

## **Part Two**

# **Data Interpretation and Application in Structural Biology**

**Xiaobing Zuo (NCI-Frederick)**

**Alex Grishaev (NIDDK)**

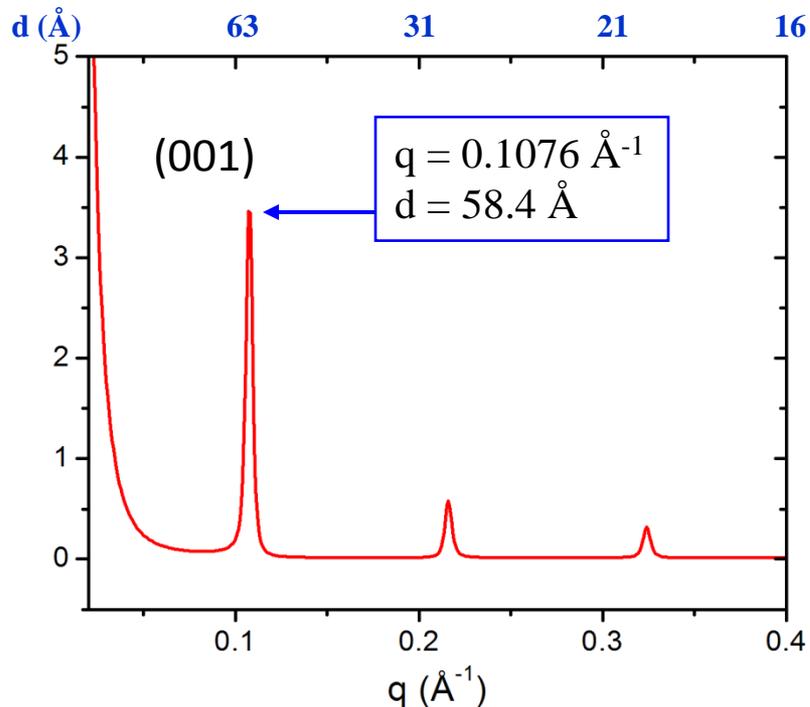
**Jinbu Wang (NCI-Frederick)**

# 1. X-ray scattering profile and embedded structural information

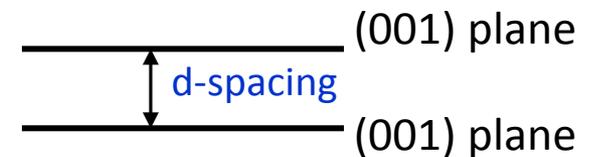
- d-spatial resolution
- hierarchical structural information
- SAXS vs. WAXS
- Guinier plot
- Radius of gyration
- Molecular mass
- Porod's law
- Porod invariant
- Pair distance distribution function (PDDF)

# d-spacing/Characteristic Length/Spatial Resolution

d-spacing / characteristic length:  $d = 2\pi/q$



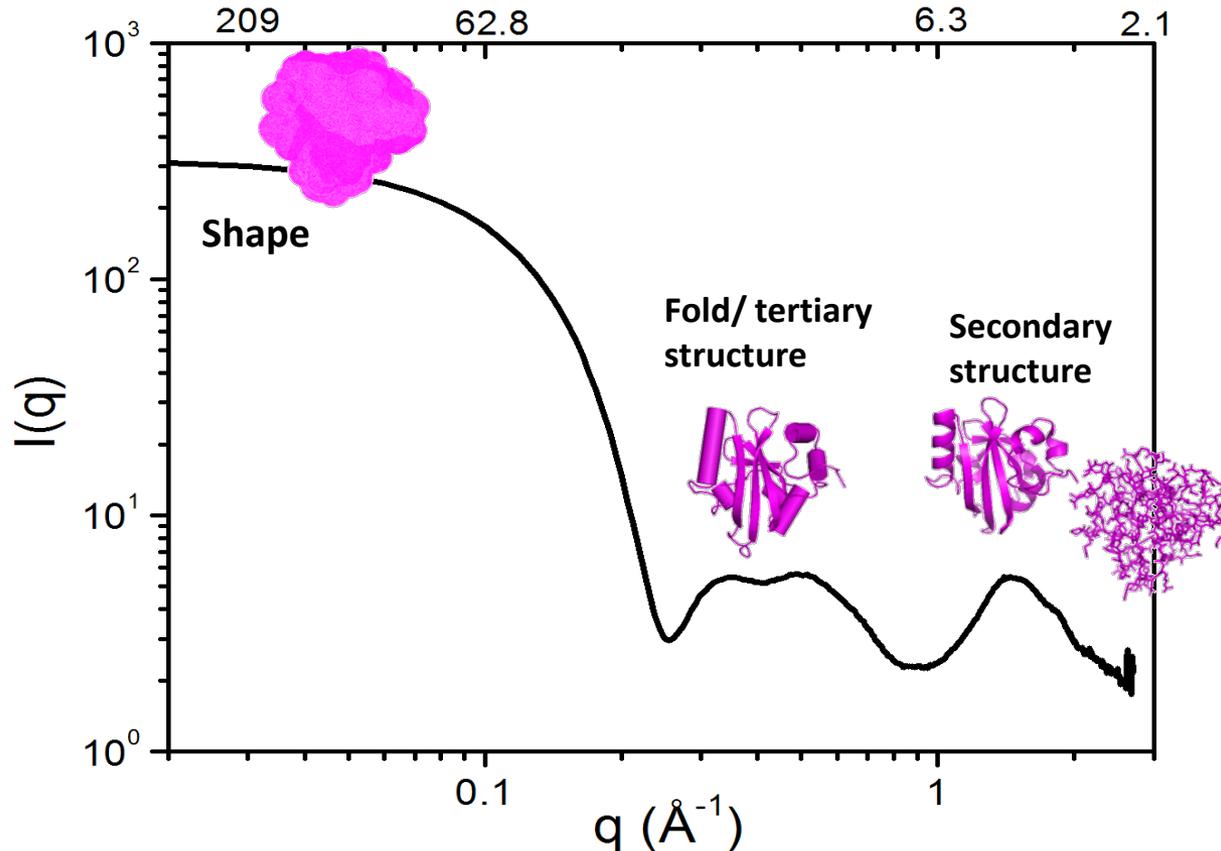
powder scattering / diffraction  
of silver behenate



- In powder diffraction pattern, peak positions represent certain characteristic distance (d-spacing) in the sample.
- A certain q in reciprocal space corresponds to a certain spatial distance ( $d = 2\pi/q$ ) or resolution in real space.

# Hierarchical structural information

characteristic length / spatial resolution:  $d = 2\pi/q$



- In various  $q$ -region, we view the molecule at different scale/resolution.
- In different  $q$ -region, scattering data show different levels of resolution on the structural details.

# SAXS vs. WAXS Data profile

➤ SAXS  $q$  range: 0 - 0.2~0.4  $\text{\AA}^{-1}$ ; WAXS:  $>0.2 \text{\AA}^{-1}$

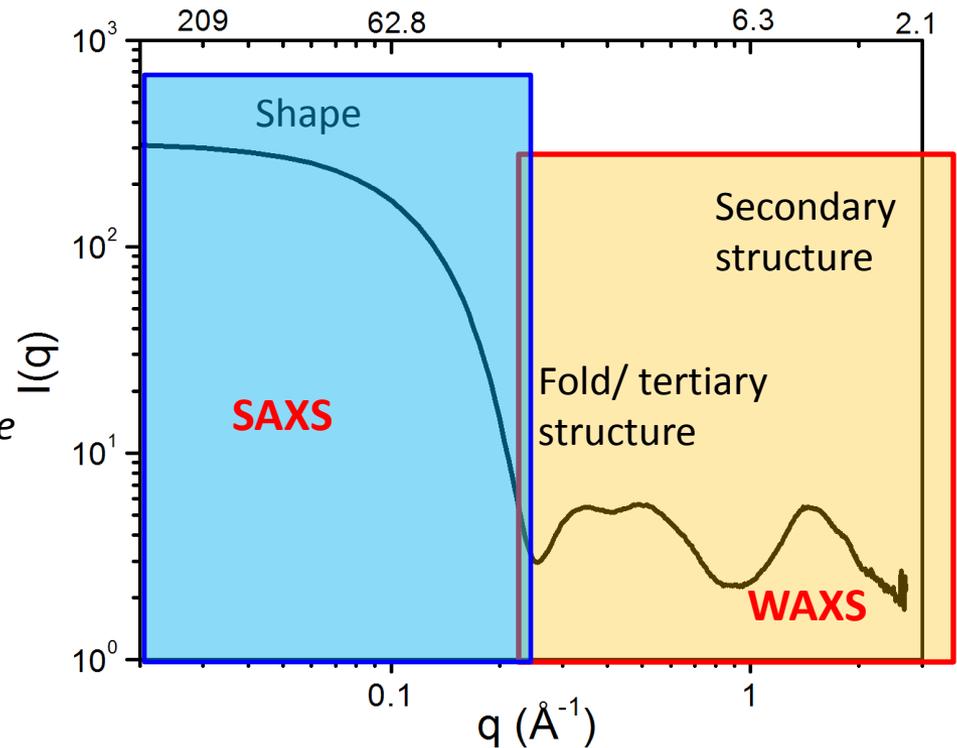
## ➤ Structural Information:

### ➤ SAXS:

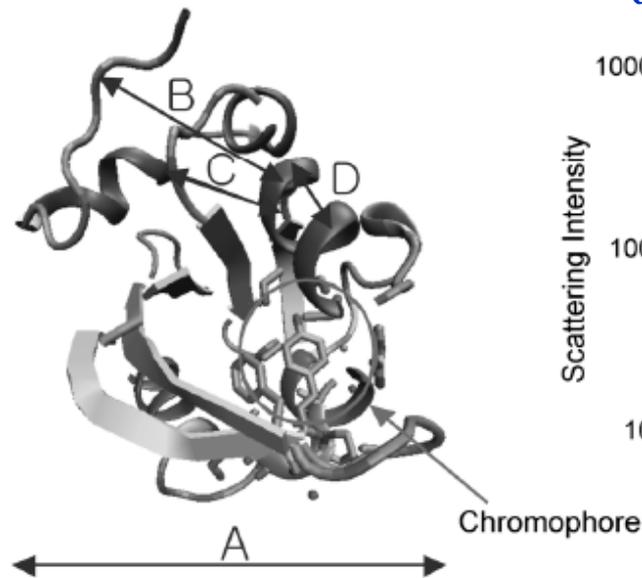
- *Size, shape, MW*
- *Conformation*
- *Inter-particle interactions*
- *Molecular envelope*

### ➤ WAXS:

- *Fingerprints of internal structure*

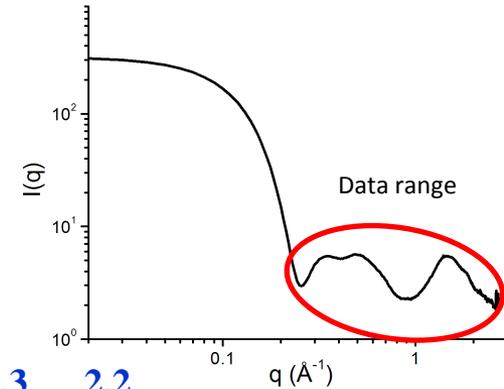
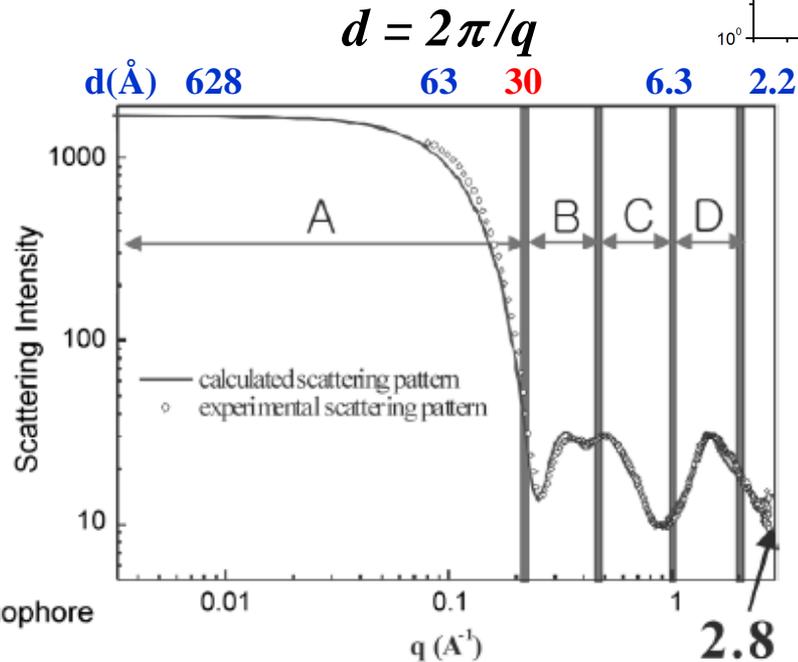


# WAXS feature assignment/structural mapping



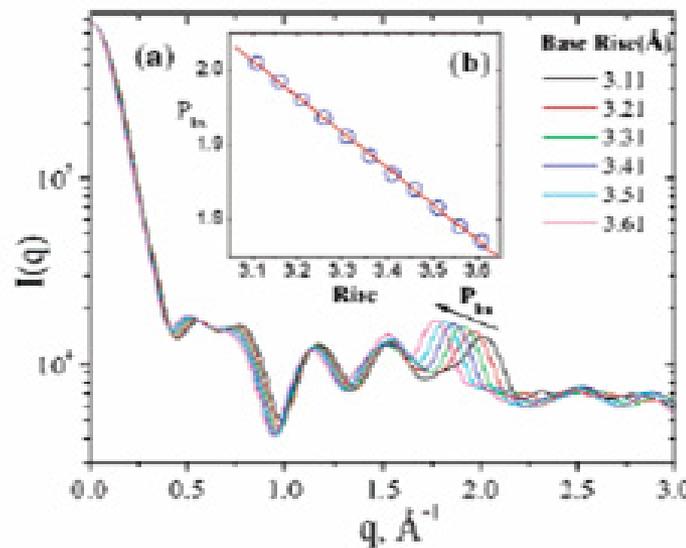
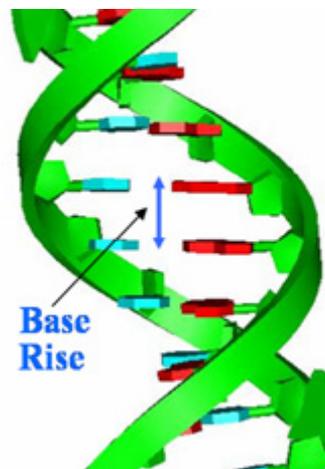
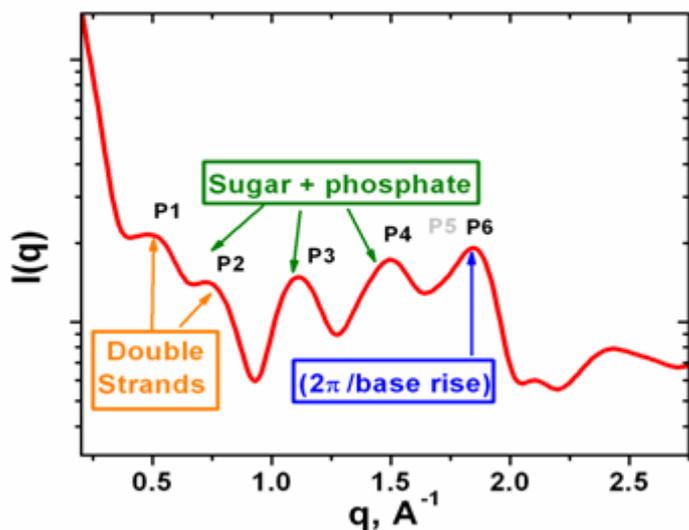
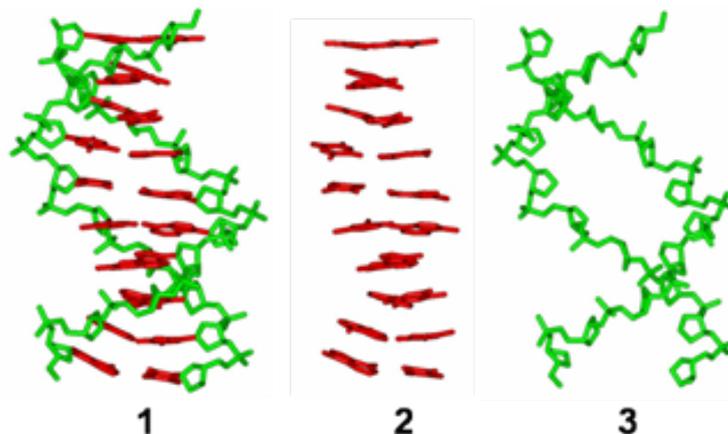
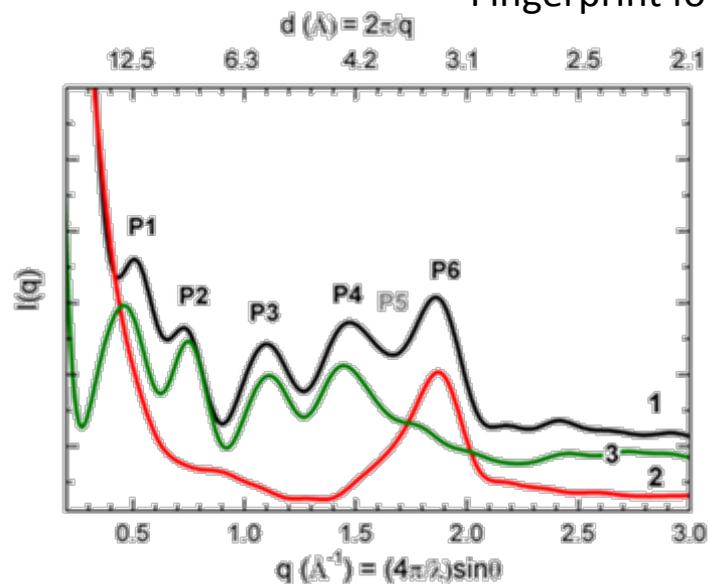
**Photoactive Yellow Protein (PYP)**

- SAXS provides information of size, shape, interparticle interaction, etc;
- WAXS fingerprints higher resolution structural characters.
- If coordinates available, further assignment/structural mapping is possible.
- For highly regular molecules, it is possible to extract structural information from waxes



# WAXS example 1: "fingerprints" for DNA conformation

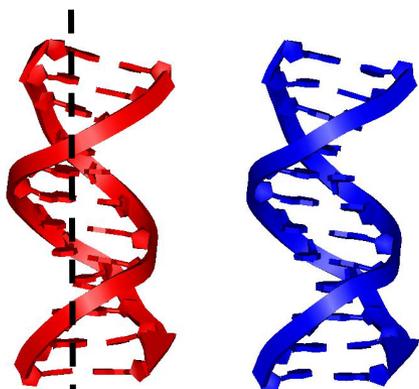
Fingerprint for canonical B-form Duplex DNA:



# WAXS example 2: Resolve Ambiguity in Structure Determination

## Dickerson DNA

Crystal:



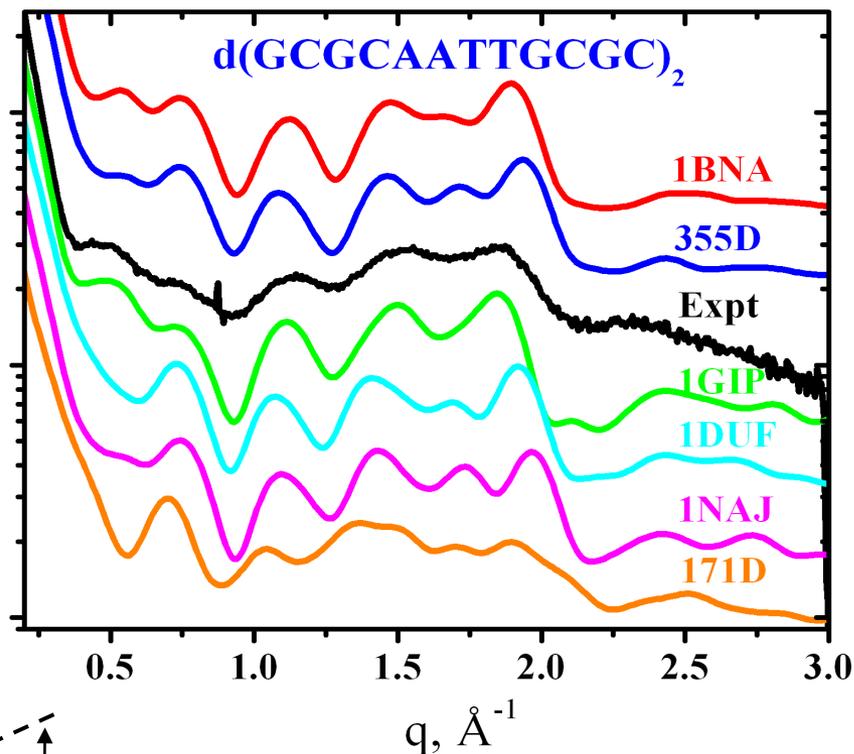
1BNA

355D

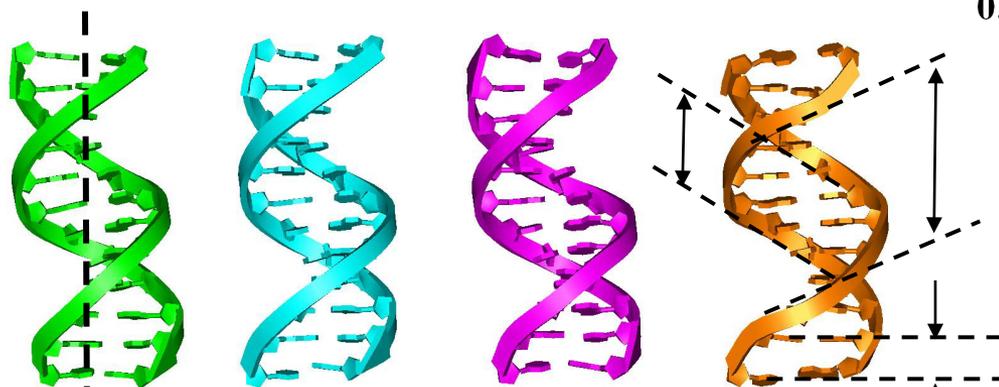
FT

WAXS patterns summarize molecular structure: distance resolved atom pair correlation

$I(q)$



NMR Solution:



1GIP

1DUF

1NAJ

171D

- Structures differ in details (linearity, twist, packing)
- Demonstrate ambiguity of determining structure
- WAXS can distinguish

# Guinier equation

Scattering intensity can be expanded in powers of  $q^2$ :

$$I(q) = I(0) \left[ 1 - \frac{R_g^2 q^2}{3} + kq^4 + \dots \right]$$

When  $q \rightarrow 0$ ,

$$I(q) \cong I(0) \exp\left(-\frac{R_g^2 q^2}{3}\right)$$

$qR_g < 1.3$  for globular;

