Buffer subtraction

Subtraction of the scattering contribution from the solvent allows to obtain the scattering data corresponding to scattering from biomacromolecules of interest. Previously averaged data of the buffer is subtracted from averaged sample data and subtracted curve is used for data merging further analysis.

At P12 the buffer subtraction is performed automatically by the data processing pipeline. Manually this could be performed using program DATOP as a stay alone version or as is implemented in the primusqt interface in ATSAS package.

In this example averaged buffer SAXS curve will be subtracted from averaged sample curve and concentration-normalized using *primusqt* interface:

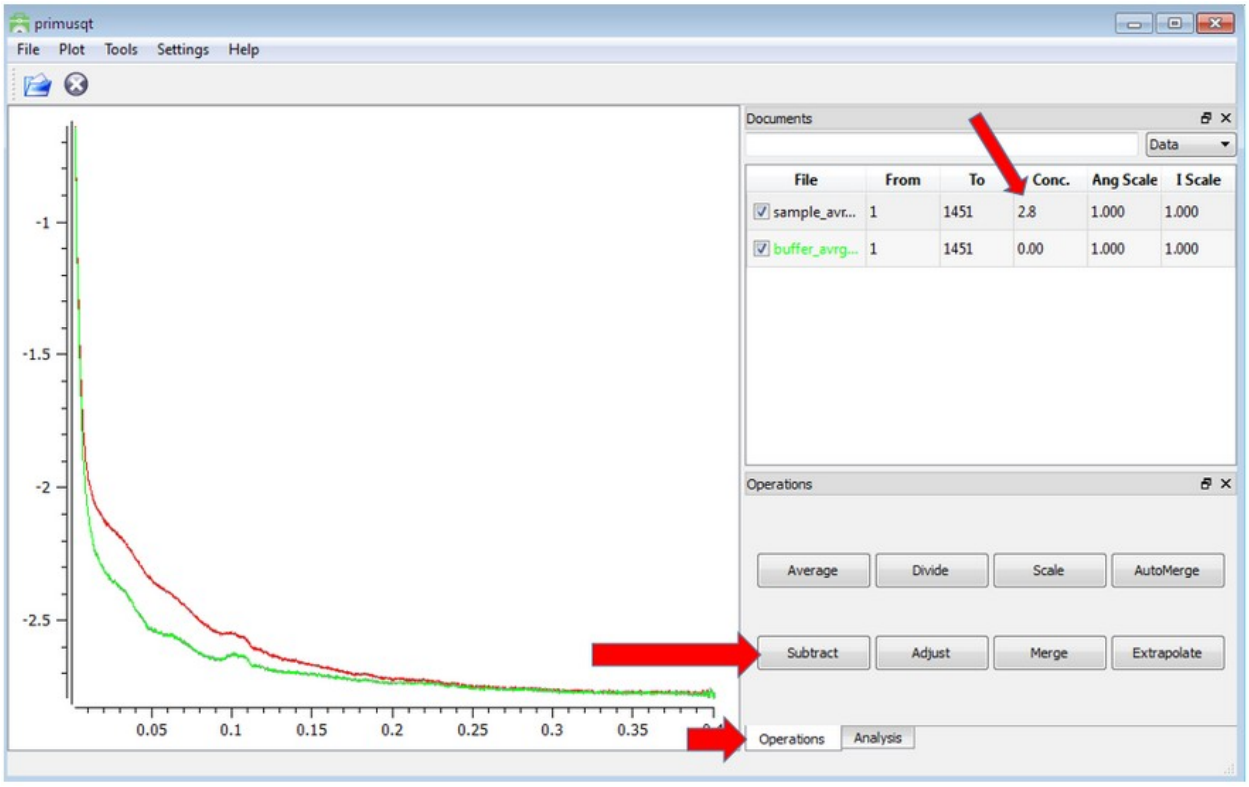

Figure 1. *Buffer subtraction and concentration normalization.* Averaged scattering curves of the sample and buffer plotted in primusqt interface. Subtraction is performed by clicking the "Subtract" button on the "Operation" tab. Resulting subtracted data could be concentration-normalized, by adding the concentration value [mg/ml] in "Conc." column of the "Document" window.

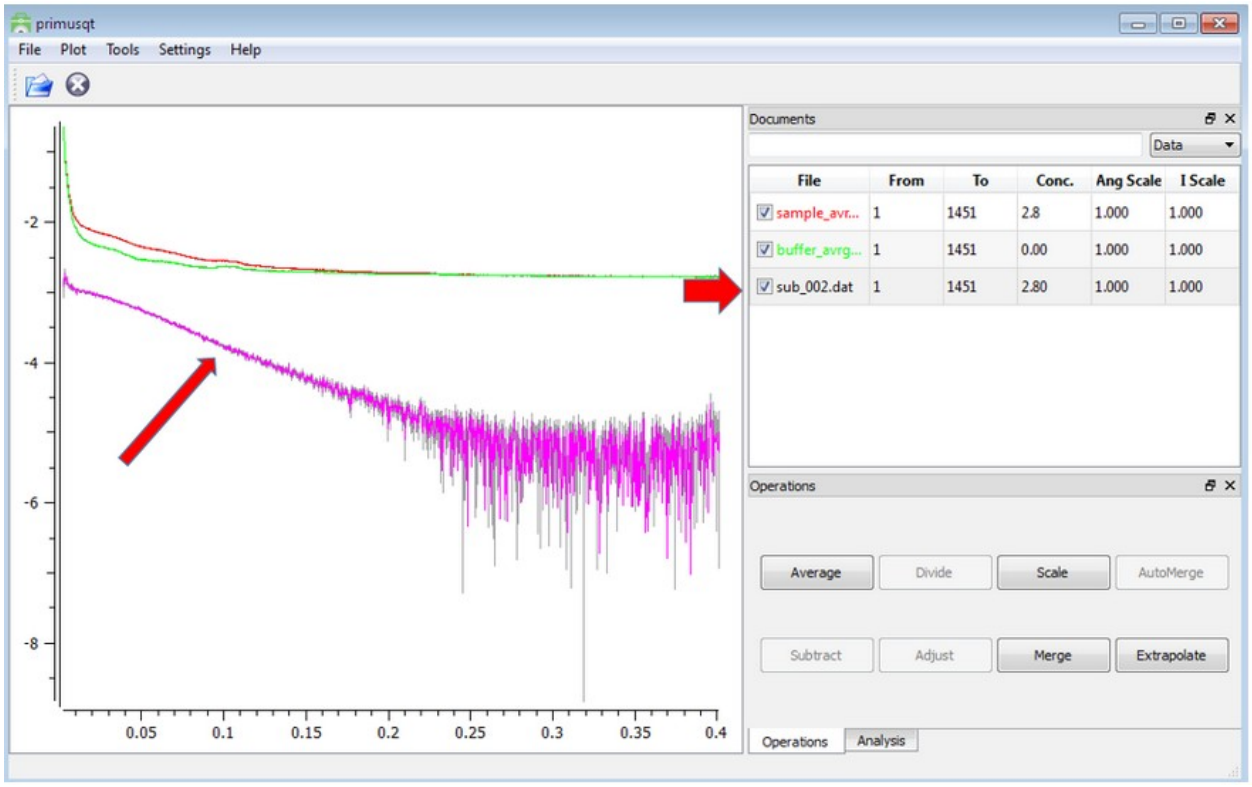

Figure 2. *Buffer subtraction.* Subtracted dataset is created and saved with automatic file name prefix "sub". Subtracted concentration-normalized data correspond to scattering from biomacromolecules or complex of interest and could be used for data merging of further analysis.

Data merging

Scattering data obtained from samples of different concentration, exposure times or angular range could be merged to obtain dataset with optimal signal:noise ratio.

Generally, scattering data measured from highly concentrated samples are less noisy in higher angles, but scattering at low angles could be affected by concentration effects as repulsion or attraction. On the other hand data from low concentration samples contains higher noise level at higher angular range, but not affected by concentration effects at low angles, see Fig. 1. Usually, optimal angular range is selected for each concentration and merged subsequently. By merging the buffer-subtracted, concentration-normalized scattering data one can obtain optimal data set for further analysis.

Scattering data from concentration series could be extrapolated to zero concentration to obtain scattering data corresponding to infinite dilution. Data extrapolated to zero concentration are free of any scattering contribution due to interparticular interaction. Extrapolation to zero concentration make sense only when no concentration dependent changes in oligomeric state (i.e. *Rg*) are observed.

The selection of the optimal angular range could be performed “manually”, based on subjective criteria of noisiness of individual user or by using available programs (e.g. ALMERGE or SAXS Merge) for automated merging based on statistical methods.

In the first example (Fig. 1-3) two datasets from samples at different concentration will be manually merged, in the second example (Fig. 4-5) datasets from concentration series will be extrapolated to zero concentration, both using the *primusqt* interface:

