THE BEGINNER'S GUIDE TO SAXS DATA PROCESSING AND ANALYSIS

by

ERIK HENDERSON

(Under the Direction of Jeffery Urbauer)

ABSTRACT

The popularity of small angle x-ray scattering (SAXS) has increased dramatically over the last decade. Despite an influx of researchers to the field, there is still no comprehensive guide for processing data available. This presents a large barrier to researchers who are not experienced structural biologists, or are otherwise unfamiliar with SAXS data processing. This guide is a complete walkthrough for small angle scattering data processing using the ATSAS software package. The purpose of this guide is: to provide researchers who have no experience in SAXS data processing with an easy to follow document that will help introduce them to the field and allow them to correctly process data. The specific focus is on processing SAXS and WAXS data sets to determine structural information and generate 3D models.

INDEX WORDS: SAXS, WAXS, SAS, Small Angle, Wide Angle, X-ray, Scattering, Structural Biology, Data processing, Data Analysis, Beginner, Guide, Neutron, Biochemistry, Molecular Biology, Molecular Modeling, Molecule, Envelope, ATSAS, DAMMIN, PRIMUS, GNOM, DAMMIF, AutoRg, Rmax, Dmax, P(r) distribution, Guinier, Kratky Plot

THE BEGINNER'S GUIDE TO SAXS DATA PROCESSING AND ANALYSIS

by

ERIK HENDERSON

B.S., The University of Georgia, 2008

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

MASTER OF SCIENCE

ATHENS, GEORGIA

2011

THE BEGINNER'S GUIDE TO SAXS DATA PROCESSING AND ANALYSIS

by

ERIK HENDERSON

Major Professor:	Jeffery Urbauer
Committee:	John Rose William Lanzilotta Louise Wicker

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2011

TABLE OF CONTENTS

Page

Chapter

Chapter 1

Introduction and Background

1.1 About this Guide

The purpose of this guide is to give novice users a concise document that will take them from the beginning to the end of small angle x-ray scattering (SAXS) data processing in a step-by-step manner. This guide will focus on the "how" of data processing more than the "why". Screenshots, with the important areas highlighted and numbered according by step, will be presented for each stage of the data processing. The simple explanations and step by step directions are intended to allow someone with limited experience in SAXS data processing sit down with their data set and produce a reasonable structure.

When I began processing SAXS data I had no previous experience with data processing. As could be expected, my learning curve was very high. I read the literature, downloaded every tutorial I could find, and consulted the forums when I got stuck. It didn't take me long to realize that there was something missing from the literature. Even though the European Molecular Biology Laboratory (EMBL) website provides manuals for the All Thats Small Angle Scattering (ATSAS) software suite (http://www.embl-hamburg.de/biosaxs/software.html), and groups like Bioisis (http://www.bioisis.net/tutorial) provide small tutorials on various aspects of SAXS data processing. The EMBL manuals give definitions and brief explanations for the software prompts, but do not specifically explain how to use the software. This is useful information once you understand the basic data processing steps, but it is not geared towards the novice user. The Bioisis tutorials are very helpful, but their scope is limited to only a few steps of processing SAXS data. A detailed, step-by-step guide to processing SAXS data is missing from the field. Hopefully this guide will help to fill that gap. This guide is not meant to review or advance the current literature in the field, nor is it meant to be a comprehensive explanation of the theories behind small angle scattering. For a thorough review of SAXS theory, read *Structural Analysis by Small-Angle X-Ray and Neutron Scattering*, by Demitri Svergun [3]. The purpose of this guide is to provide novice users with a complete, easy-to-follow, step by step tutorial of SAXS data processing. The focus will be on practice rather than theory. Each step will be accompanied by screenshots and simple explanations of the steps. Limited background information will be presented to help the user understand why each step is necessary. The background provided will not be exhaustive and should not be substituted for more compete reviews in the field [1-4]. Within the suggested reading portion (Section 1.5) of the guide I recommend review articles which are easy to understand and provide detailed information on SAXS theory.

1.2 Introduction

In recent years the use of small angle scattering (SAS) experiments to study biological molecules has gained significant popularity (for a review, see reference 4). The three most common forms of small angle scattering (SAS) are: small angle x-ray scattering (SAXS), wide angle x-ray scattering (WAXS), and small angle neutron scattering (SANS).

SAXS, and the closely related WAXS, are techniques that measure x-ray scattering from proteins in-solution. The data obtained from SAXS/WAXS experiments can provide information on the molecular weight and oligomeric state of a protein. The scattering profile can also be used to generate a 3D structural envelope of a protein. SAXS/WAXS is often used to validate crystal structures or to determine structural details of proteins that can't be easily crystallized. The setup is very similar in concept to x-ray crystallography. An x-ray source produces a beam of x-rays which is focused through a series of pin hole filters and then passed through the sample. When x-rays interact with a protein in solution it causes a portion of the beam to change direction or scatter. Beyond the sample is a detector which records the scattering pattern. SAXS data alone is low resolution, between 10-50Å. WAXS data has a higher resolution range and can be used to investigate some tertiary structural elements. Simultaneous collection of SAXS/WAXS data allows for a significant increase in resolution over SAXS alone. The only difference between the experimental setup for SAXS and WAXS is the scattering angle. The range of scattering angles that are recorded is controlled by changing the distance of the sample from the detector. During WAXS experiments the detector is placed closer to the sample to collect higher angle (and higher resolution) scattering data.

SANS is fundamentally very similar to other small angle scattering techniques (SAXS/WAXS) [2]. The difference is that instead of scattering x-rays, SANS relies on neutron scattering. This means that the scattering profile is a result of the nuclear scattering as opposed to the electron scattering. In neutron scattering experiments, the scattering factor of hydrogen is much more irregular than it is with x-ray scattering. Hydrogen has a negative coherent scattering factor as well as an incoherent scattering factor [9]. These properties cancel out the positive scattering factors of other atoms in neutron density maps [9]. To increase the resolution, proteins are deuterated to provide contrast labeling. Deuterium has a positive scattering factor that greatly increases the scattering intensity of the protein compared to the buffer [9]. This increased separation makes buffer subtraction possible. Deuteration is important for more than just buffer subtraction. By labeling just one component of a protein complex it is possible to determine where that protein binds as well as the overall binding conformation [3]. The hydrogen in the unlabeled proteins make them “invisible” in the density map, and they are subtracted away with the buffer leaving only the deuterated sample [3]. These data can be combined with SAXS/WAXS data of the entire complex to show where the protein of interest is bound.

1.3 History

The technique of using x-ray diffraction of molecules in solution to determine structural characteristics has been around since the 1930s [4]. Originally, the technique was used for more material science than biological applications. The French scientist André Guinier demonstrated one of the first practical applications of small angle x-ray scattering with his research on aluminum copper alloys [6]. These studies led to the discovery of the relationship between the x-ray scattering intensity and particle size. Scientists soon realized that, in addition to information about the size and shape of particles, scattering profiles also contained information about the internal structure [6]. Despite this knowledge, limitations in data processing restricted the information obtained from SAXS experiments to overall size, shape, and level of internal order of the particle.

In the 1960s, x-ray crystallographic work on biological molecules led to an increased interest in other structural techniques. Small angle scattering became a popular technique to determine the size and shape of samples which were difficult to crystallize. One of the major constraints on the expansion of SAXS experimentation was the difficulty of data analysis. Most of the data analysis at the time was limited to determining the radius of gyration and the volume of the protein [4]. Major advances in structural work with small angle scattering would not come until the 1970s, when synchrotron x-ray sources became available [6]. Researchers combined information from SAXS and SANS experiments to determine that DNA wraps around the nucleosome core particles over 20 years before a high resolution crystal structure was completed [4]. The same techniques were used to map the position of proteins and RNA subunits in the ribosome [4]. A complete map of the prokaryotic 30S subunit, and a partial map of the 50S subunit were completed. The partial map of the 50S subunit they produced predated the first crystal structure [10] of the subunit by nearly a decade [4]. Despite advances in the capabilities of small angle scattering, data analysis was still the limiting factor. The data analysis was so complex, that more often than not it would require experts from several fields to draw significant conclusions from the results [4].

More powerful computers and advances in data processing software have made generating structural envelopes from scattering profiles a common and relatively straightforward process [4]. Biological small angle scattering is a rapidly growing field. The ease and flexibility of performing experiments, in conjunction with recent advances in higher resolution scattering, is attracting more researchers to the field each year. Small angle scattering techniques are expanding beyond the traditional SAXS/WAXS experiments and are beginning to take advantage of some of the unique characteristics of these experiments. The following section describes some of the most recent advances in the field.

1.4 Future Directions

Higher resolution small angle scattering techniques such as WAXS have allowed researchers to monitor structural changes that are beyond the scope of SAXS only experiments. The higher resolution range has made SAS experiments more useful for structural studies of proteins. It would be impractical to describe all of the advances in the field. To give you an idea of the type of work being done, I would like to briefly mention two techniques I personally found very interesting.

There are many difficult steps involved in solving a protein structure using x-ray crystallography. The first technique I want to mention uses SAXS/WAXS to help solve one of the toughest problems in x-ray crystallography. Even after the researcher has a well diffracting crystal, there are still some significant challenges to solving the structure. Solving the phasing problem is one of those problems. The most common phasing techniques require heavy atoms to be incorporated into the protein during production [8]. When the structure of a homologous protein is already available the initial phasing can be done with molecular replacement [8]. The limitations to both of these techniques has led researchers to search for alternative methods. One approach uses SAXS envelopes to conduct solvent flattening [11]. Envelope-based phasing uses a molecular model from a SAXS experiment to determine the phase. Specialized software uses the SAXS model as a template to locate the molecular envelope within the crystal unit cell [8]. The reliability of this technique has always been limited by the low resolution of SAXS models. In 2009, a study in the Journal of Applied Crystallography demonstrated that combining SAXS and WAXS data allowed for more accurate envelope-based phasing. The higher resolution WAXS data was "essential for locating the molecular envelope in the crystal unit cell" [8]. Envelope-based phasing is unlikely to replace more traditional methods of phasing in the near future, but it is a prime example of the increasing role for SAXS/WAXS in structural biology.

Optical spectroscopy and nuclear magnetic resonance (NMR) are two of the most common techniques used to study the dynamic conformational changes of a protein in solution [7]. The second technique I would like to introduce uses WAXS to monitor conformational changes in proteins on a nanosecond time scale. Time-resolved wide angle x-ray scattering (TR-WAXS) follows changes in the WAXS scattering profile to detect tertiary and quaternary changes in the protein conformation [7]. When conformational changes occur, there is a corresponding change to the WAXS profile. In this study the initial and finial conformations had already been determined by x-ray crystallography [7]. The WAXS profiles were monitored throughout the change to help identify intermediate states. The experimental setup is similar to a traditional SAXS/WAXS experiment with a few key differences. A laser is directed at the sample chamber and is used to initiate a conformational change in the protein through photolysis.

Multiple WAXS profiles are recorded to document changes over time [7]. There are several characteristics of small angle scattering that make it unique compared to other structural techniques. These techniques represent only two examples of the novel uses researchers are finding for SAXS/WAXS experiments. As the technology improves, the use of small angle scattering in structural biology will continue to increase.

1.5 Suggested Reading

Understanding SAS can be daunting for someone who doesn't have a background in structural biology. The literature available varies greatly between easy to understand and technical mathematics based reviews. If you already have experience with x-ray crystallography then the mathematics should look very familiar. However, if you are not an experienced structural biologist I suggest you take a look at the reviews listed below. I have selected these papers because they are easy to understand and contain important background information on small angle scattering. I strongly encourage you to go beyond this small list, as this only represents a small portion of the literature available.

- *X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution* [1].
- *Small-angle scattering studies of biological macromolecules in solution* [2].
- *Structural Analysis by Small-Angle X-Ray and Neutron Scattering* [3].
- *Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls* [4].

- *Structural characterization of proteins and complexes using small-angle X-ray solution scattering* [5].

These are all very important reviews authored by very influential figures in the small angle scattering field. For these papers, I suggest referring to these reviews while working through the data processing. All three of these papers have a table of contents so it will be easy to stop and find the background information about the step you are on. The paper entitled *Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls* (reference 4)*,* is an extremely useful source for processing data. The entire focus of this paper is on processing data without making common mistakes. This review gives a detailed explanation of the controls and safeguards you should use while processing SAXS data. Even though I will address many of these controls, safeguards, and common mistakes within this guide, I highly recommend you read this paper before you begin processing data. There is no single review article that can answer all of the questions that may arise during data processing. Many of the technical aspects of the process are also not addressed in the literature. If you run into difficulties, the community forums can be a great source for help. I have personally found the Small Angle X-ray Scattering Initiative for Europe (SAXIER) forum to be very useful for answering obscure questions that may come up during data processing (www.saxier.org/forum/).

1.6 How to Use This Guide

This guide provides step-by-step instructions to processing your raw SAXS data to determine physical values and to generate 3D SAXS envelopes. The design of this guide allows it to stand alone, and it is possible to begin data processing with no outside knowledge and obtain a structure. However, to ensure that you get the best possible learning experience I suggest you take a different approach. Don’t skip right ahead to the data processing. Each section will present a brief background on the steps you are about to do. Take the time to read through these introductory portions of this guide. If you are already familiar with SAXS or similar structural techniques then most of the information presented here will be a review. If not, hopefully these sections will help introduce you to some of the common terms in the field. Regardless of your experience level, take the time to read the 5 papers listed in the previous section. These papers encompass a range of topics from data processing to theory. It is impossible for me to present every caveat you may run into while processing your data. Understanding the theory behind each of the steps in greater detail will help you make educated decisions when you encounter an obstacle not covered in the guide. Even a cursory understanding of scattering principals will allow you to interpret your data with higher confidence. In addition, before you begin processing your experimental data, I strongly suggest you try processing some sample data sets.

Proper data processing is crucial to getting usable results from SAXS data. The nature of SAXS experiments makes them highly prone to error. The quality of your results can be impacted by any number of factors ranging from the protein’s unique structure to the individual sample’s quality. Even the components of the buffer can significantly impact the results of a SAXS experiment. With all of these potential sources of error already present, it is important to make sure you are not introducing additional errors through incorrect data processing. The two data sets that are used throughout this guide have been provided to serve as a hands on example. Both are SAXS/WAXS data for the protein lysozyme at two different concentrations. To help hone your data processing skills, use the sample data to recreate the results presented in the guide. The samples are not meant to represent perfect sample quality. These samples came from one of the first data sets our lab collected, and they demonstrate some data quality issues common to SAXS experiments. Determining the quality of a sample is a crucial step in evaluating experimental data. Taking the time to work through the sample data will help you learn how to identify problems in your own data.

After using the sample data to master the basics of data processing you can move on to your own experimental data. It is important to remember that the shape of your protein may not fall in the same class as lysozyme. This can have a dramatic effect on how the correct graphs and plots will look. If your protein has a different shape (or an unknown shape) don't rely on the lysozyme example plots as a guide. Make sure that the plots you produce resemble the ideal plots (discussed in detail later) for your protein's shape. Once you have finished processing the sample data and your first data set, look back at each step and ask yourself the following questions:

- Is the sample quality good enough to trust my results?
- Have I performed appropriate controls on my data?
- Do my graphs and plots match the expected results?
- Does my structural envelope make sense based on what is known about my protein?
- Can I reproduce these results with the same sample?
- Can I reproduce these results with a different concentration sample?
- Are the results from the data alone or did I over-parameterize?

- Are deviations from the expected result because of my unique protein, or is it the result of an error in data processing?

While it may seem rudimentary, this simple exercise can help identify obvious errors, and will help train you to keep these things in mind while processing future data. These are just a few of the questions you should ask yourself when you are looking at the processed data, and the all of the specific details of data evaluation will be discussed later.

At some point you will probably ask yourself a question that you will not be able to find the answer to. That is why it is important to have access to other, more experienced researchers. Despite gaining popularity within the scientific community over the last decade, biological small angle scattering (BioSAXS) is still a relatively small field. Consequently, finding someone at your institution with extensive knowledge in SAXS data processing may be difficult. The online forum community for SAXS can provide a novice user with answers to common questions, or it can serve as a sounding board for ideas from advanced users. Regardless of your background, if you are not already a member of a small angle scattering forum I highly recommend you join one. If you don't already have a preferred forum, I recommend the Small Angle X-ray Scattering Initiative for Europe (SAXIER) forum. The SAXIER forum has an active community of patient and knowledgeable members from all levels of experience. BioSAXS may be a small field, but there is still no need for you to reinvent the wheel while you learn. During data processing, it is common to get a repeatable result that looks abnormal or is otherwise unexpected. In many cases, you will not be the first person to have encountered a problem, and often the answer may already be on the forum. Keeping current with existing posts and posting questions of your own can save you alot of time and effort. Also, being active in the community can go beyond helping you overcome immediate problems it can also help you stay current with standard practices for generating and reporting your results.

Unlike x-ray crystallography there are very limited standards for SAXS data processing and reporting that are universally enforced by reviewers. The problem is a result of how SAXS data is published in many journals. When a protein has been well studied, SAXS data alone does not often provide enough novel information for publication on its own. Instead, higher resolution structural data is often the focus of the paper, and SAXS data is provided as a supplement. In turn, reviewers focus on the higher resolution techniques and the SAS data can sometimes be overlooked. The experimental controls and the data processing results used to support the SAS results are often not included. The lack of standardized controls makes it difficult to evaluate results from different research groups with any level of consistency. As the capabilities and popularity of SAS expand, the standards for processing and reporting data will solidify. In the mean time, staying active in the forums and current in the literature will help you understand what the community considers acceptable.

Chapter 2

Sample Preparation for SAS Experiments

SAXS/WAXS experiments themselves are very short. Data collection at a synchrotron source can be completed in a just few seconds. The most critical and time consuming step in any SAS experiment is the sample preparation. The nature of SAS relies on essentially one dimensional data to generate three dimensional images. For these structures to be reliable it is crucial that potential sources of error be eliminated. The most direct way of doing this is to carefully prepare samples which will give the highest quality data possible. While this brief chapter is certainly not intended to be comprehensive, we will discuss some universal sources of error and how their effects can be minimized through careful sample preparation. For additional information on sample preparation many of the beamline websites offer guides to designing buffer conditions and handling samples.

2.1 Sample Purity

Scattering intensity is exponentially dependent on the molecular weight of the protein. If a small amount of high molecular weight protein contaminant is present in the sample it can significantly distort the results. To help prevent this it is important to ensure that samples are composed of only the target protein in a single oligomeric state. There are a wide variety of protein purification techniques available. The specific technique used will depend on your lab's specific needs and the protein being purified. It is important to remember that the more pure the sample is, the higher quality the scattering data will be. Poorly purified samples will often result in data which is not useable. Often times you will not discover that the data is poor until you begin data processing. Time on a synchrotron source is not unlimited, so proper sample preparation is essential to avoid wasted trips. There are numerous ways to determine sample purity. For those labs who may have limited wet lab access, one simple way to test protein purity is to overload an SDS PAGE gel with increasing amounts of sample. A pure sample will only have a band corresponding to the molecular weight of your protein. The presence of multiple bands as you increase the concentration indicates the presence of contaminants. There are no stead fast rules for the amount of protein to use for these experiments, but in general your highest concentration lane should include 1mg of protein when possible. The exact amount of sample needed will depend on the protein concentration and will need to be determined by the lab.