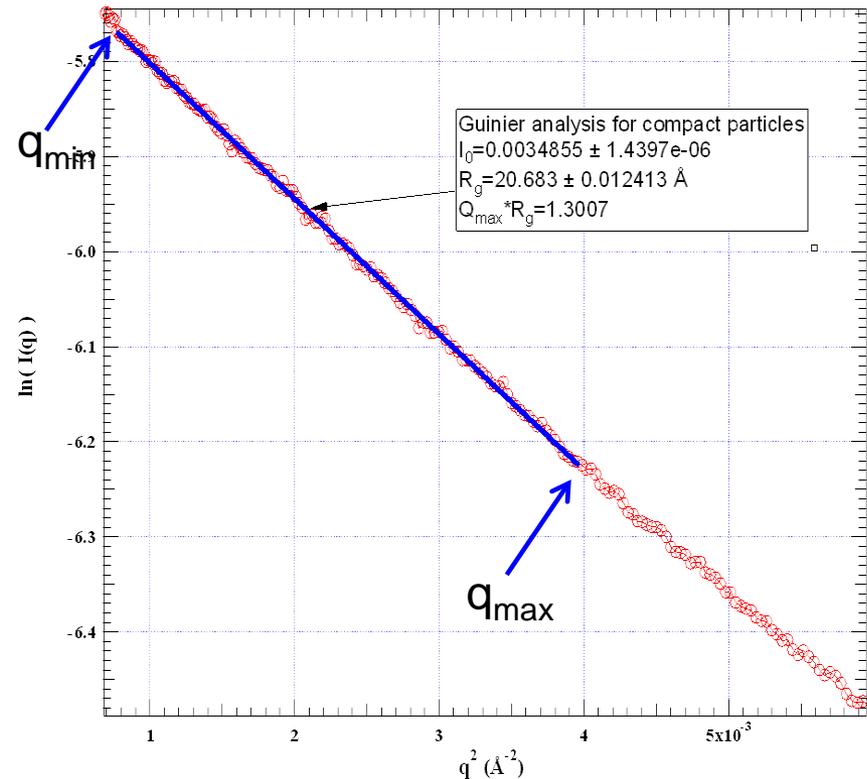
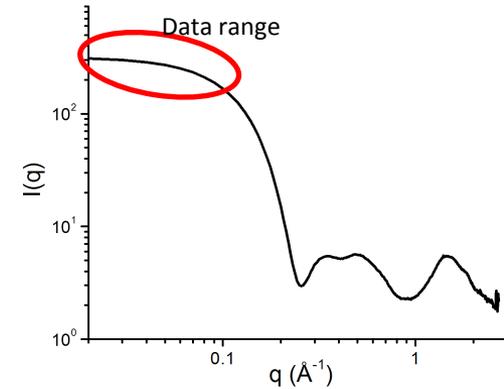
$qR_g < 0.8$  for elongated

$I(0)$ : forward scattering

$R_g$ : radius of gyration

$$I(q) = \frac{2I(0)}{q^4 R_G^4} (q^2 R_G^2 - 1 + e^{-q^2 R_G^2})$$

$qR_g < 1.4$  for elongated



To get reliable Guinier plot /  $R_g$  analysis:

- $q_{\min} \leq \pi/D_{\max}$
- $q_{\max} * R_g < 1.3$  for globular;  $< 0.8$  for elongate
- Multiple ( $\geq 5$ ) data points in linear fashion

# Radius of gyration of some homogenous bodies

$R_g$ : radius of gyration

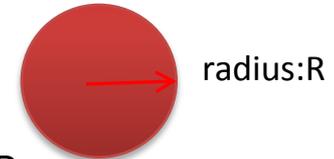
$$R_g^2 = \frac{\int \Delta\rho(r)r^2 dV}{\int \Delta\rho(r)dV}$$

$$R_g^2 \approx \frac{\sum_j \Delta n_j r_j^2}{\sum_j \Delta n_j}$$

excess electrons

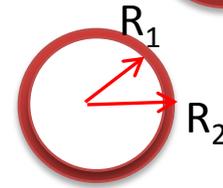
Sphere  
(radius R)

$$R_g^2 = (3/5)R^2$$



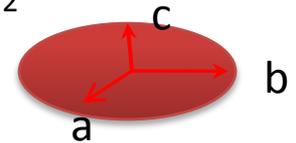
Hollow sphere  
(radii  $R_1 < R_2$ )

$$R_g^2 = (3/5) \frac{R_2^5 - R_1^5}{R_2^3 - R_1^3}$$



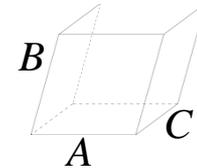
Ellipsoid  
(semi-axes  $a, b, c$ )

$$R_g^2 = (1/5)(a^2 + b^2 + c^2)$$



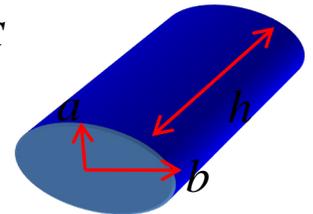
Parallelepiped  
(edge length A, B, C)

$$R_g^2 = (1/12)(A^2 + B^2 + C^2)$$



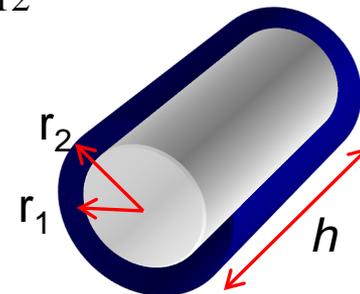
Elliptic cylinder  
(semi-axes  $a, b$ ; height  $h$ )

$$R_g^2 = \frac{a^2 + b^2}{4} + \frac{h^2}{12} = R_c^2 + \frac{h^2}{12}$$



Hollow cylinder  
(height  $h$ , radii  $r_1, r_2$ )

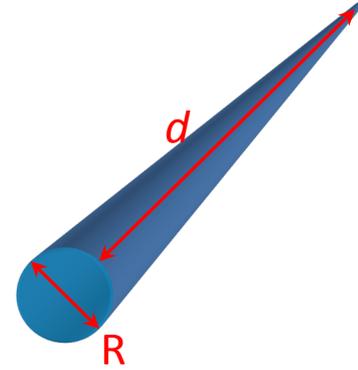
$$R_g^2 = \frac{r_1^2 + r_2^2}{2} + \frac{h^2}{12}$$



# Rod-like and lamellar particles

Elliptic cylinder  
(semi-axes  $a, b$ ; height  $h$ )

$$R_g^2 = \frac{a^2 + b^2}{4} + \frac{h^2}{12} = R_c^2 + \frac{h^2}{12}$$



Rod/needle-like particle:  $d \gg R$

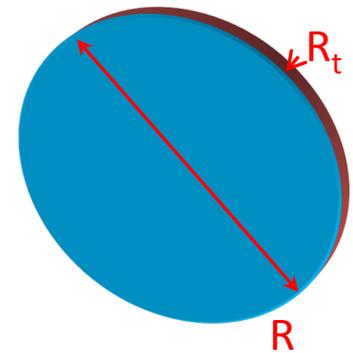
$$qI(q) = I_c(q) = I_c(0) \exp\left(-\frac{R_c^2 q^2}{2}\right)$$

2-d analog of  $R_g$ : cross-section  $R_c^2 = \frac{\int \Delta\rho_c(r) r^2 dr}{\int \Delta\rho_c(r) dr}$

Lamellar/disk/sheet-like particle:  $R \gg R_t$

$$q^2 I(q) = I_t(q) = I_t(0) \exp(-R_t^2 q^2)$$

1-d analog of  $R_g$ : thickness  $R_t^2 = \frac{\int \Delta\rho_t(r) r^2 dr}{\int \Delta\rho_t(r) dr}$



# Forward scattering $I(0)$ measures molecular weight

Atomic apparent form factor  
/ contrast :

$$A_j(q) = f_j(q) - g_j(q)$$

↑  
atomic form  
factor in vacuum
←  
form factor of  
excluded solvent

$$I(q) = \sum_j \sum_k A_j A_k \frac{\sin(qr_{jk})}{qr_{jk}}$$

$$I(0) = \sum_j \sum_k A_j A_k \frac{\sin(qr_{jk})}{qr_{jk}} = \left( \sum_j A_j(0) \right)^2$$

excess electrons

$$\sum_j A_j(0) = \left( \sum_j Z_j - V\rho_s \right) = \left( \sum_j Z_j \right) \left( 1 - \frac{\rho_s}{\rho_m} \right) \propto MW$$

average electron densities of  
solvent( $\rho_s$ ) and the molecule( $\rho_m$ )

$$I(0) \propto C_{(mol/ml)} (MW)^2$$

$$I(0) \propto C_{(mg/ml)} MW$$

$$MW \propto \frac{I(0)}{C_{(mg/ml)}}$$

Solution X-ray scattering uses mass concentration!

# Determination of molecular mass

$$MM_p = I(0)_p / c_p \frac{MM_{st}}{I(0)_{st} / c_{st}}$$

MM or MW: molecular mass/weight

$I(0)$  is the most precise parameter that can be extracted from SAXS data. It is very important that it is free from structure factor effects. Concentration series should be recorded when using  $I(0)$  to extract MW.

$I(0)$  can be calculated using either Guinier fit or  $P(r)$  transform (GNOM).

Concentration has to be known accurately (5-10%). UV absorption works best.

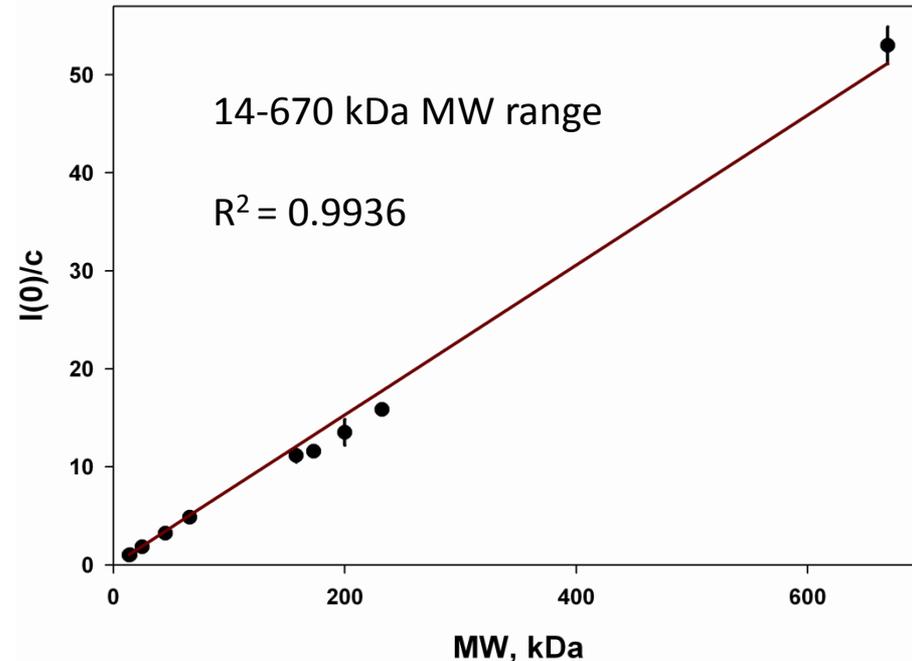
The data for a set of proteins from 14 to 670 kDa are shown.

The method assumes a fixed density for the proteins. Effective partial specific volume was calculated to be  $0.7425 \text{ cm}^3 \text{ g}^{-1}$

The errors in MW with secondary standard (protein) or water calibration **should not exceed 10%**.

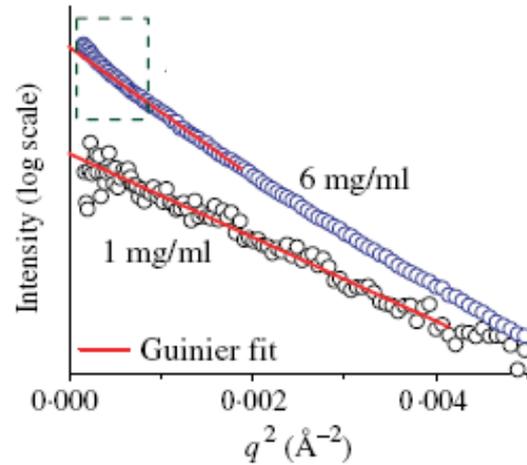
Standard should have **same nature** of the molecules to determine, and with close MW. Using multiple standards are suggested.

[Mylonas, E., Svergun, D. \(2007\) Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. J. Appl. Cryst. 40, s245-s249.](#)



# Molecular weight or $I(0)/c$ for determine aggregation state

Monomer vs dimer:

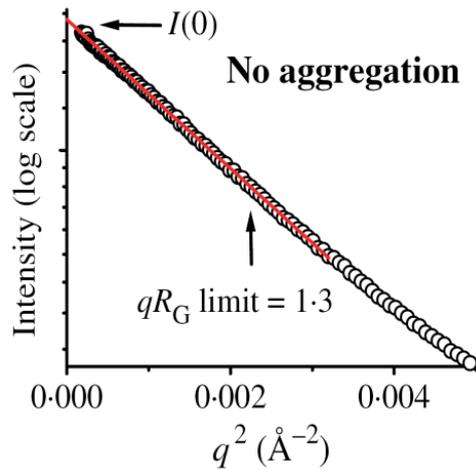


Sample	$R_G$	$I(0)/c$
6 mg/ml	30 $\text{\AA}$	94
1 mg/ml	22 $\text{\AA}$	40

Geometric size of an aggregate grows slower than its mass.  
 $I(0)$  More sensitive than  $R_g$ !

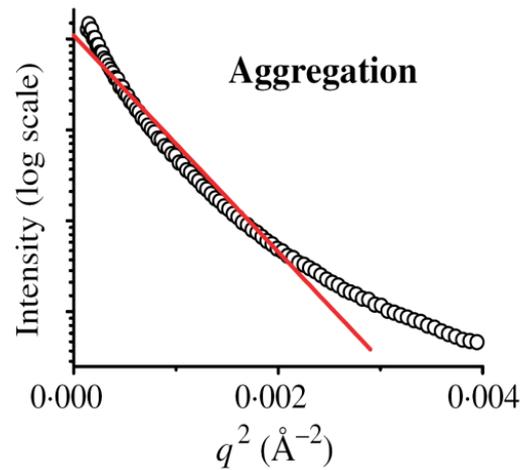
# Guinier Plot: sample dispersion

## Normal / linear



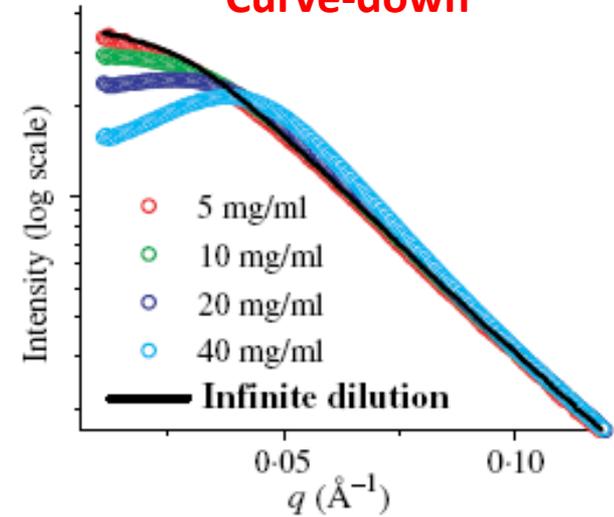
Mono-dispersed

## Curve-up



Poly-dispersed  
aggregates

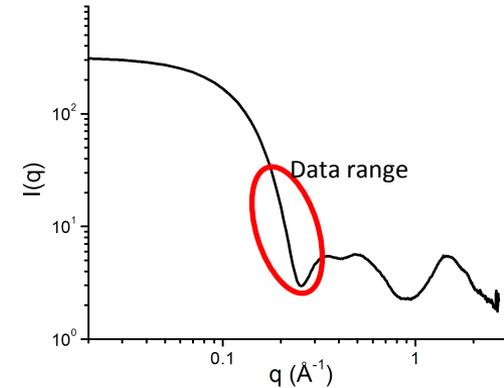
## Curve-down



Repulsion /  
Structure factor

# Porod's law

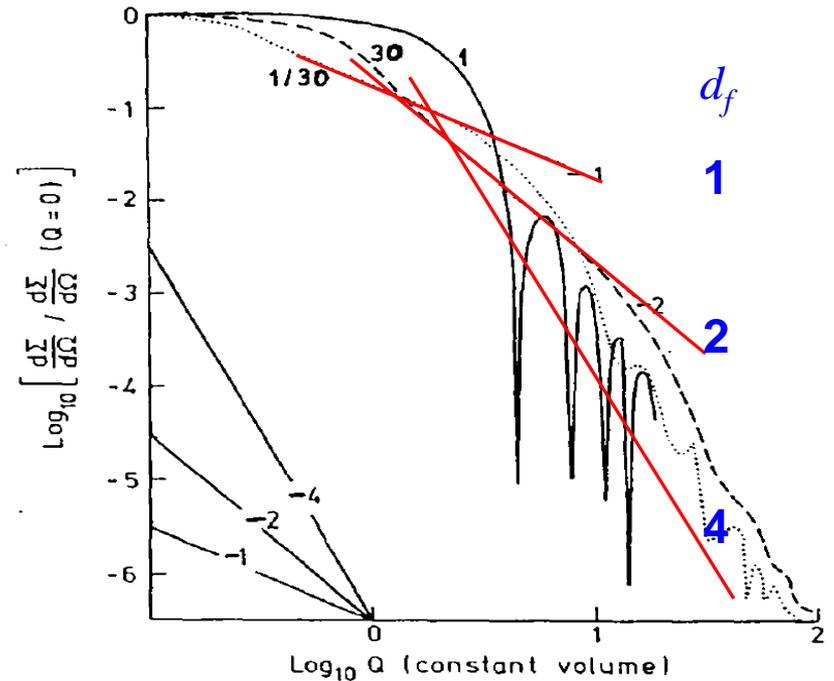
Higher q values contains molecular shape information



$$I(q) \propto q^{-d_f}$$

$d_f$  degree of freedom

$d_f = 1$  rod-like  
 $d_f = 2$  lamellar  
 $d_f = 4$  sphere

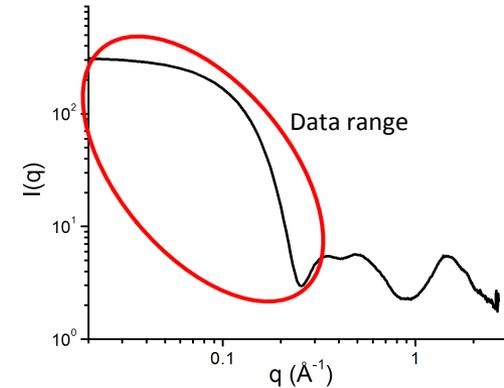


- Assumes uniform density from the scatterer
- Break down when atomic resolution information contribute significantly

# Porod Invariant

$$2\pi^2\gamma(0) = 2\pi^2 \left\langle \int_V \Delta\rho(\bar{u})\Delta\rho(\bar{u} + \bar{r}_{=0})d\bar{u} \right\rangle_{\Omega} \xrightarrow{\text{Homogenous particle}} 2\pi^2 \int_V \Delta\rho(\bar{u})\Delta\rho(\bar{u})d\bar{u} = 2\pi^2(\Delta\rho)^2V$$

$$p(r) = r^2\gamma(r) = \frac{r^2}{2\pi^2} \int_0^\infty q^2 I(q) \frac{\sin qr}{qr} dq \xrightarrow{r=0} \gamma(0) = \frac{1}{2\pi^2} \int_0^\infty q^2 I(q) dq$$



Porod Invariant Q:

$$Q \equiv \int_0^\infty q^2 I(q) dq = 2\pi^2 (\Delta\rho)^2 V$$

$$I(0) = (\Delta\rho V)^2$$

$$V = \frac{2\pi^2 I(0)}{Q}$$

The integration of  $q^2 I(q)$  over all  $q$  range is a constant ( $Q$ ), which only depends on the property of the molecule under study.

Calculation of particle volume does not require absolute data scaling. The accuracy of the derived volume varies depending on the shape and s/n of the data and is a “soft” number. Quality deteriorates above  $q_{\max} \sim 0.2 \text{ \AA}^{-1}$ . Very inaccurate for highly asymmetric particles.

# Pair distance distribution function

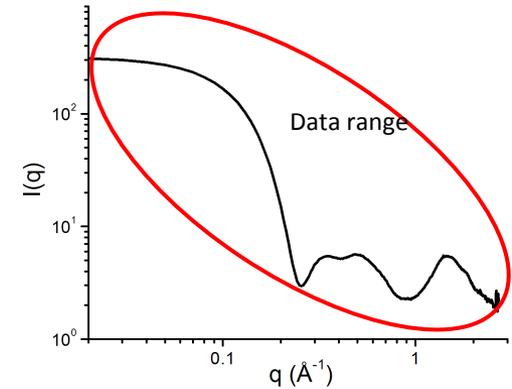
Scattering data encodes structural information/pair distances, let's find out the pair distances from the scattering data:

X-ray scattering amplitude  
of an object in solution:

$$A(\vec{q}) = \int_V \Delta\rho(\vec{r}) \exp(i\vec{q} \cdot \vec{r}) d\vec{r}$$

$$\Delta\rho(\vec{r}) = \rho(\vec{r}) - \rho_s$$

(contrast)



$$I(q) = \langle A(\vec{q}) A^*(\vec{q}) \rangle_{\Omega} = \left\langle \int_V \int_V \Delta\rho(\vec{r}) \Delta\rho(\vec{r}') \exp(i\vec{q} \cdot (\vec{r} - \vec{r}')) d\vec{r} d\vec{r}' \right\rangle_{\Omega} = 4\pi \int_0^{D_{\max}} r^2 \gamma(r) \frac{\sin qr}{qr} dr$$

autocorrelation function

$$\gamma(r) = \left\langle \int_V \Delta\rho(\vec{u}) \Delta\rho(\vec{u} + \vec{r}) d\vec{u} \right\rangle_{\Omega}$$

Pair distance distribution function(PDDF/PDF/p(r)):

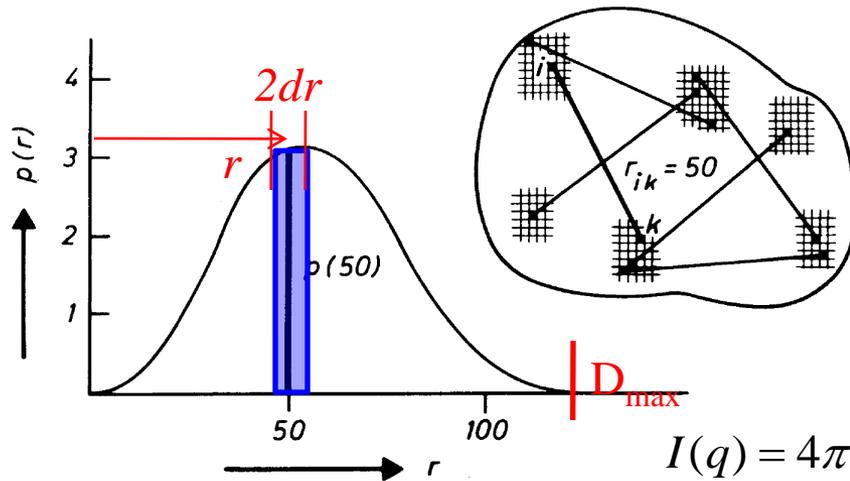
$$p(r) \equiv r^2 \gamma(r) = \frac{r^2}{2\pi^2} \int_0^{\infty} q^2 I(q) \frac{\sin qr}{qr} dq$$

$$I(q) = 4\pi \int_0^{D_{\max}} p(r) \frac{\sin qr}{qr} dr$$

$p(r)$  and  $I(q)$  linked by Fourier transform!