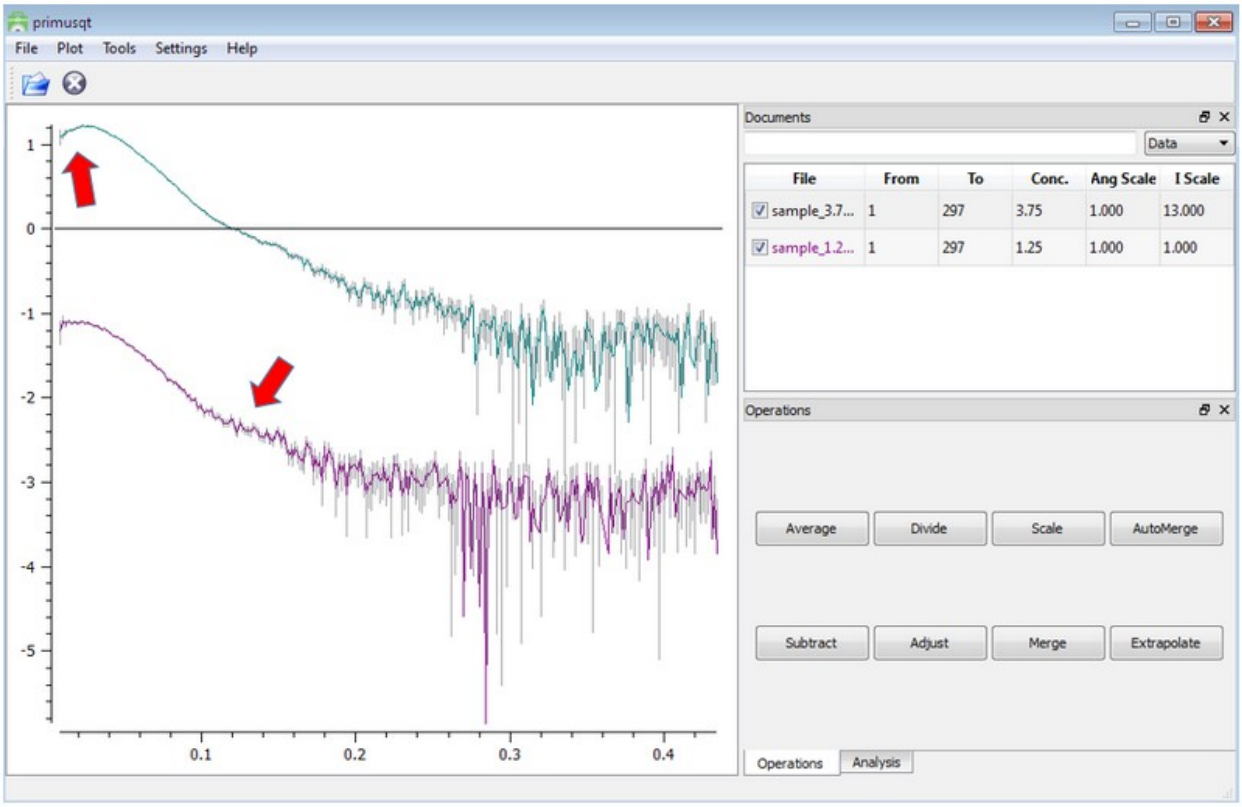

Figure 1. *Concentration effect and subjective evaluation of the noise level in the scattering data*. Buffer-subtracted concentration-normalized scattering curves of protein sample measured at two different concentration (3.75 and 1.25 mg/ml) plotted in primusqt interface. The higher concentration curve is shifted upwards (intensity is multiplied by scaling factor 13.0) to clearly illustrate the noise level. The scattering curve of high concentration sample is affected by interparticle repulsion effect exhibited as characteristic decay of intensity at low angles. The scattering curve of the low concentration sample (magenta) is not affected by interparticular repulsion, but is noisier (less smooth and with higher error), particularly in the higher angular range.

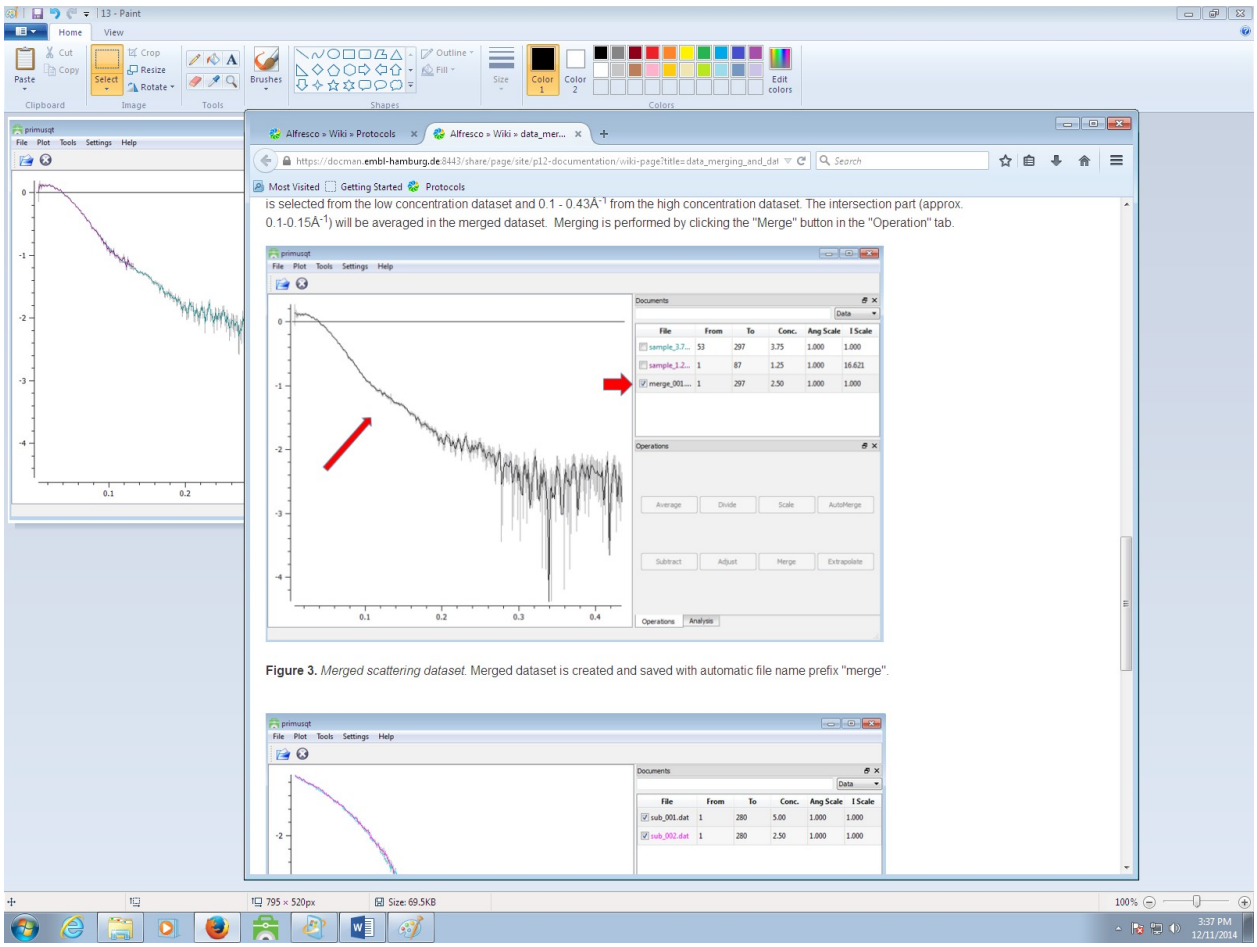

Figure 2. *Selection of optimal angular range of scattering data prior data merging*. By changing the data-point number in the "From" and "To" column in the "Document" window, the optimal angular range for each dataset is selected. In this case approx. min - 0.15Å^{-1} is selected from the low concentration dataset and 0.1 - 0.43Å^{-1} from the high concentration dataset. The intersection part (approx. 0.1-0.15Å^{-1}) will be averaged in the merged dataset. Merging is performed by clicking the "Merge" button in the "Operation" tab.

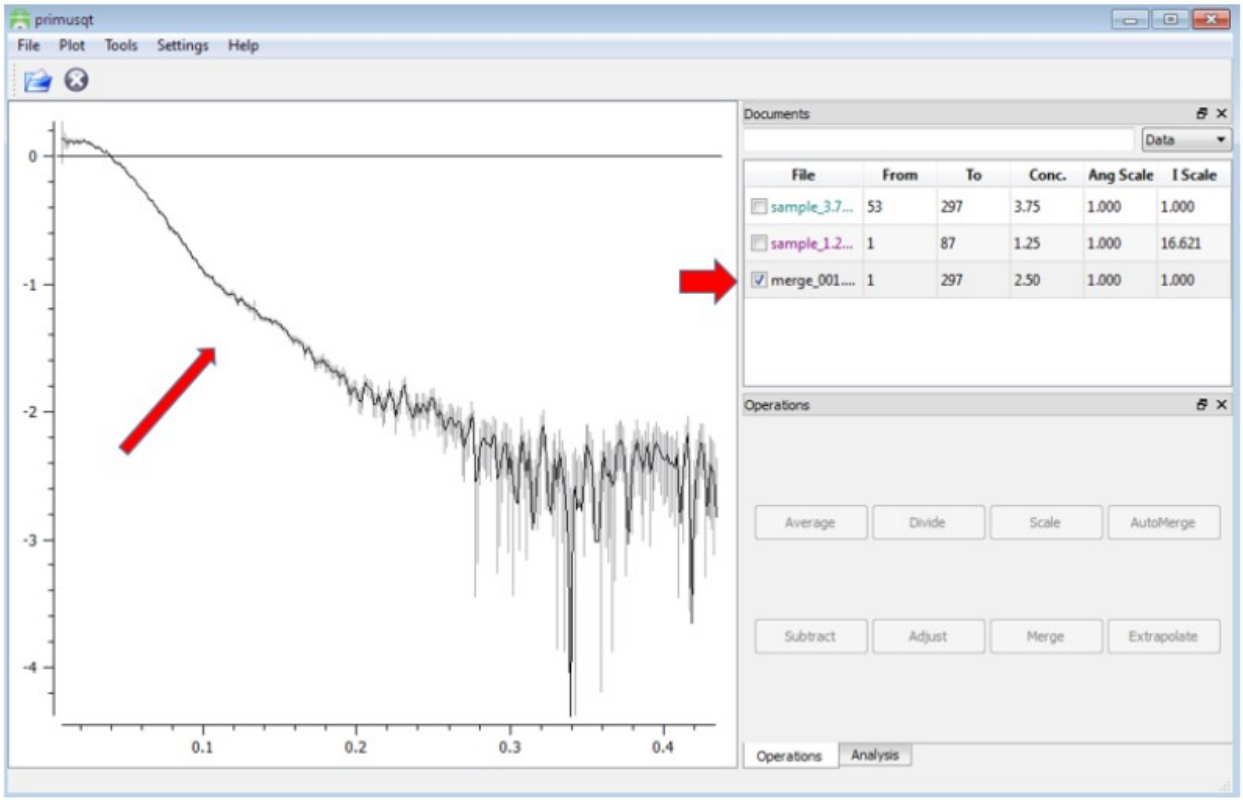

Figure 3. *Merged scattering dataset.* Merged dataset is created and saved with automatic file name prefix "merge".