2.2 Buffer Selection

All SAS experiments are compared with a buffer blank. This is standard procedure and is the only way that scattering data can be interpreted from the noise of buffer components. The nature of SAS demands exact buffer matches to ensure that buffer subtraction removes the most noise possible from the scattering data. The scattering from a target protein is not significantly greater than that of the background [4]. This means that preparing a similar buffer to the sample for a blank is not sufficient for SAS experiments. Once the sample has been concentrated and determined to be monodisperse, the sample product should be dialyzed against the final buffer. In many cases this will be the same buffer that the protein is already in. Dialysis is the best way to obtain an identical sample. Once your protein has been concentrated, dialyze it in the same buffer it is already in. The buffer blank should be taken from that final dialysis buffer to ensure that the two samples are identical. Beyond concerns about buffer subtraction, there are several components that may be necessary for protein purification, but have a large scattering intensity. Detergents are one of these components. Many detergents have a large scattering signal, and in some cases the detergent signal can wash out the signal from the target protein. When possible SAS samples should be prepared without detergents. However in some cases detergents may be necessary to prepare monodisperse samples. The SIBYLS beamline has produced a list of detergents ranked according to their suitability to SAS samples (http://bl1231.als.lbl.gov/saxs_protocols/index.php). This list can be found on their website, and is a great resource when determining buffer content. Salt concentration can also impact the scattering data. Generally salt concentrations should be kept to the minimum concentration needed for a stable sample. However, according to the SIBYLS beamline, concentrations up to 1M have been used with good results [12]. It is important to note that aside from detergents, buffer scattering is significantly less crucial than sample monodispersity. This means that if high salt concentrations are required for a sample to be monodisperse the sample should be prepared with high salt [12]. In general, samples should be prepared in the simplest buffer (lowest concentration/ least components) that allows for a monodisperse and stable sample at a series of protein concentrations.

2.3 Monodispersity

The most significant factor to consider when preparing samples for SAS experiments is the monodispersity of the sample. Aggregation is the most common problem seen at the beamlines [12], and it can result in meaningless data. The first step in preventing aggregation is in the sample preparation. The buffer the protein is in should support a monodisperse sample that is stable in solution. Often times a monodisperse sample can become aggregated over time. In some cases this can be prevented by preparing the samples immediately before they are shipped to the beamline. Some extreme cases may require concentrating the sample at the beamline immediately prior to the experiment. There are several ways to determine monodispersity of a sample, the most common techniques used for SAXS samples are; native gel electrophoresis and dynamic light scattering. One of these should always be done before each SAS experiment. In fact, some beamlines may request some evidence of monodispersity [12].

2.4 Concentration

To determine information about the molecular weight of a sample an exact concentration is required. Calculating the molecular weight from scattering data can serve as an important control for cases where the molecular weight is already known. It is important to note that some buffer components can effect the perceived concentration of a sample. DTT can be oxidized when exposed to oxygen, and this change in oxidative state can result in an altered absorbance (at 280nm) reading when determining concentration using UV spectroscopy. Differential rate of oxidation between the sample and buffer can result in an inaccurate concentration reading. Also, it is important to conduct SAS in a concentration series to determine any concentration dependent aggregation or interparticle interference. Concentration series should consist of at least three points ranging between 1mg/mL and 10 mg/mL. The three concentration points will depend on the stability and monodispersity of the protein at a given concentration.

2.5 Designing a SAS Experiment

There are a great number of factors that go into planning a SAS experiment. The most important thing to remember is that the quality of your data determines the reliability of your models. If you choose not to utilize the appropriate controls then your results may not be accurate. Specific guidelines for experimental setup and controls have been thoroughly outlined in a 2010 review appropriately titled: *Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls* [4]. Taking the additional steps recommended in this article can save you alot of time and effort down the road.

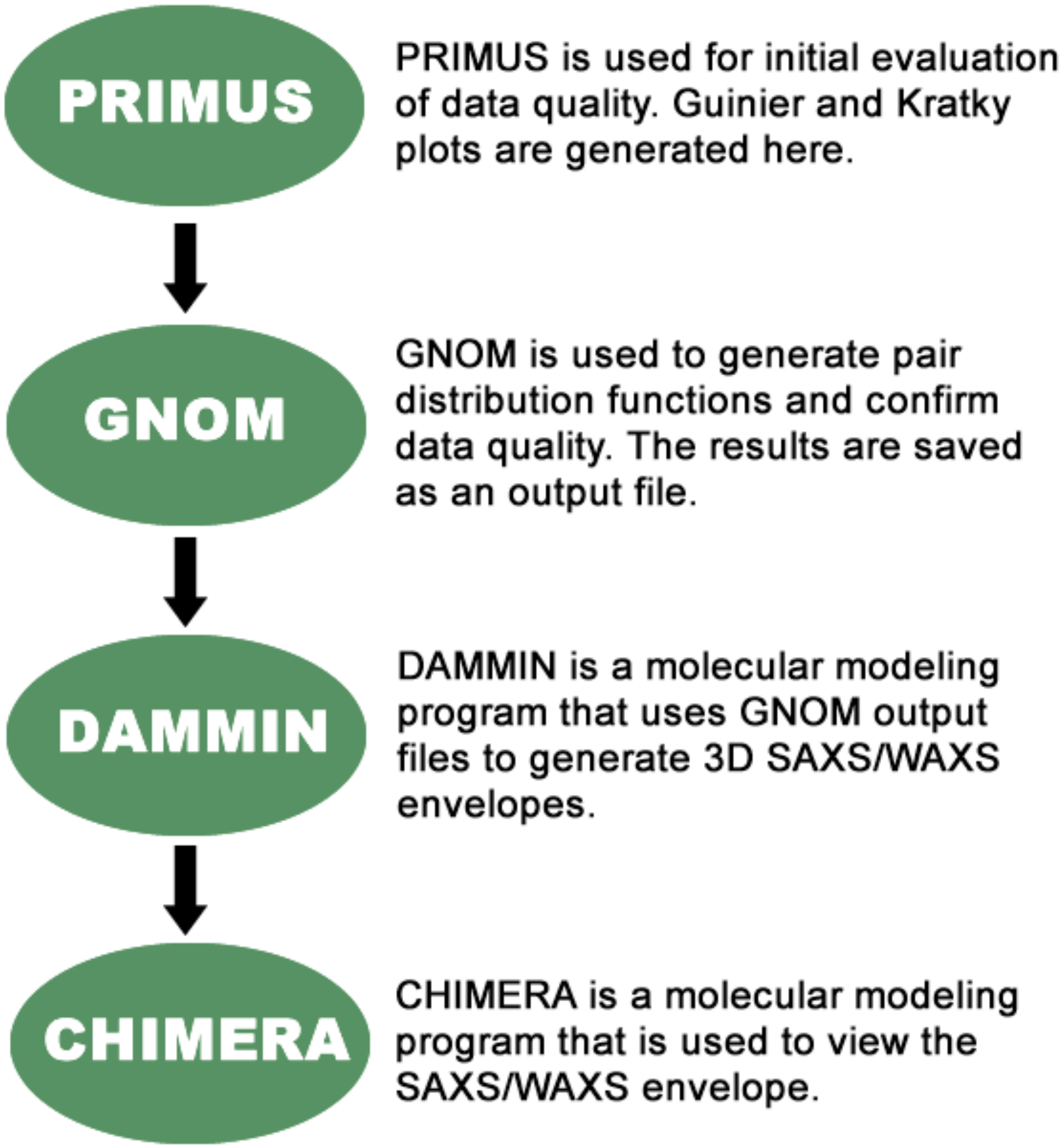
Chapter 3

Initial Data Analysis

In this section we will discuss how to determine the quality of your data as well as the steps needed to begin data processing. All data processing will be done within the ATSAS software package (http://www.embl-hamburg.de/biosaxs/software.html), and an overview of data processing can be found in Figure 1. We will not address the steps of buffer subtraction and data reduction. These steps are normally done at the beamline or before beginning data processing. If you would like more information on these topics, Bioisis (www.bioisis.net/tutorial) provides a series of useful tutorials on the topic.

3.1 PRIMUS

The first program we will be using in the ATSAS software suite is called PRIMUS [13]. This program is useful for plotting your data and determining the data quality. When using PRIMUS the standard form of the data will be a plot of I(q) vs q, where I is the intensity of the scattering pattern and q ($q=(4\pi \mathrm{Sin}\theta)/\lambda$) is the amplitude of the scattering vector or momentum transfer [1-4]. This plot is equivalent to the intensity as a function of the scattering angle [1-4]. The general shape of this plot can give some basic information on the quality of the data, however it can often be difficult to draw conclusions from this alone. Guinier plots are the first quality control method used to determine data quality. Examples of good and bad data can be seen in Figure 2.



PRIMUS is used for initial evaluation of data quality. Guinier and Kratky plots are generated here.

GNOM is used to generate pair distribution functions and confirm data quality. The results are saved as an output file.

DAMMIN is a molecular modeling program that uses GNOM output files to generate 3D SAXS/WAXS envelopes.

CHIMERA is a molecular modeling program that is used to view the SAXS/WAXS envelope.

Figure 1: Overview of SAXS data processing. This figure shows a flow chart for the basic steps in SAXS data processing. Detailed explanations of each step are given in the text.

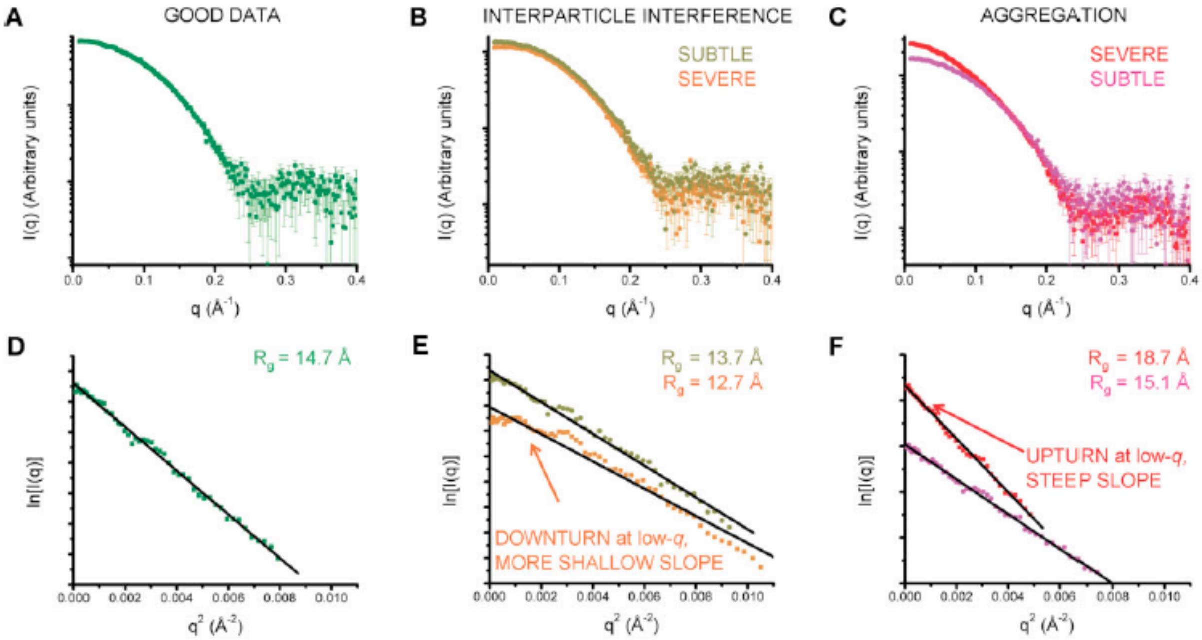
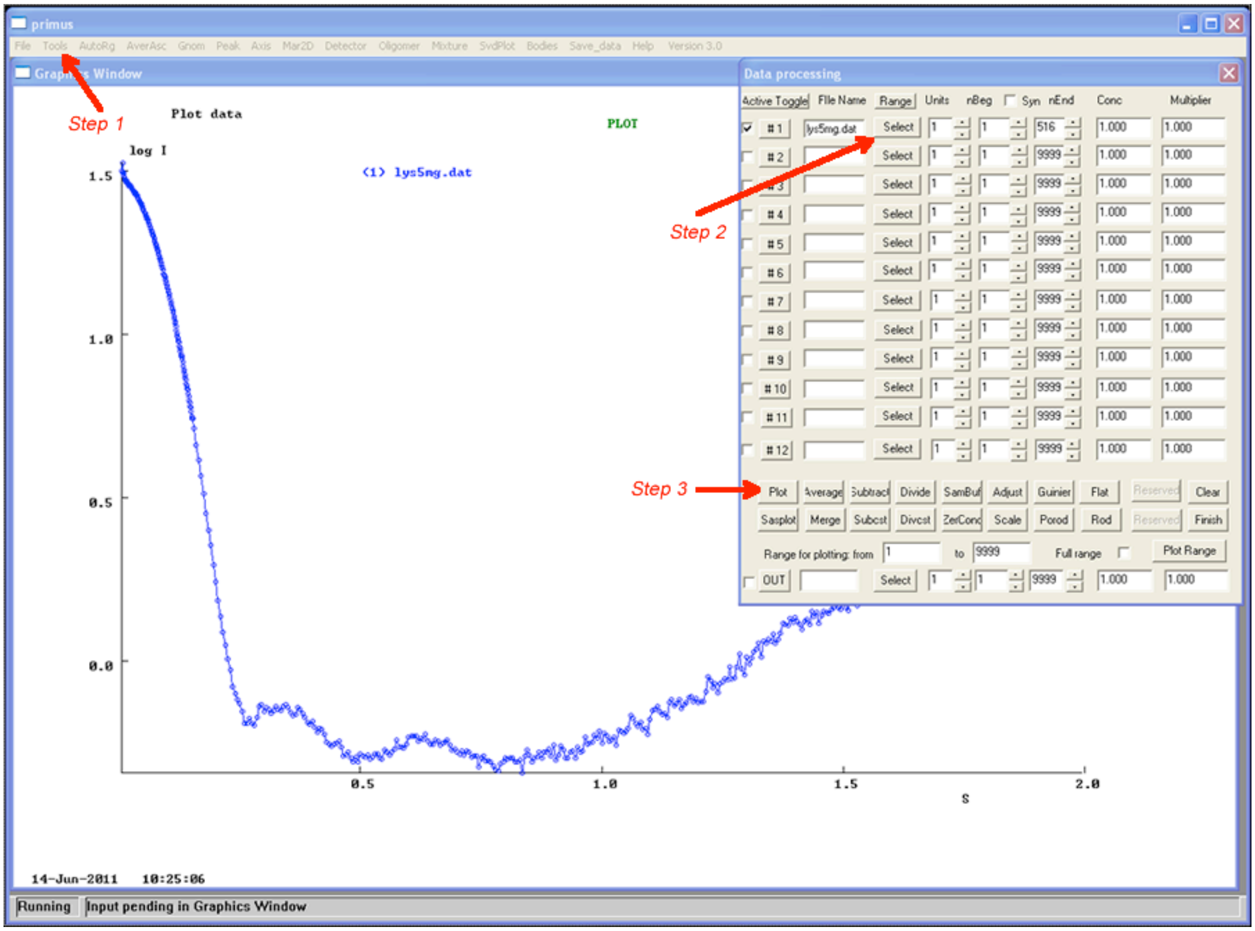


Figure 2: A-C are examples of good and bad data when viewed as a plot of I(q) vs q. **(A)** shows data that is free from aggregation and interparticle interference. This is confirmed by the Guinier plot of this data shown in **(D)**. **(B)** shows data that has interparticle interference. This is not easily visible in the I(q) vs q plot, but is evident in the Guinier plot. **(E)** which shows a decrease in I(0) (downturn at low q) and a corresponding decrease in Rg. **(C)** shows data which is aggregated. Severe aggregation is noticeable in the I(q) vs q, and is confirmed by the Guinier plot **(F)**. The increase in I(0) (upturn at low q) and the increase in Rg are both characteristic of aggregation.

Figure taken from Jacques, D.A, Trewhella, J., Protein Science. 2010 (**19**): 642-657

3.2 Plotting Data in PRIMUS

Figure 3 shows the PRIMUS startup screen. The first step is to click on the Tools menu as shown (Step 1: Figure 3). When you do this the data processing window will load. Note: do not close this window until you are ready to exit PRIUMS because you will have to restart the program to re-open it. Once you have the data processing window open you will click on the "select" button (Step 2: Figure 3). This opens a standard "open" window. Select the data set you would like to begin with, and click open. Click the "plot" button in the lower left corner or the plot range button in the lower right corner to display your data range (Step 3: Figure 3). When you are initially plotting the full range there is no difference between these options, but if you want to plot a limited range of data points you must click the "plot" button in the lower left corner. One thing you may notice is different from what you have encountered before is the units. In PRIMUS, the momentum transfer, 'q', is shown as 's'. Functionally there is no difference. For now we will focus on one data set at a time, but as you move on you can load multiple data sets and plot them simultaneously to check for differences. This is especially useful when you are looking at concentration series, or if you would like to merge two data sets. To do this simply press select on the #2 row and select your second data set (Step 1: Figure 4). When plotting multiple sets, the check boxes on the left hand side determine which set will be plotted. When working with multiple data sets you need to plot each set individually to conduct specific data analyses (i.e. Guinier, Kratky, etc.) otherwise only the lowest numbered data set plotted will be used. The data shown here as an example is SAXS/WAXS data for lysozyme at 5 mg/mL concentration. WAXS data has a different profile from SAXS data. Pure SAXS data is