# Pair distance distribution function

$$p(r) = \frac{r^2}{2\pi^2} \int_0^\infty q^2 I(q) \frac{\sin qr}{qr} dq$$



$$p(r) \sim \sum_{|\bar{r}_j - \bar{r}_k| < r+dr} 1 \times \Delta n(\bar{r}_j) \times \Delta n(\bar{r}_k) \times r^2$$

excess electrons of atom  $j$  over solvent

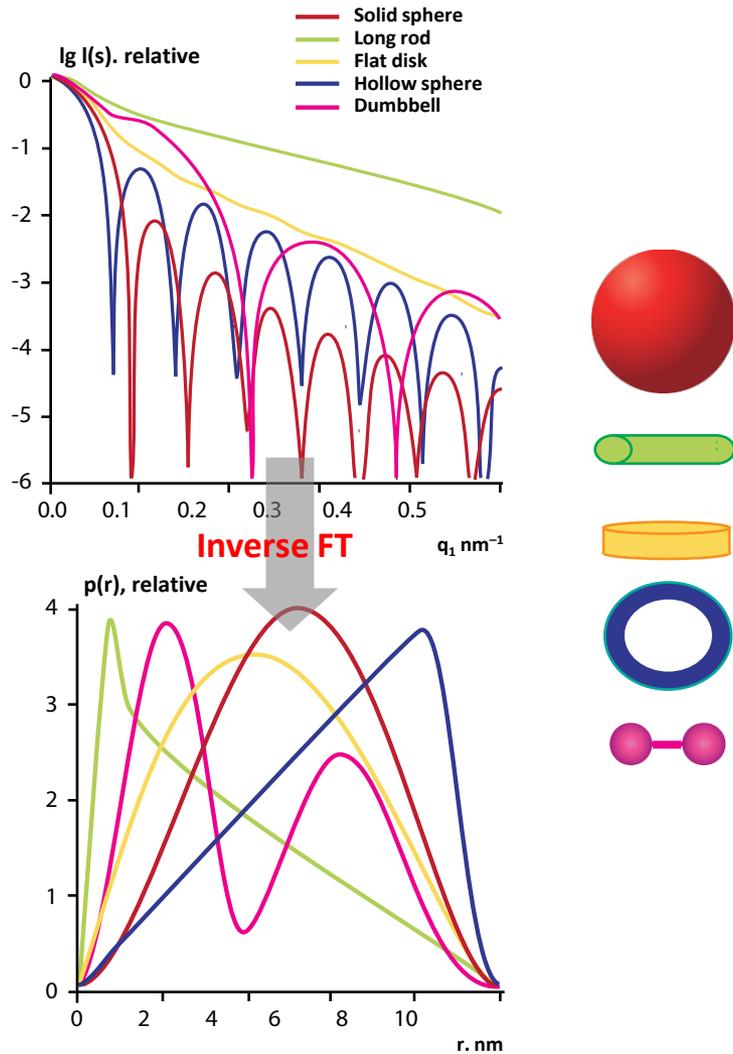
$$I(q) = 4\pi \int_0^{D_{\max}} p(r) \frac{\sin qr}{qr} dr$$

$$I(0) = 4\pi \int_0^{D_{\max}} p(r) dr$$

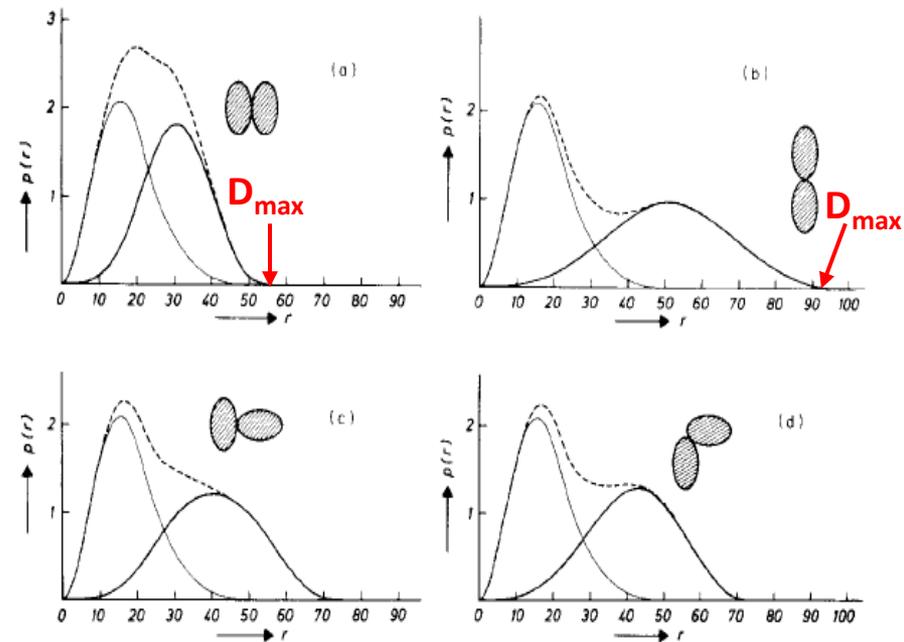
- The PDDF of a molecule is the (net-charges and distance) weighted atom-pair distance histogram.
- Can be used to determine  $D_{\max}$ ,  $I(0)$ , shape, etc.

# PDDF of various shapes

Both scattering profile and PDDF may be characteristic for shape/conformation determination.  
Objects with similar views in the reciprocal space may be very different in the view of real space, visa versa.



## dimer with various conformations



The various dimeric conformations change  $p(r)$  and  $D_{\max}$  as well.

## Calculate pair distance distribution function / $p(r)$ from SAXS

Obtaining PDDF from SAXS is an ill-condition problem because limited  $q$  range.  $p(r)$  calculated from direct integration is severely distorted by the  $q$  truncation.

$$p(r) = \frac{r^2}{2\pi^2} \int_0^\infty q^2 I(q) \frac{\sin qr}{qr} dq$$

Program GNOM is an indirect Fourier transform program with perceptual criteria: for example: smoothness, stability, absence of systematic deviations, etc.

The default parameters optimized for globular shape.

# GNOM

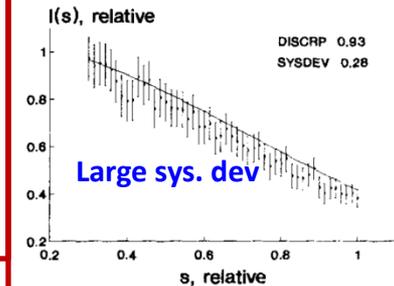
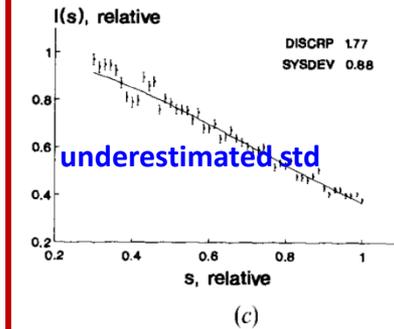
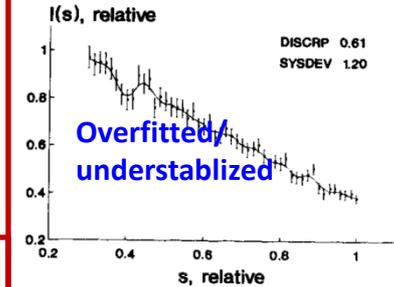
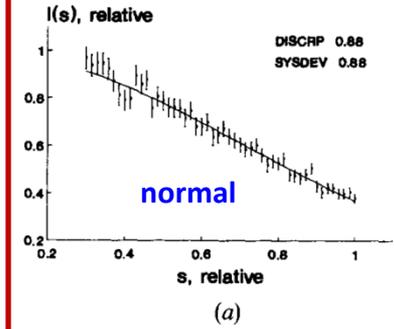
Principles of operation:

1. Minimal oscillations in the fitted  $P(r)$
2. Non-negativity of the  $P(r)$ .
3. Proper shape for the central part of  $P(r)$ .
4.  $p(0)=0$  and  $p(D_{max})\sim 0$ .
5. Highest stability of the regularized solution
6. Good data fit quality
7. Minimal systematic deviations of the  $I(q)$  fit  
(highest number of sequential residuals changing sign)

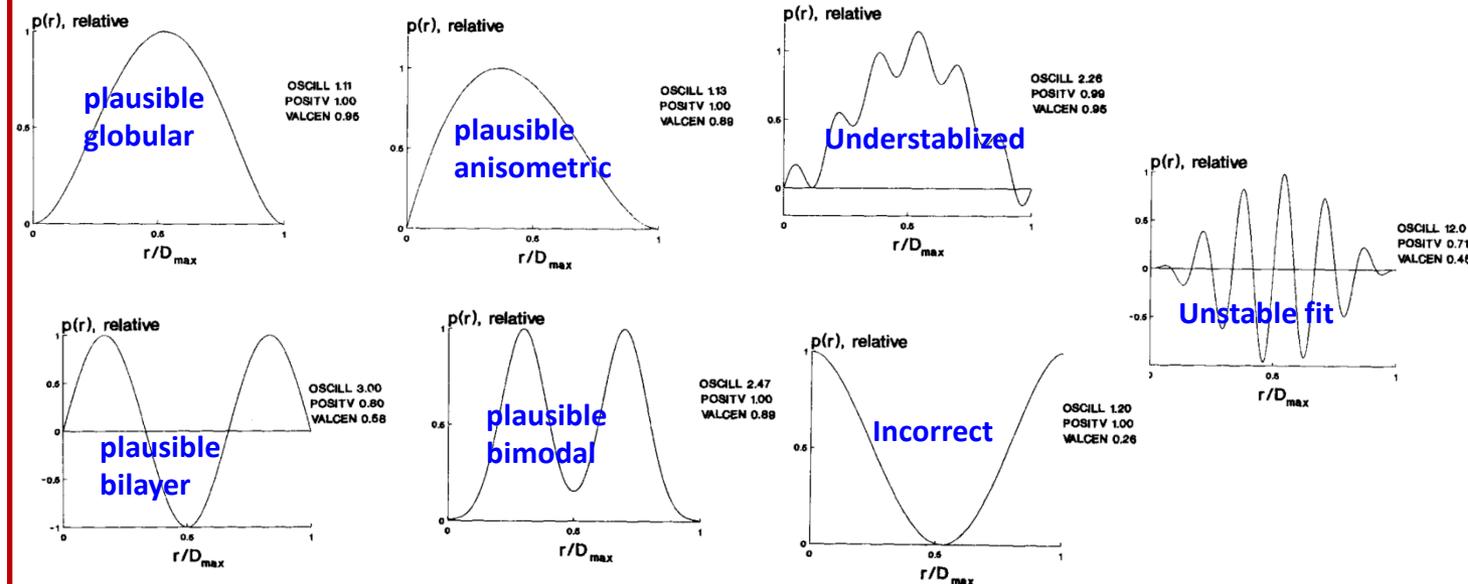
Information obtained from fit:

1. PDDF
2.  $D_{max}$
3.  $R_g$
4.  $I(0)$
5. Data extrapolation  $q \rightarrow 0$

## SAXS fitting



## PDDF profiles



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

# GNOM operation

Working directory: C:\Documents and Settings\Zuox\Desktop\workshop\demo\_data\gnom\  
nom\  
File to be opened: test\_30.dat

```
Output file          [ gnom.out    ] : gtest30.out
No of start points to skip [ 0      ] :
Run title: 02.16440.01489
Number of points in the run is 100
Input data, second file [ none    ] :
No of end points to omit [ 0      ] :
Total number of input data points read is 100
Angular range as read: from 0.00300 to 0.30000
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) : 150
Kernel-storage file name [ kern.bin ] :
Experimental setup (0/1/2) [ 0     ] :
Evaluating design matrix. Please wait...
```

Using 0 for Synchrotron data

Dmax

Evaluating stabilizer matrix. Please wait ...

The measure of inconsistency AN1 equals to 0.1657E+00

Alpha	Discrp	Oscill	Stabil	Sysdev	Positv	Valcen	Total
0.1597E+02	0.6934	1.1855	0.0086	0.3434	1.0000	0.8706	0.71410

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	0.693	1.186	0.009	0.343	1.000	0.871

Estimate 1.000 0.980 0.995 0.000 1.000 0.646

Angular range : from 0.0030 to 0.3000  
Real space range : from 0.00 to 150.00

Highest ALPHA (theor) : 0.416E+04 JOB = 0  
Current ALPHA : 0.160E+02 Rg : 0.518E+02 I(0) : 0.211E+01

Total estimate : 0.714 which is A REASONABLE solution

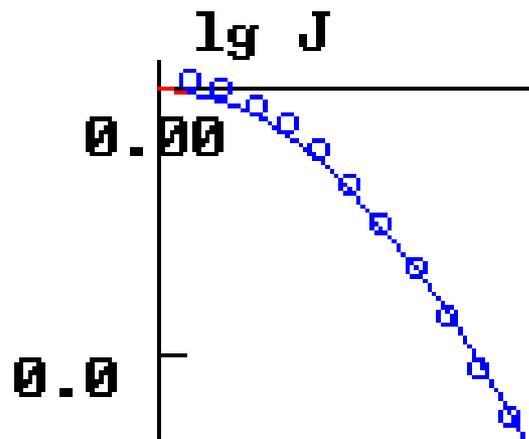
=== Select one of the following options ===

```
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 :

Default parameters are optimized for globular shape.

Screen capture from GNOM program

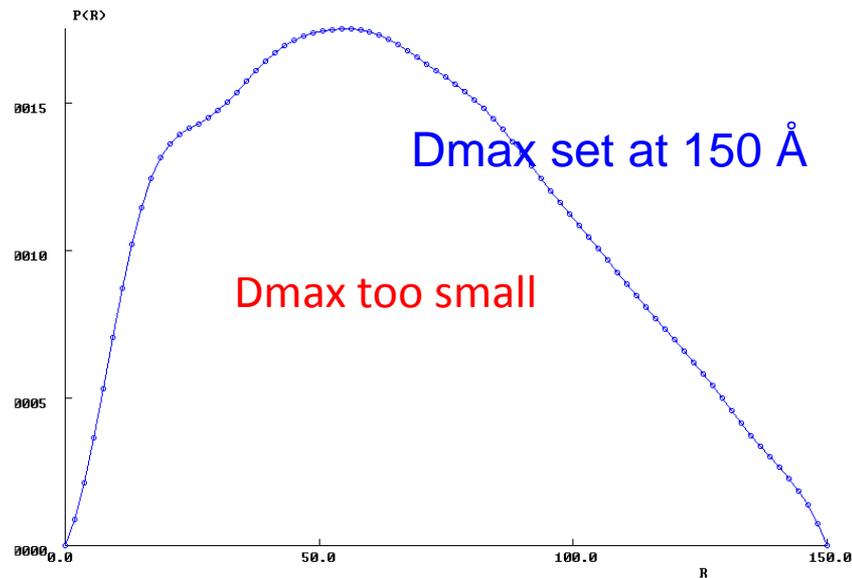
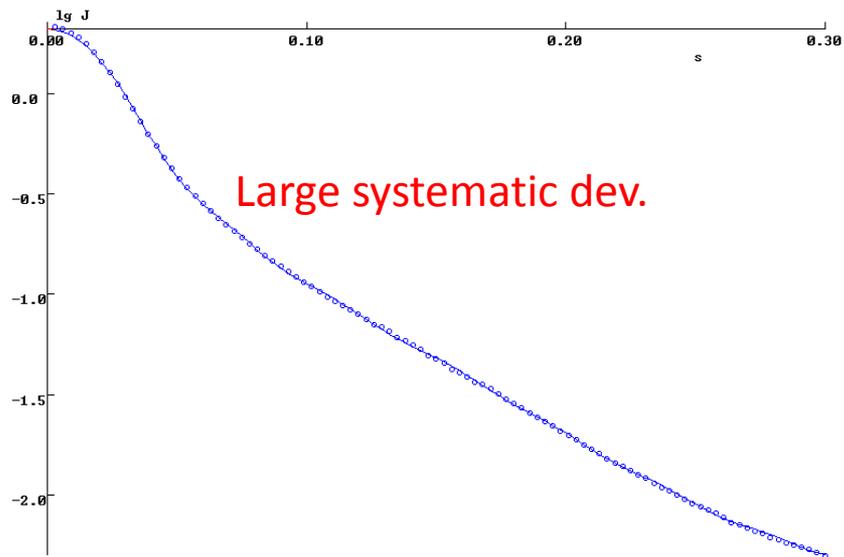


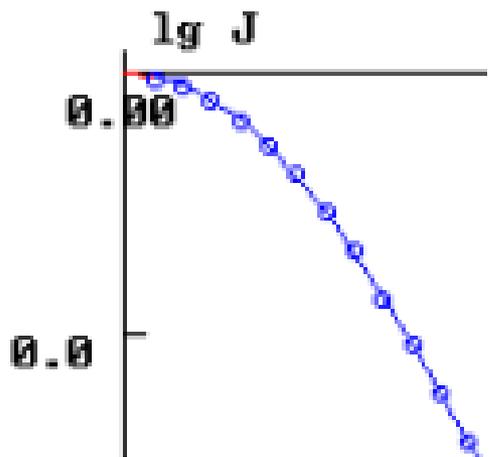
incorrect PDDF fit

$D_{\max}$  is under-estimated

Input file(s) : test\_30.dat \*\*\* JOB = 0  
 Reciprocal space: Rg = 51.81 ,  $\langle I(0) \rangle = 0.2106E+01$

Input file(s) : test\_30.dat \*\*\* JOB = 0  
 Real space: Rg = 51.89 ,  $\langle I(0) \rangle = 0.2106E+01$





incorrect PDDF fit

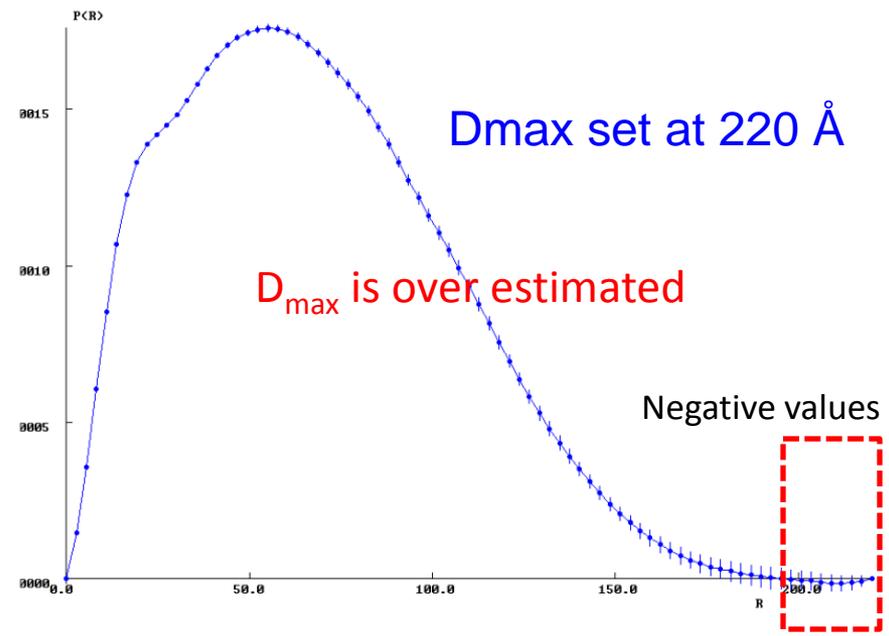
$D_{\max}$  is over-estimated

Input file(s) : test\_30.dat \*\*\* JOB = 0  
 Reciprocal space:  $R_g = 54.19$  ,  $\langle I \rangle = 0.2187E+01$

Input file(s) : test\_30.dat \*\*\* JOB = 0  
 Real space:  $R_g = 54.34 \pm 0.814$  ,  $\langle I \rangle = 0.2187E+01 \pm 0.1725E-01$



SAXS data fitting looks ok



$D_{\max}$  set at 220 Å

$D_{\max}$  is over estimated

Negative values

# Obtaining correct PDDF fit

Input file(s) : test\_30.dat \*\*\* JOB = 0  
 Reciprocal space: Rg = 54.23 I(0) = 0.2188E+01

```

Next data set      (Yes/No/Same) [ y      ] :
Input data, first file [ test_30.dat ] :
Output file        [ gtest.out  ] :
No of start points to skip [ 0      ] :
Run title: 02.16440.01489
Number of points in the run is 100
Input data, second file [ none    ] :
No of end points to omit [ 0      ] :
Total number of input data points read is 100
Angular range as read: from 0.00300 to 0.30000
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) [ 190    ] :
Kernel-storage file name [ kern.bin ] :
Experimental setup (0/1/2) [ 0     ] :
Evaluating design matrix. Please wait...
    