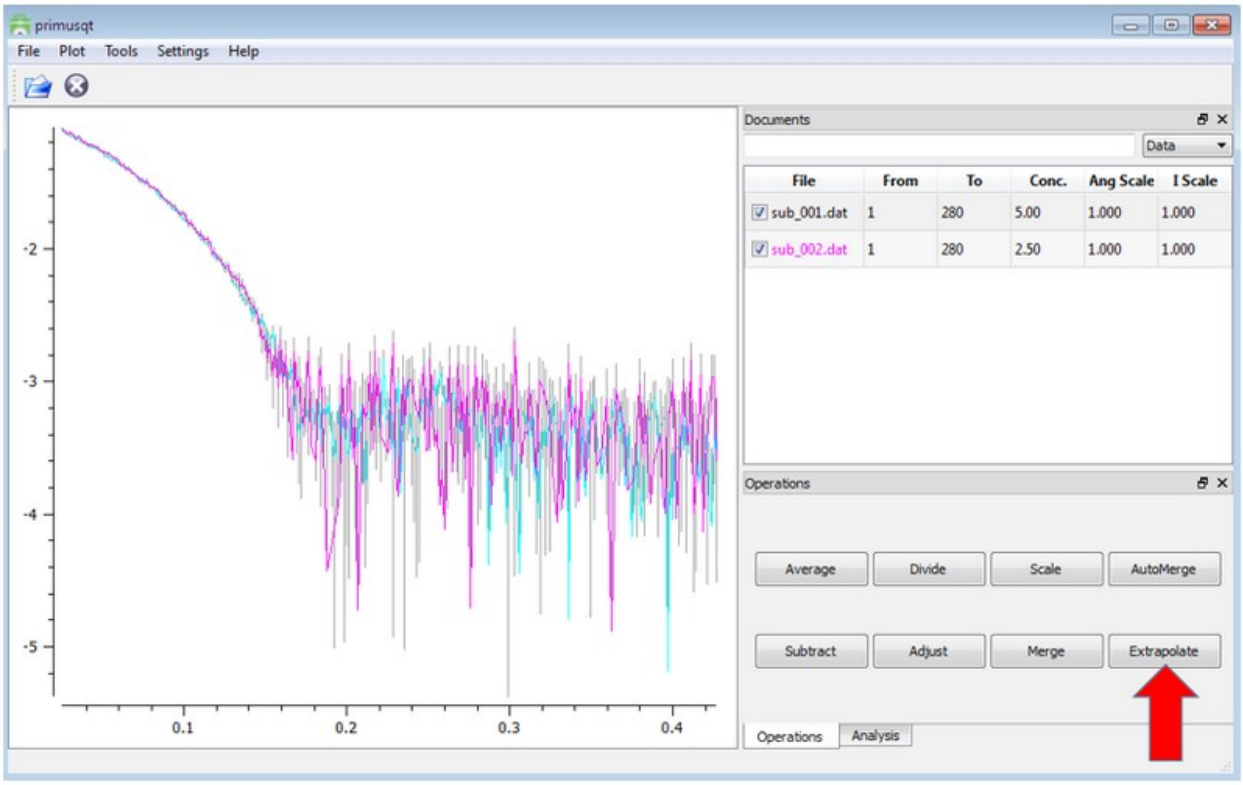

Figure 4. *Extrapolation to infinite dilution.* Two buffer-subtracted, concentration-normalized

datasets opened in primusqt interface. The extrapolation to zero concentration is performed by clickong the "Extrapolate" button in the "Operation" tab.

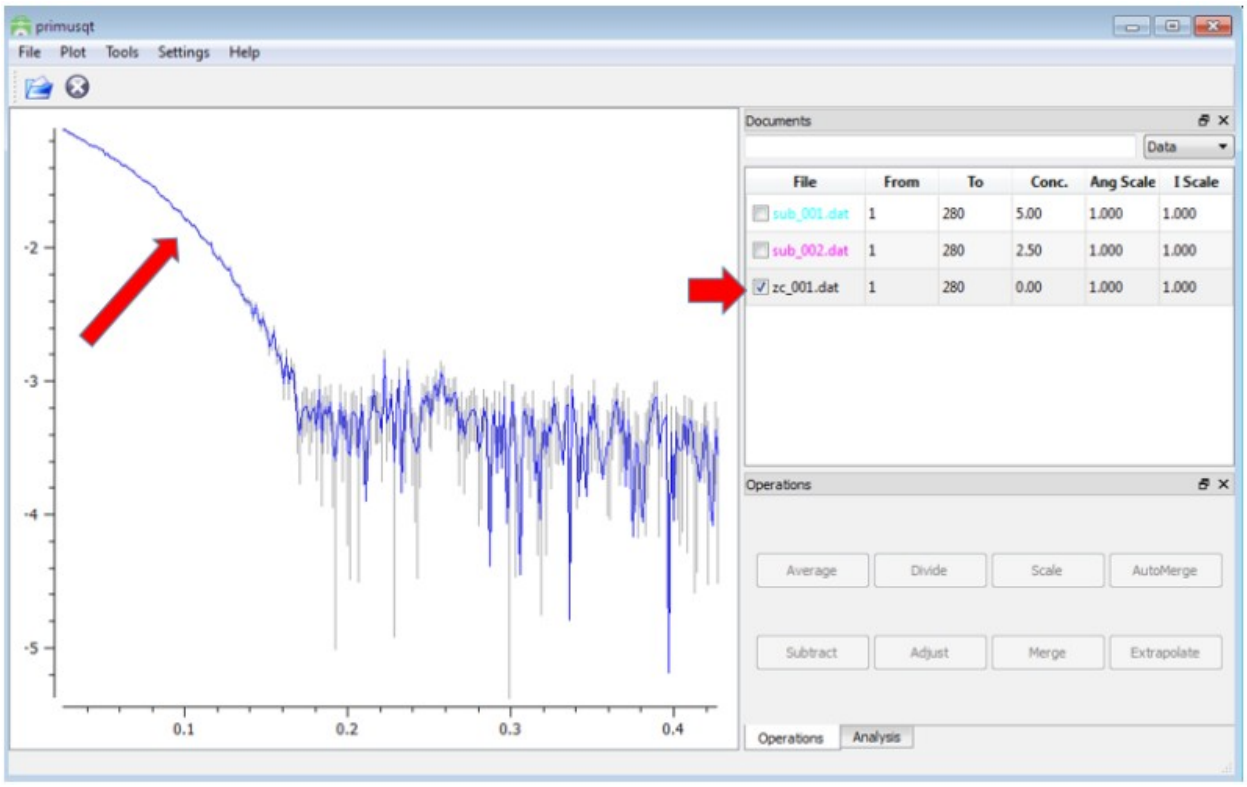

Figure 5. *Extrapolation to infinite dilution*. The dataset extrapolated to zero concentration is created and saved with default file name prefix "zc". Scattering data extrapolated to infinite dilution are suitable for further analysis.

Guinier analysis

Guinier analysis is one of the first steps in SAXS data evaluation, following initial data processing steps as averaging, buffer-subtraction and concentration-normalization, etc.. Guinier analysis provide information as radius of gyration of the particle, sample condition (monodispersity, aggregation, repulsion) and forward scattering intensity, which is proportional to molecular weight of the biomacromolecule.

André Guinier showed that in very low angles the intensity decay is proportional to radius of gyration regardless the particle shape. For monodisperse globular particles, the Guinier approximation is given by $I(q) = \exp(- Rg^2 s^2 / 3)$. Radius of gyration ($Rg$) is mechanic size parameter describing the distribution of mass of the particle. Rg could be defined as root mean square distances of the excess electron density to the center of gravity of the particle.

Guinier analysis is performed in Guinier plots, where the scattered intensity on natural logarithmic scale is plotted as a function of scattering vector square (Fig.1). In Guinier region (limited to maximal scattering vector $s<1.3/Rg$) the scattering intensity could be fitted by straight line (Fig.1). The slope of this line is proportional to particle Rg (see the Guinier approximation eq.) and by extrapolation to zero angle the forward scattering intensity is obtained ($I(0)$), see molecular weight estimation. If the Guinier plot in the Guinier zone is not linear, sample is considered to be aggregated or interacting by intramolecular repulsion (Fig.2). Scattering data from aggregated samples should not be further analyzed and attention should be focused on sample preparation. Note, linear Guinier zone is not a proof of monodispersity of the sample: oligomeric mixtures or samples of complexes containing free subunits exhibit linear Guinier behavior and medium values of Rg and $I(0)$, see polydisperse systems.

In this example the Guinier regions will be inspected using *primusqt* interface:

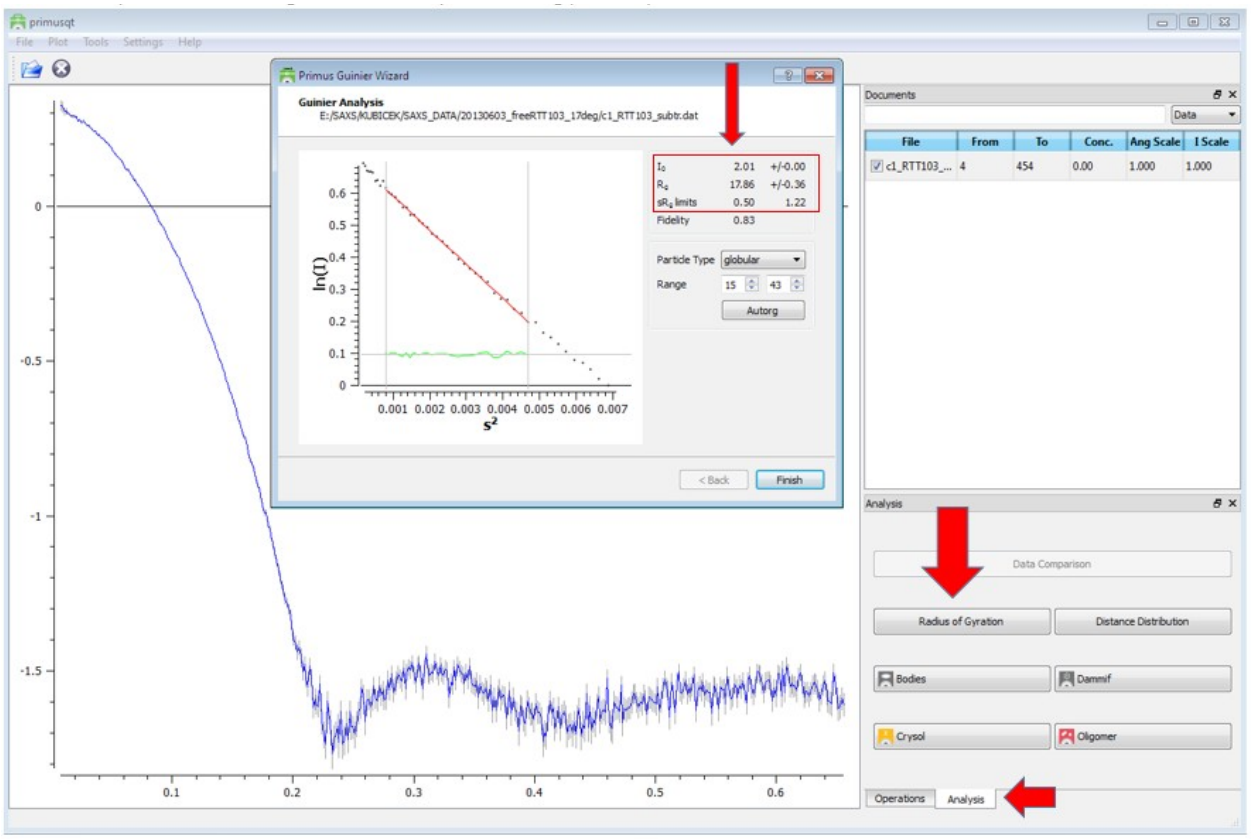

Figure 1. *Determination of Rg and I(0) by Guinier analysis.* The automated *Rg* determination procedure is performGuinier analysis is one of the first steps in SAXS data evaluation, following initial data processing steps as averaging, buffer-subtraction and concentration-normalization, etc.. Guinier analysis provide information as radius of gyration of the particle, sample condition (monodispersity, aggregation, repulsion) and forward scattering intensity, which is proportional to molecular weight of the biomacromolecule.