Step 1

Step 2

Step 3

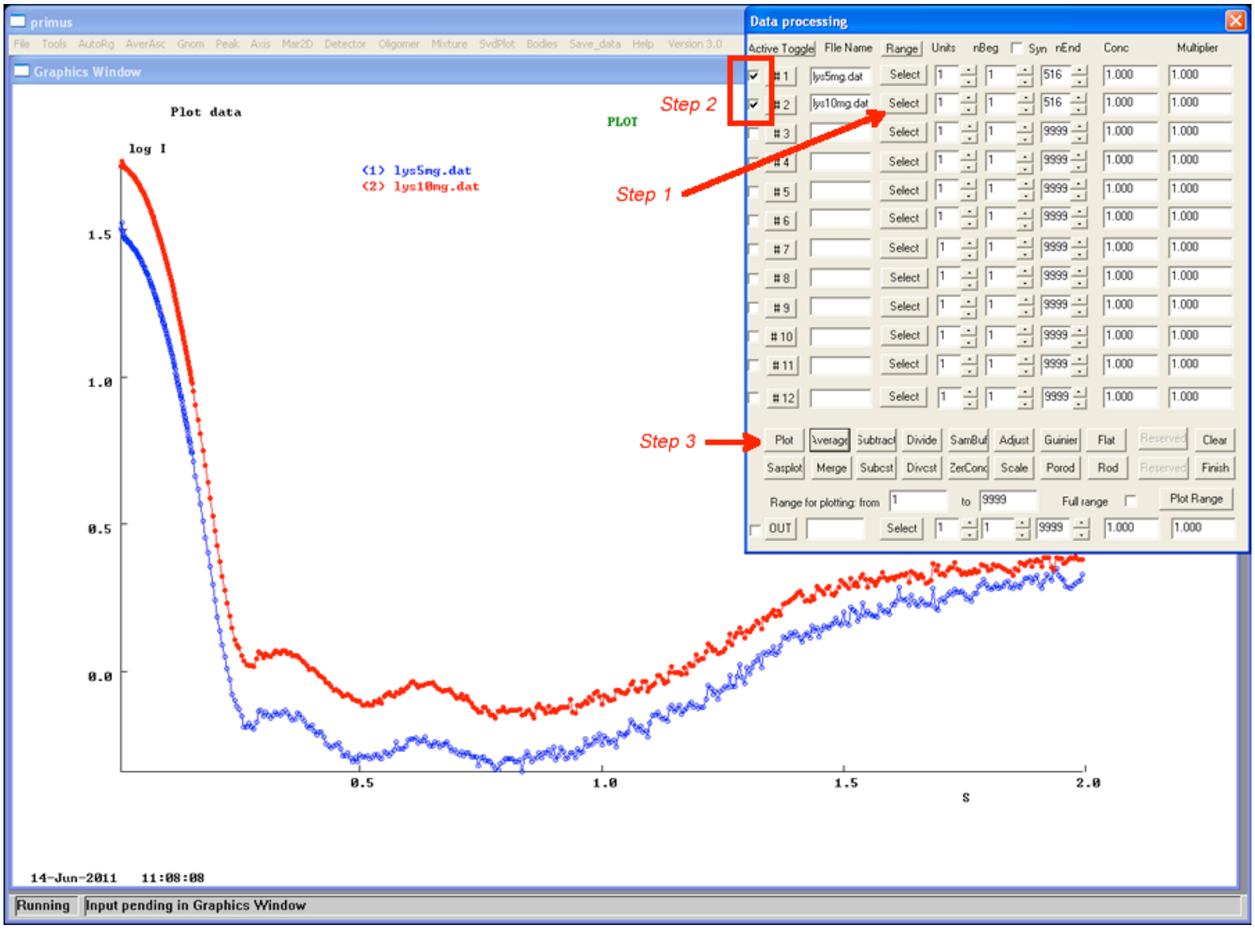
Active Toggle		File Name	Range	Units	nBeg	□ Syn nEnd	Conc	Multiplier
☑	# 1	lys5mg.dat	Select	1	1	516	1.000	1.000
□	# 2		Select	1	1	9999	1.000	1.000
□	# 3		Select	1	1	9999	1.000	1.000
□	# 4		Select	1	1	9999	1.000	1.000
□	# 5		Select	1	1	9999	1.000	1.000
□	# 6		Select	1	1	9999	1.000	1.000
□	# 7		Select	1	1	9999	1.000	1.000
□	# 8		Select	1	1	9999	1.000	1.000
□	# 9		Select	1	1	9999	1.000	1.000
□	# 10		Select	1	1	9999	1.000	1.000
□	# 11		Select	1	1	9999	1.000	1.000
□	# 12		Select	1	1	9999	1.000	1.000

Range for plotting: from 1 to 9999 Full range □

□	OUT		Select	1	1	9999	1.000	1.000

Running | Input pending in Graphics Window

Figure 3: This is the startup screen of PRIMUS with directions for plotting a data set. **Step 1:** Click on the tools window. This will bring up the data processing window (Do not close this window until you are ready to exit PRIMUS). **Step 2:** Click select from the first row in the data processing window to bring up the 'open' window. From here select the data set that you wish to plot. **Step 3:** Once you have loaded a data set, click the plot button to display the data. This appearance of this plot will vary based on the protein being studied. Additional steps are required to determine the quality of the data.


primus
File Tools AutoRg AverAsc Gnom Peak Axis Mar2D Detector Oligomer Mixture SvdPlot Bodies Save_data Help Version 3.0
Graphics Window
Plot data
PLOT
log I
(1) lys5mg.dat
(2) lys10mg.dat
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
s
14-Jun-2011 11:08:08
Running
Input pending in Graphics Window
Step 1
Step 2
Step 3
Data processing
Plot Average Subtract Divide SamBuf Adjust Guinier Flat Reserved Clear
Sasplot Merge Subcst Divcst ZerConc Scale Porod Rod Reserved Finish
Plot Range

Active Toggle	File Name	Range	Units	nBeg	□ Syn nEnd	Conc	Multiplier
☑ # 1	lys5mg.dat	Select	1	1	516	1.000	1.000
☑ # 2	lys10mg.dat	Select	1	1	516	1.000	1.000
□ # 3		Select	1	1	9999	1.000	1.000
□ # 4		Select	1	1	9999	1.000	1.000
□ # 5		Select	1	1	9999	1.000	1.000
□ # 6		Select	1	1	9999	1.000	1.000
□ # 7		Select	1	1	9999	1.000	1.000
□ # 8		Select	1	1	9999	1.000	1.000
□ # 9		Select	1	1	9999	1.000	1.000
□ # 10		Select	1	1	9999	1.000	1.000
□ # 11		Select	1	1	9999	1.000	1.000
□ # 12		Select	1	1	9999	1.000	1.000
□ OUT		Select	1	1	9999	1.000	1.000

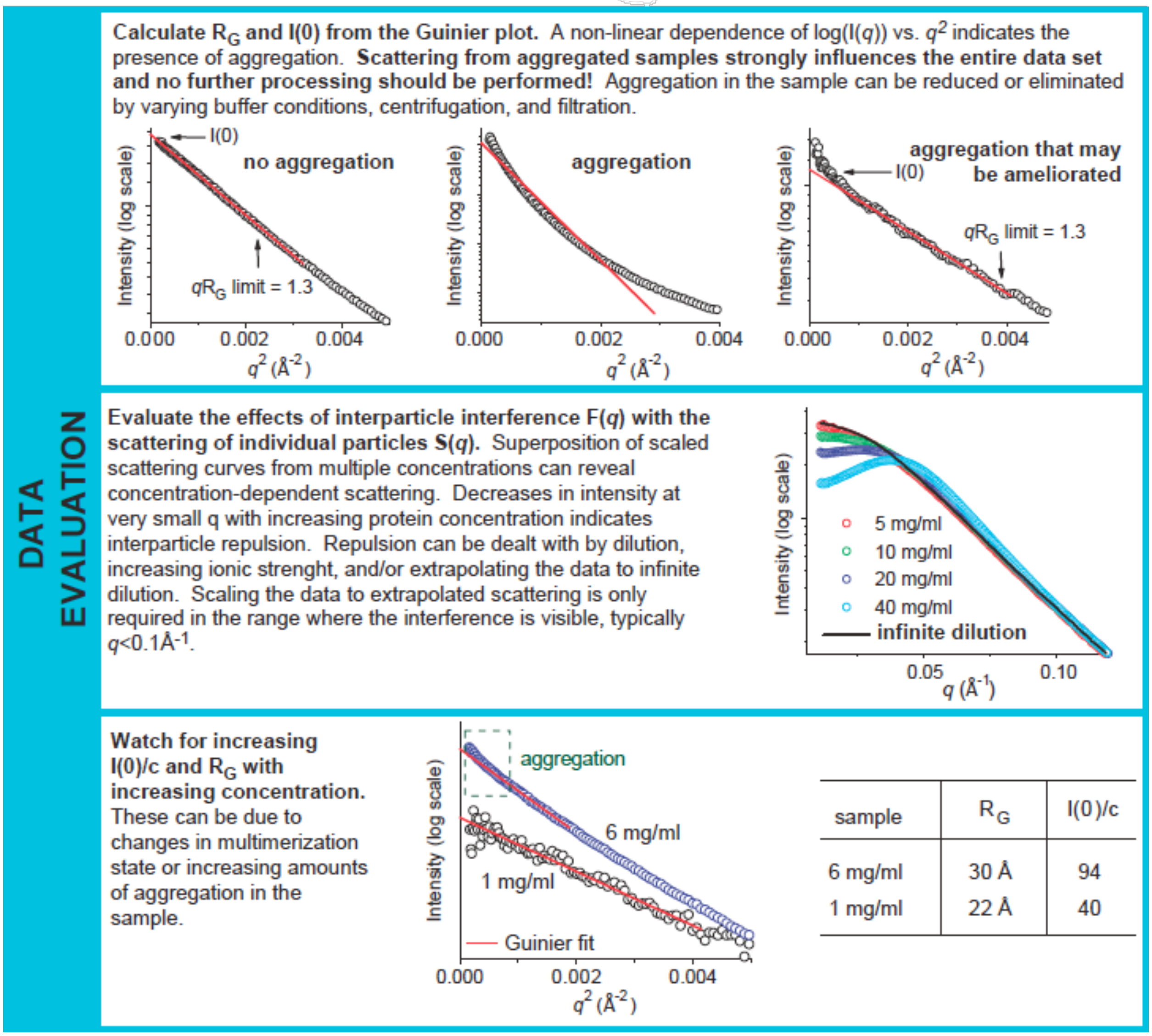
Range for plotting: from 1 to 9999 Full range □

Figure 4: Instructions for plotting multiple data sets. **Step 1:** After you have loaded your first data set (shown in Figure 3), click on the select button on the second row of the data processing window. Select the second data set that you want to load. **Step 2:** Make sure that the check boxes for both data sets are selected. If only one is selected then both data sets will not plot. **Step 3:** Plot the two data sets by clicking the plot button in the lower left corner of the tool window.

much shorter and will consist of mainly noise at high q values. WAXS data however, is defined by the high q values and the profile will be significantly longer than SAXS data alone. The large degree of noise in the higher q values for SAXS data can make structural analysis more difficult. Data truncation is sometimes needed to remove these regions. While it is possible to remove data points permanently using PRIMUS, I do not suggest this as a first choice. Instead, it is much easier to simply exclude these points from data analysis if they are impacting the generation of a reasonable P(r) distribution (see Chapter 4). Data truncation will be discussed in more detail in Chapter 4. If you do choose to remove the points in PRIMUS it is a very simple process. Load and plot the data set you wish to truncate. To adjust the length of the plotted graph you must use the plot button as opposed to the plot range option. Once the data is plotted, adjust the nEnd value until the noisy data points are removed then click the merge button. This will create a new file in the same directory as your data set that is named merge##.dat. Load this file as a normal data set to work with the truncated data. After you have plotted your data the next step is to begin data analysis with a Guinier plot.

3.3 Guinier Plots

Guinier analysis is used to evaluate the SAXS scattering data at very small scattering angles [1]. Guinier plots are graphs of the natural logarithm of the scattering intensity, I(q), versus the square of the amplitude of the scattering vector, q^2 (ln(I(q) vs q^2)[1-3]. For Guinier plots to be accurate, they must obey Guinier's Law. For globular proteins this means the upper sRg limit must not exceed 1.3 (Figure 7: Step 1)[1-4]. Guinier analysis allows for the calculation of the radius of gyration (Rg) and the intensity



Calculate R_G and I(0) from the Guinier plot. A non-linear dependence of log(I(q)) vs. q^2 indicates the presence of aggregation. **Scattering from aggregated samples strongly influences the entire data set and no further processing should be performed!** Aggregation in the sample can be reduced or eliminated by varying buffer conditions, centrifugation, and filtration.

Evaluate the effects of interparticle interference F(q) with the scattering of individual particles S(q). Superposition of scaled scattering curves from multiple concentrations can reveal concentration-dependent scattering. Decreases in intensity at very small q with increasing protein concentration indicates interparticle repulsion. Repulsion can be dealt with by dilution, increasing ionic strenght, and/or extrapolating the data to infinite dilution. Scaling the data to extrapolated scattering is only required in the range where the interference is visible, typically q<0.1Å^{-1}.

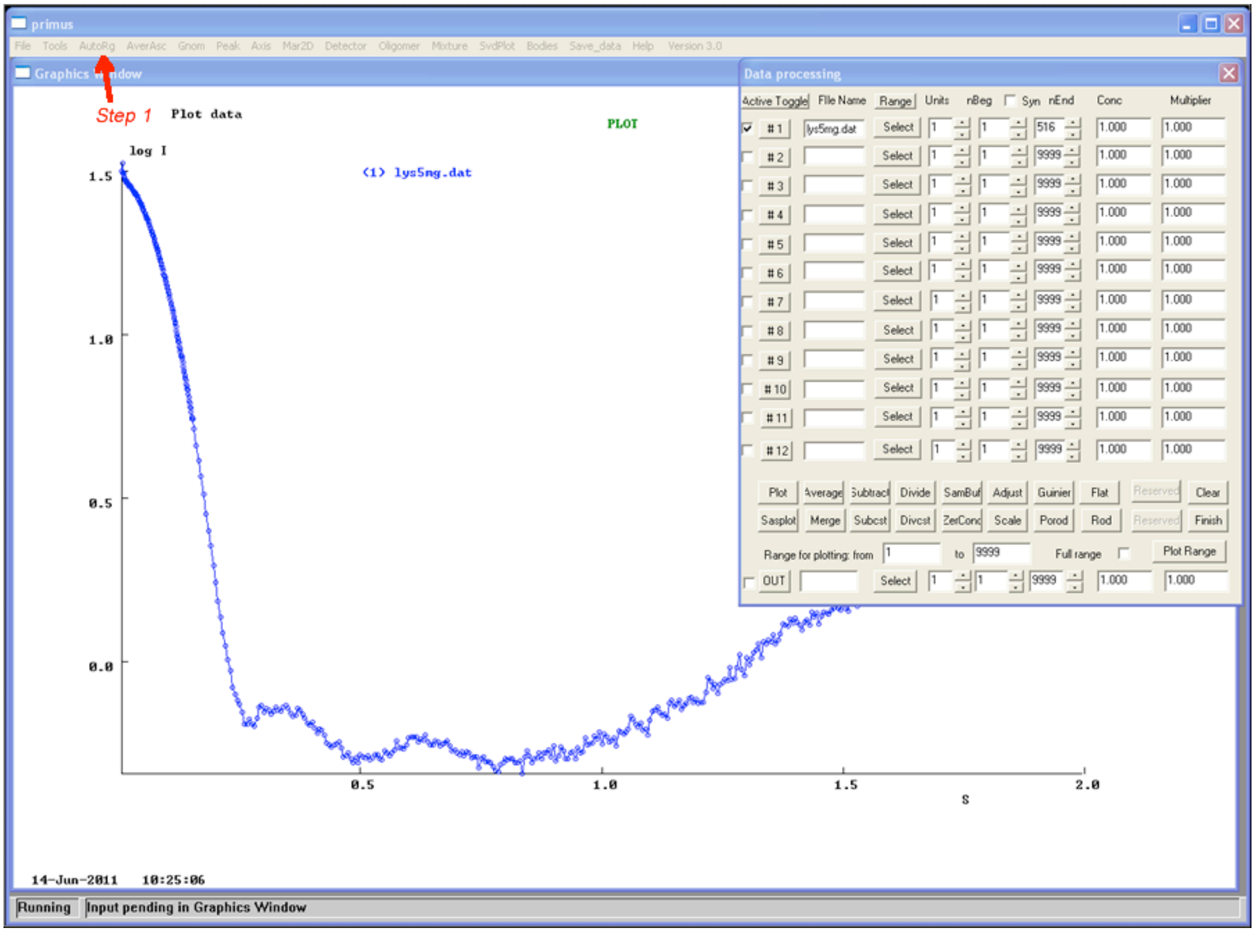
Watch for increasing I(0)/c and R_G with increasing concentration. These can be due to changes in multimerization state or increasing amounts of aggregation in the sample.

sample	R_G	I(0)/c
6 mg/ml	30 Å	94
1 mg/ml	22 Å	40

Figure 5: Evaluating data quality using Guinier plots. Guinier plots provide a quick and easy way to evaluate data quality in PRIMUS. They also provide initial approximations of the radius of gyration (Rg) and the zero scattering angle intensity (I(0)). See text for a more detailed description of Guinier plots.

Figure taken from Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. Q Rev Biophys, 2007 40(**3**):191-285

I(0) at zero angle scattering (q=0). The radius of gyration describes the volume distribution of a sample, and refers to the root-mean-squared distance of all the particles from the axis of rotation [4]. Two proteins with identical molecular weights can have different Rg values [4]. The zero angle scatting intensity is used to calculate the molecular weight of the protein, and the radius of gyration will be important for generating a valid structure. In addition to these parameters, Guinier analysis can provide important information about sample aggregation and interparticle interference (Figure 2). Guinier plots are generally shown with a line of best fit. When describing the changes that occur under different circumstances this line is used as the point of reference for movement. When samples are aggregated the slope of the plot will increase and often times points will appear consistently above the best fit line. This creates an effect commonly called a smiling guinier (Figure 5: Top panel). This observation is commonly associated with an increase in both Rg and I(0) [1,4]. Interparticle interference is another common problem. The characteristic appearance in the guinier plot is a frowning guinier (Figure 5: Middle panel). This occurs when points at low and high q values appear below the line of best fit while points in the middle are above. This is accompanied by a decrease in both Rg and I(0). Sample which are free from both aggregation and interparticle interference will have a linear guinier plot and show no change in Rg and I(0). Small changes in the Rg and I(0) values along with a linear guinier plot indicate subtle aggregation or interparticle interference [1,2,4].