```

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	0.203	1.447	0.001	0.889	1.000	0.835
-----						
Estimate	0.064	0.715	1.000	0.424	1.000	0.399

```

Angular range : from 0.0030 to 0.3000
Real space range : from 0.00 to 190.00
    
```

```

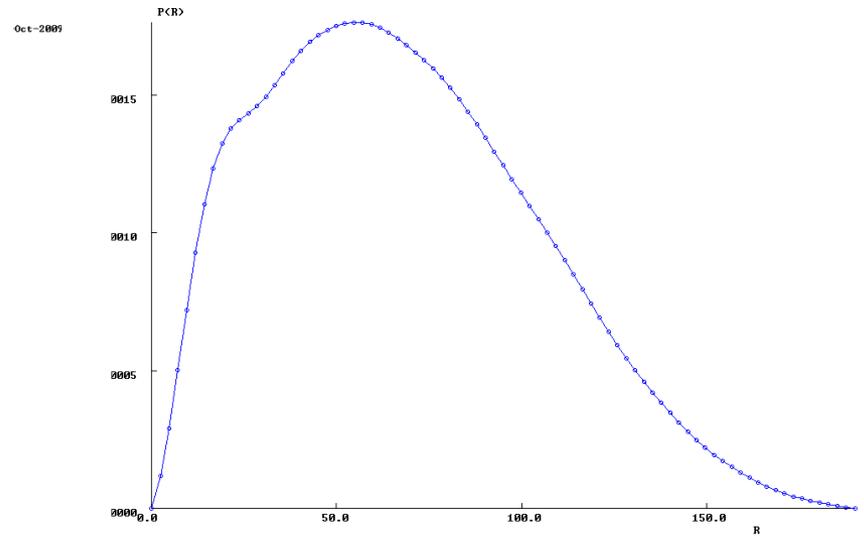
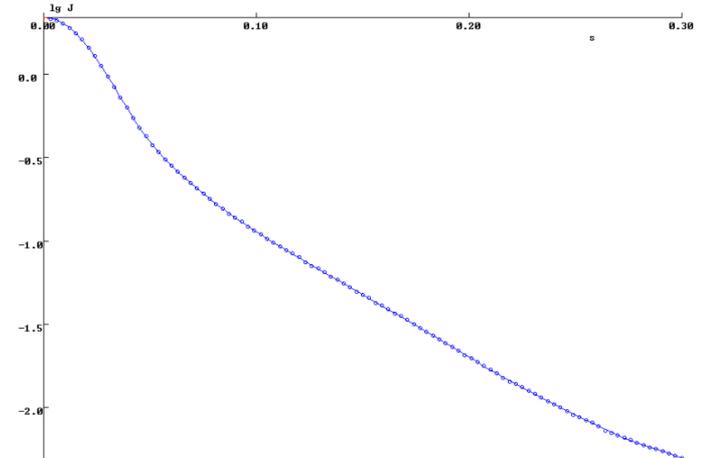
Highest ALPHA (theor) : 0.506E+04          JOB = 0
Current ALPHA         : 0.100E+01  Rg : 0.543E+02  I(0) : 0.219E+01
    