André Guinier showed that in very low angles the intensity decay is proportional to radius of gyration regardless the particle shape. For monodisperse globular particles, the Guinier approximation is given by $I(q) = \exp(- Rg^2 s^2 / 3)$. Radius of gyration (*Rg*) is mechanic size parameter describing the distribution of mass of the particle. *Rg* could be defined as root mean square distances of the excess electron density to the center of gravity of the particle.

Guinier analysis is performed in Guinier plots, where the scattered intensity on natural logarithmic scale is plotted as a function of scattering vector square (Fig.1). In Guinier region (limited to maximal scattering vector $s<1.3/Rg$) the scattering intensity could be fitted by straight line (Fig.1). The slope of this line is proportional to particle Rg (see the Guinier approximation eq.) and by extrapolation to zero angle the forward scattering intensity is obtained (*I(0)*), see molecular weight estimation. If the Guinier plot in the Guinier zone is not linear, sample is considered to be aggregated or interacting by intramolecular repulsion (Fig.2). Scattering data from aggregated samples should not be further analyzed and attention should be focused on sample preparation. Note, linear Guinier zone is not a proof of monodispersity of the sample: oligomeric mixtures or samples of complexes containing free subunits exhibit linear Guinier behavior and medium values of *Rg* and *I(0)*, see polydisperse systems.

In this example the Guinier regions will be inspected using *primusqt* interface:

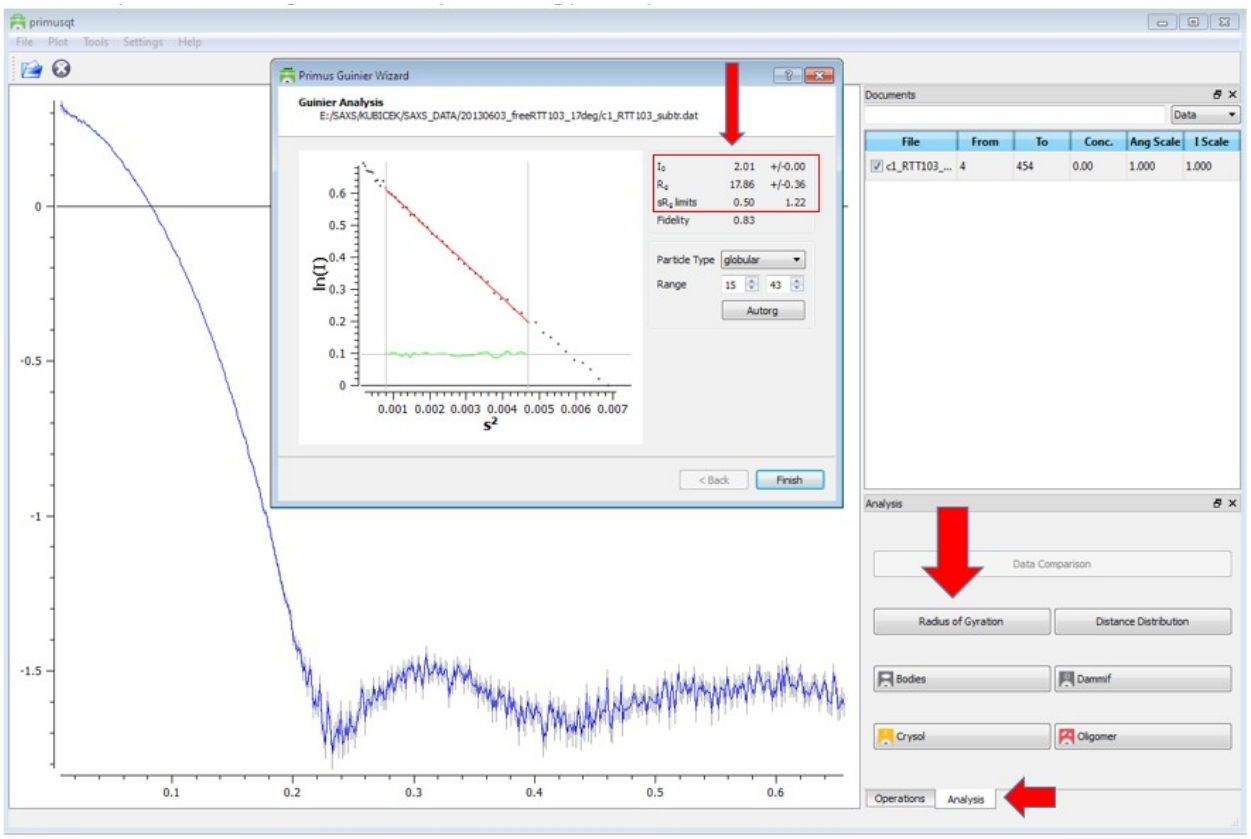

Figure 1. *Determination of Rg and I(0) by Guinier analysis*. The automated *Rg* determination procedure is performed by clicking the "Radius of Gyration" button in the "Analysis" tab. New window "Primus Guinier Wizard" appears, where the Rg and I(0) is estimated from the Guinier plot in valid scattering vector range (*sRg*<1.3)

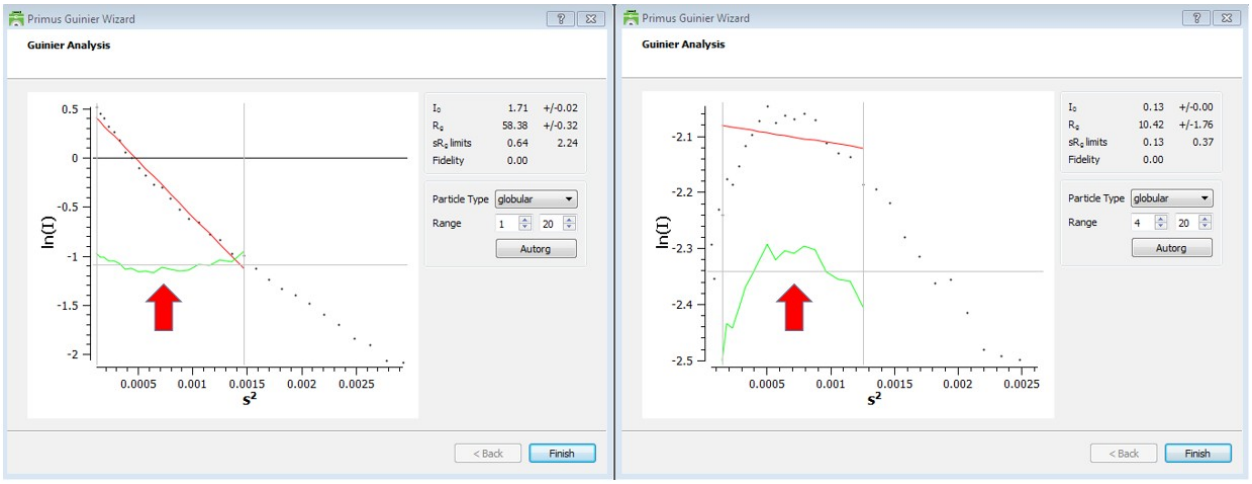

Figure 2. *Detection of aggregation and repulsion by Guinier zone inspection*. Two different datasets are plotted in "Primus Guinier Wizard". **Left:** if large aggregates are present in the samples, the typical increase of the scattering intensity is observed in low angles. This could be detected as non-linear Guinier zone, illustrated as of upswing fit residuals (green). SAXS data from samples

containing aggregates are not suitable for further analysis. Attention should be driven to improving the sample quality, as dilution series, centrifugation or altering the buffer and purification conditions. **Right:** interparticle repulsive interaction is observed as typical decrease of scattering intensity at low angles. This could be detected as downswing of fit residuals in Guinier zone. The repulsion is usually concentration-dependent at could be avoided by dilution.

ed by clicking the "Radius of Gyration" button in the "Analysis" tab. New window "Primus Guinier Wizard" appears, where the Rg and I(0) is estimated from the Guinier plot in valid scattering vector range (*sRg*<1.3)

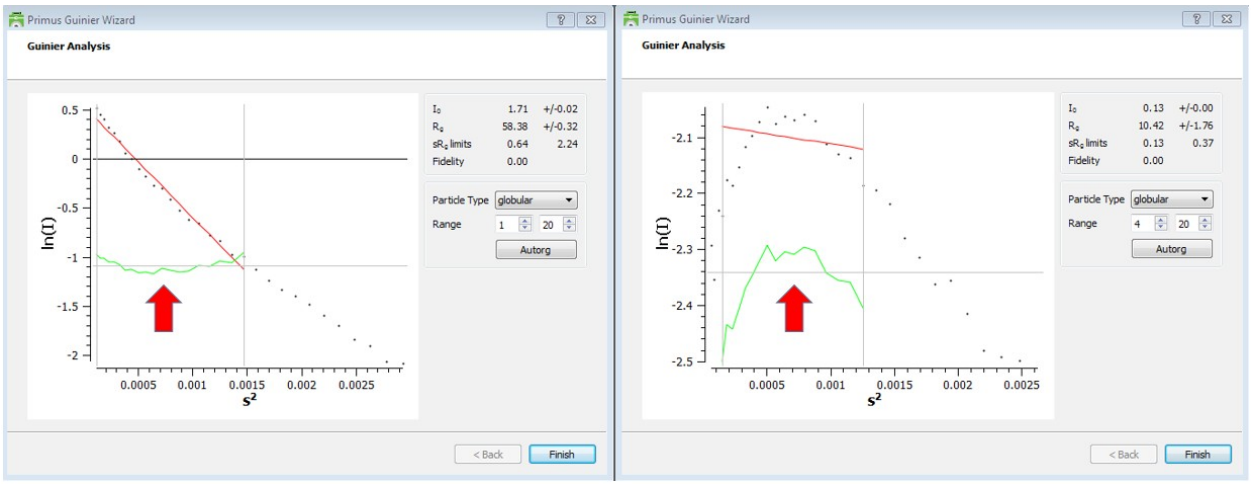

Figure 2. *Detection of aggregation and repulsion by Guinier zone inspection*. Two different datasets are plotted in "Primus Guinier Wizard". **Left:** if large aggregates are present in the samples, the typical increase of the scattering intensity is observed in low angles. This could be detected as non-linear Guinier zone, illustrated as of upswing fit residuals (green). SAXS data from samples containing aggregates are not suitable for further analysis. Attention should be driven to improving the sample quality, as dilution series, centrifugation or altering the buffer and purification conditions. **Right:** interparticle repulsive interaction is observed as typical decrease of scattering intensity at low angles. This could be detected as downswing of fit residuals in Guinier zone. The repulsion is usually concentration-dependent at could be avoided by dilution.