Active Toggle		File Name	Range	Units	nBeg	□ Syn nEnd	Conc	Multiplier
☑	# 1	lys5mg.dat	Select	1	1	516	1.000	1.000
□	# 2		Select	1	1	9999	1.000	1.000
□	# 3		Select	1	1	9999	1.000	1.000
□	# 4		Select	1	1	9999	1.000	1.000
□	# 5		Select	1	1	9999	1.000	1.000
□	# 6		Select	1	1	9999	1.000	1.000
□	# 7		Select	1	1	9999	1.000	1.000
□	# 8		Select	1	1	9999	1.000	1.000
□	# 9		Select	1	1	9999	1.000	1.000
□	# 10		Select	1	1	9999	1.000	1.000
□	# 11		Select	1	1	9999	1.000	1.000
□	# 12		Select	1	1	9999	1.000	1.000

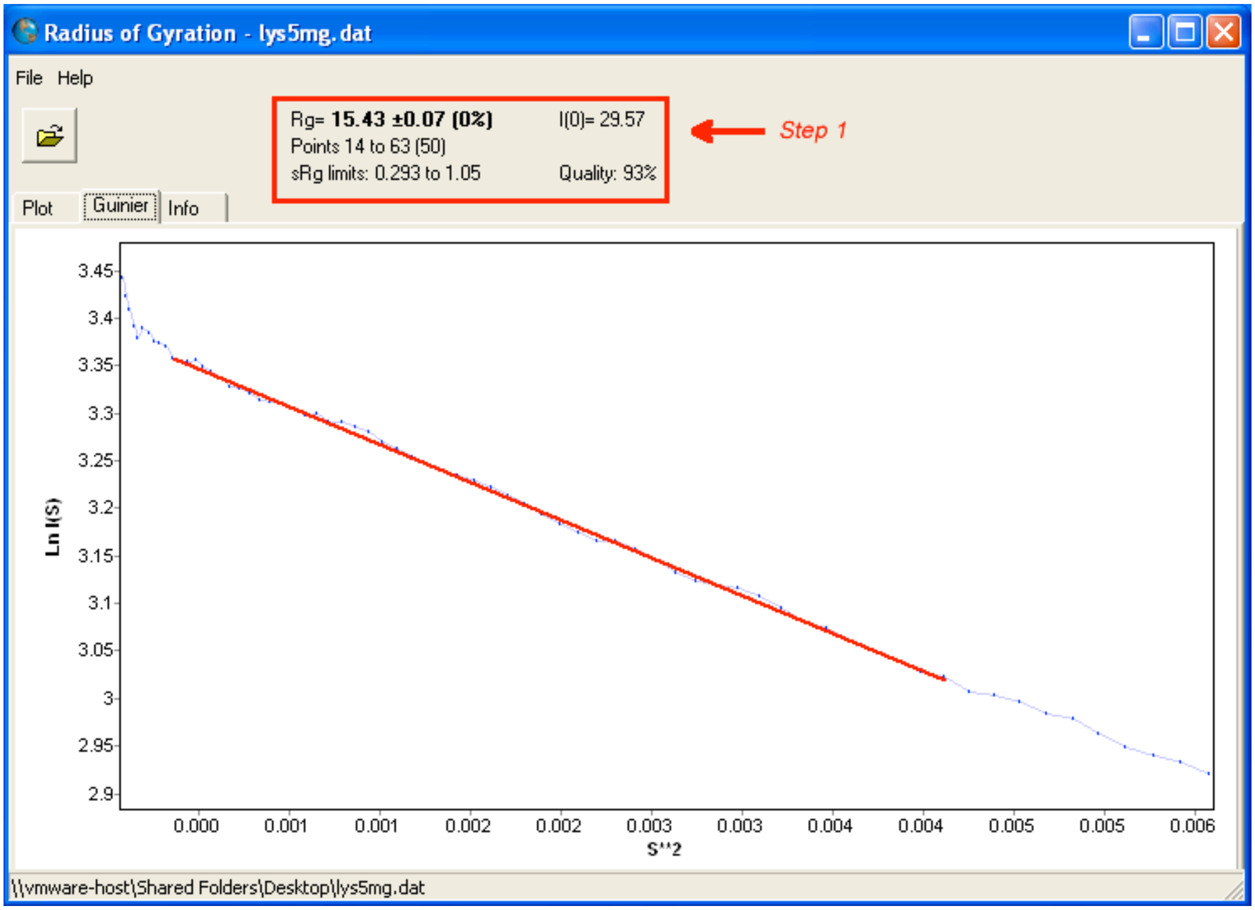
Range for plotting: from 1 to 9999 Full range □

□	OUT		Select	1	1	9999	1.000	1.000

14-Jun-2011 10:25:06

Running | Input pending in Graphics Window

Figure 6: Steps to performing an auto-Rg approximation in PRIMUS. **Step 1:** Once you have selected a data set, select the autoRg function from the menu as shown. This will bring up a small window with the results of the approximation.



Radius of Gyration - lys5mg.dat

Rg= **15.43 ±0.07 (0%)** I(0)= 29.57
Points 14 to 63 (50)
sRg limits: 0.293 to 1.05 Quality: 93%

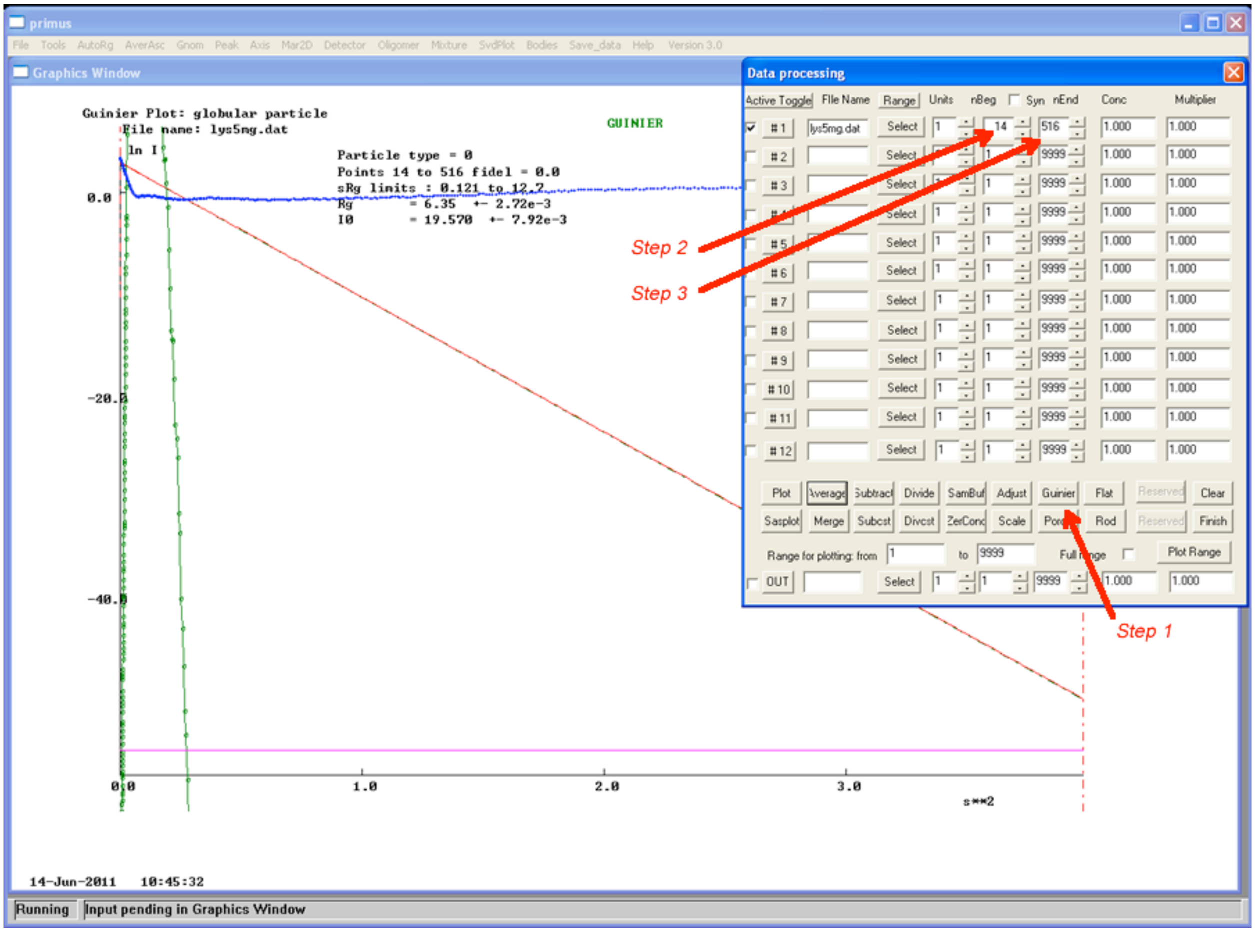
Step 1

\\vmware-host\Shared Folders\Desktop\lys5mg.dat

Figure 7: Interpreting the results of an autoRg approximation. **Step 1:** Once you have performed an autoRg the results will appear with a graph of the data points evaluated. The numerical values shown in Step 1 should be recorded. Additional information on interpreting the results can be found in the text.

3.4 Generating Guinier Plots in PRIMUS

There are two ways to generate Guinier plots in PRIMUS, it is good practice to do both. The first method uses a feature known as AutoRg which will automatically find those points which obey Guinier's Law and generate an Rg and I(0) from them. To do this, load and plot the data set you wish to evaluate. Then from the top menu bar select AutoRg (Step 1: Figure 6). This will load a pop up window with a Guinier plot and some data values listed at the top (Step 1: Figure 7). The Rg value is given with the I(0) value in the top row. This information is very useful and should be recorded. These values will be compared to future approximations to determine their accuracy and can be essential in producing an accurate P(r) distribution. The second line contains the data points that AutoRg used to generate the plot. You will notice that the evaluation does not start at the first data point. The program automatically evaluates the data and discards those points which do not obey Guinier's law. You should record the starting point. During later analysis you will want to avoid including these unreliable data points. The last line contains the sRg limits and the data quality indicator. The sRg maximum limit must be below or equal to 1.3. Data which does not fall below sRg ≤ 1.3 is considered inaccurate. For the AutoRg function these limits will always fall in the acceptable range, but when you manually generate the Guinier plot you will have to make sure the data doesn't exceed this limit. The quality rating is an indicator of how well the best fit line in the Guinier plot fits the data points. The quality of the data is determined by how linear the data points are for the evaluated range. Higher quality data which is free from aggregation and interparticle interference will be linear. In general, as data quality increases the data points in this region will be more linear. After you have recorded the


primus
File Tools AutoRg AverAsc Gnom Peak Axis Mar2D Detector Oligomer Mixture SvdPlot Bodies Save_data Help Version 3.0
Graphics Window
GUINIER
ln I
0.0
-20.0
-40.0
0.0
1.0
2.0
3.0
s**2
Data processing
Plot Average Subtract Divide SamBuf Adjust Guinier Flat Reserved Clear
Sasplot Merge Subcst Divcst ZerConc Scale Porod Rod Reserved Finish
Plot Range
Step 1
Step 2
Step 3

```
Guinier Plot: globular particle
     File name: lys5mg.dat

                    Particle type = 0
                    Points 14 to 516 fidel = 0.0
                    sRg limits : 0.121 to 12.7
                    Rg        = 6.35   +- 2.72e-3
                    I0        = 19.570  +- 7.92e-3

14-Jun-2011   10:45:32
```

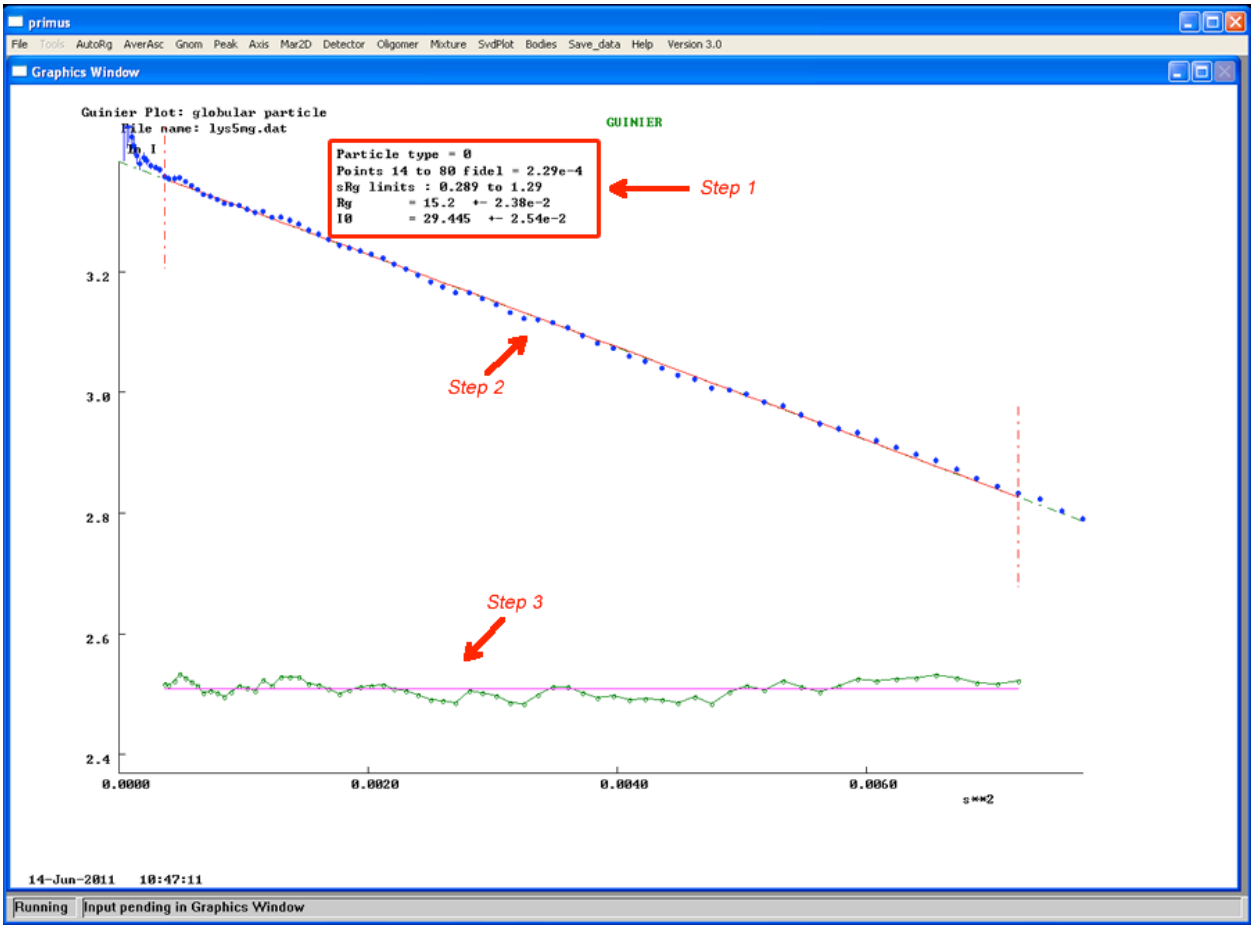
Active Toggle	File Name	Range	Units	nBeg	□ Syn nEnd	Conc	Multiplier
☑ #1	lys5mg.dat	Select	1	14	516	1.000	1.000
□ #2		Select	[illegible]	1	9999	1.000	1.000
□ #3		Select	1	1	9999	1.000	1.000
□ #4		Select	1	1	9999	1.000	1.000
□ #5		Select	1	1	9999	1.000	1.000
□ #6		Select	1	1	9999	1.000	1.000
□ #7		Select	1	1	9999	1.000	1.000
□ #8		Select	1	1	9999	1.000	1.000
□ #9		Select	1	1	9999	1.000	1.000
□ #10		Select	1	1	9999	1.000	1.000
□ #11		Select	1	1	9999	1.000	1.000
□ #12		Select	1	1	9999	1.000	1.000

Range for plotting: from 1 to 9999 Full range □

□ OUT		Select	1	1	9999	1.000	1.000

Running | Input pending in Graphics Window

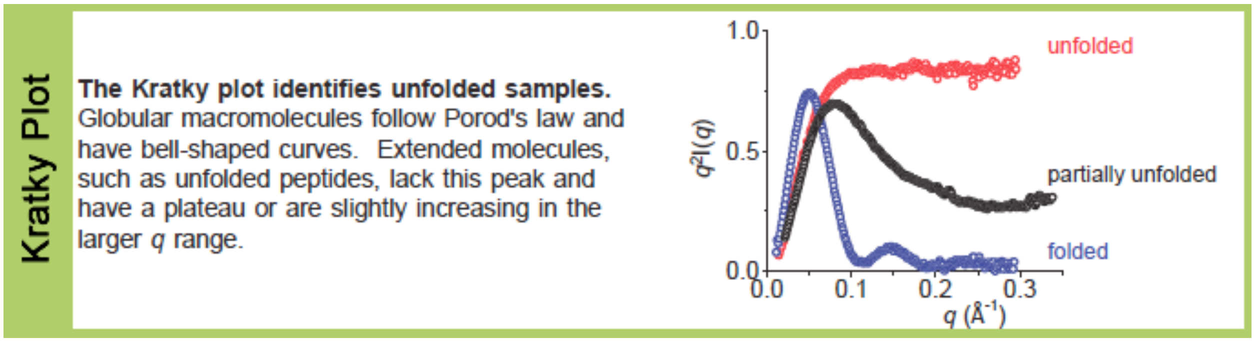
Figure 8: Manually generating a Guinier plot. **Step 1:** After you have loaded your data set, click the Guinier button on the data processing window. **Step 2:** Adjust the nBeg value to exclude the points that do not follow Guinier’s Law. This value can be found by completing an AutoRg evaluation (see Figure 7). **Step 3:** Adjust the nEnd value until the upper sRg limit is less than or equal to 1.3.


primus
File Tools AutoRg AverAsc Gnom Peak Axis Mar2D Detector Oligomer Mixture SvdPlot Bodies Save_data Help Version 3.0
Graphics Window
Guinier Plot: globular particle
File name: lys5mg.dat
ln I
GUINIER
Step 1
Step 2
Step 3
3.2
3.0
2.8
2.6
2.4
0.0000
0.0020
0.0040
0.0060
s**2
14-Jun-2011 10:47:11
Running
Input pending in Graphics Window

```
Particle type = 0
Points 14 to 80 fidel = 2.29e-4
sRg limits : 0.289 to 1.29
Rg          = 15.2   +- 2.38e-2
I0          = 29.445  +- 2.54e-2
```

Figure 9: Evaluating manual Guinier plots. **Step 1:** The important data and errors can be found at the top of the graph (red box). These values should be recorded for comparison while generating a P(r) distribution (Chapter 4). **Step 2:** Examine the profile of the plotted data to identify any trends. **Step 3:** The horizontal plot can be easier to identify trends in the data. For a more detailed explanation of identifying trends in data can be found in the text.

values from the plot (or generated a screen shot of the plot) close the window. The next step will be to generate a Guinier plot manually. To do this select and plot the data you want to evaluate, then click the Guinier button at the bottom of the data processing window (Step 1: Figure 8). A new graph will appear in the graphics window. Initially the plot will attempt to incorporate all of the data points. You will need to narrow the range. To do this first you will need to change the nBeg value to the first data point evaluated with AutoRg, this way you exclude any data which does not obey Guinier's law (Step 2: Figure 8). Now you can adjust the nEnd point until the sRg upper limit is in the acceptable range (≤ 1.3) (Step 3: Figure 8). Using a manual approximation, it is normally possible to include more data points in the plot than with the AutoRg approximations. Once you have reached an appropriate plot with in the sRg limits, record the Rg and I(0) values (Step 1: Figure 9). These should be very similar to those generated with AutoRg. One of the advantages to using the manual plot is the graph at the bottom of the graphics window. The green line of data points with a pink line of best fit (Step 3: Figure 9). If the data is close to linear you should not observe any trends in the data points (i.e. a series of points either above or below the line of best fit). This graph is generally used to determine if a Guinier plot is “smiling” or “frowning”. Smiling graphs will have a series of data points below the line of best fit, with points above the line on either end, forming a “smile”. Frowning graphs will exhibit the opposite. These trends can indicate aggregation and interparticle interference respectively. As previously described, changes in the values of Rg and I(0) in response to changes in concentration can also indicate problems in sample quality.



The Kratky plot identifies unfolded samples. Globular macromolecules follow Porod's law and have bell-shaped curves. Extended molecules, such as unfolded peptides, lack this peak and have a plateau or are slightly increasing in the larger *q* range.

Figure 10: Kratky plots give information concerning the folded or unfoldedness of a protein. A more detailed description of Kratky plots can be found in the text.