```

Total estimate : 0.657 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 :



$D_{max}$  error: 5~10% of  $D_{max}$

Put  $p(r_{max})=0$  "No" first, give  $r_{max}$  a number in range  $2R_g-6R_g$ , and see where  $p(r_{max})$  ends. Then change  $r_{max}$ . Alpha is another very important parameter affects the fitting, tune alpha to have a stable fit. Alpha should not be too small ( $>0.1$ )

# $R_g$ , $I(0)$ and data extrapolation

Reciprocal space:  $R_g = 56.77$  ,  $I(0) = 0.2265E+01$   
Real space:  $R_g = 57.07 \pm 0.862$   $I(0) = 0.2266E+01 \pm 0.2212E-01$

Reciprocal space parameters:  
Directly obtained from Guinier plot

Real space parameters:

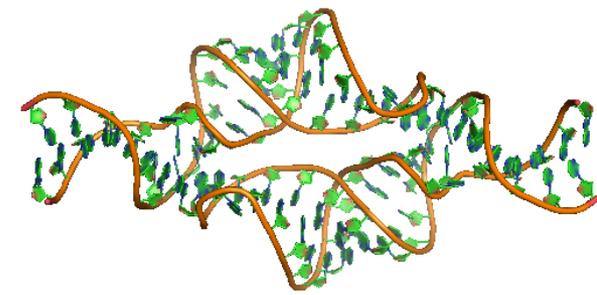
$$I(0) = 4\pi \int_0^{D_{\max}} p(r) dr$$

$R_g$  also calculated from  $p(r)$ , may differ from that calculated from Guinier plot

## data extrapolation

S	J EXP	ERROR	J REG	I REG	
0.0000E+00				0.2266E+01	} Extrapolated data
0.5569E-03				0.2265E+01	
0.1114E-02				0.2263E+01	
0.1671E-02				0.2259E+01	
0.2228E-02				0.2253E+01	
0.2784E-02				0.2247E+01	
...					
0.1058E-01				0.2010E+01	
0.1114E-01				0.1985E+01	
0.1169E-01				0.1959E+01	
0.1225E-01				0.1932E+01	
0.1281E-01				0.1904E+01	
0.1337E-01				0.1875E+01	
0.1392E-01				0.1846E+01	
0.1448E-01				0.1817E+01	
0.1504E-01	0.1797E+01	0.2309E-01	0.1786E+01	0.1786E+01	
0.1559E-01	0.1752E+01	0.2324E-01	0.1756E+01	0.1756E+01	
0.1615E-01	0.1729E+01	0.2264E-01	0.1725E+01	0.1725E+01	
0.1671E-01	0.1693E+01	0.1664E-01	0.1693E+01	0.1693E+01	
0.1726E-01	0.1658E+01	0.1873E-01	0.1662E+01	0.1662E+01	
...					

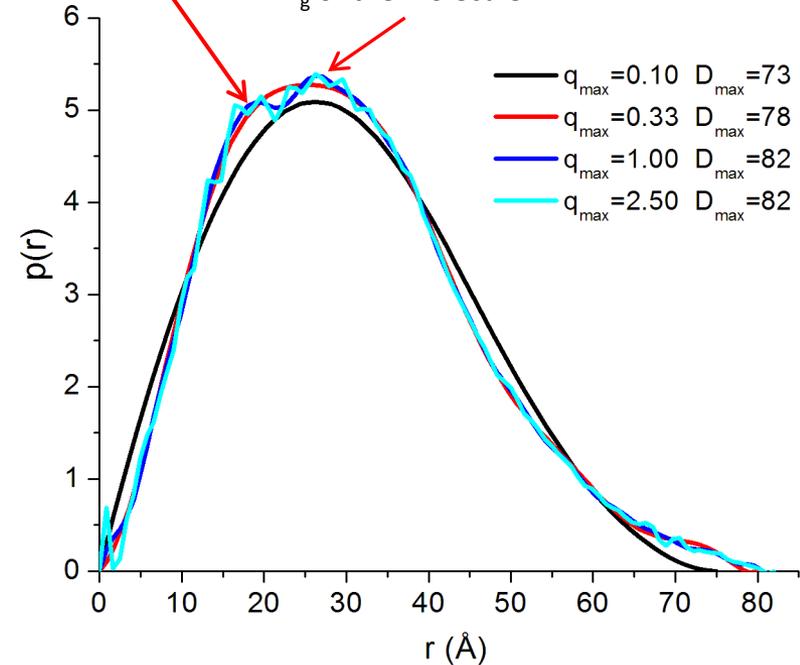
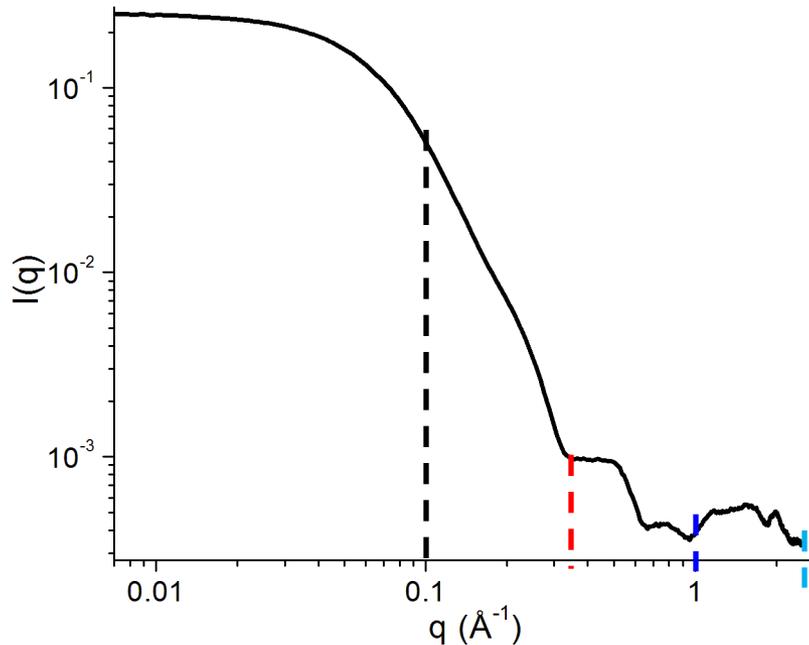
# PDDF at various resolutions



Diameter of A-RNA duplex

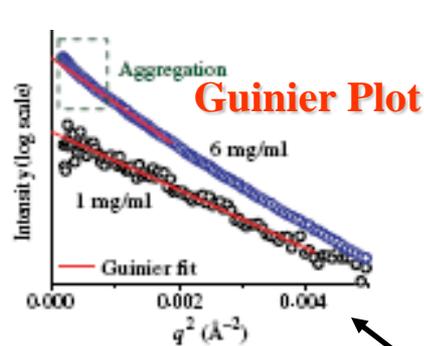
$R_g$  of the molecule

tectoRNA dimer

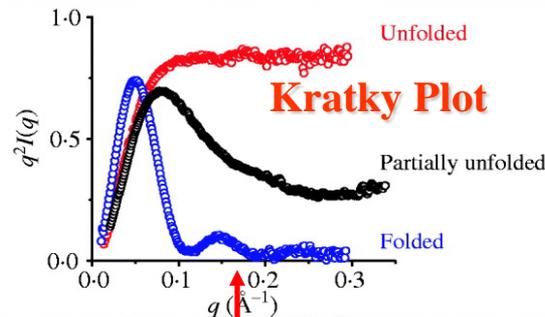


Pddf/ $p(r)$  obtained from certain  $q$ -range reflects the resolution defined by  $q_{\max}$ .  $D_{\max}$  may also be affected. High enough  $q_{\max}$  needed for PDDF used in structure/conformation analysis.

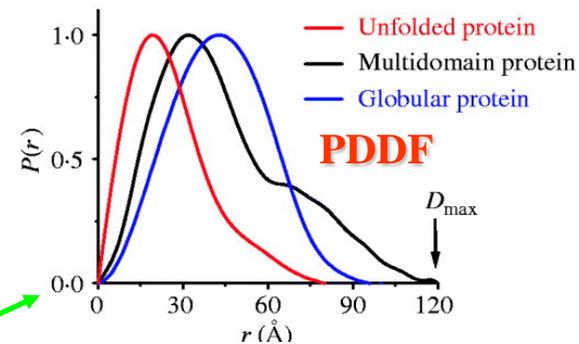
# Other SAXS regions, structural information and application



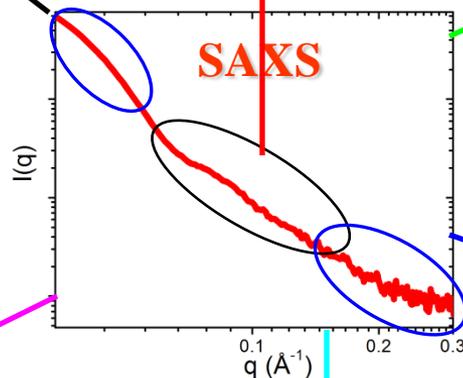
Overall size:  $R_g$   
 Molecular weight:  $I_0$   
 Aggregation, Hydration,  
 Ion distribution, etc



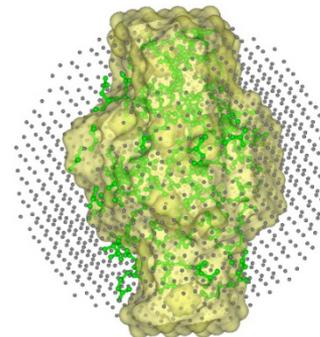
Folded / unfolded conformation



Structural Info in Real Space:  
 $D_{max}$ , Shape,  $R_g$ , etc

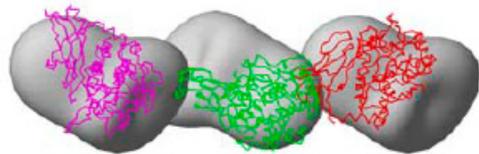


**Bead-/DR-model Structure**



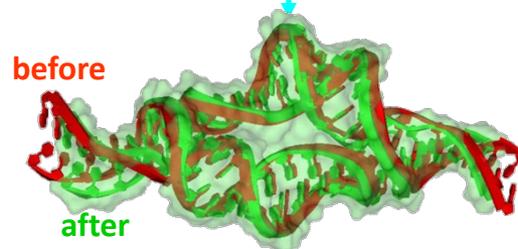
Low Resolution Structures

**Rigid-body Modeling**



Complexes Reconstructed Using  
 sub-units w/ Known Structures

**Global Restraints**



High Resolution Structures with  
 Accurate Global Shape

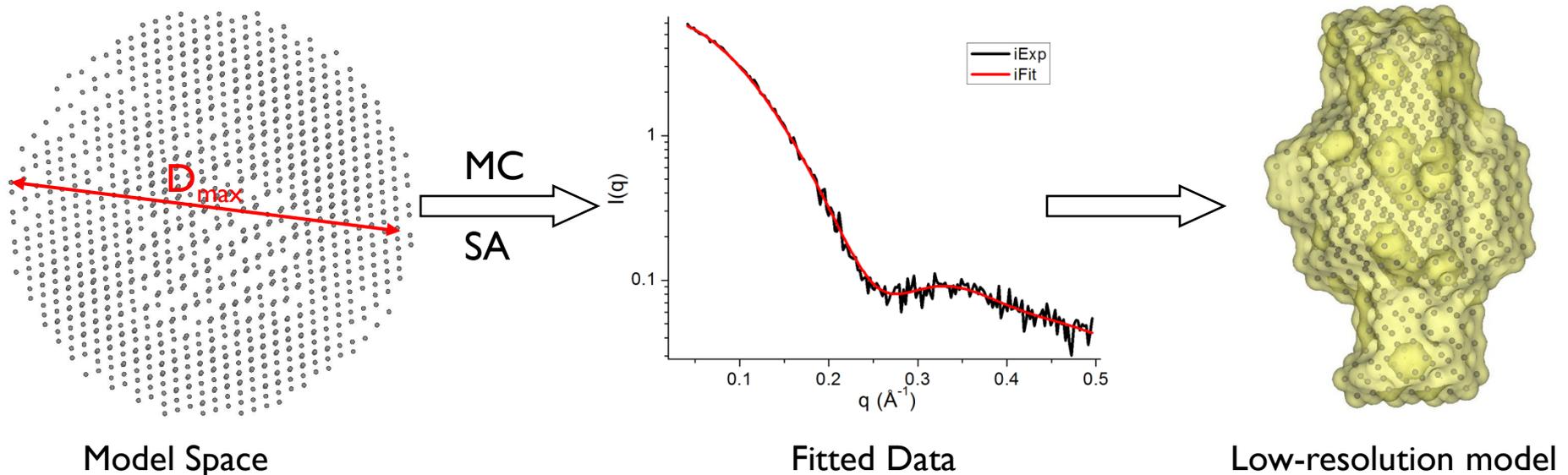
## **2. Molecular envelope and shape**

- Low-resolution model/shape reconstruction
- Molecular conformation
- Contrasting (SAXS and SANS)

# 3D shape reconstructions from SAXS data: a general idea

Obtaining 3D shapes from 1D SAXS data is an ill-defined problem that can be solved by regularizing the fitted models.

Imposing prior restraints on the fitted models such as non-negativity and compactness/connectivity greatly increases solution stability.



## Available Programs:

- Genetic Algorithm: DALAI\_GA (1998)
- Simulated Annealing: DAMMIN (1999), GASBOR (1999)
- Monte Carlo: saxs3d (1999)
- Monte Carlo: LORES (2005)

# Molecular shape reconstructions with DAMMIN

Penalty function:  $f(X) = \chi^2 + \alpha P(X)$

Regularizer P(X) penalizes loose structures

$$P(X) = 1 - \langle C(N_e) \rangle$$

$$C(N_e) = 1 - [\exp(-0.5N_e) - \exp(-0.5N_c)]$$

Fits *ab initio* low-resolution molecular shapes from SAXS/SANS data only.

Points to remember:

Input to DAMMIN = output of GNOM (do not modify the file in any way!).

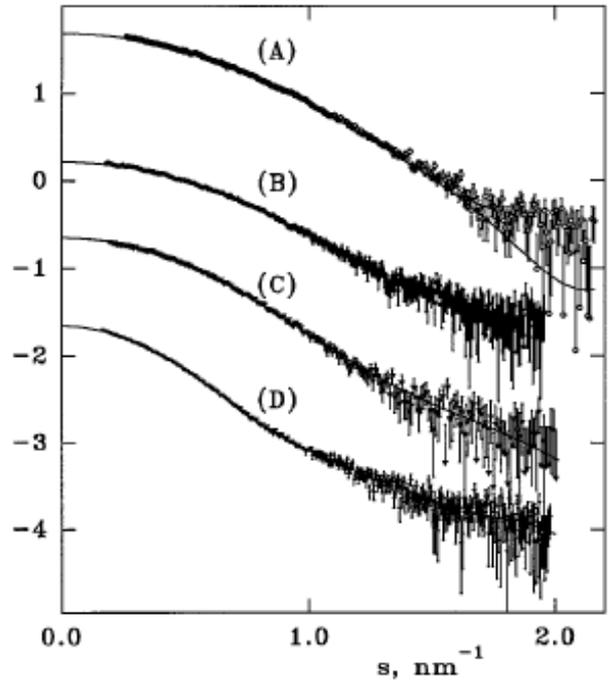
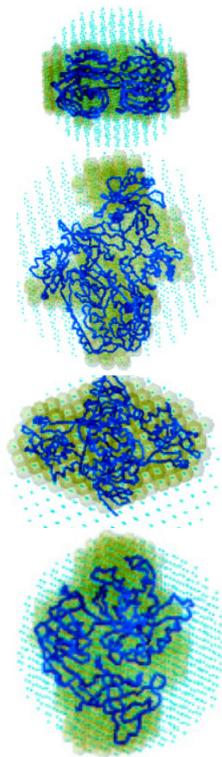
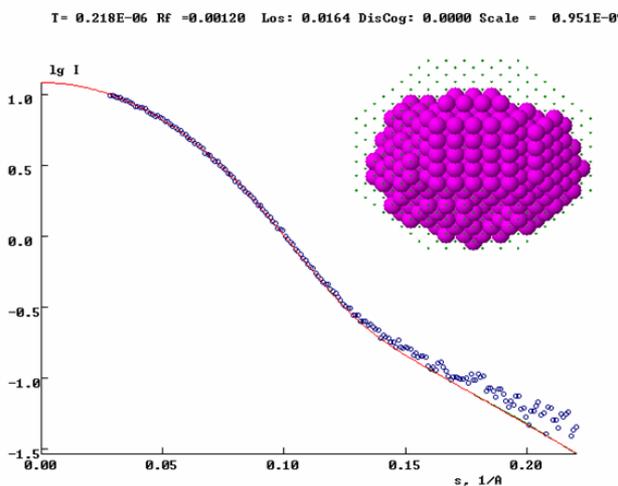
$q_{max} < \sim 0.2-0.3 \text{ \AA}^{-1}$ ,  $p/q_{max}$  = distance uncertainty

Since molecule is modeled with uniform density, a constant is subtracted from  $I(q)$ .

Single phase modeling, not applicable to RNA/protein or DNA-protein complexes, etc.

Point symmetry/particle anisometry and relative orientations of the symmetry and anisometry axes can (*and should*) be specified when known!

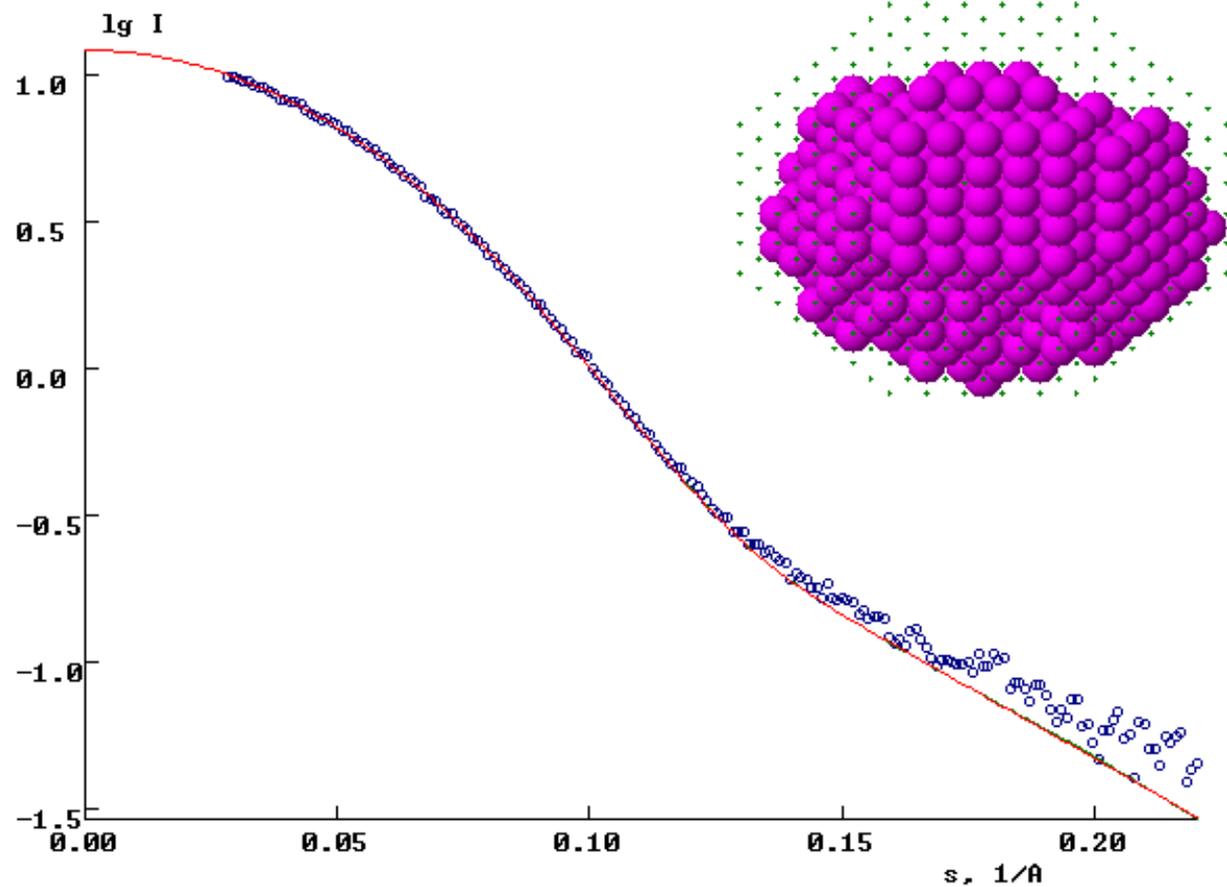
Not applicable to natively unfolded proteins.



Svergun, D. et al. (1999) Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing. Biophys. J. 80, 2946-2953.

# DAMMIN example

T= 0.218E-06 Rf =0.00120 Los: 0.0164 DisCog: 0.0000 Scale = 0.951E-09



15-Aug-2006 17:01:27

Gnom file : 022\_78a.out

## DAMAVER: superimposition and averaging of *ab initio* models

DAMAVER program:

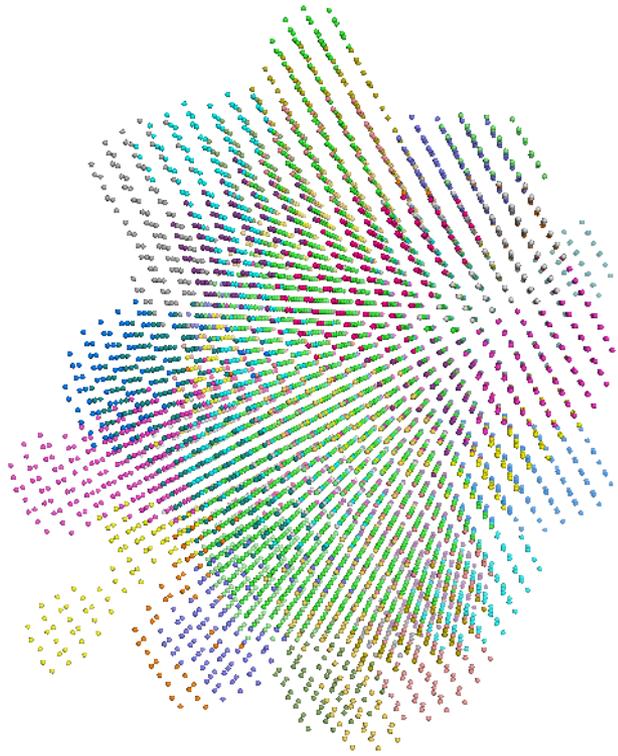
- (1). Align any two models, calculate normalized spatial discrepancy (NSD).
- (2). Choose the one with smallest total NSD to the rest of the models as the reference.
- (3). Align all models onto the reference model.
- (4). Remap beads from all models onto the packed grid, mark the occupancy.
- (5). The grids with non-zero occupancies are chosen to generate a final consensus model with the volume equal to the average excluded volume of all the models.

Important parameters:

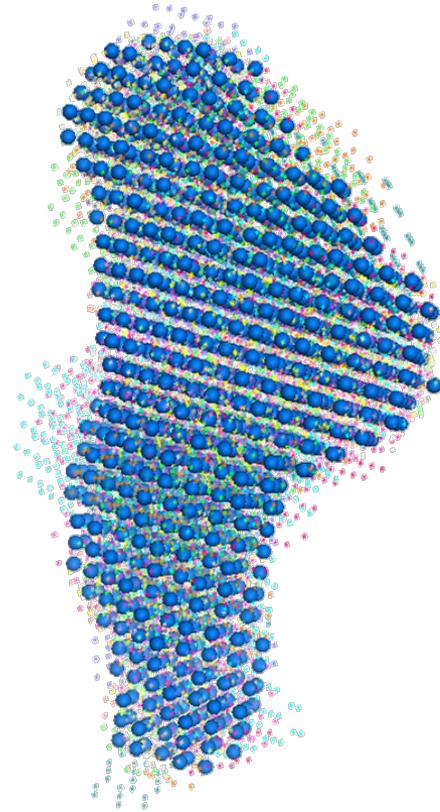
Average NSD: < 0.7 - excellent; ~1.0 - acceptable; > 1.5 - needs attention.

Load all re-aligned model, check consistency. Trust your eyes!

# DAMAVER: example



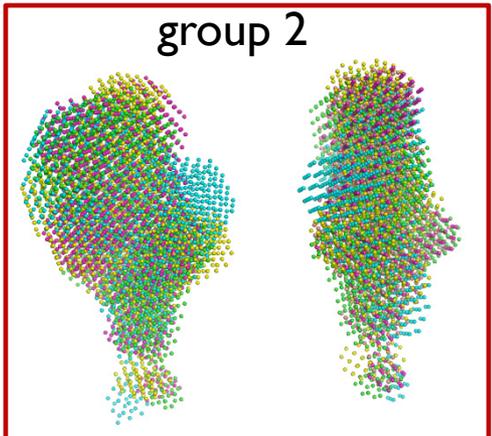
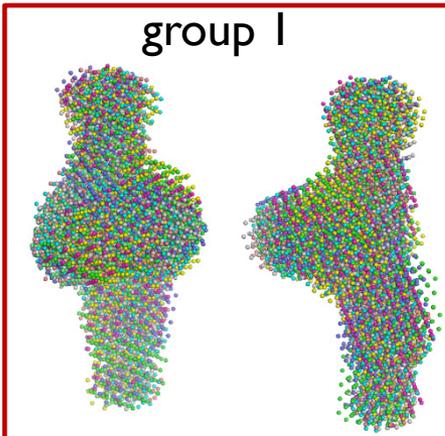
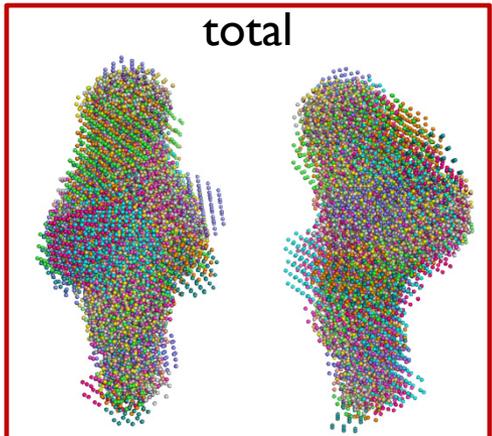
Before alignment and averaging



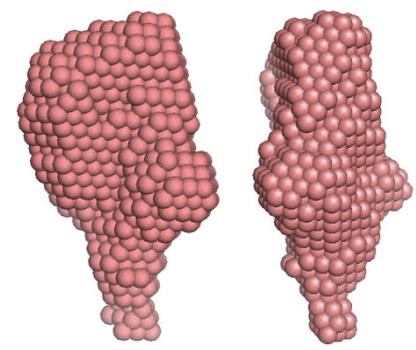
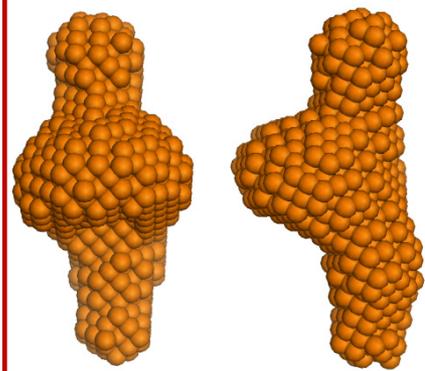
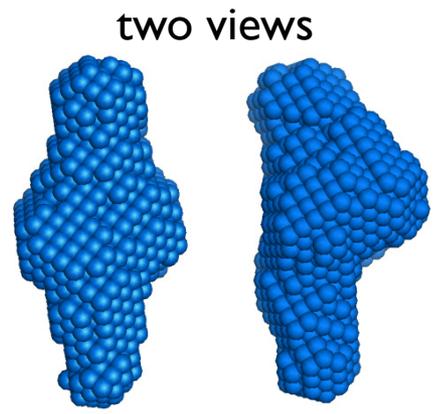
After alignment and averaging

# Dammin ab initio reconstructions: multiple solutions

Bead models



Averaged bead model



NSD population

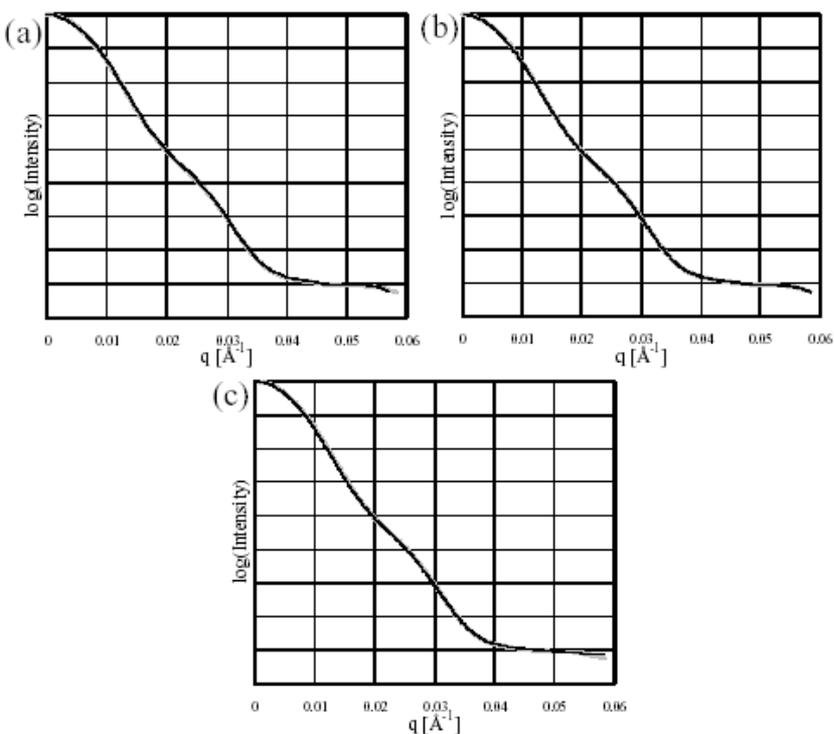
$0.72 \pm 0.08$

$0.51 \pm 0.01$   
2/3

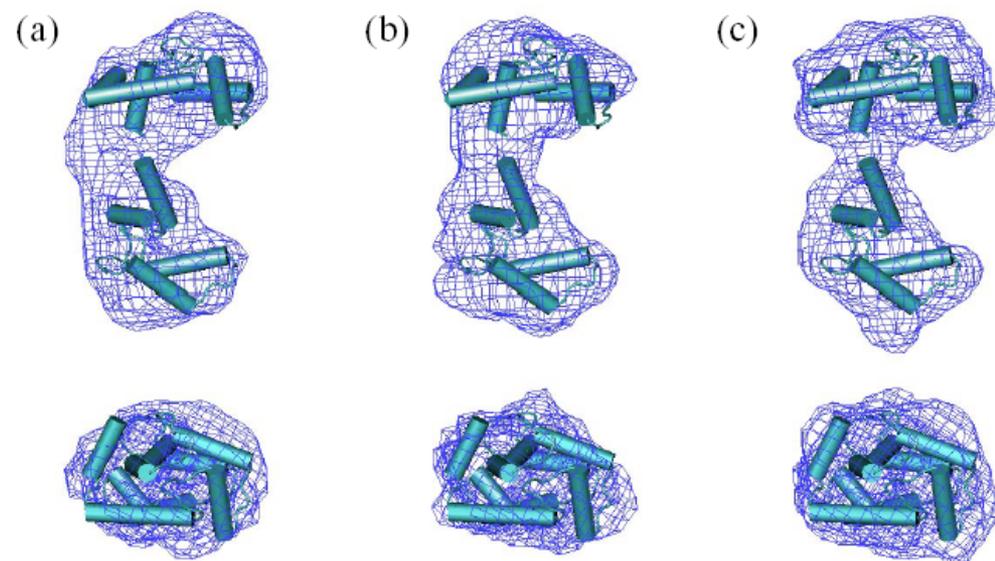
$0.71 \pm 0.07$   
1/3

Solution degeneracy is quite common. Check individual solutions visually, divide them into groups according to shape and average separately. Use prior information to discard those that do not agree with it.

# Reconstruction accuracy – DAMMIN vs DALAI\_GA vs SAXS3D



**Figure 1** The input and fitted scattering curve of troponin-C. Bold grey line corresponds to the smoothed scattering curve from the output file of *GNOM*. And solid line of (a) *dammin*, (b) *dalai\_ga* and (c) *saxs3d* is one of the results for a bead radius of 3 Å, respectively.



**Figure 4** Superposition of envelope from averaged models of troponin-C calculated by three algorithms on 1TNX. Envelope from averaged models by (a) *dammin*, (b) *dalai\_ga* and (c) *saxs3d*, respectively. Wire frame of envelope was created with 3 Å of beads radius using *Situs* program package and cartoon shows 1TNX. The bottom pictures are the view rotated 90° around horizontal axis of upper pictures. They are plotted using program *VMD* (Humphrey *et al.*, 1996). A model of *saxs3d* with single run looked very different from 1TNX but when the models are averaged, *saxs3d* showed the best agreement with 1TNX.

Accuracies are comparable.

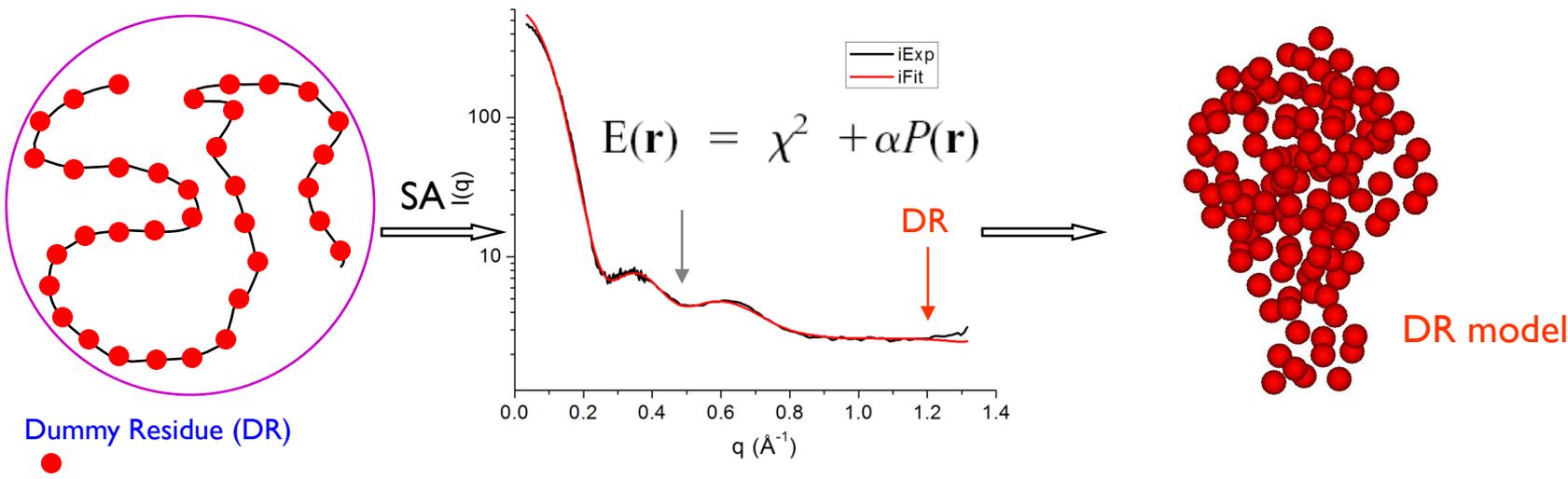
All three methods could be tried but DAMMIN is the most popular by far.

Takahashi, Y. et al. (2003) Evaluation of three algorithms for *ab initio* determination of three-dimensional shape from one-dimensional solution scattering profiles. *J. Appl. Cryst.* 36, 549-552.

DALAI\_GA: Chacon P., et al. (1998) Low resolution structures of proteins in solution retrieved from X-ray scattering with genetic algorithm. *Biophys. J.* 74, 2760-2775.

SAXS3D: Walther D., et al. (2000) Reconstruction of low resolution 3D density maps from 1D SAXS data in solution. *J. Appl. Cryst.* 33, 350-363.

# GASBOR: a dummy-residue approach



GASBOR uses quasi-realistic “average residue” form factor and fits to higher  $q_{\max}$

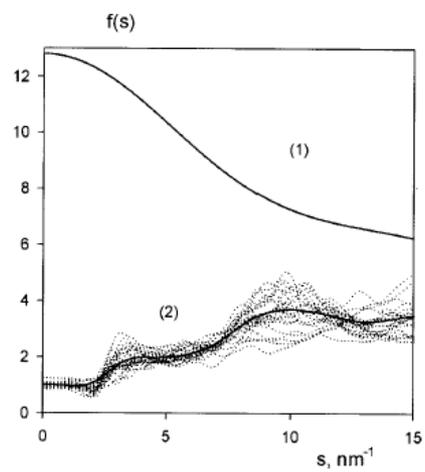
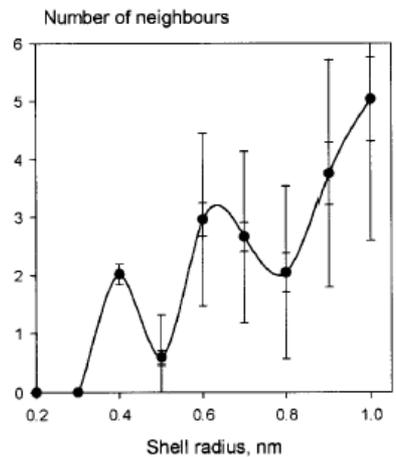