Molecular weight wstimation

Most straightforward way to estimate the molecular weight is to use the relation between molecular weight (*MW*) and Porod volume given by: $MW[\text{kDa}] \approx Vp[\text{nm}^3]*0.625$. The factor 0.625 called "magic number" is known from experimental praxis. This approach used for scattering data from well folded monodisperse protein solutions results in *MW* estimation with error less than 20%. Such a precision is sufficient for rapid estimation of the oligomeric state or to distinguish the complex formation from mixture of its subunits. Similar approximation for nucleic acids is given by $MW[\text{kDa}] \approx Vp[\text{nm}^3]$.

Another way to estimate the *MW* of protein of interest is to use the Guinier extrapolation of forward scattering intensity *I(0)* of protein standard, as BSA (bovine serum albumin) or lysozyme. The *MW* estimation of the protein is given by $I(0)_{protein}/I(0)_{standard} \approx MW_{protein}/MW_{standard}$. This approach requires two SAXS measurements and precise concentration determination of the protein and standard solution.

Another tool for rapid *MW* estimation is *SAXS MoW*, available as a web service. SAXS Mow algorithm uses Porod volume to estimate *MW*. As a input serves the scattering data on a relative scale in form of *P(r)* function file (*.out) obtained from program *GNOM*, see pair-distance distribution function. Usually, this approach results in *MW* estimation with error less than 10%.

In this example, the *MW* of the protein of interest (expected *MW*=12 kDa) will be estimated using the forward scattering *I(0)* of the protein standard (Fig. 1-3) and compared with the *MW* obtained from Porod volume using the $MW \approx Vp*0.625$ approximation (Fig. 4):

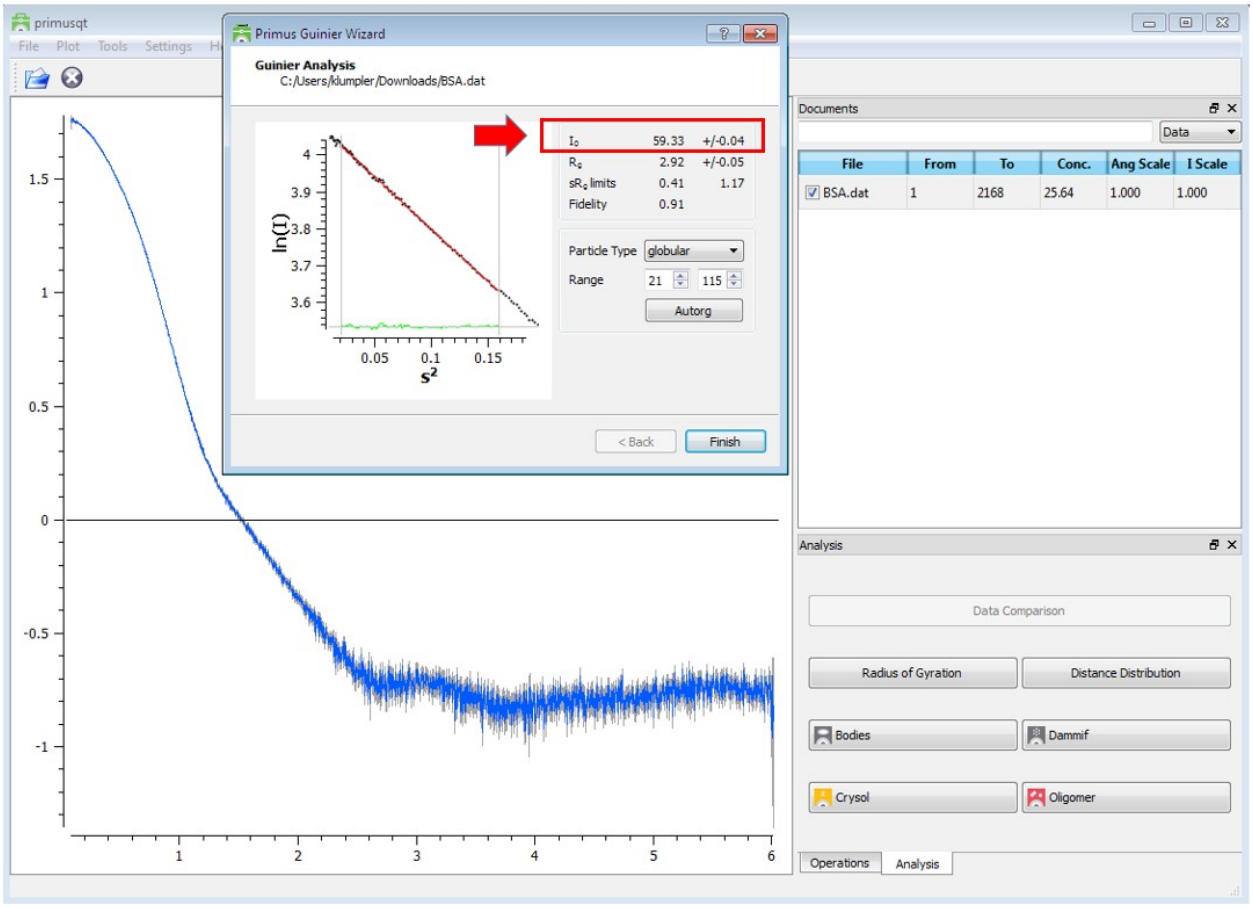

Figure 1. *The Guinier extrapolation of forward scattering of protein standard.* The forward scattering of the buffer-subtracted, concentration-normalized scattering data from BSA (c_{BSA}=25.64 mg/ml; in in 50 mM HEPES, KCl 50 mM) was extrapolated using Guinier analysis. The extrapolated value *I(0)*=59.33 is highlighted in red.

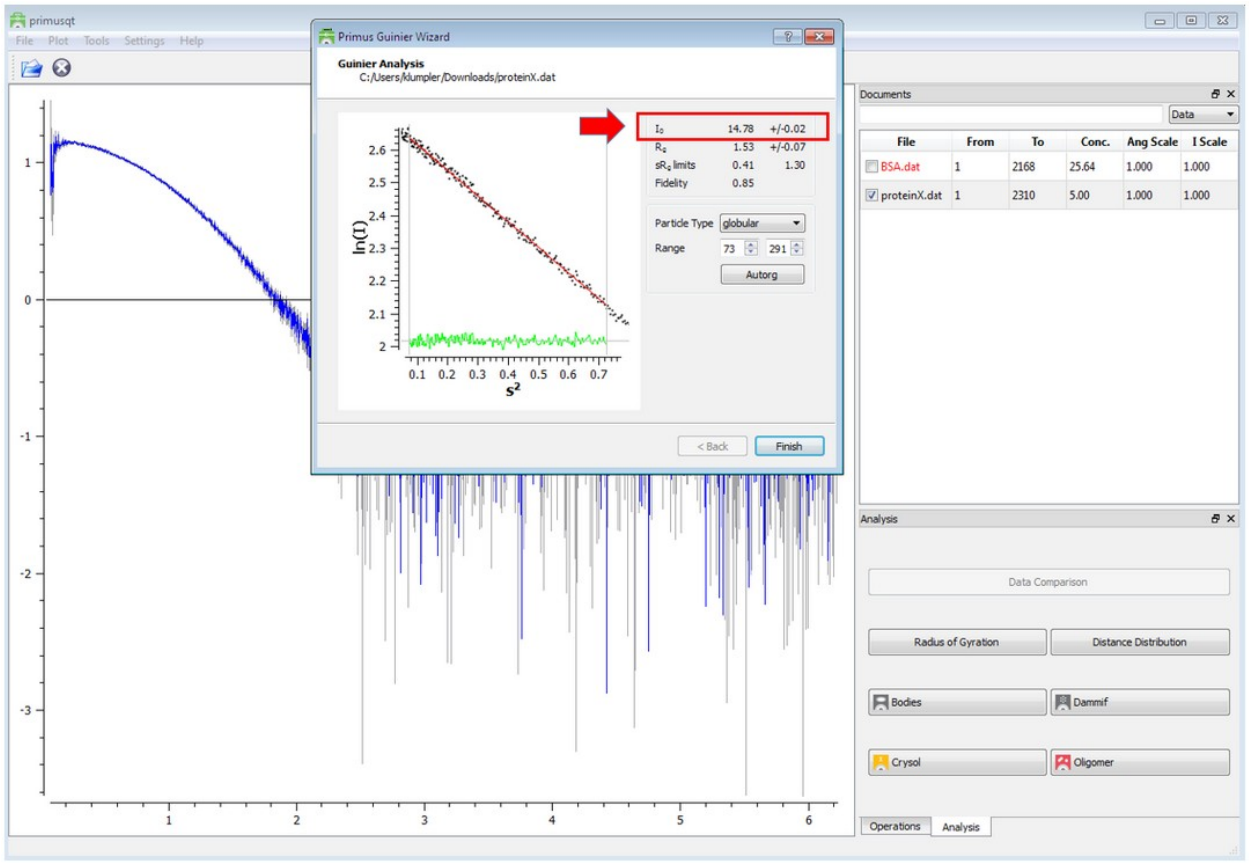

Figure 2. *The Guinier extrapolation of forward scattering of protein of interest.* The forward scattering of the buffer-subtracted, concentration-normalized scattering data from protein of interest ($c_{PROTEIN}$=mg/ml) was extrapolated using Guinier extrapolation. The extraplated value *I(0)*=14.78 is highlighted in red.