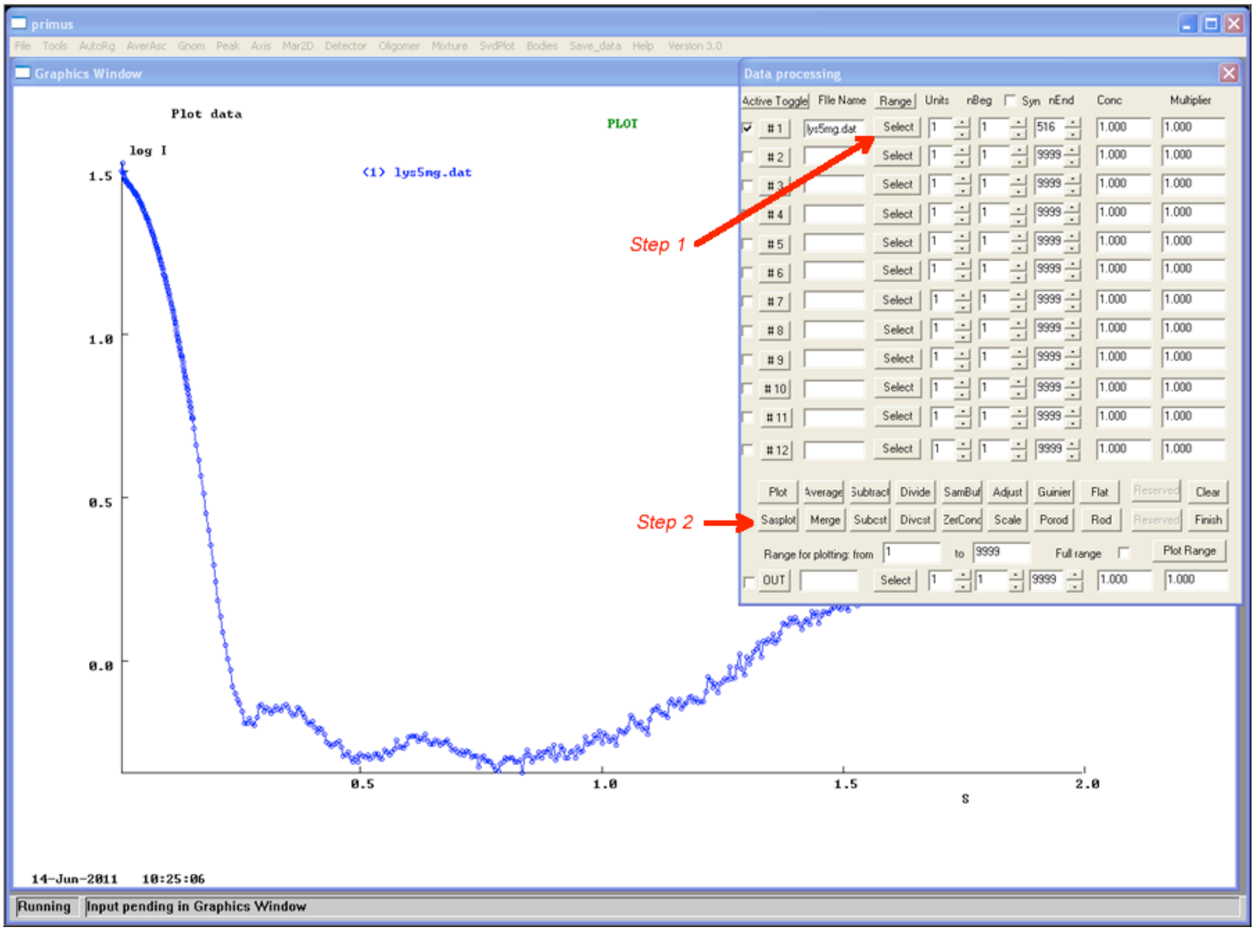
Figure taken from Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A., Q Rev Biophys. 2007 40(**3**):191-285

3.5 Kratky Plot Analysis

A Kratky plot (I(q)*q2 vs q) is used to confirm the folded or unfolded state of the sample [1]. This information does not give information about the monodispersity of the sample, but can be useful for confirming that the protein is in the correct folded state. The Kratky plot can help provide information as to the dynamic state of the protein [1]. Tightly folded proteins will behave differently from highly dynamic or unfolded proteins. Figure 10 shows the common results for Kratky plot analysis according to the protein folded state. A folded protein will have a sharp peak at low q values, and then the plot will return to near zero (blue). Completely unfolded proteins (red) will have a sharp rise at low q values but will maintain a plateau appearance. Partially unfolded protein (black) will have a broader peak at low q values than folded proteins but will remain at a higher level at high q values and not return to zero. This looks similar to the unfolded state. The information from a Kratky plot can be used to confirm that your sample is in the expected conformation. It also allows you to make a rough approximation on how dynamic a protein is in solution.

3.6 Generating Kratky Plots through PRIMUS

Kratky analysis can be done directly through PRIUMS. To start, load and plot the data that you would like to analyze (Step 1: Figure 11). Then select the SASPLOT button from the data processing menu (Step 2: Figure 11). This will launch a second window with a log(I) vs s plot. Select the Symbols option from the top menu (Step 1: Figure 12), this will change the data points to circles. This step is not required and is only for aesthetics. Now click the view drop down menu and select the option Y*X^2 : X, this generates the Kratky plot (Step 2: Figure 12). Please refer to the introductory text and the referenced reviews for more information on interpreting this plot (Step 3: Figure 12). Please note that if you are using SAXS/WAXS data it may be necessary to truncate the data to $s \leq 0.5$ or the plot may be scaled in a way which is hard to interpret.


primus
File Tools AutoRg AverAsc Gnom Peak Axis Mar2D Detector Oligomer Mixture SvdPlot Bodies Save_data Help Version 3.0
Graphics Window
Plot data
PLOT
log I
(1) lys5mg.dat
1.5
1.0
0.5
0.0
0.5
1.0
1.5
2.0
s
14-Jun-2011 10:25:06
Running
Input pending in Graphics Window
Data processing
Step 1
Step 2
Plot Average Subtract Divide SamBuf Adjust Guinier Flat Reserved Clear
Sasplot Merge Subcst Divcst ZerConc Scale Porod Rod Reserved Finish
Plot Range

Active Toggle	File Name	Range	Units	nBeg	□ Syn nEnd	Conc	Multiplier
☑ #1	lys5mg.dat	Select	1	1	516	1.000	1.000
□ #2		Select	1	1	9999	1.000	1.000
□ #3		Select	1	1	9999	1.000	1.000
□ #4		Select	1	1	9999	1.000	1.000
□ #5		Select	1	1	9999	1.000	1.000
□ #6		Select	1	1	9999	1.000	1.000
□ #7		Select	1	1	9999	1.000	1.000
□ #8		Select	1	1	9999	1.000	1.000
□ #9		Select	1	1	9999	1.000	1.000
□ #10		Select	1	1	9999	1.000	1.000
□ #11		Select	1	1	9999	1.000	1.000
□ #12		Select	1	1	9999	1.000	1.000
□ OUT		Select	1	1	9999	1.000	1.000

Range for plotting: from 1 to 9999 Full range □

Figure 11: Generating a Kratky plot from your data. **Step 1:** Select and load the data set that you wish to analyze. **Step 2:** Click the Sasplot button on the data processing window to bring up the data in a second window.

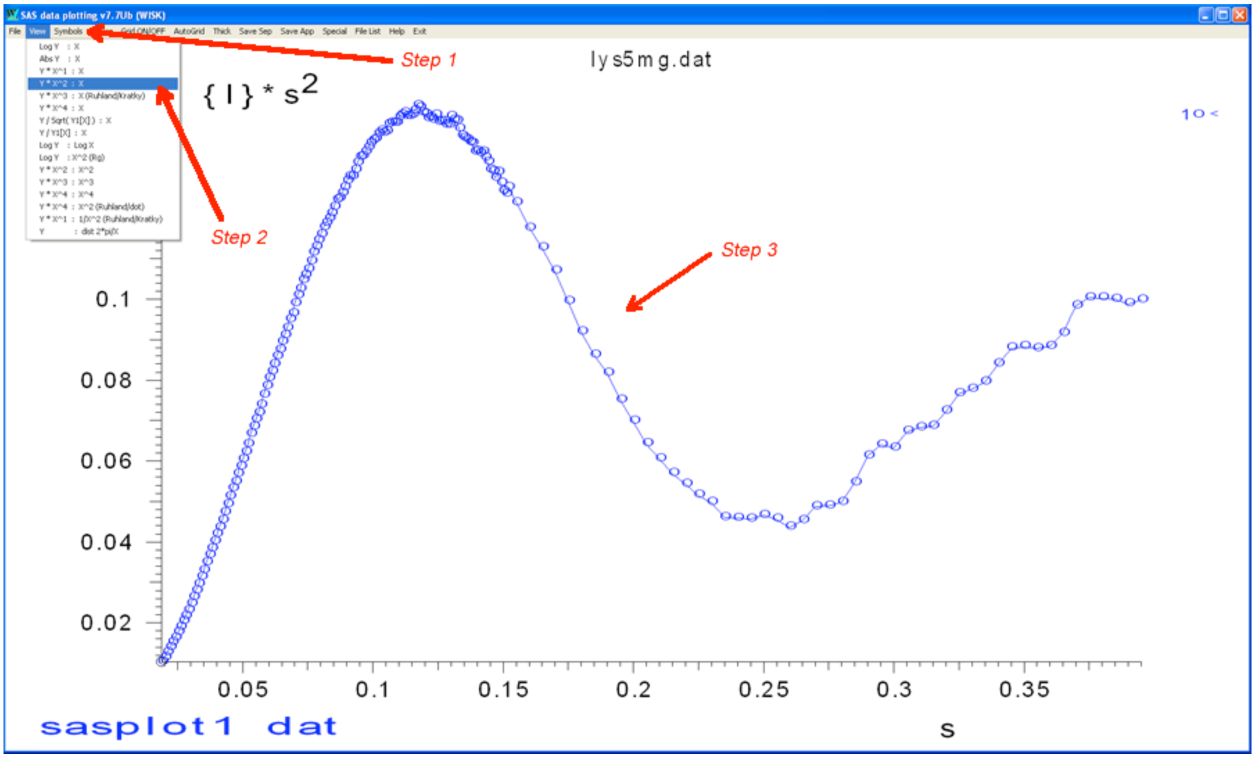


Figure 12: Data manipulation within Sasplot to generate a Kratky plot. **Step 1:** Click on Symbols in the top menu, this will change the appearance of the data points to circles. **Step 2:** Change the data view to **Y*X^2 : X** this will replot the data to form a Kratky plot. **Step 3:** Evaluate the results of the Kratky plot as described in the text.

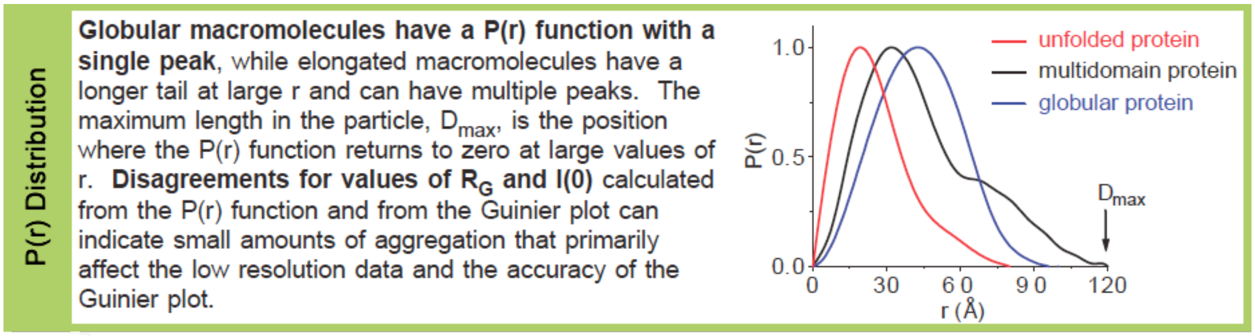
Chapter 4

Generating P(r) Distributions and Preparing for Structural Modeling

The next stage of data analysis will focus on the generation of a P(r) distribution. This step utilizes a second program from the ATSAS software package; GNOM [14]. GNOM will generate P(r) distributions as well as several other plots of varying importance. It will also generate an output file that is used to generate a structure. The P(r) distribution can be used to determine the quality of the sample as well as a general description of the overall structure of the protein. The description within this guide is not meant to be comprehensive, and should not be taken as such. The referenced material will provide a much more detailed description of the theory and mathematics behind the topics we discuss here.

4.1 Interatomic Distance Distributions Function, P(r)

Also known as the pair distance or vector length distance function, the P(r) distribution relates the position of electrons within the scattering sample [1-3]. Similar to the Patterson function from x-ray crystallography, the P(r) distribution is a Fourier transformation of the scattering profile (I(q) vs q plot) commonly shown as P(r) vs r. The major difference between P(r) distributions and Patterson functions is that P(r) distributions are radially averaged [1]. This means they do not provide information about vectors between scattering particles [1]. There are several important pieces of



P(r) Distribution

Globular macromolecules have a P(r) function with a single peak, while elongated macromolecules have a longer tail at large r and can have multiple peaks. The maximum length in the particle, D_{max}, is the position where the P(r) function returns to zero at large values of r. **Disagreements for values of R_G and I(0)** calculated from the P(r) function and from the Guinier plot can indicate small amounts of aggregation that primarily affect the low resolution data and the accuracy of the Guinier plot.

Figure 13: Pair distribution functions, P(r), relates the position of electrons within a scattering sample. For more information about assigning a Dmax, see text.

Figure taken from Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. Q Rev Biophys. 2007, 40(**3**):191-285

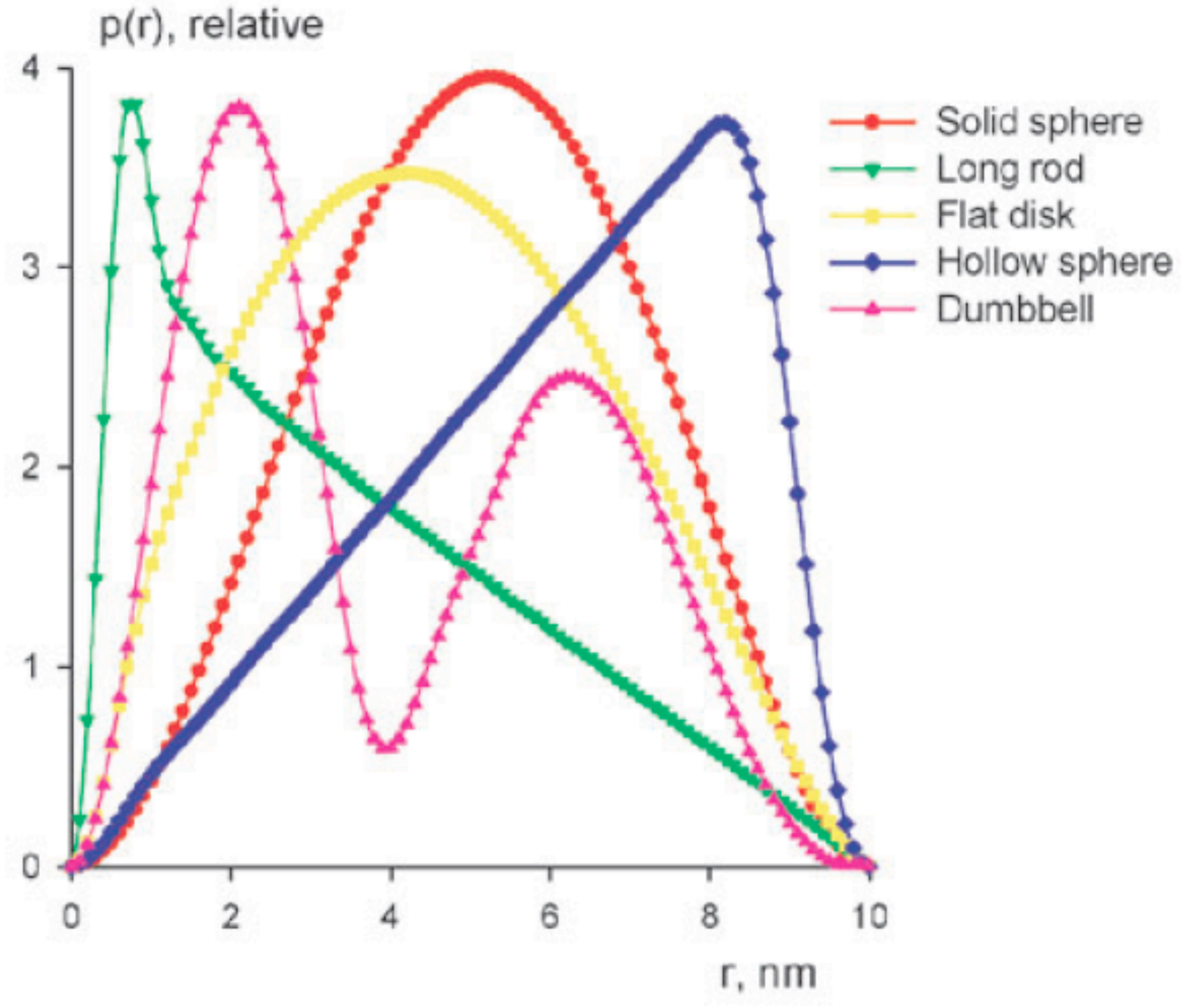
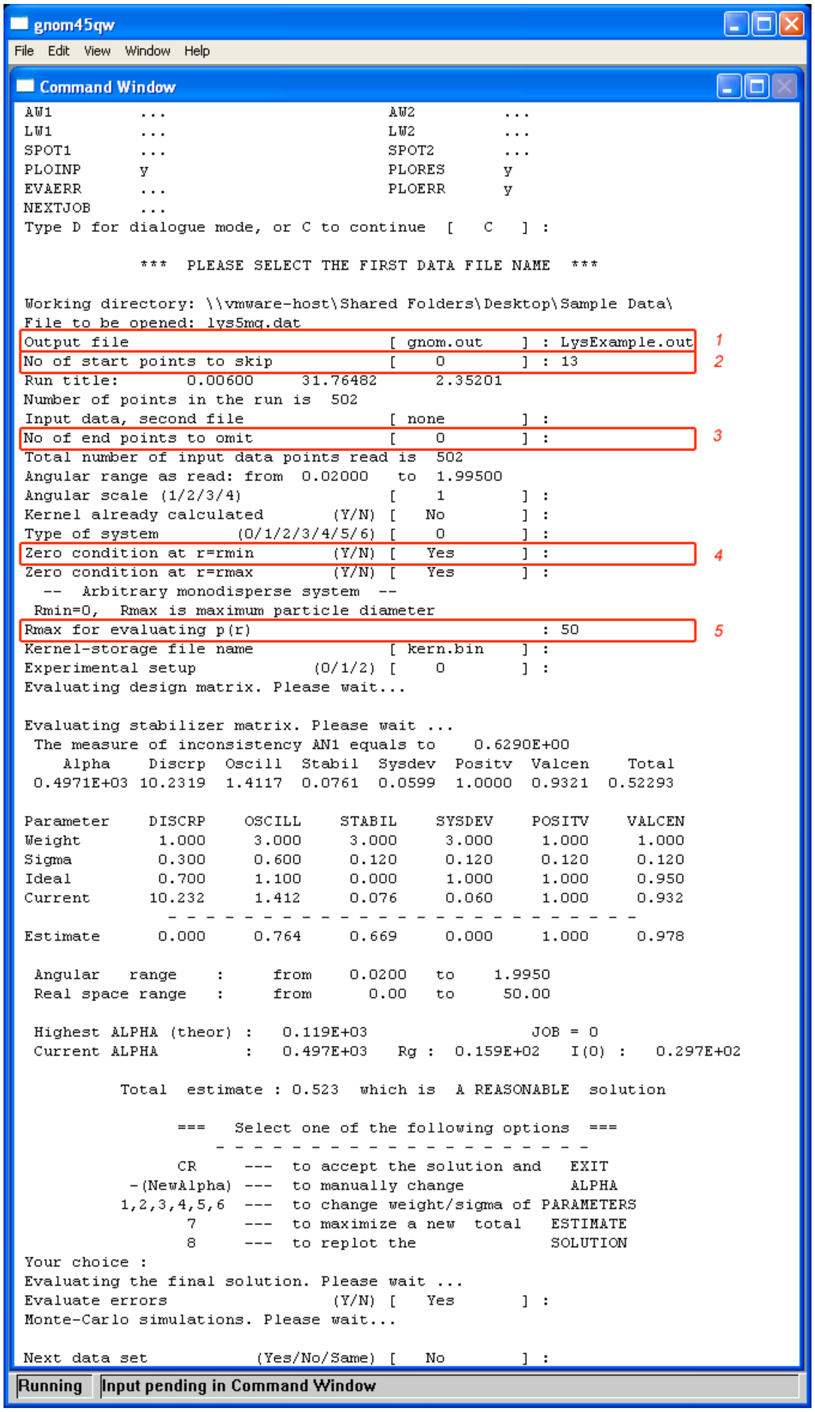


Figure 14: Ideal P(r) distributions colored according to protein shape. For more information on identifying the correct shape for a P(r) distribution see text.