FIGURE 2 Averaged form factor of a residue (1) and the average correction factor (2). Dotted curves represent individual correction functions for the proteins in Fig. 1.

## Regularizer:

$$P(\mathbf{r}) = \sum_k [W(R_k)(N_{DR}(R_k) - \langle N(R_k) \rangle)]^2 + G(\mathbf{r})$$



$$+ [\max\{0, (|\mathbf{r}_c| - r_0)\}]^2$$

FIGURE 3 Histogram of an average number of C<sub>α</sub> atoms in 0.1 nm thick spherical shells around a given C<sub>α</sub> atom. Smaller error bars: variation of the averaged values over all proteins; larger error bars: averaged variation within one protein.

# Low resolution density reconstructions with GASBOR

Points to remember:

Fits *ab initio* low-resolution molecular shapes from SAXS data only.

Input to GASBOR = output of GNOM (do not modify the file in any way!).

Higher  $q_{\max}$  data are required ( $q_{\max}=0.4$  at least)—output models have higher apparent resolution.

Not applicable to multiphase systems.

Number of residues has to be specified, does not perform well for >1000 a.a.

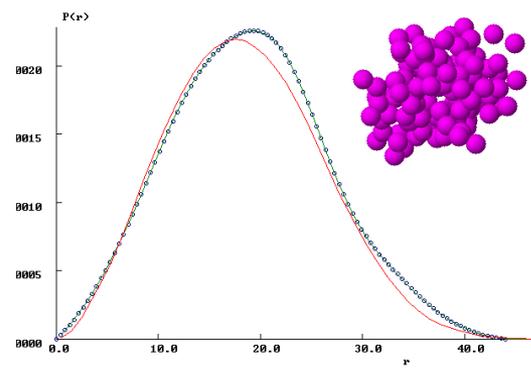
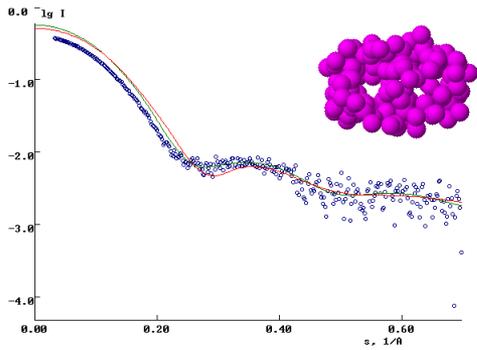
P(r) and I(q) versions exist.

Point symmetry/particle anisometry can (*and should*) be specified!

Not applicable to natively unfolded proteins (pool of models)

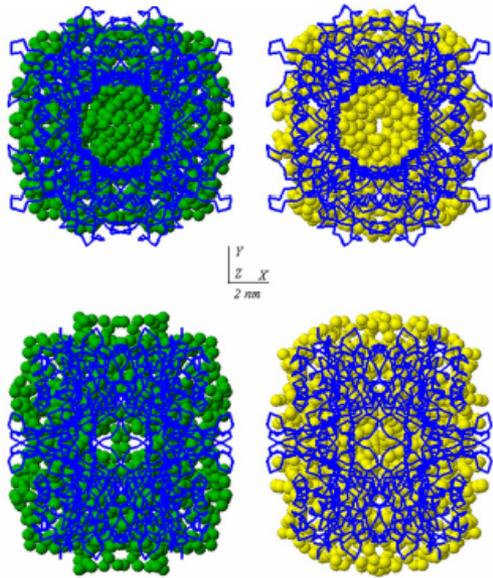
Applicability to RNA is questionable (regularizer is protein-based).

# Real vs. Inverse Space Modules of GASBOR



Log file : C:\01ex\NIHNSAXS\Talks\Ntrecht\_08\tutorial\DataAnalysis\gasbor\junk.log  
26-May-2008 05:41:02

gnom\_lsc\_07.out  
t\_08\tutorial\DataAnalysis\gasbor\junk.log



*urate oxidase:*

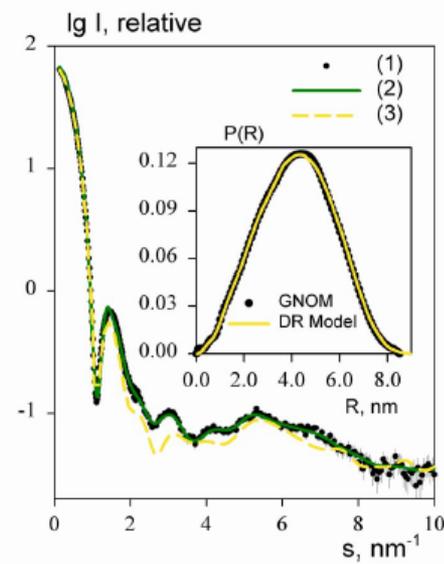


Figure 5

Atomic model of urate oxidase (blue  $C_{\alpha}$ -chain) superimposed with the *ab initio* DR models displayed as spheres. Green and yellow models are obtained by fitting in reciprocal and real space, respectively. The bottom view is rotated by 90° counterclockwise around X axis.

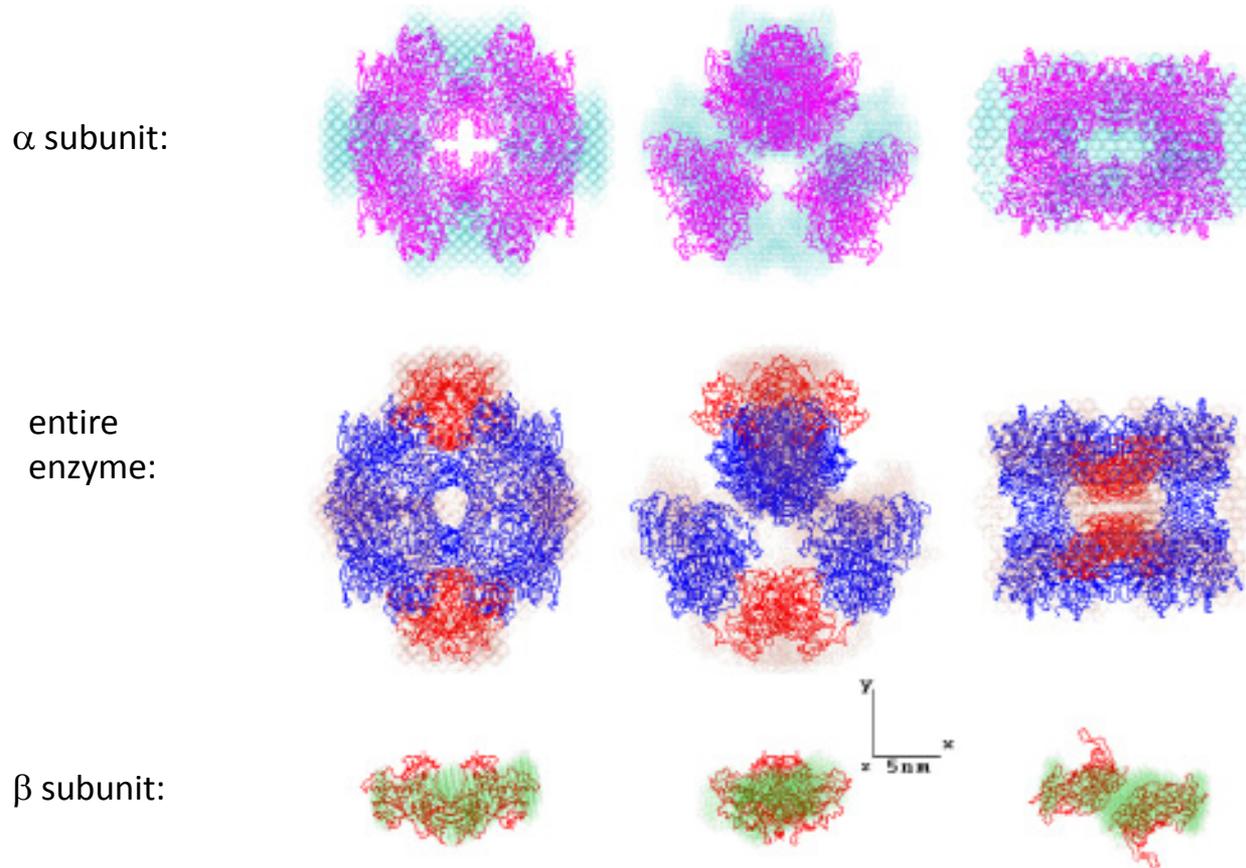
Figure 4

X-ray scattering from urate oxidase (1) and scattering from the DR models: reciprocal space fitting model (2), model obtained by fitting in real space (3). The insert displays the experimental distance distribution function of UOX computed by GNOM (dots) and that of the real space constructed model (full line).

Inverse space mode is ~4 times slower than real space mode.  
 Accuracies are comparable, both should be run.  
 Inverse space fitted models do a better job of reproducing I(q) profiles.

# Reconstruction Accuracy – GASBOR

Glts homoenzyme:



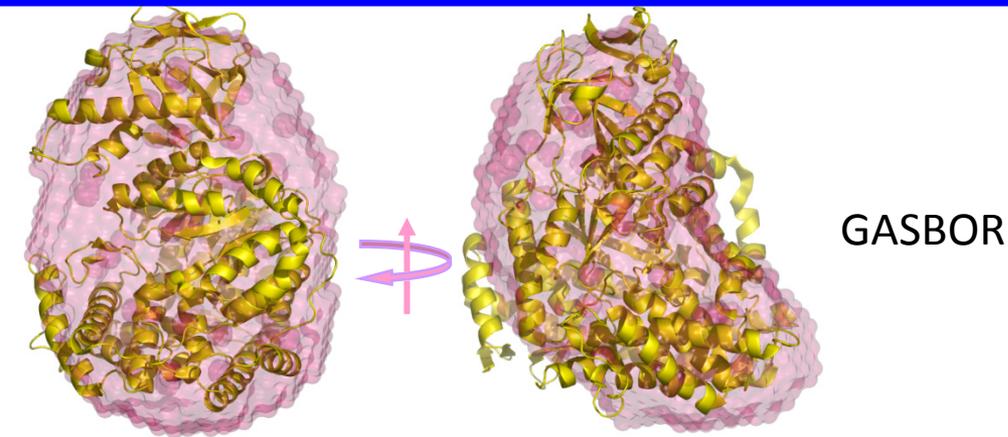
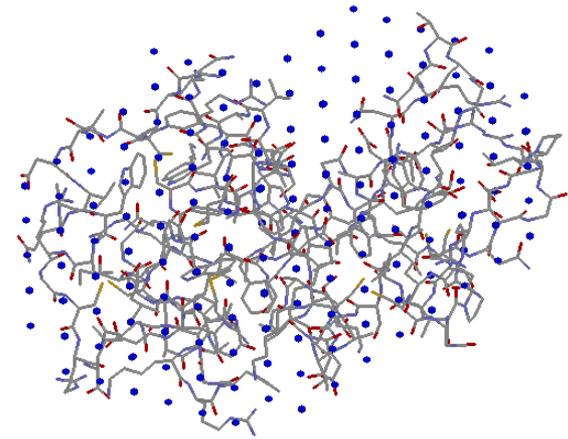
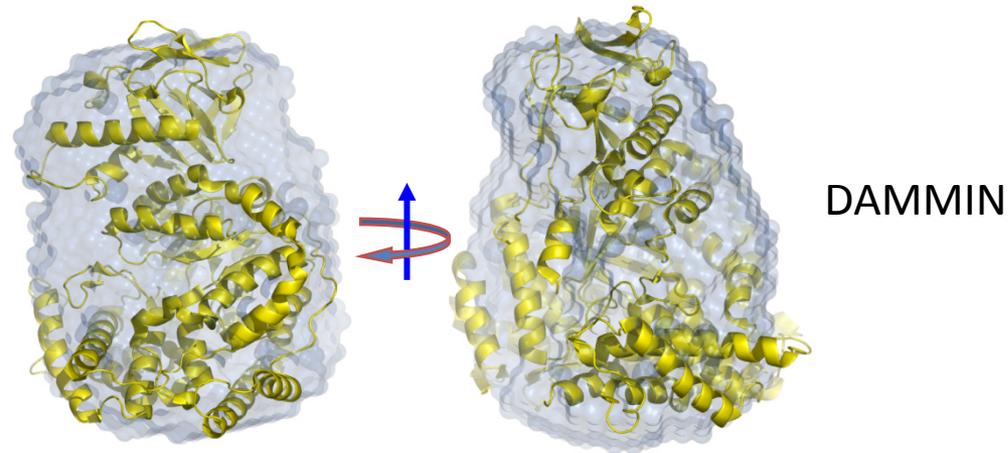
Superposition of the Ca traces with low-resolution reconstructions shows that agreement can be far from quantitative. Fitting high-resolution structures to low-res reconstructions is much easier than finding precise positions of the individual domains within the reconstructions.

# Reconstruction accuracy – DAMMIN vs GASBOR

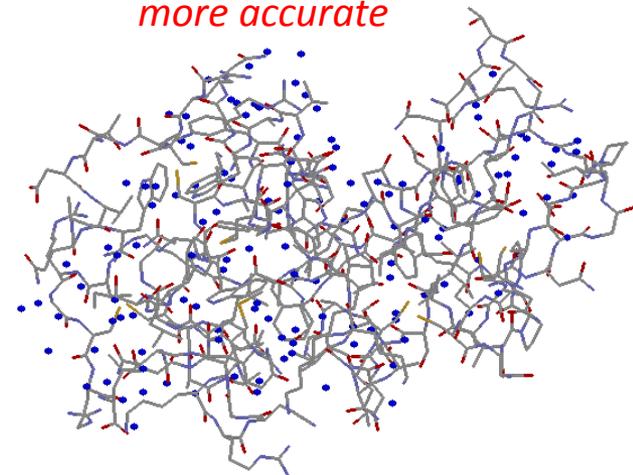
*MSG*

*Lysozyme*

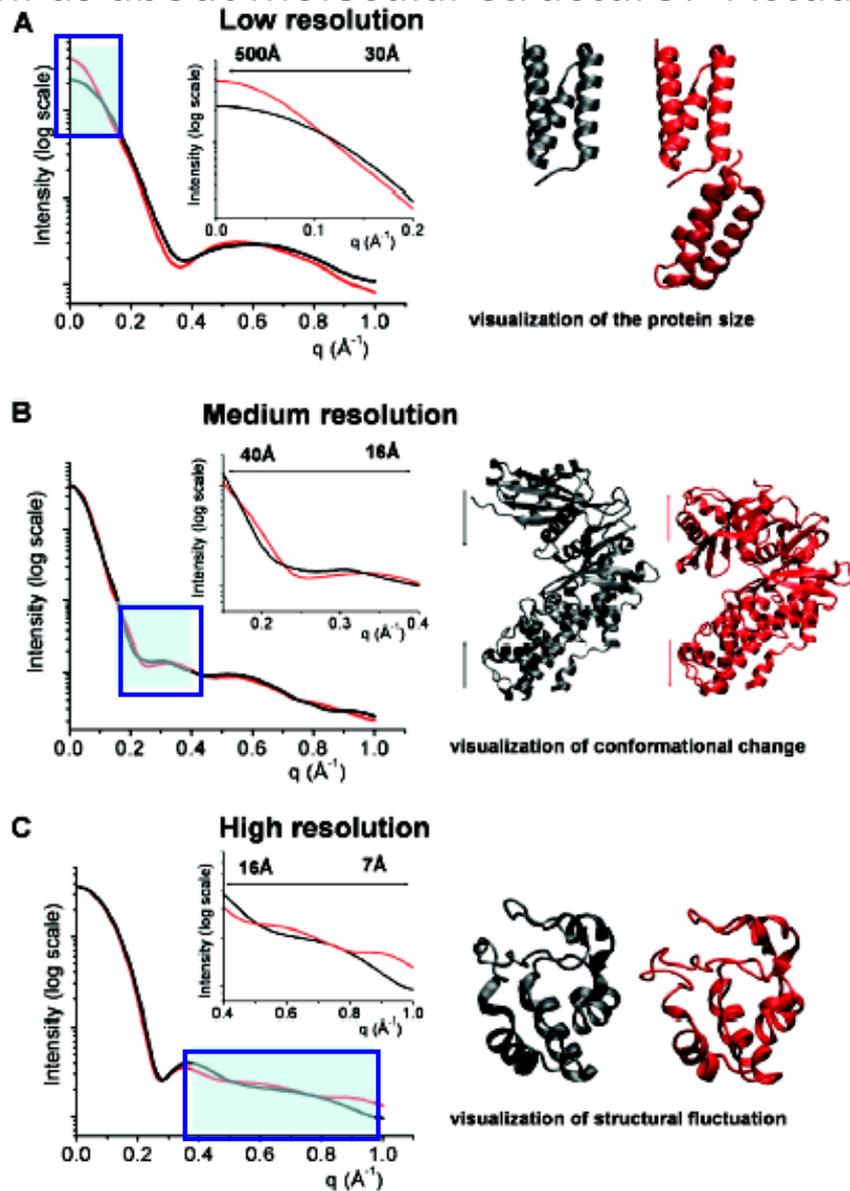
*more accurate*



*more accurate*

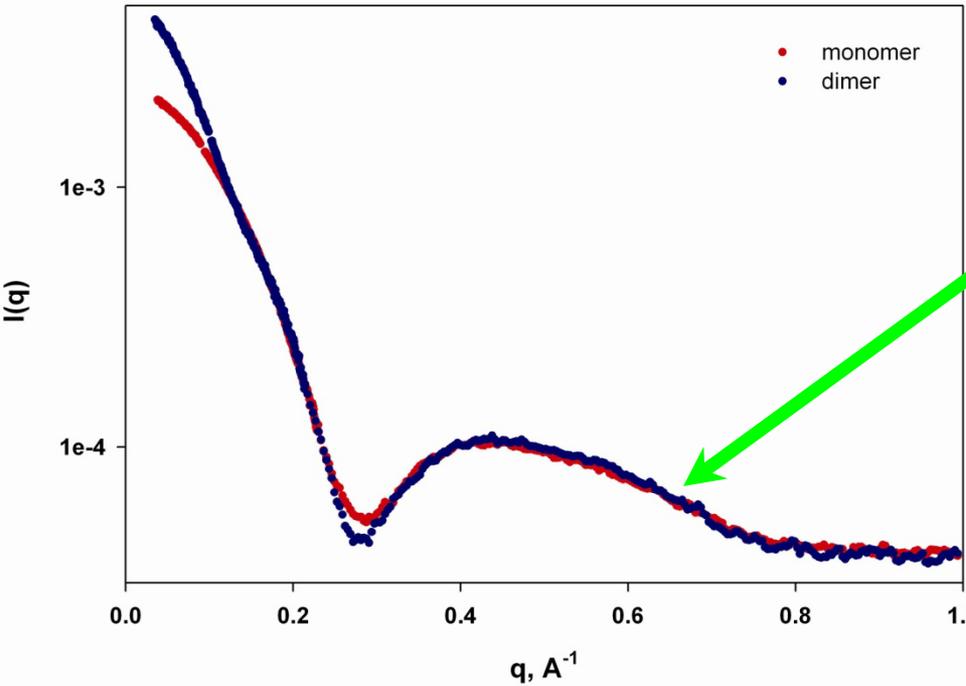


What can SAS data tell us about molecular structure? Actually, a lot...



A diverse range of structural characteristics can be captured by solution scattering. **Resolution range, signal/noise, and prior structural information** will determine how much it reveals. These 3 factor define the information content of the solution scattering data.

# Molecular size from the low-angle scattering data: monomer vs dimer



High-res. individual domain structure is identical for the monomer and dimer

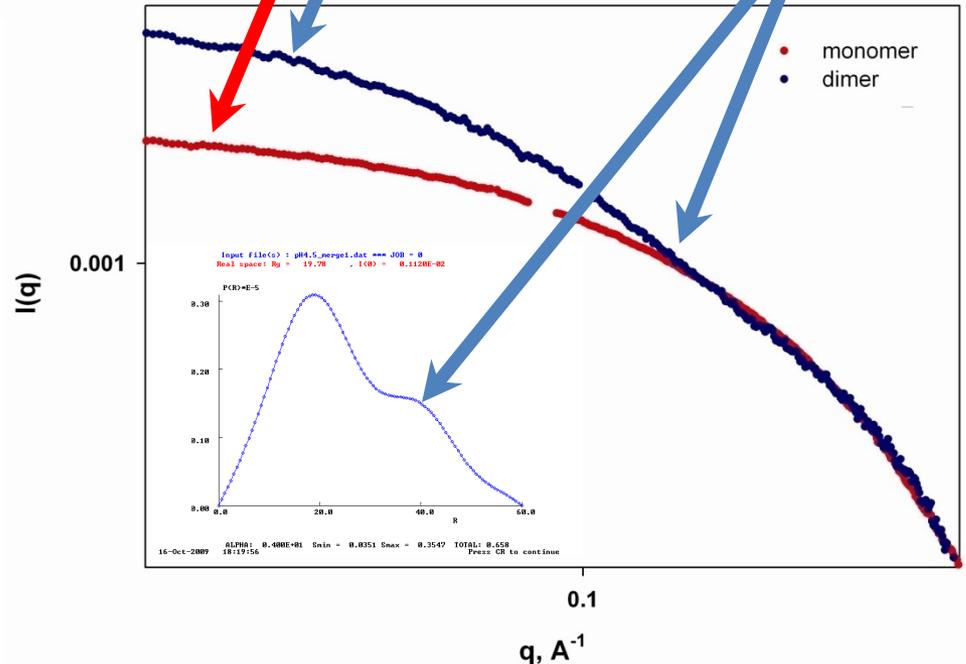
$R_g = 15.3 \text{ \AA}$  for the monomer

$R_g = 19.3 \text{ \AA}$  for the dimer

dimer domain separation is  $\sim 35 \text{ \AA}$

Very little prior structural information in needed for this P(r)- and Guinier-based analyses!

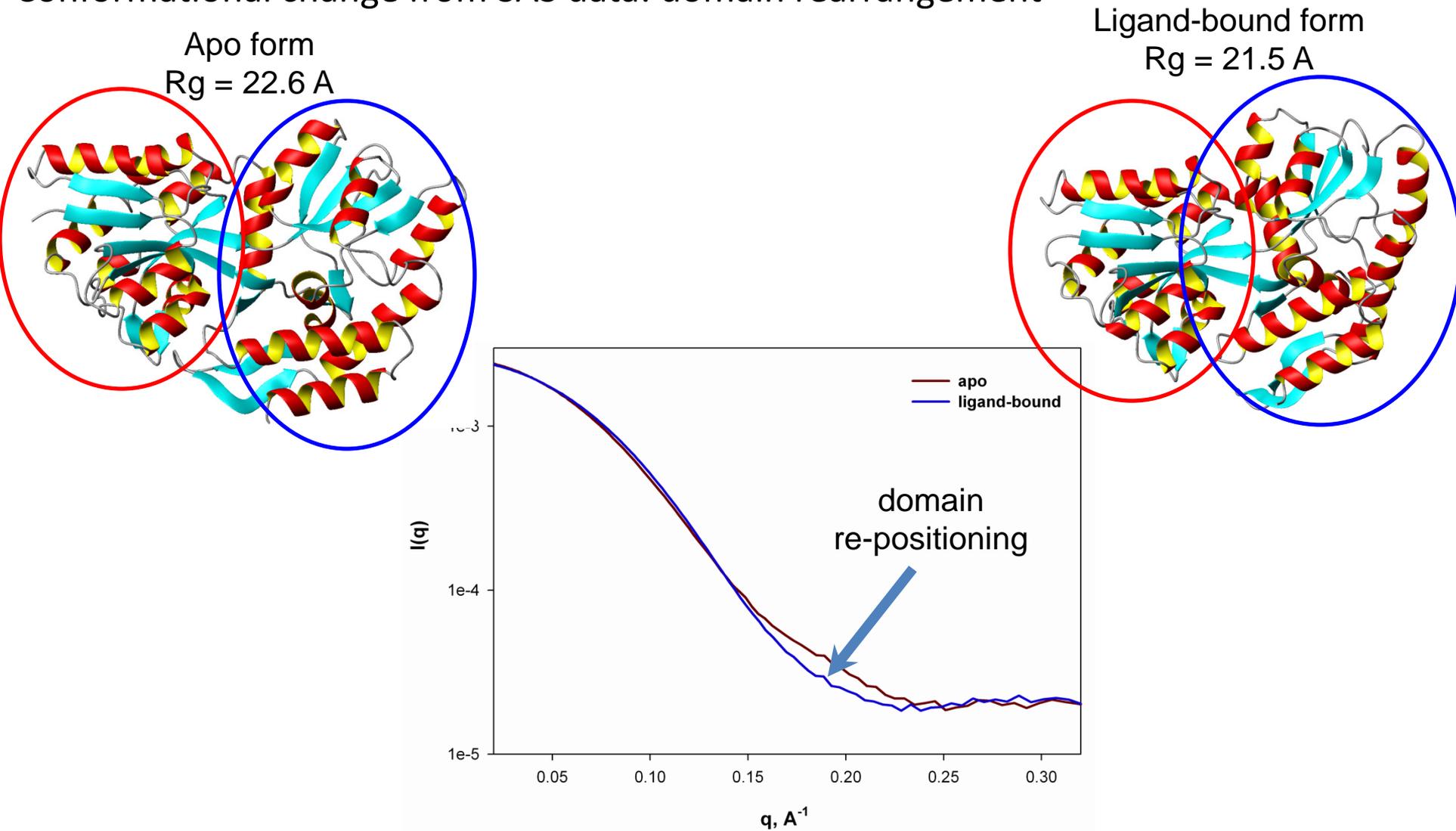
Low-res density reconstructions operate comfortably in this regime.



Input file(s) : p04.5\_merget.dat \*\*\* JOB = 0  
 Real space:  $R_g = 19.29$  ,  $ICR = 0.1120E-02$   
 PCRE-5  
 0.30  
 0.20  
 0.10  
 0.00  
 0.0 20.0 40.0 60.0  
 r

ALPHA: 0.400E+01 Swin = 0.8251 Smax = 0.2547 TOTAL: 0.658  
 16-Oct-2009 18:19:56 Press CR to continue

# Conformational change from SAS data: domain rearrangement

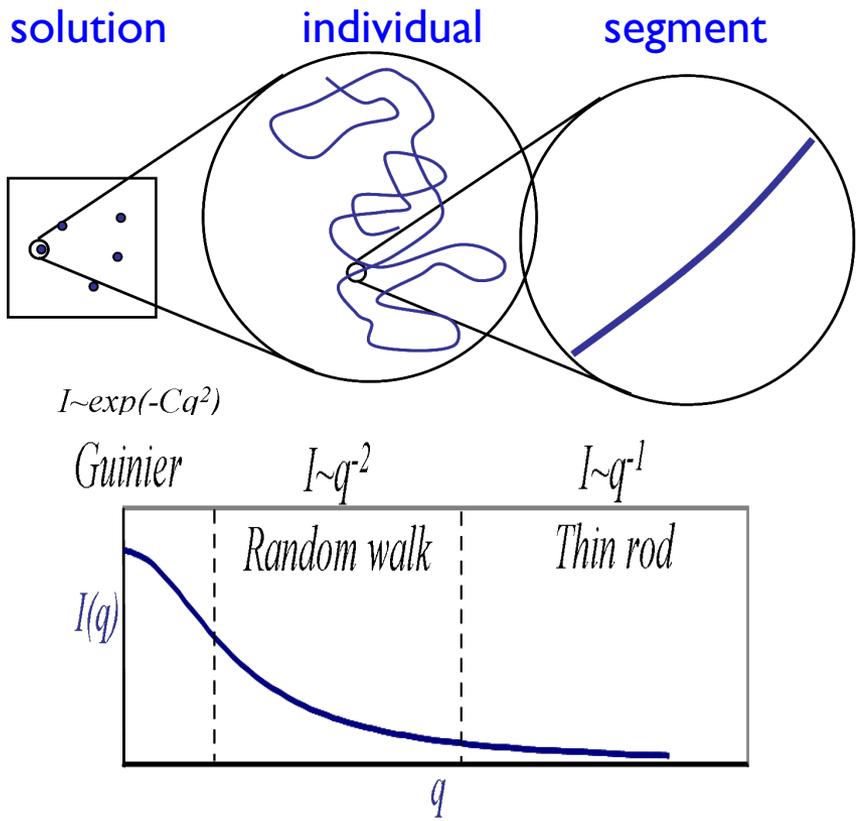


Precise prior structural information and additional restraints are very useful here!

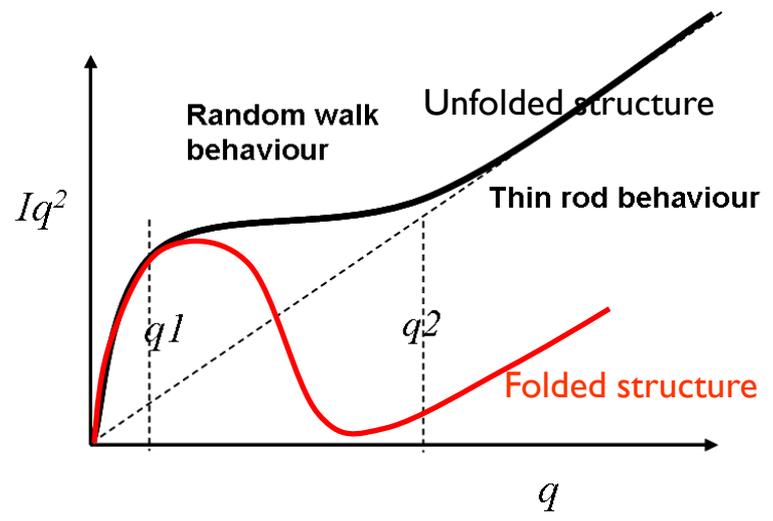
This situation is more challenging for low-res density reconstructions.

NMR-SAXS refinement often operates in this regime.

# Impact of flexibility on SAS data: Kratky plot

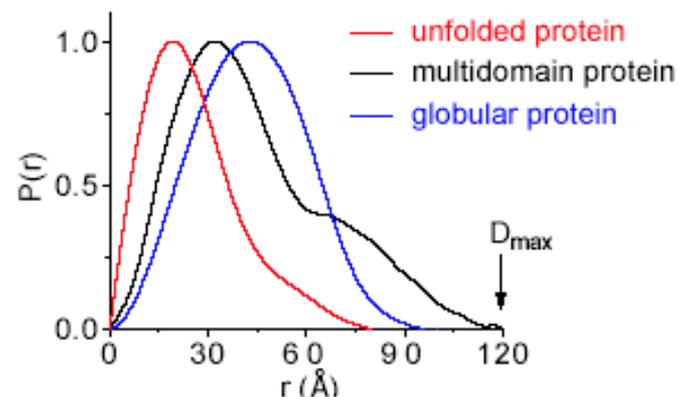


Kratky plot:  $I(q)q^2$  vs  $q$

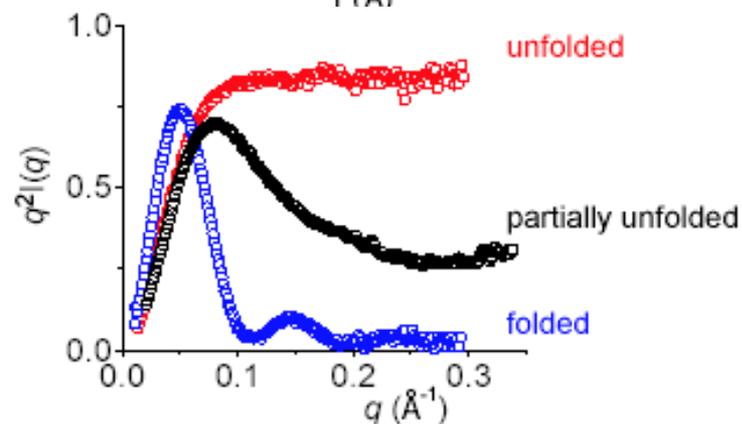


# Can disorder/flexibility to be reliably detected from SAXS data?

**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.



**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.



When flexibility is present  $D_{\max}$  from  $P(r)$  transforms is underestimated.

Features in Kratky plot argue against disorder.

Significant polydispersity of sizes leads to extremely narrow Guinier region at very low  $q$ .

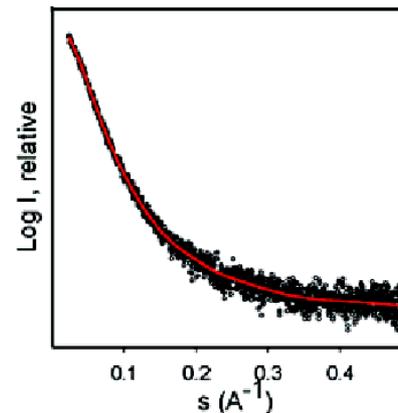
EOM is a useful tool for interpreting data affected by disorder.

- Heller, W. (2004) Influence of multiple well-defined conformations on small-angle scattering of proteins in solution. *Acta Cryst.* D61, 33-44.  
Bernado et al. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* 129, 5656-5664.  
Wang, Y et al. (2008) Small-Angle X-ray Scattering of Reduced Ribonuclease A: Effects of Solution Conditions and Comparisons with a Computational Model of Unfolded Proteins. *J. Mol. Biol.* 377, 1576-1592.

# Ensemble Optimization Method (EOM): Ensemble fitting for flexible systems

N-member ensemble is reconstructed that reproduces the observed scattering data.

Inputs:  $I(q)$  data, protein sequence, individual rigid domains structures (if any).



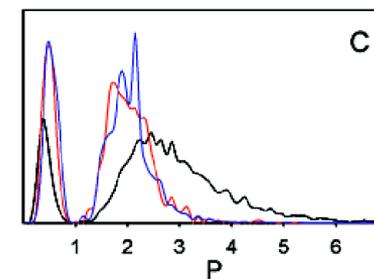
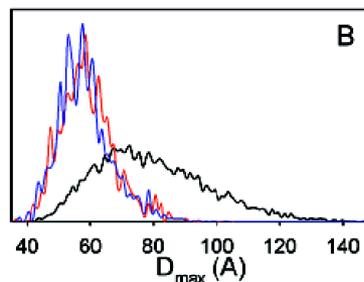
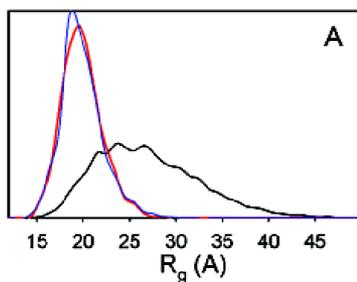
$$I(s) = \frac{1}{N} \sum_{n=1}^N I_n(s)$$

Starting pool of conformers can be either program-generated or user-specified.

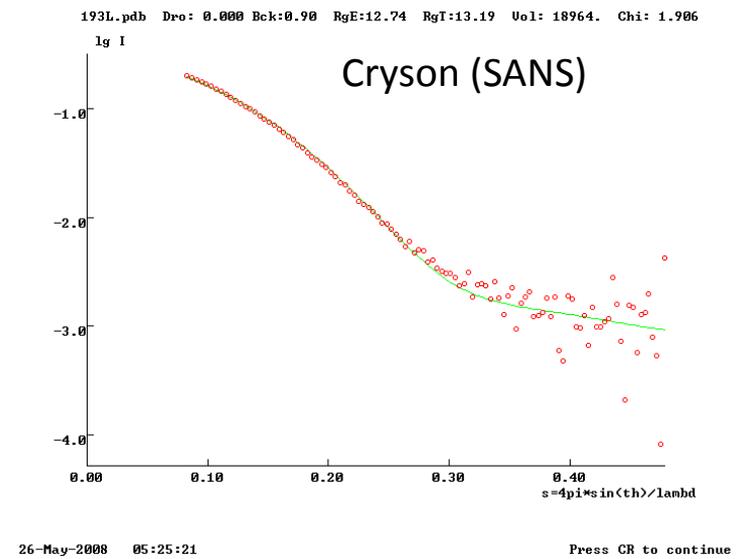
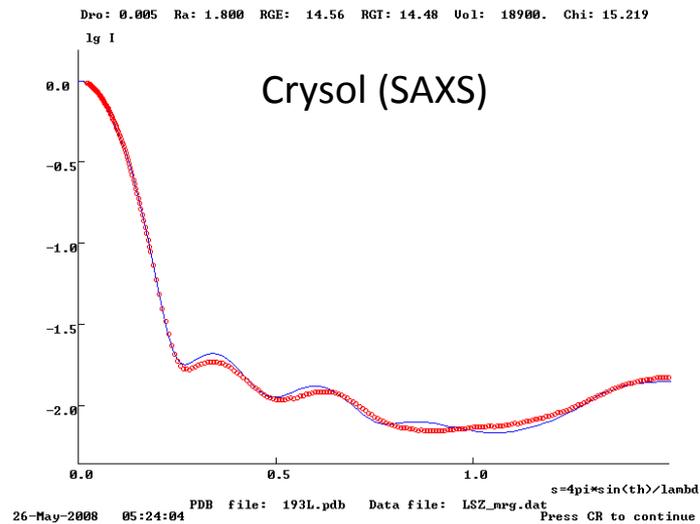
Applicable for both natively unfolded systems (assumed to be fully flexible) or rigidly held domains connected by flexible linkers.

Multiple scattering curves from deletion mutants can be fitted simultaneously.

Program outputs are :best-fitting ensemble members and distributions of  $R_g$ ,  $D_{max}$ , and anisotropy parameters



# Crysol/Cryson: evaluation of agreement between SAXS/SANS data and atomic models



These programs predict the  $I(q)$  data from the atomic coordinates or optimize the fit between predicted and experimental data by optimizing the following adjustable parameters:  
*overall scale, average atomic displaced solvent multiplier, total excluded volume, and contrast of the bound solvent layer.*

Cryson takes into account experimental beam collimation parameters and wavelength divergence.  
Highest level of calculations works best: 50 for the harmonics order and 18 for Fibonacci grid order.  
Bound solvent treatment by these methods is inaccurate for highly anisometric or hollow structures.  
Data predicted above  $1.0\text{-}1.5 \text{ \AA}^{-1}$  may not be accurate.

Svergun, D. et al. (1995) CRY SOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28, 768-773.

Svergun, D. et al. (1998) Protein hydration in solution: Experimental observation by X-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95, 2267-2272.

# SolX: Solution X-ray scattering simulation software

## solX:

- Designed for biomolecules and supramolecules with a flexible molecular dictionary
- Coordinate-based calculations of solution x-ray scattering curve, PDDF, anomalous scattering data
- Recognizes common biomolecules including proteins, nucleic acids and complexes.
- Easy to extend to user-defined molecules such as ligands.

expanding recognizable molecular scope by extending atom conversion map (SolxAtomMap.txt):