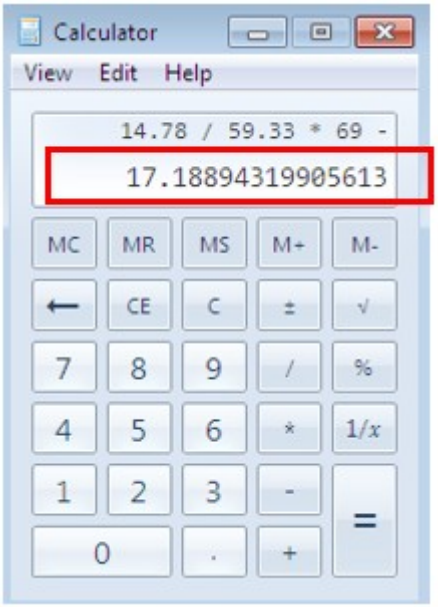

Figure 3. *The MW estimation using the forward scattering I(0) of protein standard*. *MW* of the protein standard BSA is 69kDa. By solving the approximation

$I(0)_{protein}/I(0)_{standard}=MW_{protein}/MW_{standard}$, the *MW* of the protein of interest is estimated: *MW*≈17.2 kDa. Expected MW of this protein is 12 kDa.

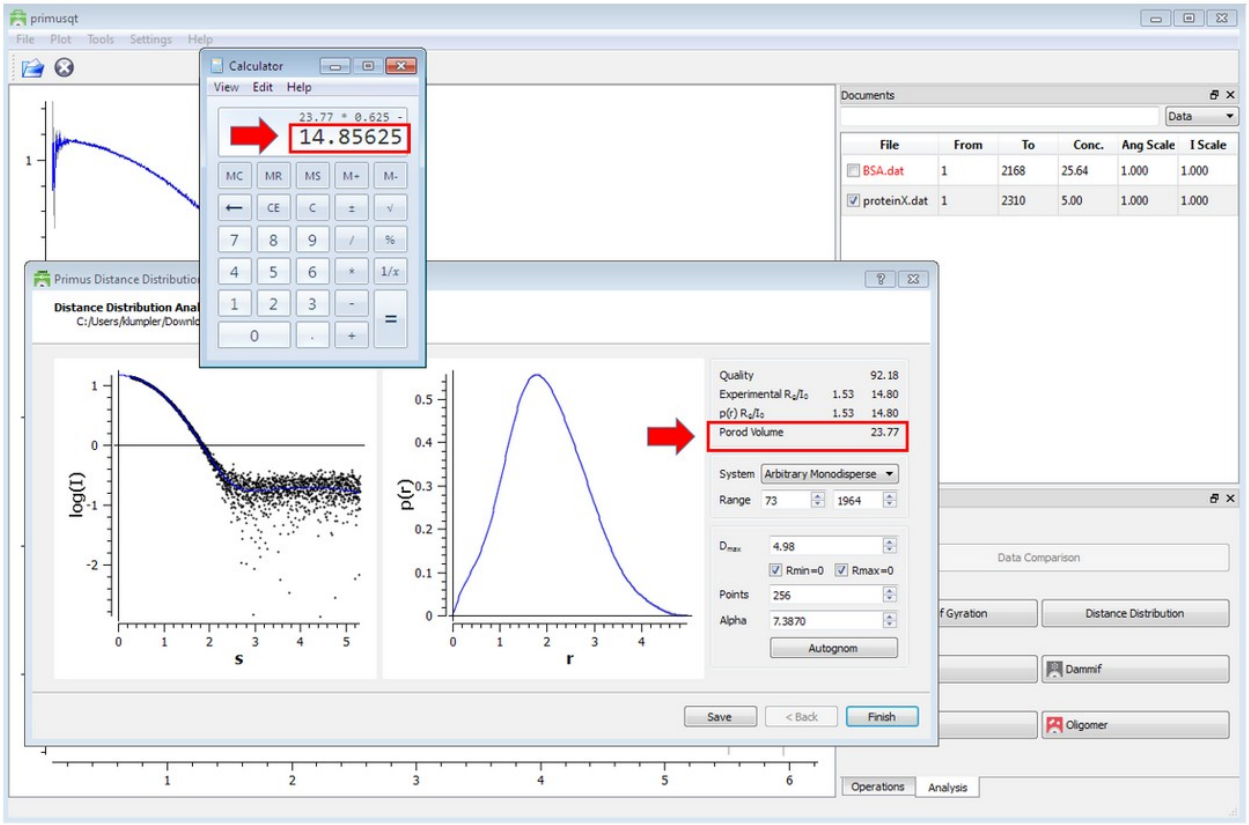

Figure 4. *The MW estimation using the Porod volume.* The Porod volume *Vp*=23.77 Å^3 of the protein of interest was determined using *primusqt* interface. By solving the *MW*≈*Vp**0.625 approximation, the *MW*=14.9 mg/ml of the protein of interest is estimated.

Kratky analysis

Flexibility and compactness of the biomacromolecule could be qualitatively evaluated by inspection of Kratky plot, where $s^2 I(q)$ is plotted as a function of s (Fig. 2.). The scattering intensity of a compact, globular particles decay proportionalyy to s^{-4}, what could be observed as a bell-shaped curve in the Kratky plot. Scattering intensity of unfolded macromolecules as intrinsically disordered proteins (IDP) decays slower, e.g. random chain proportionally to s^{-2}, what could be observed in Kratky plot as plateau followed by monotonic increase. Scattering intensity of partially unfolded macromolecules as multi-domain proteins with flexible linkers exhibits intermediate behavior in Kratky plot (Fig. 2).

Estimation of the folding state by inspection of Kratky plot is routine step of SAXS data. Kratky plot analysis is used for detection of flexibility, in folding/unfolding experiments, etc. Note, scattering intensity of rigid but elongated particles decay slower (proportionally up to s^{-1}), thus "flexible-like" shape of scattering data in Kratky plot should be considered as indication, rather than the proof of flexibility.

In this example three typical behavior of scattering data in Kratky plot will be illustrated using *primusqt* interface:

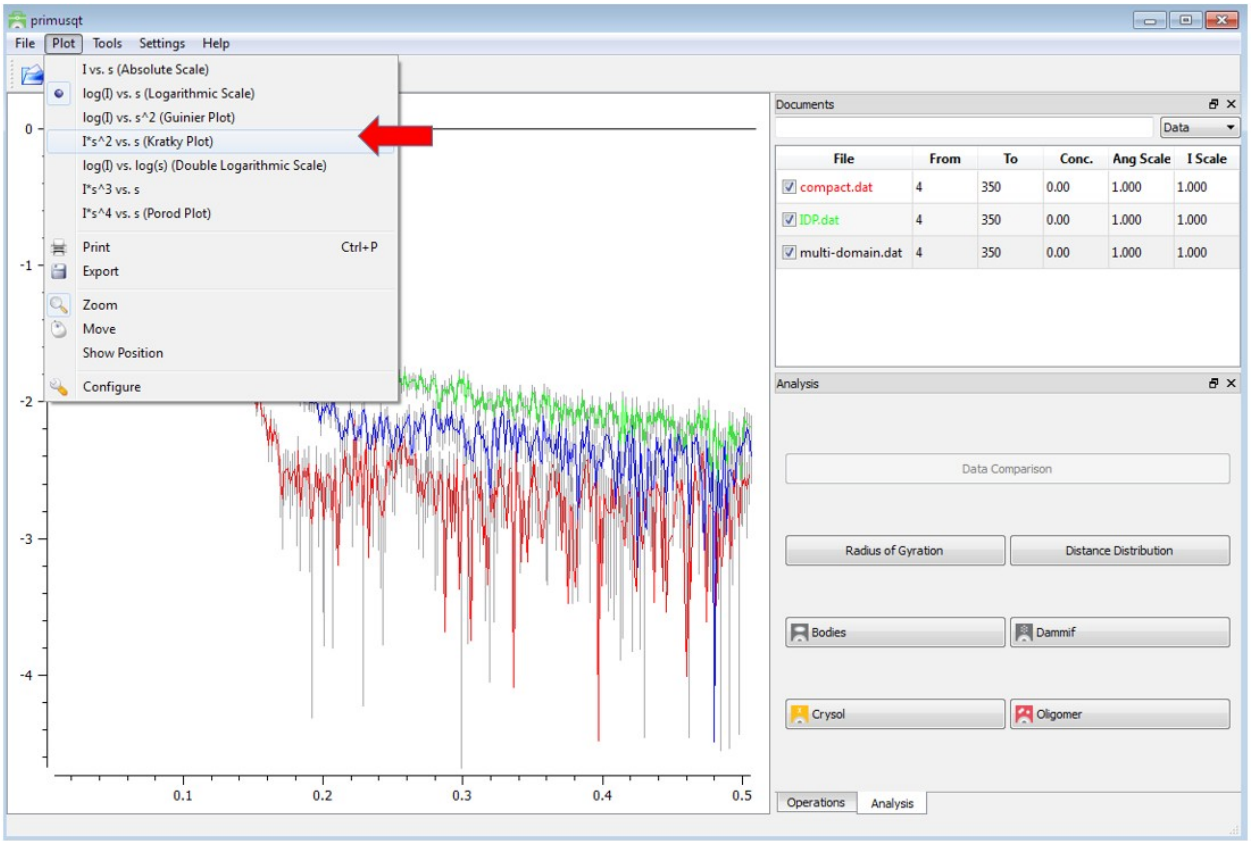

Figure 1. *SAXS datasets of rigid, flexible and multi-domain protein with flexible linkers.* Re-plotting scattering data into Kratky plot is performed by selecting the *"I*s^2 vs. s (Kratky plot)"* from the

“Plot” option in the top menu.