Figure taken from Svergun, D.I. and M.H.J. Koch, Rep Prog Phys, 2003 66:1735-82 information that can be obtained from the P(r) distribution. The general overall structure of a protein can be determined by the shape of the P(r) distribution [5]. Large conformational variations appear as distinct graph patterns on the plot. Figure 14 shows ideal curves for each protein shape. For the purposes of this guide we will be preparing data for a solid sphere. All descriptions of ideal graphs will be in reference to this. In addition to structural information, both the Rg and I(0) can be calculated from the P(r) distribution. The advantage to this is that it includes the entire data range instead of the limited range used by Guinier plots. This is commonly called a ‘real space’ approximation and may be more accurate. However, there is a potential source of error in both the structural calculations and the real space approximations for I(0) and Rg. When generating P(r) distributions with real data the scattering intensity is not directly Fourier transformed to form the plot [1,3,4,5]. Instead, it is often an indirect Fourier transformation based on a user entered value, Dmax. Dmax represents the longest linear distance across a protein, or in simpler terms the maximum width of the protein. In practical terms, Dmax is the point where P(r)=0 when r > 0. P(r) distributions are constrained to zero at r=0 and then Dmax assigned by the user. A graphical example of this can be seen in Figure 13. If a sample is properly buffer subtracted and an appropriate Dmax is chosen, a P(r) can give accurate real space approximations and be used to generate reliable 3D models. However, when inappropriate values of Dmax are used, or the sample quality is low the P(r) distribution can provide inaccurate information. When samples are aggregated the P(r) distribution will not approach zero in a smooth way. Assigning an artificially high Dmax (7-8*Rg) can provide some useful information. Under these conditions the P(r) distribution may have several peaks at high r values and never smoothly approach zero [4]. While failure to reach zero can identify aggregated samples, approaching zero at artificially small values can also provide some useful information [1,2]. If the P(r) reaches zero abruptly it most likely indicates that the chosen Dmax is too small and should be increased [4]. If the plot approaches zero in a smooth way (as seen in Figure 12) at artificially small values (below 3-4*Rg) it can indicate interparticle interference [4]. Often you can determine that a Dmax was not correct based on the approximations that are generated from it. If the Rg value is not relatively close to the Rg from the autoRg value then the Dmax is inappropriate. Many times with samples that are partially or fully unfolded there is a large degree of variability in the P(r) distribution. This results in several Dmax values that can produce good looking P(r) plots but are not necessarily accurate. One way to help prevent confusion is to use the Guinier approximations for I(0) and Rg as a guide. Details of how to do this are be outlined in section 4.2. Samples that are partially aggregated can be more reliably evaluated using P(r) distributions because aggregation has a larger effect on the small range of data used by Guinier plots.

4.2 Generating P(r) Distributions in GNOM

After completing the initial data analysis in PRIMUS, the second stage of data processing occurs in the program GNOM [14]. This program generates P(r) distributions which are saved as output files for use in modeling software. GNOM does not have a full graphic interface and it is operated by a series of text prompts. This guide will not cover every prompt in detail, instead, only those prompts that are required to generate a functional P(r) distribution and output file will be described. *If a prompt is not described*



```
 AW1          ...                         AW2          ...
 LW1          ...                         LW2          ...
 SPOT1        ...                         SPOT2        ...
 PLOINP       y                           PLORES       y
 EVAERR       ...                         PLOERR       y
 NEXTJOB      ...
 Type D for dialogue mode, or C to continue  [   C    ] :

             ***  PLEASE SELECT THE FIRST DATA FILE NAME  ***

 Working directory: \\vmware-host\Shared Folders\Desktop\Sample Data\
 File to be opened: lys5mg.dat
 Output file                               [ gnom.out    ] : LysExample.out
 No of start points to skip                [     0       ] : 13
 Run title:        0.00600     31.76482       2.35201
 Number of points in the run is  502
 Input data, second file                   [ none        ] :
 No of end points to omit                  [     0       ] :
 Total number of input data points read is  502
 Angular range as read: from  0.02000   to  1.99500
 Angular scale (1/2/3/4)                   [     1       ] :
 Kernel already calculated           (Y/N) [   No        ] :
 Type of system            (0/1/2/3/4/5/6) [     0       ] :
 Zero condition at r=rmin            (Y/N) [   Yes       ] :
 Zero condition at r=rmax            (Y/N) [   Yes       ] :
   --  Arbitrary monodisperse system  --
  Rmin=0,  Rmax is maximum particle diameter
 Rmax for evaluating p(r)                                  : 50
 Kernel-storage file name                  [ kern.bin    ] :
 Experimental setup                (0/1/2) [     0       ] :
 Evaluating design matrix. Please wait...

 Evaluating stabilizer matrix. Please wait ...
  The measure of inconsistency AN1 equals to    0.6290E+00
     Alpha     Discrp  Oscill  Stabil  Sysdev  Positv  Valcen    Total
  0.4971E+03 10.2319  1.4117  0.0761  0.0599  1.0000  0.9321  0.52293

 Parameter     DISCRP    OSCILL    STABIL    SYSDEV    POSITV    VALCEN
 Weight         1.000     3.000     3.000     3.000     1.000     1.000
 Sigma          0.300     0.600     0.120     0.120     0.120     0.120
 Ideal          0.700     1.100     0.000     1.000     1.000     0.950
 Current       10.232     1.412     0.076     0.060     1.000     0.932
               - - - - - - - - - - - - - - - - - - - - - - - - - -
 Estimate       0.000     0.764     0.669     0.000     1.000     0.978

  Angular    range    :     from    0.0200   to    1.9950
  Real space range    :     from      0.00   to     50.00

  Highest ALPHA (theor) :    0.119E+03                JOB = 0
  Current ALPHA         :    0.497E+03    Rg :  0.159E+02   I(0)  :   0.297E+02

           Total  estimate : 0.523  which is  A REASONABLE  solution

                 ===   Select one of the following options  ===
                     - - - - - - - - - - - - - - - - - - - -
                 CR      ---  to accept the solution and   EXIT
            -(NewAlpha)  ---  to manually change           ALPHA
           1,2,3,4,5,6   ---  to change weight/sigma of PARAMETERS
                  7      ---  to maximize a new  total    ESTIMATE
                  8      ---  to replot the               SOLUTION
 Your choice :
 Evaluating the final solution. Please wait ...
 Evaluate errors                     (Y/N) [   Yes       ] :
 Monte-Carlo simulations. Please wait...

 Next data set               (Yes/No/Same) [   No        ] :
```

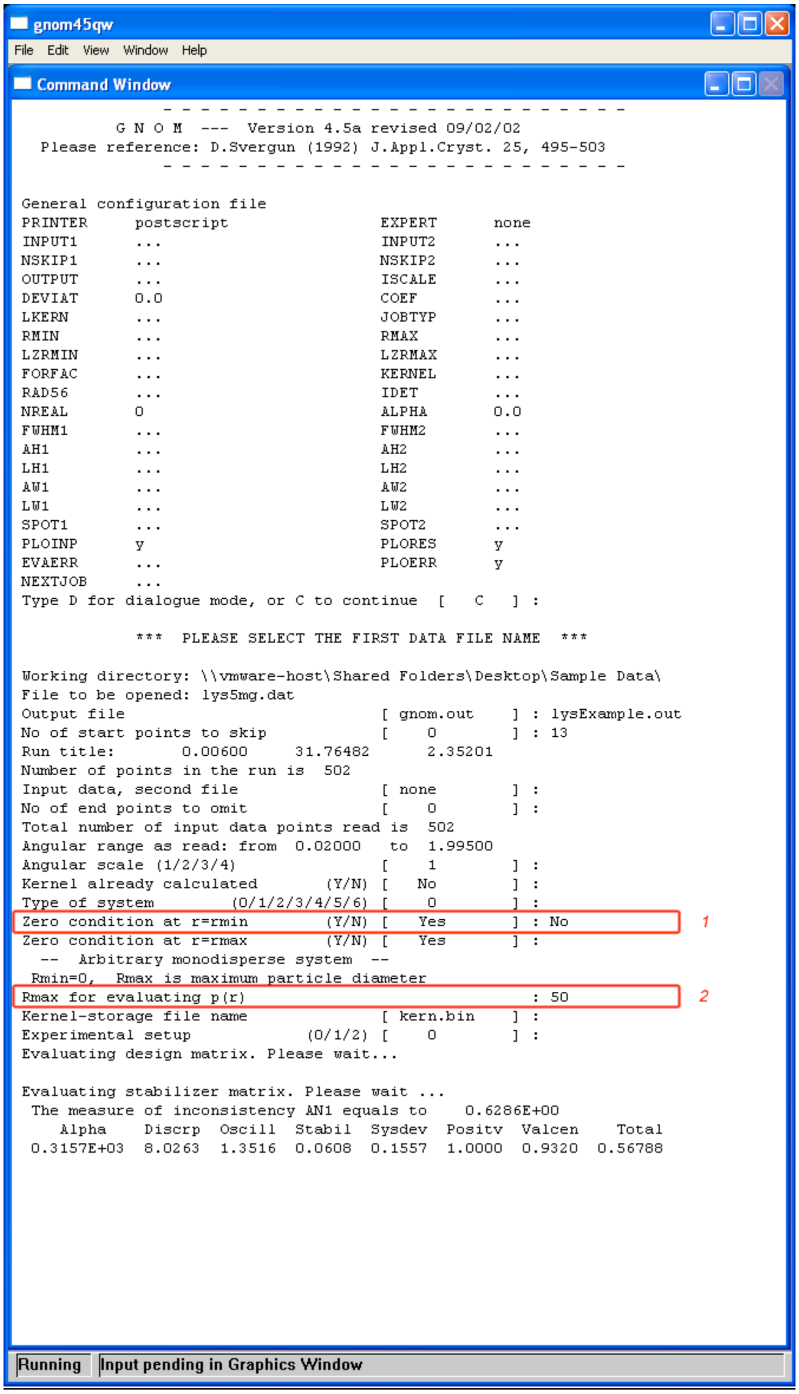
Figure 15: General GNOM walkthrough. **Line 1:** Name the output file. **Line 2:** Designate the number of start points to skip according to the AutoRg results. **Line 3:** Enter the amount of points you wish to truncate from the end of the data set. **Line 4:** Constrain the graph to zero at r=0. Removing this restriction is used to check the buffer subtraction. **Line 5:** Enter a Rmax to evaluate the graph at. This value should be 3-4*Rg. For more detailed descriptions of the GNOM prompts see text.

in this guide then the default entry will be sufficient for most data sets. To learn more about some of the advanced options, please refer to the GNOM manual provided on the EMBL website (http://www.embl-hamburg.de/biosaxs/software.htm). There you will find a detailed description of each prompt and the choices you are given.

Generating a P(r) distribution can be a somewhat difficult task. Often data can be flexible enough to accommodate a number of distinct results. Determining which P(r) distribution will work best is often a series of trial and error. Many different conditions will must considered before you decide which one is the best option. In many cases it is wise to generate several output files from GNOM. All of these conditions can be processed into structures. The conditions that produce the best structure can be refined to give more detail. Before you begin trying to generate a P(r) distribution, you should first run some simple conditions to help further evaluate the quality of the data. I will discuss each of these initial runs in detail below, but first we will walk through the prompts to help you familiarize yourself with how the program works.

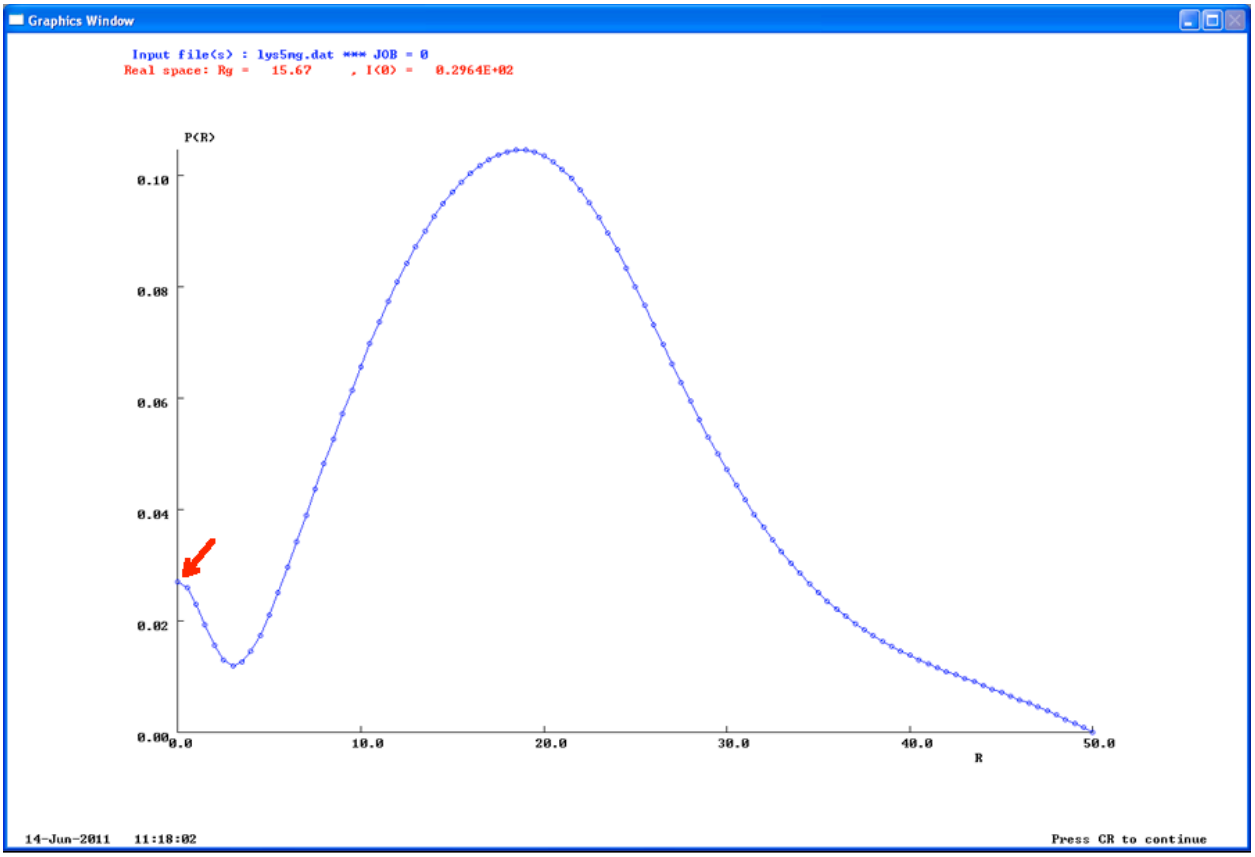
General Walkthrough

Upon opening the program press return to open a data file. After selecting your data file you will be prompted to provide a name for your output file (Line 1: Figure 15). Please note that for the file to be given the proper filename extension you must include ".out" after your filename. After naming your output file, you will be prompted to skip start points (Line 2: Figure 15). You should enter the points which were excluded by AutoRg for not obeying Guinier's law. Next you will be prompted to enter a second data file, this can be skipped unless you need to manually merge two data files. The following prompt asks how many endpoints you would like to omit (Line 3: Figure 15). In many cases, especially when dealing with SAXS data, it will be necessary to omit some of the endpoints which are badly distorted by noise. In general, the less data you remove the better, but including bad data can distort your results. There is no hard and fast rule for data truncation, normally it is a matter of guess and check. You will truncate points and evaluate the resulting P(r) distribution. The next prompt is for the angular scale. This refers to the units used to describe the scattering data. This value is determined at the time of data collection. If your data was collected by a collaborator and sent to you, or completed by a "mail-in" program at a national beamline be sure to ask what form the data was collected in. Often your sample will be returned with a form that describes hte file names and other general information. These sheets normally describe the angular units the data were collected in. For many beamlines the standard scale will be in angstroms, and option 1 (default) will be sufficient. The details of each of these options can be found in the GNOM manual (http://www.embl-hamburg.de/biosaxs/software.htm). The "kernel already calculated" and "type of system" prompts should be kept at the default settings of 'no' and '0' respectively. The next two prompts concern assigning r=0 at the minimum and maximum of the P(r) distribution (Line 4: Figure 15). When you are generating P(r) distributions these settings should be set to yes, however, there are some data quality tests that involve removing these restrictions. Specifically, removing the r=0 at 0 restriction will allow the user to evaluate if the buffer subtraction has been done correctly.



```
               - - - - - - - - - - - - - - - - - - - - - - - - -
          G N O M  ---  Version 4.5a revised 09/02/02
  Please reference: D.Svergun (1992) J.Appl.Cryst. 25, 495-503
               - - - - - - - - - - - - - - - - - - - - - - - - -

General configuration file
PRINTER     postscript                EXPERT      none
INPUT1      ...                       INPUT2      ...
NSKIP1      ...                       NSKIP2      ...
OUTPUT      ...                       ISCALE      ...
DEVIAT      0.0                       COEF        ...
LKERN       ...                       JOBTYP      ...
RMIN        ...                       RMAX        ...
LZRMIN      ...                       LZRMAX      ...
FORFAC      ...                       KERNEL      ...
RAD56       ...                       IDET        ...
NREAL       0                         ALPHA       0.0
FWHM1       ...                       FWHM2       ...
AH1         ...                       AH2         ...
LH1         ...                       LH2         ...
AW1         ...                       AW2         ...
LW1         ...                       LW2         ...
SPOT1       ...                       SPOT2       ...
PLOINP      y                         PLORES      y
EVAERR      ...                       PLOERR      y
NEXTJOB     ...
Type D for dialogue mode, or C to continue  [   C   ] :

            ***  PLEASE SELECT THE FIRST DATA FILE NAME  ***

Working directory: \\vmware-host\Shared Folders\Desktop\Sample Data\
File to be opened: lys5mg.dat
Output file                           [ gnom.out    ] : lysExample.out
No of start points to skip            [    0        ] : 13
Run title:       0.00600     31.76482      2.35201
Number of points in the run is  502
Input data, second file               [ none        ] :
No of end points to omit              [    0        ] :
Total number of input data points read is  502
Angular range as read: from  0.02000   to  1.99500
Angular scale (1/2/3/4)               [    1        ] :
Kernel already calculated       (Y/N) [   No        ] :
Type of system          (0/1/2/3/4/5/6) [    0        ] :
Zero condition at r=rmin        (Y/N) [   Yes       ] : No
Zero condition at r=rmax        (Y/N) [   Yes       ] :
  --  Arbitrary monodisperse system  --
 Rmin=0,  Rmax is maximum particle diameter
Rmax for evaluating p(r)                              : 50
Kernel-storage file name              [ kern.bin    ] :
Experimental setup            (0/1/2) [    0        ] :
Evaluating design matrix. Please wait...

Evaluating stabilizer matrix. Please wait ...
 The measure of inconsistency AN1 equals to    0.6286E+00
    Alpha    Discrp  Oscill  Stabil  Sysdev  Positv  Valcen    Total
 0.3157E+03  8.0263  1.3516  0.0608  0.1557  1.0000  0.9320  0.56788
```