```
// Valine (VAL)
ATOM OXT OH VAL PRT
ATOM NT NH2 VAL PRT
ATOM N NH VAL PRT
ATOM CA CH VAL PRT
ATOM C C VAL PRT
...
```

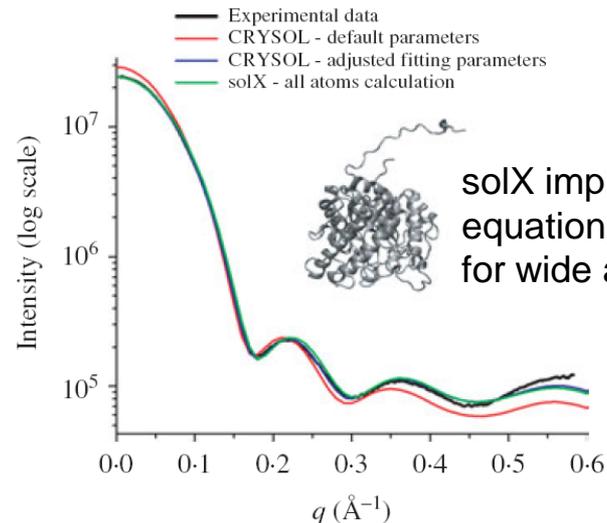
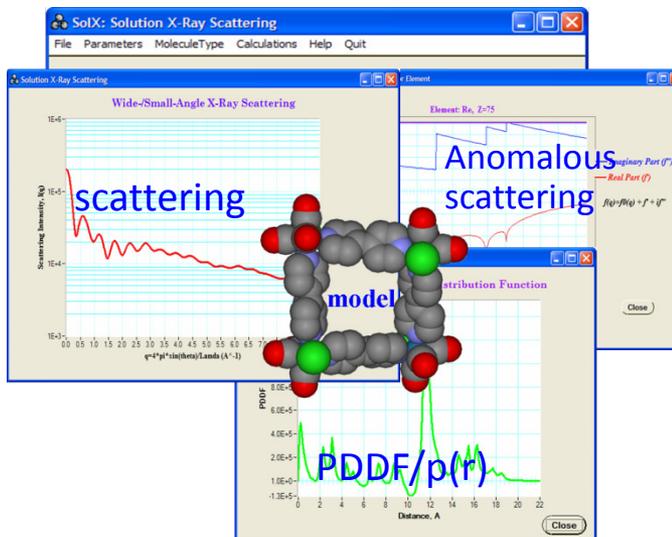
Atoms in regular biomolecules