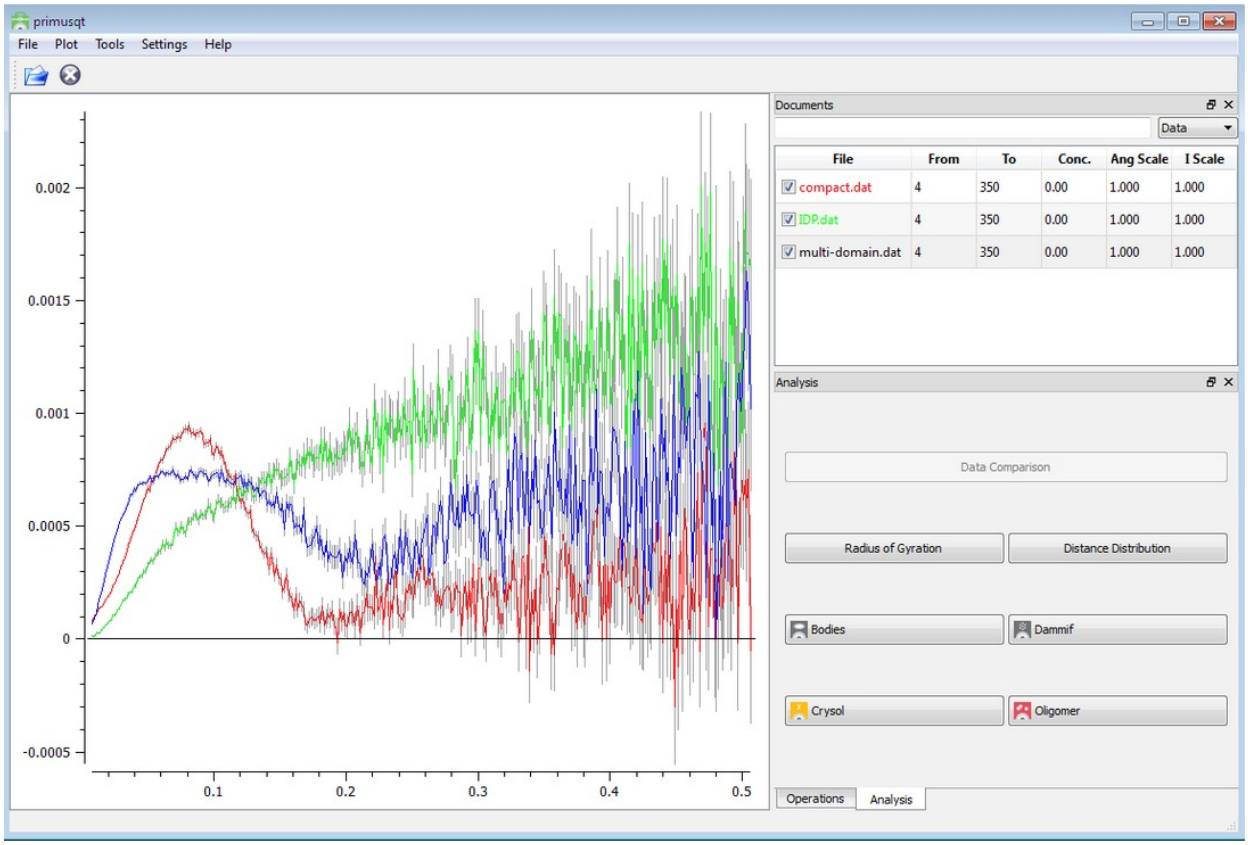

Figure 2. *Detection of protein flexibility by inspection of Kratky plot*. The Kratky plot of the well folded, compact protein (red) exhibits a clearly defined maxima in the bell-shaped curve. Flexible polypeptide chain of IDP (green) doesn’t exhibit this clear maxima, rather the plateau following by increase with higher s values. The Kratky plot of the multi-domain protein complex with flexible linkers (blue) exhibits intermediate shape.

Porod volume

Volume of the studied particles could be determined from the scattering data. Günther Porod shows the asymptotic decay of the scattering intensity at high s range. The integral of $Q = \int s^2[I(s) - K]ds$ is called Porod invariant (Q), where K is a constant determined to ensure the asymptotical intensity decay proportional to s^{-4} at higher s range. The Porod invariant Q it is related to the volume (Vp) of the particle by $Vp=2\pi^2 I(0)/Q.$, where $I(0)$ is the forward scattering intensity, see Guinier analysis.

The Porod volume is informative for well folded macromolecules, while Porod volume of flexible macromolecules will appear higher than the real volume. Determined Porod volume of well folded protein macromolecules is proportional to molecular weight by $MM \approx Vp*0.625$.

The Porod value could be determined by program *DATPOROD* stay alone or as is implemented in *primusqt* interface.

In this example the Porod volume will be determined using *primusqt* interface:

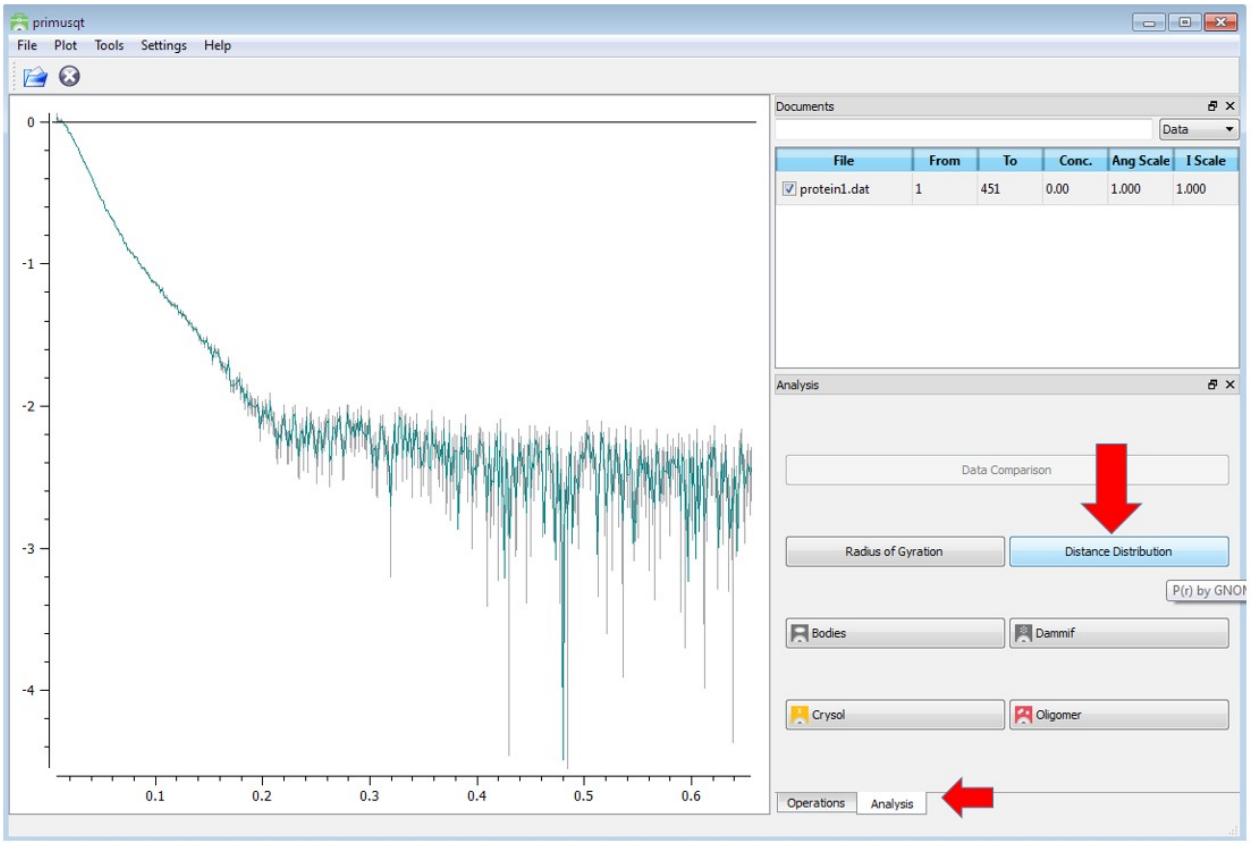

Figure 1. *Porod volume determination*. The Porod volume value is determined by clicking the "Distance Distribution" button in the "Analysis" tab.

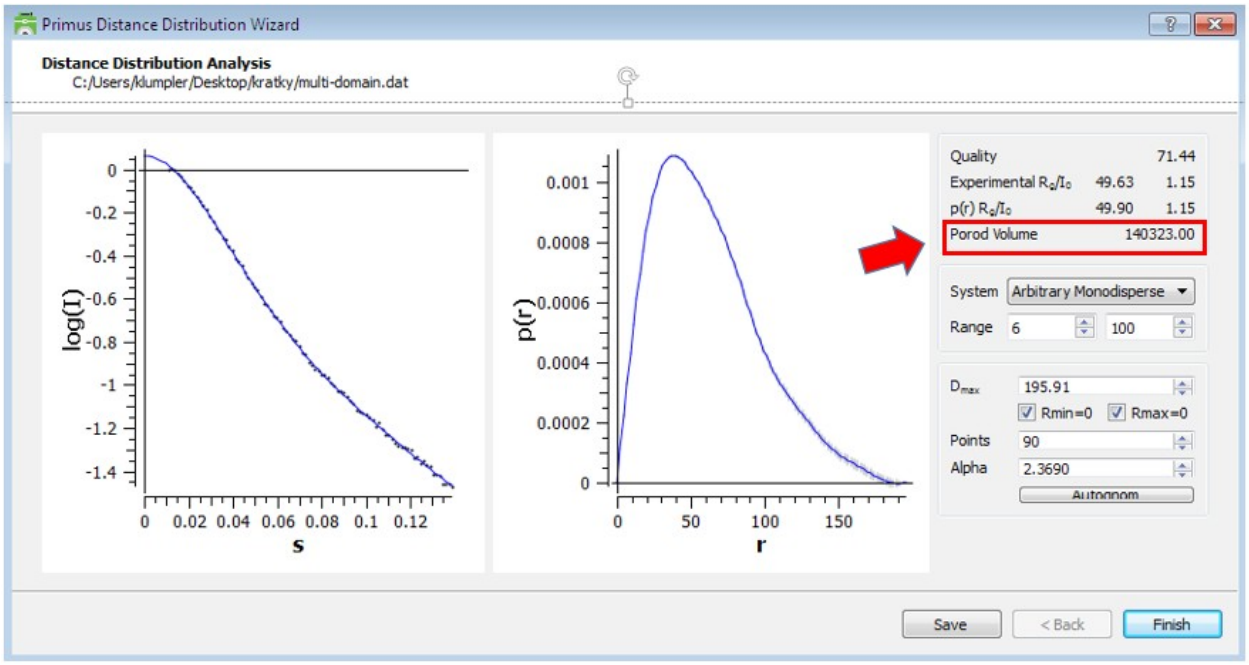

Figure 2. *Porod volume determination.* New window “Primus Distance Distribution Wizard” appears. The units of the determined Porod Volume value are the same as used in the scattering data. In this case cubic Ångströms.

Pair distance distribution function

Indirect Fourier transform of the scattering data results to the pair-distance distribution function of the single macromolecule. The pair-distribution function *P(r)* describes the distribution of distances between pairs of points (electrons) within the macromolecule. By defining correct *P(r)* function the maximal chord length of the particle (*Dmax*) is obtained. The *P(r)* is used for shape restoration experiments using the *ab initio* modeling programs.