Figure 16: Evaluating buffer subtraction with GNOM. **Line 1:** Remove the zero constraint at r=0. Data with correct buffer subtraction will remain at zero without this constraint. **Line 2:** Evaluate the data at 3-4*Rg as described in the text.



```
Input file(s) : lys5ng.dat *** JOB = 0
Real space: Rg =   15.67     , I(0) =   0.2964E+02
```

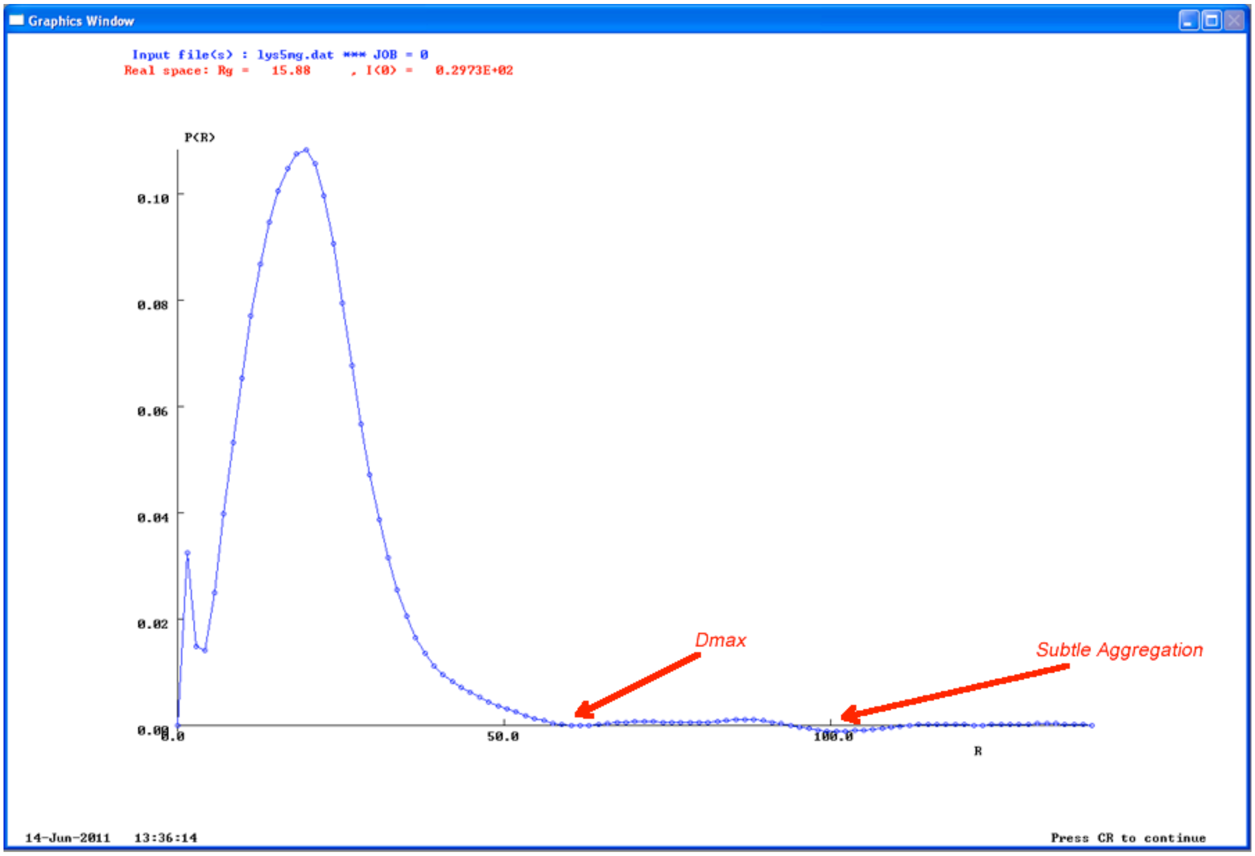
Figure 17: This figure shows the P(r) distribution generated from Figure 16. The buffer subtraction on this sample is insufficient. This can be seen from the peak at r=0. For a more detailed description of evaluating buffer subtraction, see text.

Checking Buffer Subtraction

One of the first things you should do when working with GNOM is check to see if the buffer subtraction was done correctly. Incorrect buffer subtraction can result in strange features in P(r) distributions. Load your data and follow the prompts as you normally would, but change the setting on the Rmin=0 to ‘no’ (Line 1: Figure 17). This will no longer constrain your data to zero at q=0. Set the Rmax value to about 3-4*Rg (Line 2: Figure 16). This initial approximation is just to determine buffer subtraction, so the Rmax value does not need to be optimized. A properly buffer subtracted sample should be very close to zero at q=0 with no restrictions. Deviations from this indicate that the buffer was either over or under subtracted. If your P(r) distribution starts above zero then the buffer was not sufficiently subtracted, and if it is below zero then the buffer was over subtracted [4]. When these errors are small it will not prevent further data analysis, but it should be noted and corrected in future experiments. The data shown here has insufficient buffer subtraction to demonstrate this common occurrence and how it will effect your plots during data analysis (Figure 17).

Checking Aggregation

At this point in the data analysis you have already done some initial data quality analysis using PRIMUS. GNOM allows you to further characterize the quality of your data using P(r) distributions. Aggregation is extremely important to identify. It can have a significant impact on your data analysis and heavily aggregated samples should not be used to draw structural conclusions. The specifics of identifying aggregation and interparticle interference are detailed more in the introductory portion of this guide. To



```
Input file(s) : lys5mg.dat *** JOB = 0
Real space: Rg =   15.88      , I(0) =   0.2973E+02
```

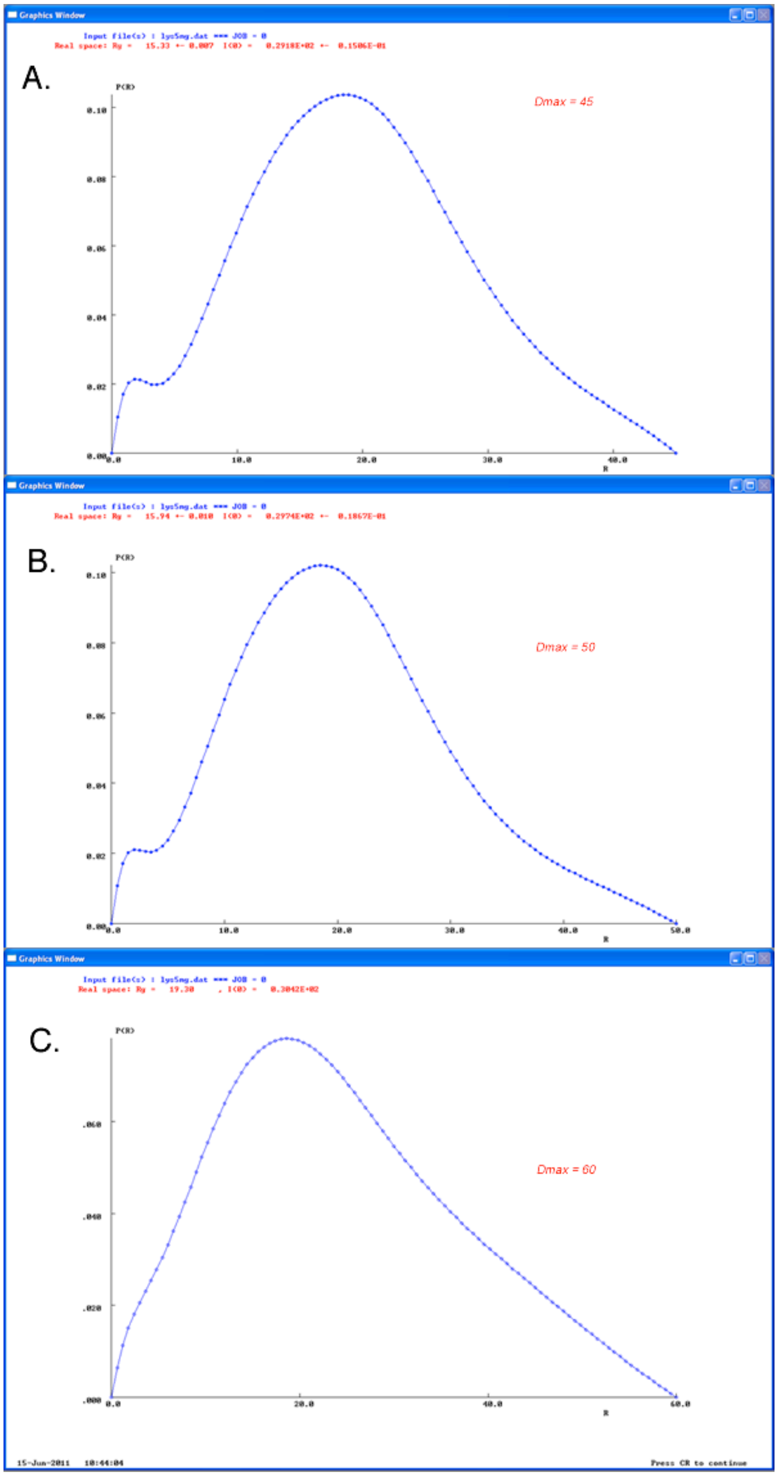
```
14-Jun-2011    13:36:14                                   Press CR to continue
```

Figure 18: Generating a P(r) distribution with a large Rmax can help to evaluate the aggregation in a sample. The true Rmax will be the point where the graph approaches zero in a smooth (low slope) way. Signs of aggregation will appear as additional peaks beyond the Rmax point. Well behaved, non-aggregated data will stay at zero after the Rmax.

begin, load your data and proceed through the prompts as normal. When you reach the Rmax=0 change this option to no, then when choosing an Dmax choose a larger than normal value so you can evaluate changes over a longer range. A good starting range for this is about 7*Rg. The resulting P(r) distributions should curve down to zero with a smooth profile (Figure 18). If the plot never reaches zero, or if there are multiple peaks at high r values this can indicate aggregation. In Figure 17, the plot returns to zero and stays fairly constant. There are some minor deviations, which could indicate subtle aggregation. The place where the plot reaches zero indicates Rmax (Dmax). This number may not prove to be the best value for generating a P(r) distribution, but it should give you a better starting point. Again please see the introductory text on P(r) distributions and the reviews cited in this guide for more detailed explanations.

P(r) Distributions

One of the most important parts of SAXS data analysis is generating a reliable P (r) distribution. This process is normally characterized by a significant amount of guess and check. The two major parameters you will have to decide on are the amount of data to truncate, and the Dmax value (Rmax). The first step is to load your data and name your output file. When you are removing points from the beginning of the data set, you should start with those that were excluded from the low q ranges in the Guinier plot. For example, if your AutoRg used points 14-50 you should exclude the first 13 points. After declining to load a second data set you will be asked to remove points from the end of the data set. For now don't remove any data, the first concern will be to determine where the right Dmax value lies. After you are in the ball park you can come back and start tweaking with data truncation. The next prompt which requires a non default option is the Rmax value. This value determines the maximum linear distance across your protein and can have the largest impact on the profile of the P(r) distribution. There are no steadfast rules in picking a Rmax, but a good place to start is to multiply the Rg value obtained from the Guinier plots by 3 or 4. It doesn't matter if you start high or low, but you should check a variety of values to be sure you are using the best one. Now come the most subjective portion of the entire data analysis; you need to look at the P(r) distribution and decide if it is appropriate for your protein. To begin, you should look back at Figure 14 for the ideal P(r) distribution for your protein's shape. Next, you should look at the P(r) distribution you just generated, specifically the Rmax value. The plot should approach zero in a smooth concave line (Figure 18-B). If your plot approaches zero with a very steep angle the Rmax is probably too small. Try replotting with a larger Rmax value. If your plot reaches zero before the Rmax value try decreasing the Rmax. It is important to identify features which are abnormal for your protein's shape. For example, if you know your protein is globular but the P(r) distribution shows a second peak at high r values, which is characteristic of a dumbbell shaped protein, then you know your plot is not correct under the current conditions. If you generate a plot which is inappropriate for your sample, the structure be inaccurate. In the example of a globular protein with a dumbbell shaped P(r) distribution, the structural mapping software would try and add a dumbbell shape. This results in a "chicken bone" effect where a large region protrudes from the bulk of the protein to try and accommodate the data from the P(r) distribution. When you are adjusting the Rmax value try small increments. For example, if you started with a Rmax of 45 (Figure 18-A),



```
Input file(s) : lys5mg.dat === JOB = 0
Real space: Rg =   15.33 +- 0.007  I(0) =   0.2918E+02 +-  0.1506E-01
```

```
Input file(s) : lys5mg.dat === JOB = 0
Real space: Rg =   15.94 +- 0.010  I(0) =   0.2974E+02 +-  0.1867E-01
```

```
Input file(s) : lys5mg.dat === JOB = 0
Real space: Rg =   19.30     , I(0) =   0.3042E+02
```

```
15-Jun-2011   10:44:04
```

Figure 18: Generating a suitable P(r) distribution. **A.** Rmax of 45 gives a plot that approaches zero with in a semi-smooth slope, but the slope is still too high and there is an artificial curve to zero close to Rmax. **B.** Rmax of 50 gives a very smooth curve that looks very close to the ideal for the protein. This is the best solution of the three shown. **C.** Rmax of 60 gives a broad peak with a high slope and no real curve to zero.

try 50 (Figure 18-B). If that plot still has a steep slope as it approaches zero, try 60 (Figure 18-C). If the plot looks good, try working backwards by increments of 1. Once you find the Rmax that looks best you can try to change it by fractions (i.e. 46.1). The goal is to find a plot that looks good and gives appropriate results. Once you have a plot which has the general features expected for your protein and approaches Rmax in a smooth concave fashion then you can begin to focus on trimming data. I generally begin with increments of 50 data points. The goal is to find the least amount of data that you can cut to generate a good looking P(r) distribution. Once you find a good range, you can try and do small refinements to the Rmax to optimize the graph. When you think you have generated your final P(r) distribution, make sure that the Rg and I(0) values are similar to the ones generated by the Guinier plot. You can expect the values to be different slightly. The P(r) distribution is a real space approximation which uses the entire data set, while the Guinier plot only uses a small portion of the data. If you have several P(r) distributions which look good, try saving them as different file types and generating structures from all of them.

Chapter 5

Generating Structural Models

There are many software programs for molecular modeling that can generate 3D structures from SAXS data. The ATSAS software package includes several programs that can be used to obtain structures from SAXS data. The specific approach to modeling varies between software packages, and there are certainly advantages and disadvantages to each modeling technique. For the purposes of this guide we will focus our attention on software that uses a common, easy to understand technique known as *ab initio* modeling. The two most commonly used programs for *ab initio* modeling are DAMMIN and DAMMIF [4,5].

5.1 Overview of Molecular Modeling

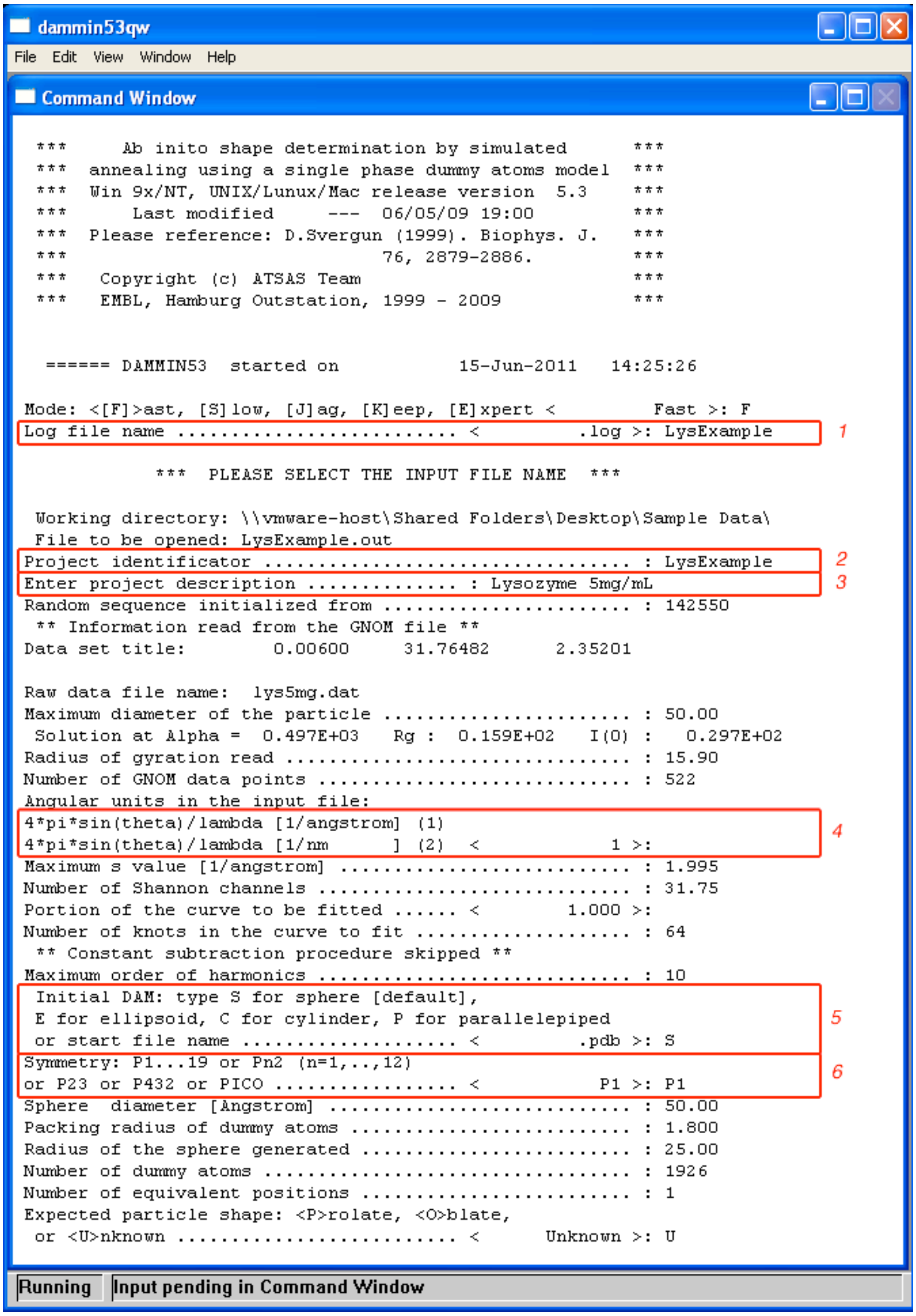
SAS data processing can be very subjective, and generating a 3D structure from a one dimensional scattering profile leaves a lot of room for error. Most advanced *ab initio* modeling software relies on a technique known as automated bead-modeling [5]. This process is very complex and the description here is only meant to give you a rough understanding. In simple terms, the modeling software represents the protein as a collection of tightly packed beads inside a container. The size of the container represents the maximum volume of the protein [5]. In DAMMIN, the volume of the protein is determined by the Rmax value you assigned in GNOM. Using a process known as simulated annealing the software determines the 3D structure by trying to minimize the overall error value. Initially particles (beads) are assigned at random to be either protein or solvent [5]. The software uses this initial approximation of the protein structure as template to refine the structure. Working with one bead at a time, the software reverses the original random assignment. If the change in particle assignment reduces the overall error value then the program keeps it, otherwise, it reverts to the assignment from the initial approximation [5]. To eliminate unrealistic solutions and help avoid obvious errors, the software requires the beads to be connected to one another and the overall structure must be compact. For more information on automated bead modeling please refer to the review by Haydyn and Svergun [5].

