ml.gm^-2"
    echo -e " A2 (GNOM) = $A2_gnom mol.ml.gm^-2"
    fi
    fi
    echo "-------------------------------------------------"
    fi | tee -a ${sample}${conc}_conc_stats.txt
done
```