In ideal case of monodisperse solution of not interacting homogenous particles, the pair-distance distribution function is related to scattering intensity by $P(r)=r/2\pi^2 \int[sI(s)\ sin\ (sr)]ds$, where *I* is scattering intensity, *s* is scattering vector and *r* is distance in real space. To solve this equation the precise scattering intensity measurement in angular range from zero to infinity is needed. In practice, the scattering intensity is measured in limited angular range and containing inherent statistical and systematic errors. The Indirect Fourier methods were developed to overcome this problems using regularized scattering data and iterative parameterization. By definition the *P(r)* function starts smoothly from zero at *p(0)* and should terminate smoothly to zero at *r=Dmax*. Deviation from zero value at *p(0)* could be caused by incorrect background-subtraction. Not smooth ends of *P(r)* function and/or multiple peaks and minima could be sign of incorrectly estimated *Dmax* (Fig. 3).

At P12 is the *P(r)* function determined automatically by the data processing pipeline. *P(r)* could be determined "manually" using program *GNOM* or *DATGNOM* stand alone or as it is implemented in the *primusqt* interface.

In first example the *P(r)* function is determined (Fig. 1-2), in the second example the incorrectly determined *P(r)* function is shown (Fig. 3). Both using the primusqt interface:

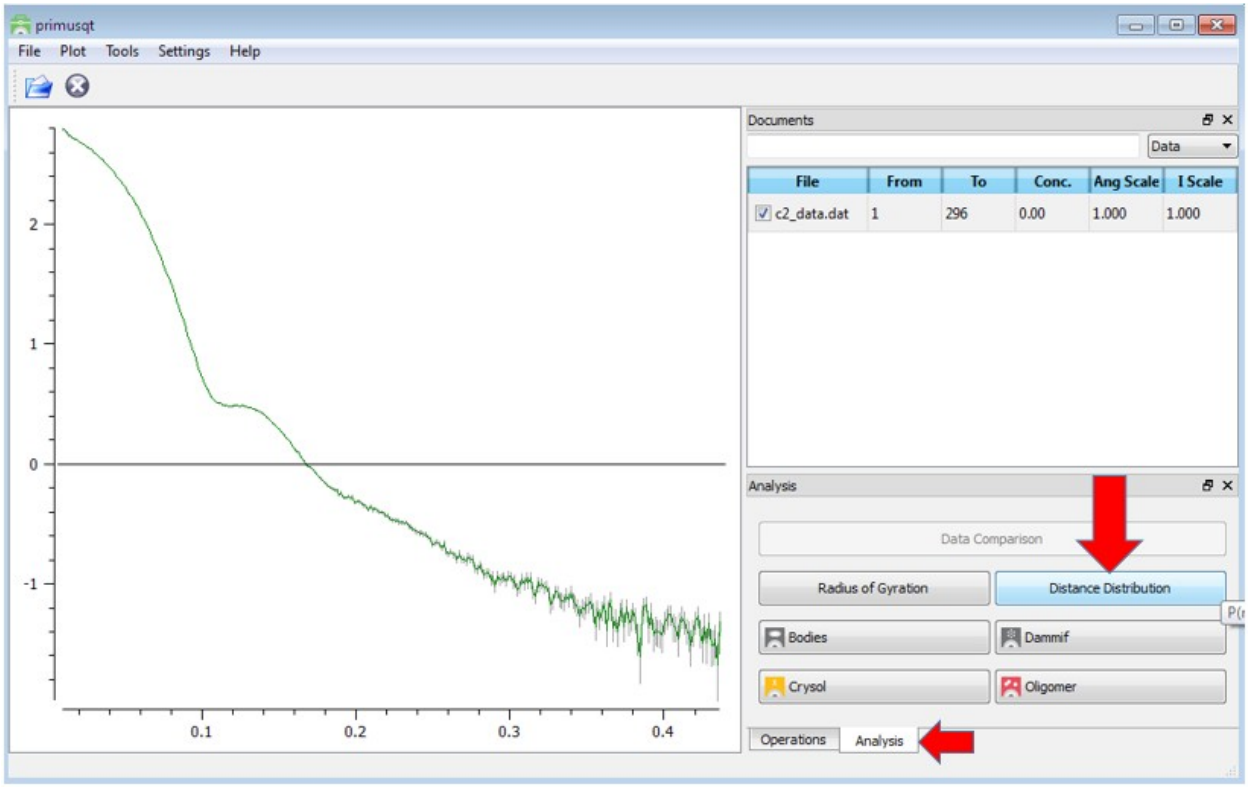

Figure 1. *Pair-distance distribution function*. Buffer-subtracted, concentration-normalized scattering data opened in the *primusqt* interface. The automatic *P(r)* function estimation is performed clicking the "Distance Distribution" button in "Analysis" tab.

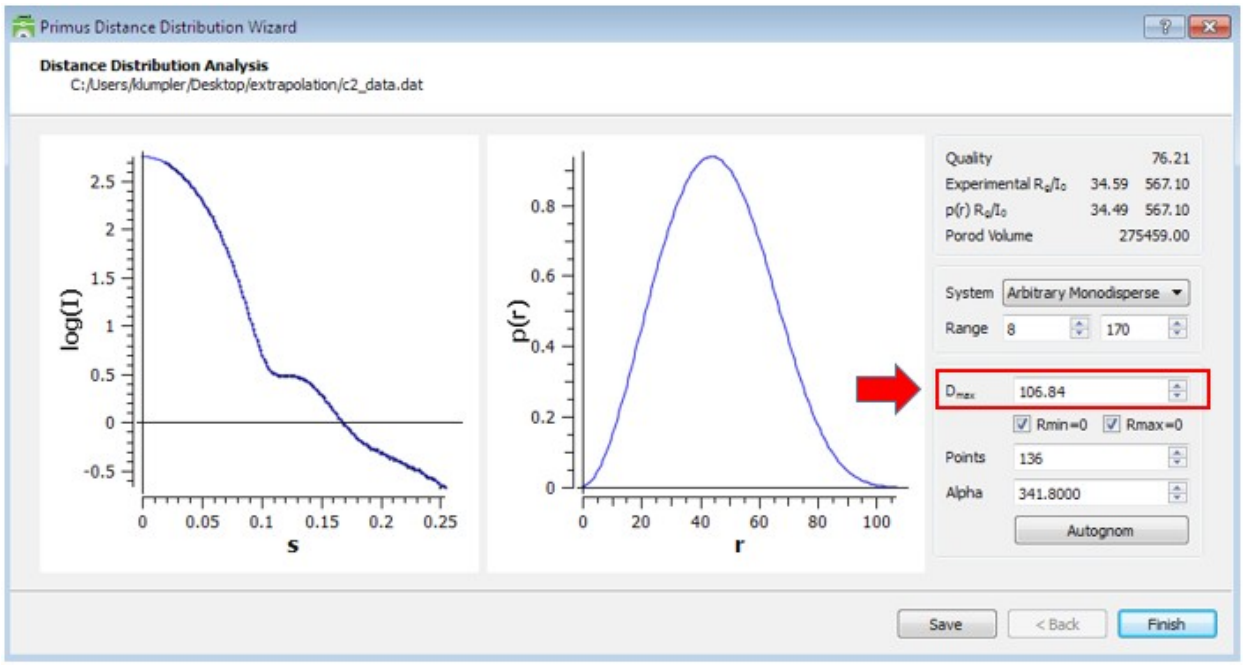

Figure 2. *Pair-distance distribution function*. New window "Primus Distance Distribution Wizard" appears, where the P(r) function is plotted on the right side. In the plot on the left the fit of the Fourier transform of the determined P(r) to experimental data is shown. Determined Dmax of the particle is highlighted. Estimation of the Dmax value could be interactively changed and the P(r) and back-fit to the experimental data is updated in real-time. The P(r) function file is saved with automatic file extension ".out". This output file is subsequently used for the shape restoration experiments by ab initio methods. The units of the *Dmax* value are the same as used in the scattering data. In this case Ångströms.

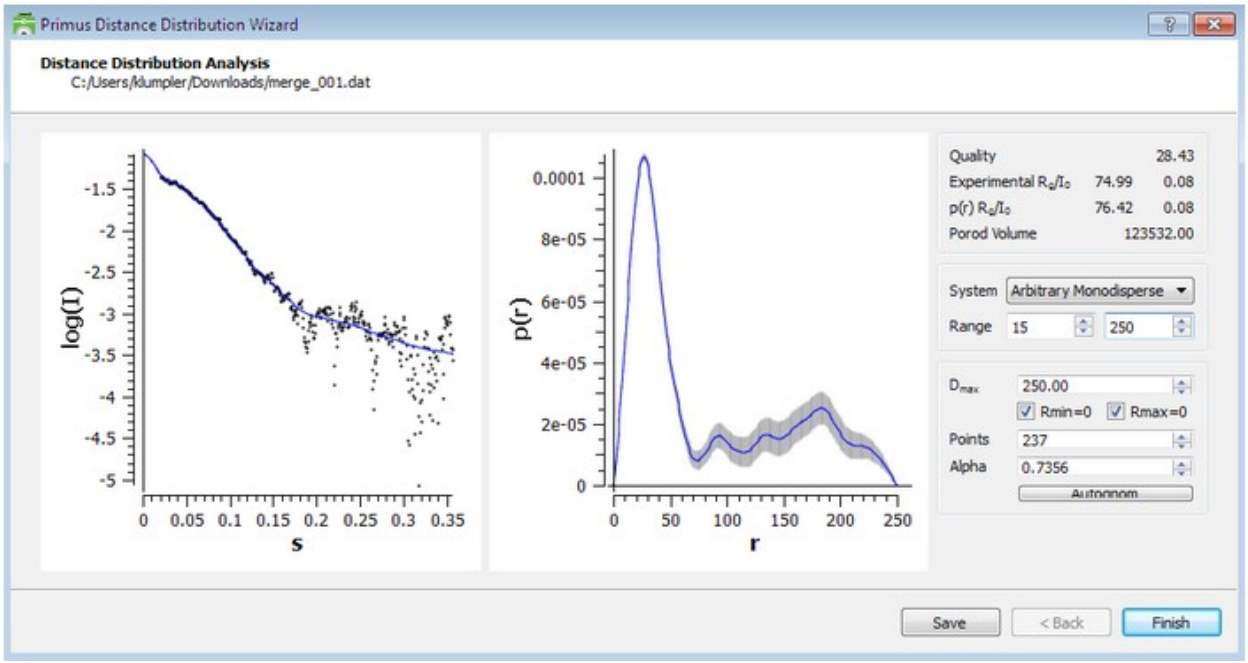

Figure 3. *Incorrectly determined P(r) function*. Analysis of scattering data of poor quality or by manual under/over estimation of Dmax could result in incorrect not smooth *P(r)* function with multiple peaks and minima. Such a *P(r)* function should be discarded and not used in further analysis.