5.2 DAMMIN & DAMMIF

DAMMIN (Dummy Atom Model Minimization) and the newer edition DAMMIF (the F stands for fast) are both included in the ATSAS software package [15,16]. The most significant difference between the two programs is how each constrains the protein volume. DAMMIN defines the volume from the Rmax value used to generate the GNOM output file. DAMMIF does not constrain the volume during simulated annealing. This allows the protein to be mapped outside of the constraints of Rmax, and can be helpful when the Rmax value has been slightly underestimated [5]. In practical terms, the use of the two programs is similar but not identical. DAMMIN allows for many more prompt entries, while DAMMIF is relatively simple. If you understand how to use DAMMIN then using DAMMIF is very easy, for this reason we will use DAMMIN for the examples in this guide.

5.3 Generating PDB Models from GNOM Output Files

DAMMIN has a similar layout to GNOM, and begins with a prompt asking to name the log file (Figure 19). Be sure that your name is unique to this sample and the output file you are using. As a personal preference I give the log files the same name as the GNOM output file I am using so that I do not have confusion. Next, you will be prompted to load your .out file from GNOM and name the output files from DAMMIN. You will then be able to enter a short description of the data. Normally, I use this to record the protein name and the concentration of the sample. This information will be stored in the log file, which is useful for looking back to determine exactly what was used to generate a structure. The next prompt will ask you to identify the angular units the data was collected in. It is important to find out this information about your data. For a more detailed description of identify the angular units of your data please refer to Chapter 3. The default units are angstroms, and under normal circumstances SAXS data will be collected in this format. You have already been asked to enter this data in GNOM. If you find out that your units were not what you thought, you must also redo the output file to reflect the correct units. The next non-default prompt will ask you to define the starting shape of the molecule. The default here is a sphere, and if you do not know the overall shape of your protein, this option will be the most forgiving. There are several other shape options that you can try if you think your protein shape would be better suited with them. For more information on these shapes, please consult the DAMMIN manual on the EBML website. The next step will be to predict symmetry. This is a step that should be taken cautiously. If you know for certain that your protein is a dimer, then assigning a P2 symmetry will yield much better structural results. As with



```
 ***     Ab inito shape determination by simulated      ***
 ***  annealing using a single phase dummy atoms model  ***
 ***  Win 9x/NT, UNIX/Lunux/Mac release version  5.3    ***
 ***      Last modified      ---  06/05/09 19:00        ***
 ***  Please reference: D.Svergun (1999). Biophys. J.   ***
 ***                       76, 2879-2886.               ***
 ***   Copyright (c) ATSAS Team                         ***
 ***   EMBL, Hamburg Outstation, 1999 - 2009            ***


  ====== DAMMIN53  started on           15-Jun-2011   14:25:26

Mode: <[F]>ast, [S]low, [J]ag, [K]eep, [E]xpert <         Fast >: F
Log file name .......................... <          .log >: LysExample

            ***  PLEASE SELECT THE INPUT FILE NAME  ***

 Working directory: \\vmware-host\Shared Folders\Desktop\Sample Data\
 File to be opened: LysExample.out
Project identificator ................................. : LysExample
Enter project description .............. : Lysozyme 5mg/mL
Random sequence initialized from ...................... : 142550
 ** Information read from the GNOM file **
Data set title:         0.00600      31.76482      2.35201

Raw data file name:  lys5mg.dat
Maximum diameter of the particle ...................... : 50.00
 Solution at Alpha =  0.497E+03    Rg :  0.159E+02   I(0) :   0.297E+02
Radius of gyration read ............................... : 15.90
Number of GNOM data points ............................ : 522
Angular units in the input file:
4*pi*sin(theta)/lambda [1/angstrom]  (1)
4*pi*sin(theta)/lambda [1/nm      ]  (2)   <             1 >:
Maximum s value [1/angstrom] .......................... : 1.995
Number of Shannon channels ............................ : 31.75
Portion of the curve to be fitted ...... <         1.000 >:
Number of knots in the curve to fit ................... : 64
 ** Constant subtraction procedure skipped **
Maximum order of harmonics ............................ : 10
 Initial DAM: type S for sphere [default],
 E for ellipsoid, C for cylinder, P for parallelepiped
 or start file name ................... <          .pdb >: S
Symmetry: P1...19 or Pn2  (n=1,..,12)
or P23 or P432 or PICO ................. <            P1 >: P1
Sphere  diameter [Angstrom] ........................... : 50.00
Packing radius of dummy atoms ......................... : 1.800
Radius of the sphere generated ........................ : 25.00
Number of dummy atoms ................................. : 1926
Number of equivalent positions ........................ : 1
Expected particle shape: <P>rolate, <O>blate,
 or <U>nknown .......................... <      Unknown >: U
```

Figure 20: General DAMMIN walkthrough. **Line 1-3:** Name the output file and enter a project description for the data set being used. **Line 4**: Identify the angular units that the data set was collected in. This must be the same response used for GNOM. **Line 5:** Select the general shape of the protein if known. If unknown use the default sphere setting. **Line 6:** If your protein has internal symmetry select it here. If symmetry is unknown then use the 'P1' setting for an initial approximation. Detailed descriptions of each prompt can be found in the text.

any parameterization, if you are unsure, you should not include a symmetry assignment at first. If the resulting structure appears to have symmetry, you can then go back and try to redo the structure with a symmetry assignment. After selecting unknown for the expected particle shape (default) the program will begin the simulated annealing. This process can take anywhere from 10 minuets to several hours depending on the speed of the computer it is being run on. To ensure that it doesn't take excessively long, I suggest not using the computer for other things while the annealing process is running. After the program is finished you will be left with two pdb files. One will contain a valid structure and the other will contain only a few points. In general, the ###1.pdb will contain the correct file. To view these structures you can use any molecular modeling software. In this guide we will discuss how to view your model in the free software Chimera [17].

5.4 DAMAVER

SAS data processing can be very subjective, and generating a 3D structure from a one dimensional scattering profile leaves alot of room for error. In general, automated bead modeling is a fairly accurate way to convert 1D data into 3D structures, but the process is not without problems. The annealing process determines the structural envelope by trying to minimize the overall error value. When there are solutions with similar error values, the software will not favor one over another. As a result, the software will always produce multiple solutions from a single set of parameters. The solutions that are generated will begin to fall into distinct classes based on the selections the software made. Often times, some of the structural classes will be obviously incorrect and can be discarded. If you only generate two major classes, and one is obviously incorrect then your job will be easy. However, in almost all cases there will be several classes of solutions that could be accurate. To determine the most likely solution we will use DAMAVER to average the most common solution classes [18]. DAMAVER is actually a collection of programs which can be used to identify the major solution class, eliminate outliers, align structures, and generate an average model. This guide will not provide a step-by-step walkthrough of DAMAVER, but we will discuss in detail how to prepare to use the software.

To ensure that you are using the most appropriate structures for your analysis, you should generate many structures with many different P(r) variations. Once you find the structure conditions that result in the best representative structures you should generate that structure with identical conditions at least 10 separate times. This is an important safeguard against the simulated annealing procedure. Using a DAMAVER averaged structure minimizes the risk of a random artifact in the annealing process impacting the final structure. Determining which structural classes are significant and which are artifacts from the annealing process is not always possible, but in some cases some obviously incorrect solutions can be eliminated.

Automated bead modeling solutions from DAMMIN must be a compact string of connected particles with a low error value. However, neither DAMMIN or DAMAVER require the solutions to to be chemically or biologically possible. The accuracy of your averaged model is directly related to the quality of the data you use. DAMAVER uses statistics to determine which classes are relevant. If an chemically impossible solution is one of the most common results, DAMAVER will include it in the final model. Evaluating if a structure is chemically possible can help to eliminate additional classes. Superimposing an accurate crystal structure of your protein in a similar state can help determine any structural differences. Remember that structures generated from crystals will be in a low energy confirmation. SAS measures the structure of the protein in solution, so you should expect to see movement (large regions of density in the structure) around dynamic regions. After removing the obviously incorrect classes, the next step is to find the best solution by running DAMAVER. A more detailed explanation of DAMAVER can be found in the online manuel at the EMBL website.

Chapter 6

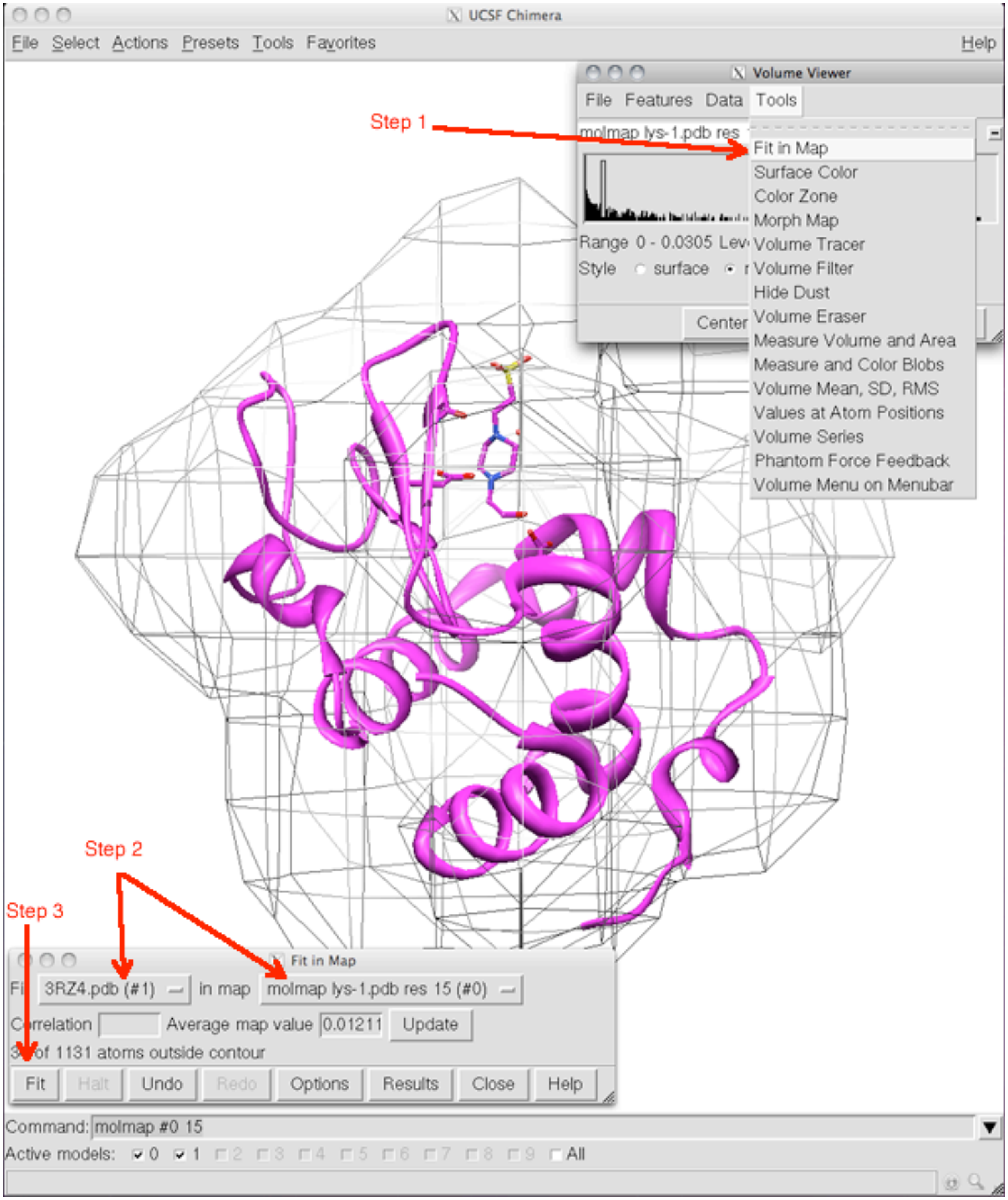
Visualizing SAXS Structures

6.1 Chimera

After you have generated a structure, the next step will be to visualize that structure and compare it with any known crystal structures. This can be done in most molecular modeling software. Chimera is freely available through the University of California, San Francisco (http://www.cgl.ucsf.edu/chimera/) [17]. This program is capable of viewing and editing molecular models, as well as creating animations. The major advantage of this software, in the context of this guide, is the ability to easily view SAXS envelopes. It is possible to do the same thing we describe here with other software, and if you have modeling program that you are familiar with feel free to use it instead. Chimera is shown here as an example because the steps to visualizing SAXS envelopes are simple and straightforward.

6.2 Visualizing SAXS Models in Chimera

When opening Chimera the first thing you need to do is load your pdb file. This can be done by selecting File --> Open and choosing the file you want to load. When you first load your structure, it should just appear as a series of lines which do not resemble normal protein structure. To view the correct structure you need to hide the ribbon model and display the envelope. To make these changes you need to select the chains



molmap lys-1.pdb res

Range 0 - 0.0305 Lev

Style ○ surface ● r

Fit 3RZ4.pdb (#1) in map molmap lys-1.pdb res 15 (#0)

Correlation Average map value 0.01211

3 of 1131 atoms outside contour

Command: molmap #0 15

Figure 21: Fitting a crystal structure from the PDB into a SAXS envelope. **Step 1:** After you have generated the envelope, load the pdb you want to fit. Use the Volume Viewer menu to select Fit in Map. **Step 2:** Using the Fit in Map menu select the pdb as #1 and the molmap as #2. **Step 3:** Click the fit button to automatically fit the two.

so you can change the display properties. To select protein chains: click Select --> Chain --> **. Once you have the chain selected you will want to turn off the ribbon and Atom/Bonds displays. To do this: click Actions --> Atoms/Bonds --> hide, then do the same for the ribbon. At this point your protein should be completely hidden from view. The remaining commands have to be entered directly into the program's command line. First, you must display the command line if it is not already visible at the bottom of the screen. To display the command line, select Tools --> General Commands --> Command Line. After you have done this, a command line will appear at the bottom of the screen. Click on the command line, and enter the following text: 'molmap #0 15'. This will display the envelope of the protein at a 15 angstrom resolution.

After generating the envelope it is common to fit a crystal structure into the SAXS envelope. To do this, load the pdb file for the crystal structure with the pdb generated by DAMMIN. Select Tool --> Fit in Map from the Volume Viewer menu (Step 1: Figure 21). This will bring up a second smaller window. Click on the box to the left of the word Fit (Step 2: Figure 21). Select the pdb of the crystal structure, then click on the second box after the words "in the map" and select the molmap of your DAMMIN structure. Finally, click the Fit box in the lower left hand corner of the menu (Step 3: Figure 21) to fit the two structures together. The software will also provide you with a numerical fit value that represents how well the two structures match. Now you can visually evaluate how well the SAXS map represents the crystal structure.

Final Thoughts

Small angle scattering is a very diverse field. There are numerous techniques that accomplish a large variety of tasks, and the applications for those techniques span across many areas of research. The more traditional aspects of SAXS and WAXS are useful to structural studies, but I don't believe they are the most useful aspects of small angle scattering. I personally think that the new WAXS based techniques that are just beginning to emerge will prove to be the most important in the long run. Traditional SAXS/WAXS experiments can quickly provide useful information, but the information can be limited to low resolution models and basic structural information. Also, this information is often available through other means. The new WAXS based techniques, such as TR-WAXS, allow for studies which were not previously possible. Structural envelopes from WAXS data are becoming more and more defined. The ability to monitor conformational changes on a very small time scale, in solution, with no limits on protein size, and under a variety of buffer conditions makes WAXS a very appealing tool for structural studies. Only time will tell how large a role small angle scattering will play in structural studies. I have confidence that no matter the role, small angle scattering will remain a powerful and highly informative structural technique.

References

1. Putnam, C.D., et al., *X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution.* Quarterly Reviews in Biophysics, 2007. **40**(3): p. 191-285.

2. Svergun, D.I. and M.H.J. Koch, *Small-angle scattering studies of biological macromolecules in solution.* Reports on Progress in Physics, 2003. **66**: p. 1735-1782.

3. Feigin, L.A. and D.I. Svergun, *Structural Analysis by Small-Angle X-Ray and Neutron Scattering*, ed. G.W. Taylor1987, Princeton: Princeton Resources.

4. Jacques, D.A. and J. Trewhella, *Small-angle scattering for structural biology-Expanding the frontier while avoiding the pitfalls.* Protein Science, 2010. **19**(4): p. 642-657.

5. Mertens, H.D.T. and D.I. Svergun, *Structural characterization of proteins and complexes using small-angle X-ray solution scattering.* Journal of Structural Biology, 2010. **172**(1): p. 128-141.

6. “Andre Guinier (1911-2000).” International Union of Crystallography (2001): n. pag. Web. 28 June 2011. <http://journals.iucr.org/a/issues/2001/01/00/es0293/es0293bdy.html>

7. Cammarata, M., et al., *Tracking the structural dynamics of proteins in solution using time-resolved wide-angle X-ray scattering*. Nature Methods, 2008. **5**(10): p. 881-886.

8. Hong, X. and Q. Hao, *Combining solution wide-angle X-ray scattering and crystallography: determination of molecular envelope and heavy-atom sites.* Journal of Applied Crystallography, 2009. **42**(2): p. 259-264.

9. Shu, F., Ramakrishnan, V., Schoenborn, B.P, *High-level expression and deuteration of sperm whale myoglobin. A study of its solvent structure by X-ray and neutron diffraction methods*. Basic Life Sciences, 1996. **64**: p 309-323.

10. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA, *The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution.* Science, **289**: p 905–920

11. Lindgvist, Y., Wang, B.C., *Structure of glycolate oxidase from spinach.* Proc Natl Acad Sci USA,1985. **82**(20): p 6855-9

12. “SAXS.” The SIBYLS Beamline. SIBYLS, n.d. Web. 6 June 2011 <http://bl1231.als.lbl.gov/saxs_protocols/index.php>

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

14. Svergun, D.I. *Determination of the regularization parameter in indirect-transform methods using perceptual criteria*. J. Appl. Crystallogr. 1992, **25**: p 495-503

15. Svergun, D.I., *Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing.* Biophys J. 1999, p 2879-2886.

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

17. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem. 2004 Oct;**25**(13):1605-12.