```
// Undetermined (UNK)
ATOM O H2O HOH PRT
ATOM Zn Zn HOH PRT
...
```

User defined heteroatoms

```
// HEM
ATOM FE Fe2 HEM PRT
ATOM NA N HEM PRT
ATOM NB N HEM PRT
ATOM NC N HEM PRT
ATOM ND N HEM PRT
...
```

User defined ligands/unusual residues



solX implements exact Debye equation, slow but more accurate for wide angle calculations

Available on request:

X. Zuo: xiaobing.zuo@gmail.com

D. Tiede: tiede@anl.gov

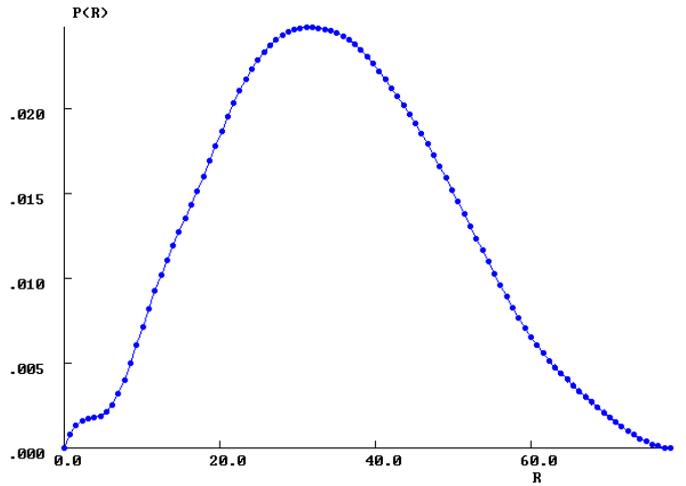
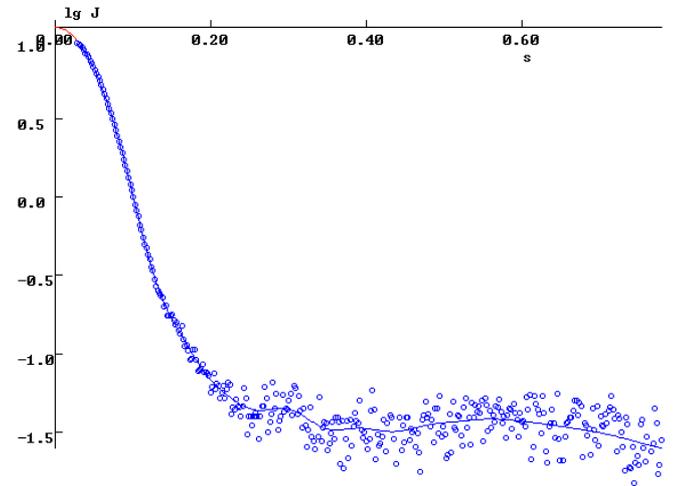
Putnam, D., et al. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quart. Rev. Biophys.* 40, 191-285.

# Running DAMMIN and GASBOR: input preparation

Both require GNOM outputs. Use  $q_{max}=0.8$  for GASBOR and  $0.2$  for DAMMIN in this example.

Input file(s) : msg\_a3334k.dat \*\*\* JOB = 0  
 Reciprocal space: Rg = 26.69 , I(0) = 0.1212E+02

Input file(s) : msg\_a3334k.dat \*\*\* JOB = 0  
 Real space: Rg = 26.64 +- 0.064 I(0) = 0.1212E+02 +- 0.3849E-01

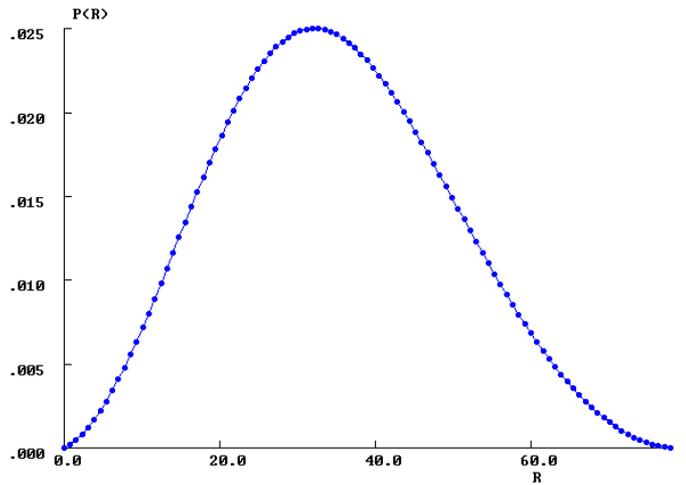
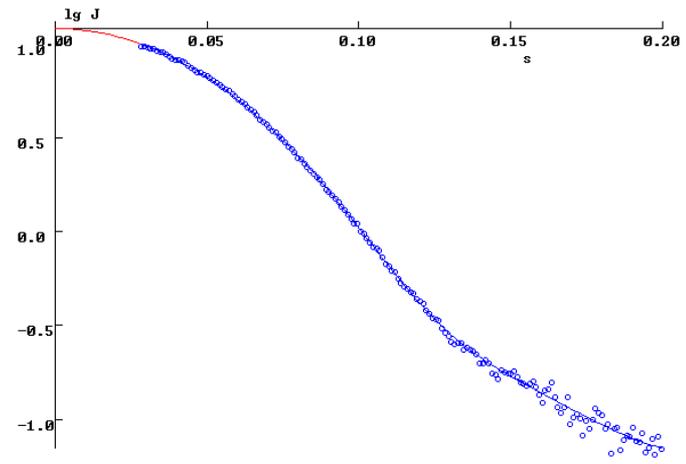


16-Oct-2009 ALPHA: 0.964E+01 Rmin = 0.00 Rmax = 78.00 TOTAL: 0.892  
 01:47:25 Press CR to continue

16-Oct-2009 ALPHA: 0.964E+01 Smin = 0.0289 Smax = 0.7797 TOTAL: 0.892  
 01:47:35 Press CR to continue

Input file(s) : msg\_a3334k.dat \*\*\* JOB = 0  
 Reciprocal space: Rg = 26.64 , I(0) = 0.1210E+02

Input file(s) : msg\_a3334k.dat \*\*\* JOB = 0  
 Real space: Rg = 26.59 +- 0.054 I(0) = 0.1210E+02 +- 0.3512E-01



16-Oct-2009 ALPHA: 0.470E+01 Rmin = 0.00 Rmax = 78.00 TOTAL: 0.746  
 01:49:20 Press CR to continue

16-Oct-2009 ALPHA: 0.470E+01 Smin = 0.0284 Smax = 0.1996 TOTAL: 0.746  
 01:49:31 Press CR to continue

# Running DAMMIN

\*\*\* Please reference: D.Svergun (1999). Biophys. J. \*\*\*  
\*\*\* 76, 2879-2886. \*\*\*

===== DAMMIN46 started on 16-Oct-2009 01:54:15

Mode: <[F]>ast, [S]low, [J]ag, [E]xpert < **Fast** >:  
Log file name ..... < .log >: 1

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

Working directory: C:\Alex\NIH\SAXS\Talks\NCI\_2009\tutorial\DataAnalysis\dammin\

File to be opened: msg\_a3334k\_dammin.out

Project identificator ..... : 1

Enter project description ..... :

Random sequence initialized from ..... : 15421

\*\* Information read from the GNOM file \*\*

Data set title: Merge of: a33000.dat a34000.dat

Raw data file name: msg\_a3334k.dat

Maximum diameter of the particle ..... : 78.00

Solution at Alpha = 0.470E+01 Rg : 0.266E+02 I(0) : 0.121E+02

Radius of gyration read ..... : 26.60

Number of GNOM data points ..... : 194

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] ..... : 0.1996

Number of Shannon channels ..... : 4.956

Portion of the curve to be fitted ..... < 1.000 >:

Number of knots in the curve to fit ..... : 20

A constant was subtracted ..... : 2.282e-2

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 >:

Symmetry: P1...6 or Pn2 (n=1,2,3,4,6) .. < **P1** >:

Sphere diameter [Angstrom] ..... : 78.00

Packing radius of dummy atoms ..... : 2.800

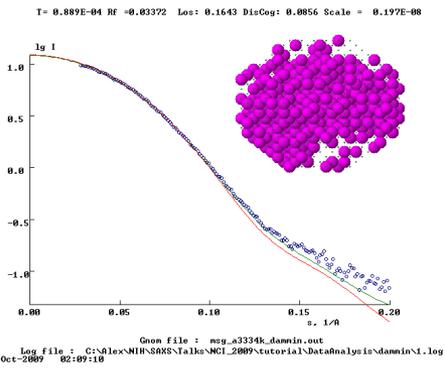
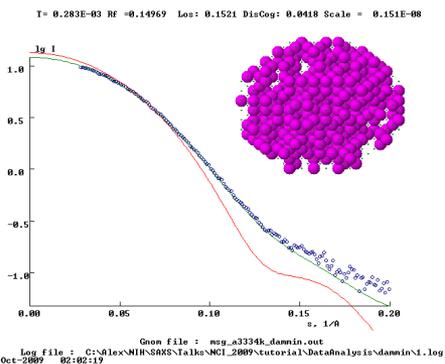
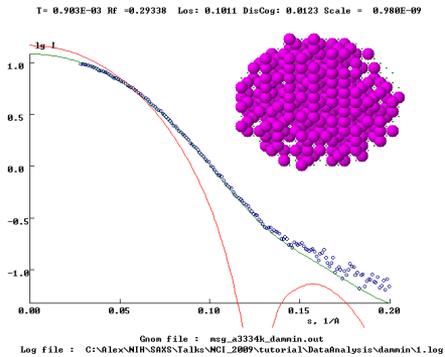
Radius of the sphere generated ..... : 39.00

Number of dummy atoms ..... : 1974

Number of equivalent positions ..... : 1

Expected particle shape: <P>rolate, <O>blate,

or <U>nknown ..... < **Unknown** >:



# GASBOR

\*\*\* M.H.J.Koch (2001) Biophys. J. 80, 2946-2953 \*\*\*

=== GASBOR Version 2.0 build 10.03.0 started on 16-Oct-2009 02:13:06

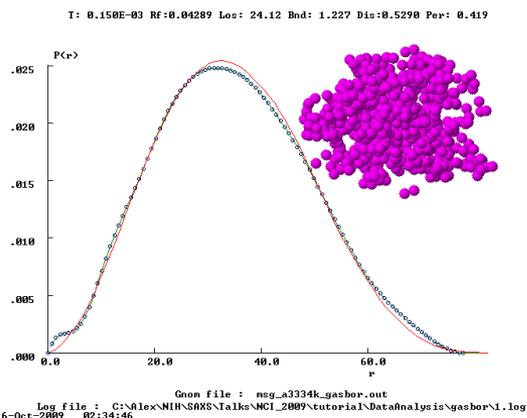
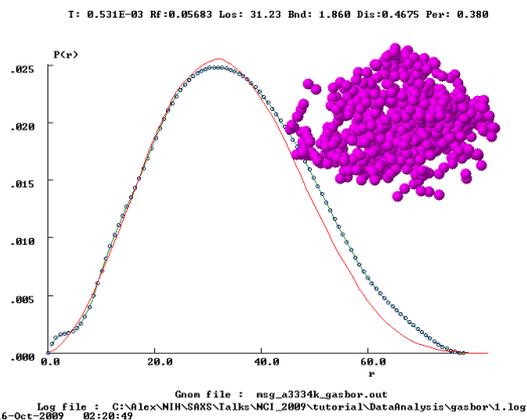
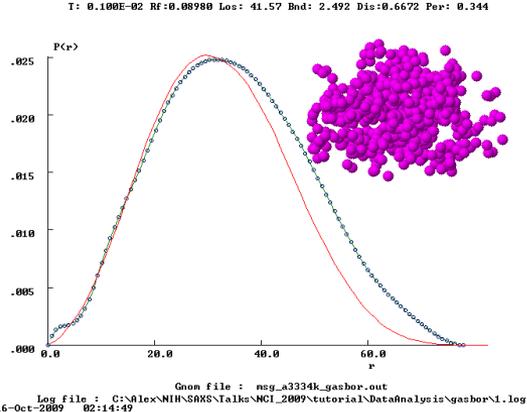
Computation mode (User or Expert) ..... < User >  
Log file name ..... < .log >: 1

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

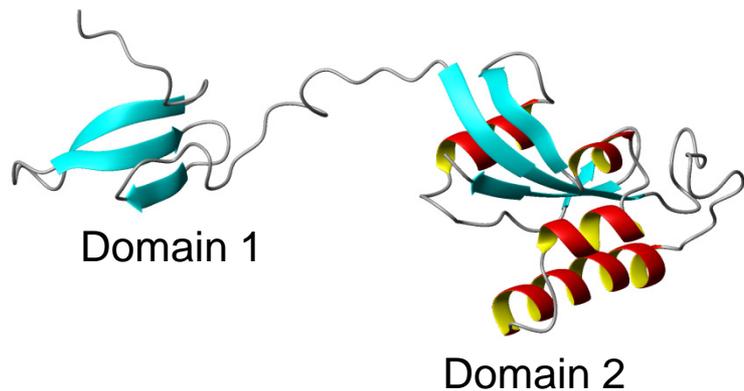
Working directory: C:\Alex\NIH\SAXS\Talks\NCI\_2009\tutorial\DataAnalysis\gasbor  
r\

File to be opened: msg\_a3334k\_gasbor.out  
Project identificator ..... : 1  
Enter project description ..... :  
Random sequence initialized from ..... : 21312  
\*\* Information read from the GNOM file \*\*  
Data set title: Merge of: a33000.dat a34000.dat

Raw data file name: msg\_a3334k.dat  
Maximum diameter of the particle ..... : 78.00  
Solution at Alpha = 0.964E+01 Rg : 0.267E+02 I(0) : 0.121E+02  
Radius of gyration ..... : 26.70  
Number of GNOM data points ..... : 379  
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] ..... : 0.7797  
Number of Shannon channels ..... : 19.36  
Portion of the curve to be fitted ..... < 1.000 >:  
Number of knots in the curve to fit ..... : 101  
Initial DRM (CR for random) ..... < ,ndb >:  
Symmetry Pn (n=1...6) or Pnm (m=1,2) ... < P1 >:  
Number of equivalent positions ..... : 1  
Number of residues in asymmetric part .. < 519 >: 723  
Fibonacci grid order ..... < 14 >:  
Number of dummy waters ..... : 611  
Excluded volume per residue ..... : 28.73  
Radius of the search volume ..... : 39.00  
Histogram penalty weight ..... : 5.000e-4  
Bond length penalty weight ..... : 5.000e-3  
Discontiguity penalty weight ..... : 5.000e-3  
Peripheral penalty weight ..... : 0.5000  
Expected particle shape: <P>rolate, <O>blate,  
or <U>nknown ..... < Unknown >:

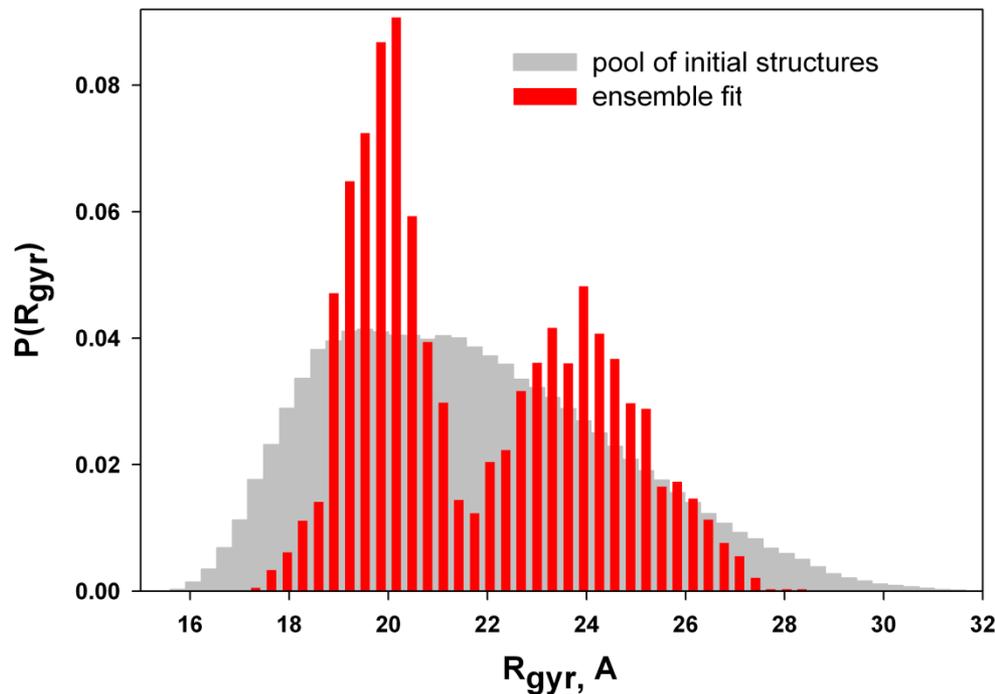
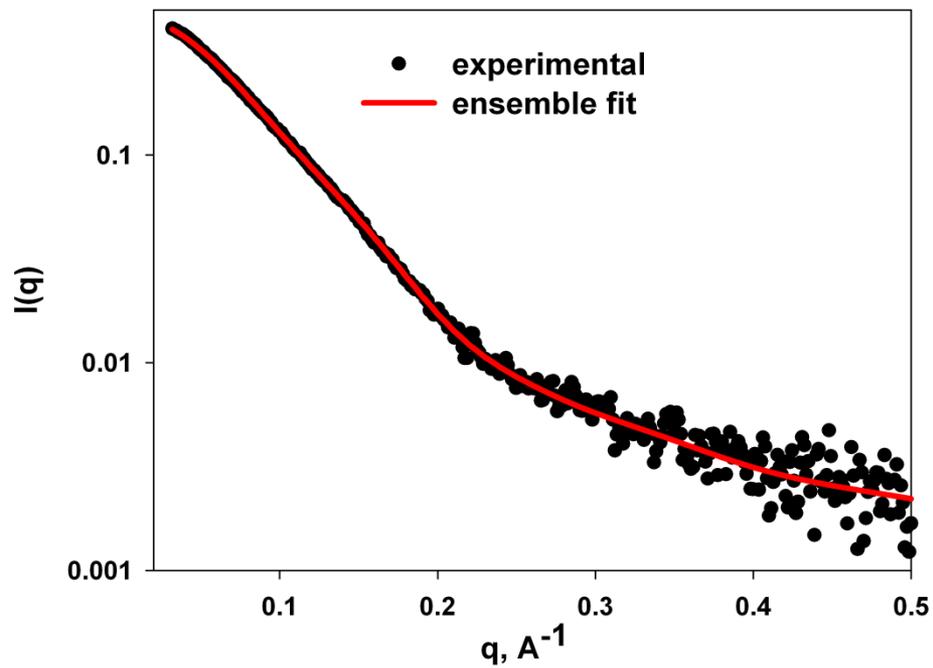


# EOM reconstruction example

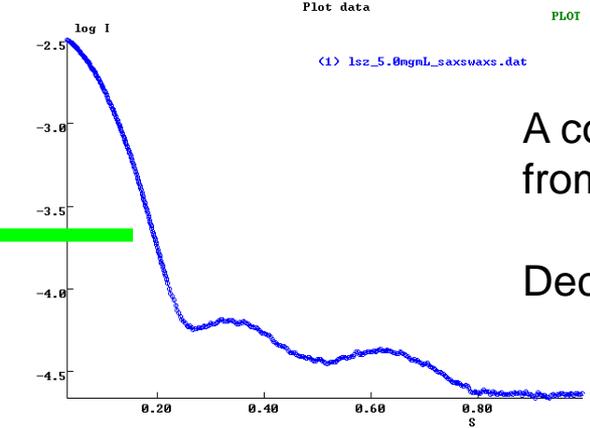
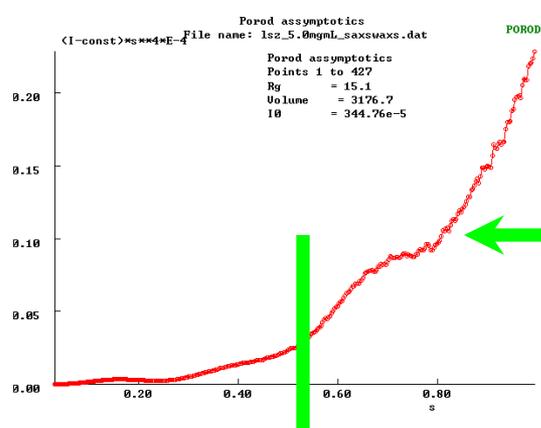


Domain 1:  
 $t_c=8.8$  ns, Pf1 Da= -7.9 Hz,  $R=0.64$

Domain 2:  
 $t_c=11$  ns, Pf1 Da = -14.5 Hz,  $R=0.54$

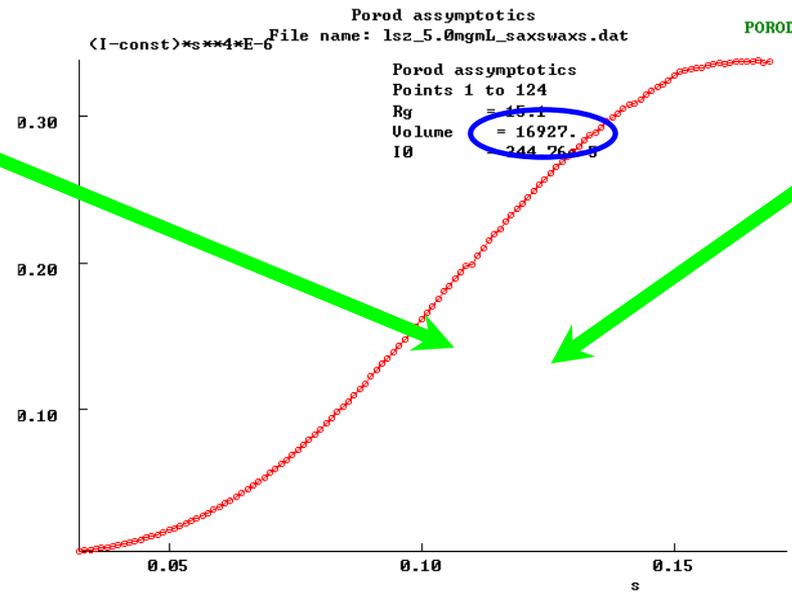
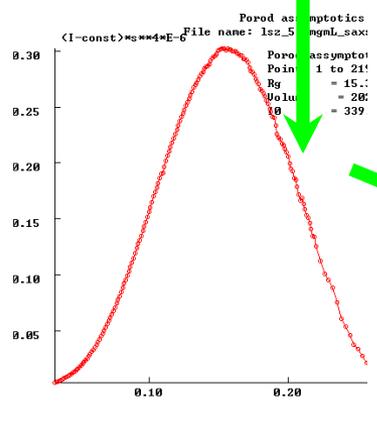
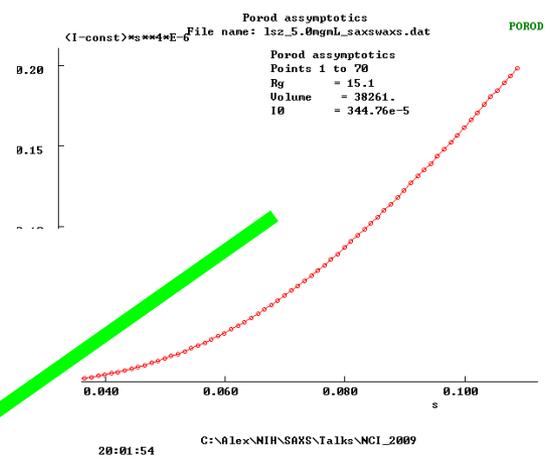


# Porod volume calculation with Primus



A constant is subtracted from the data during optimization.  
 $(I(q) - \text{const}) q^4$  vs.  $q$   
 Decrease  $q_{\text{max}}$  till high  $q$  region is flat.

$V_{\text{mol}}$  should be 17400  $\text{\AA}^3$   
 (~3% error)



# Neutron vs. X-ray scattering

**Table 1.** X-ray and neutron scattering lengths of some elements.

Atom	H	D	C	N	O	P	S	Au
Atomic mass	1	2	12	14	16	30	32	197
N electrons	1	1	6	7	8	15	16	79
$f_X, 10^{-12}$ cm	0.282	0.282	1.69	1.97	2.16	3.23	4.51	22.3
$f_N, 10^{-12}$ cm	-0.374	0.667	0.665	0.940	0.580	0.510	0.280	0.760

**Table 2.** X-ray and neutron scattering length densities of biological components.

Component	X-rays		Neutrons		
	$\rho$ (electrons nm <sup>-3</sup> )	Matching solvent	$\rho$ in H <sub>2</sub> O (10 <sup>10</sup> cm <sup>-2</sup> )	$\rho$ in D <sub>2</sub> O (10 <sup>10</sup> cm <sup>-2</sup> )	Matching % D <sub>2</sub> O
H <sub>2</sub> O	334	—	-0.6	—	—
D <sub>2</sub> O	334	—	6.4	—	—
50% sucrose	400	—	1.2	—	—
Lipids	300	—	0.3	-6.0	≈10-15%
Proteins	420	65% sucrose	1.8	3.1	≈40%
D-proteins	420	65% sucrose	6.6	8.0	—
Nucleic acids	550	—	3.7	4.8	≈ 70%
D-nucleic acids	550	—	6.6	7.7	—

For x-rays, the scattering length densities are often expressed in terms of electron density, i.e. the number of electrons per nm<sup>3</sup>; 1 electron nm<sup>-3</sup> = 2.82 × 10<sup>8</sup> cm.

Neutron atomic scattering lengths are q-independent.

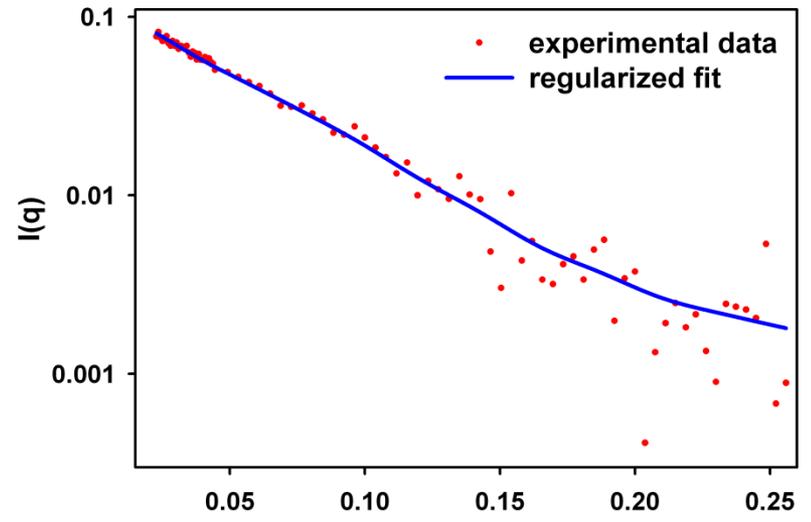
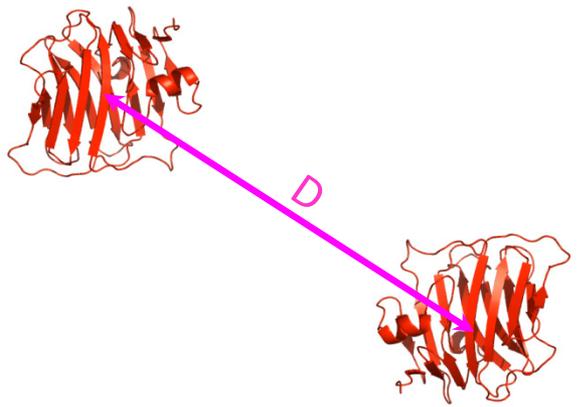
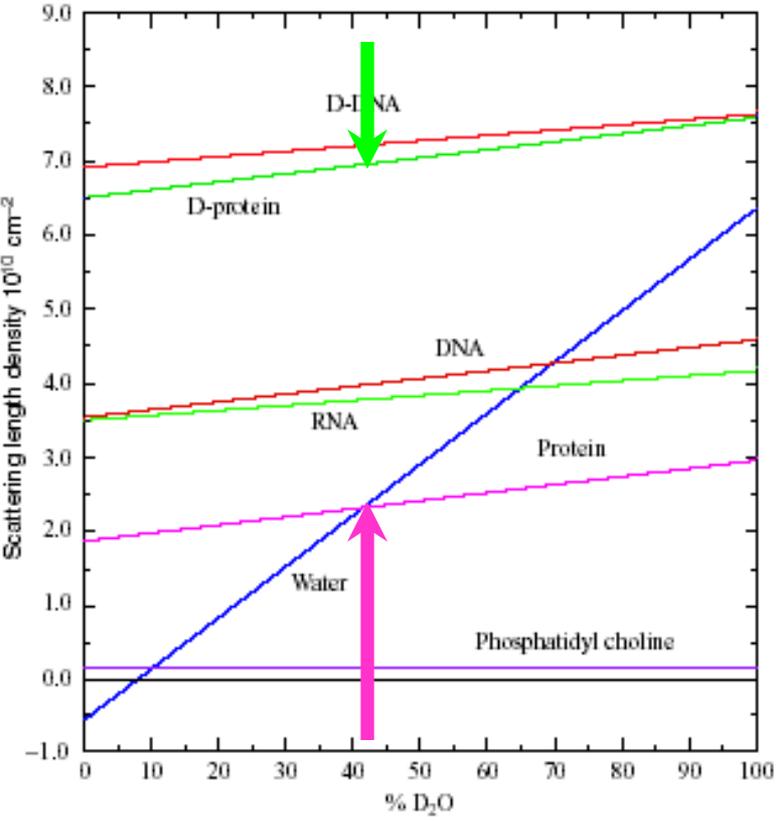
Neutron scattering is significantly affected by the hydrogen isotopes in the sample and is very different for <sup>1</sup>H vs <sup>2</sup>H.

Incoherent scattering of <sup>1</sup>H in the sample gives a strong background signal.

Neutron contrast variation is very useful for determining the structure of complexes.

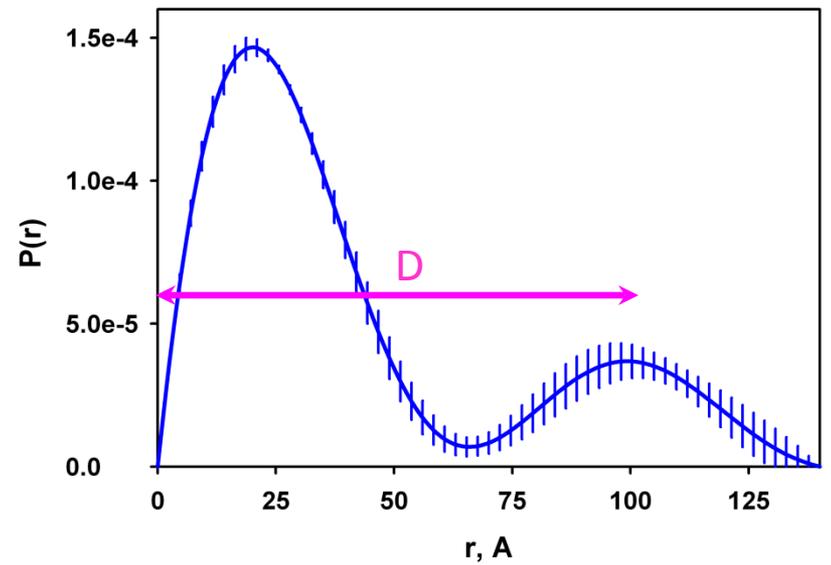
$$I(q) = \overline{\Delta\rho_1}^2 I_{11}(q) + \overline{\Delta\rho_2}^2 I_{22}(q) + \overline{\Delta\rho_1} \overline{\Delta\rho_2} I_{12}(q)$$

# Contrast-matched SANS



$q, \text{ \AA}^{-1}$

FT



## **3a. Joint Refinement with NMR data**

## Rationale for combining SAXS with NMR

- Same sample conditions.
- Small sample size: ~10-100 mL @ ~5-20 mg/mL. No isotope labeling is necessary.
- Fast data acquisition, processing and analysis (minutes on a synchrotron)
- Accurate calculation of scattering data from atomic coordinates
- Sensitivity to differences in the molecular structure as small as ~1-2 Å backbone rmsd.
- Information content and signal-to-noise ratio increase with the molecular size.
- NMR and SAXS are largely complementary in terms of the experimental restraints produced.
- Validation of the quality of the resulting structures can be done against both SAXS and NMR data.

## Potential challenges when combining SAXS with solution NMR

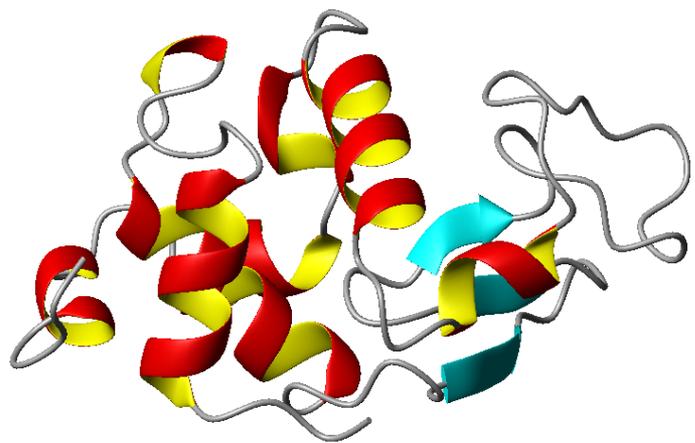
### SAXS:

- overcoming sample aggregation and radiation damage
- computational costs when fitting scattering data during MD
- polydispersity

### NMR:

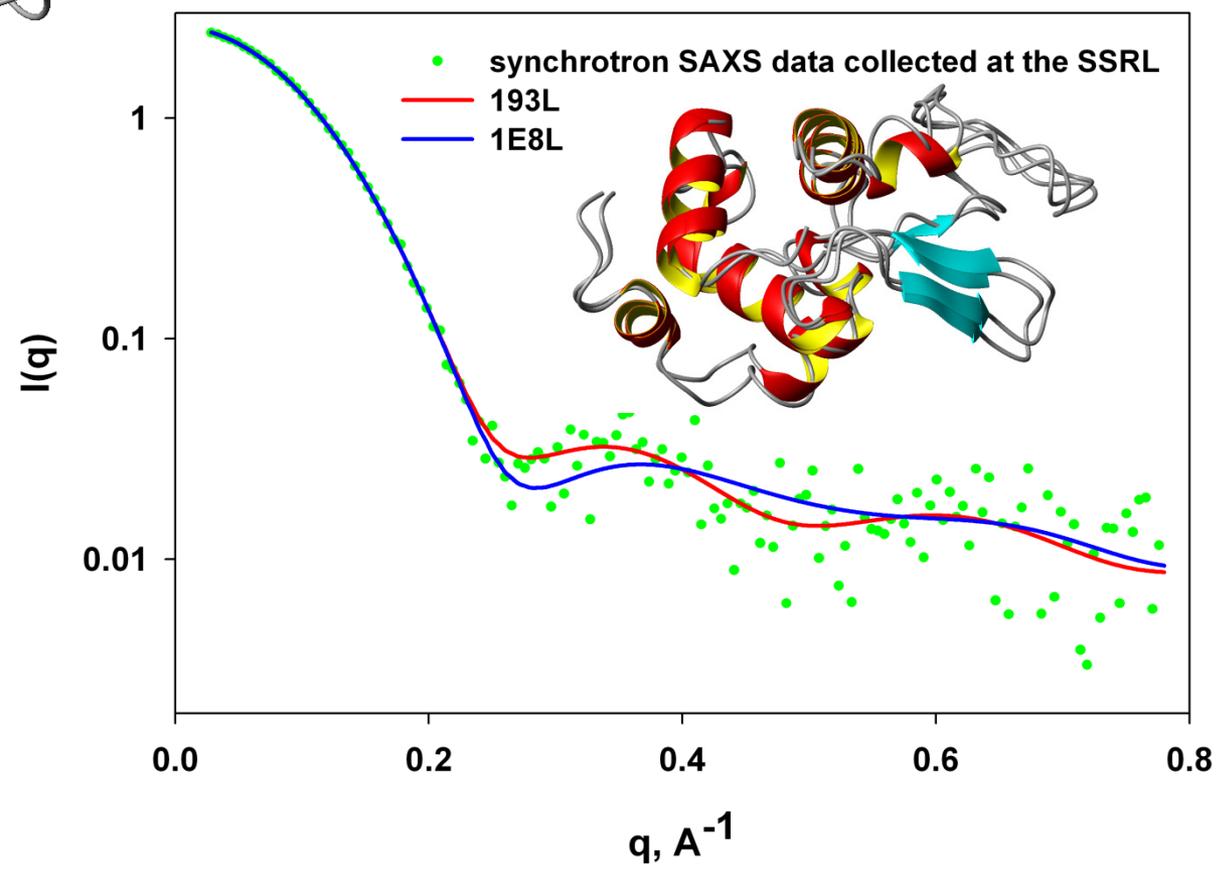
- getting restraints for large systems

# SAXS data can be sensitive to small structural differences



Hen egg white lysozyme  
MW = 14.4 kDa  
 $R_g \sim 14 \text{ \AA}$

Experimental data fits well to both  
193L and 1E8L models (1.5 Å backbone rmsd)



Predicted scattering curves differ at resolution  $> 26 \text{ \AA}$

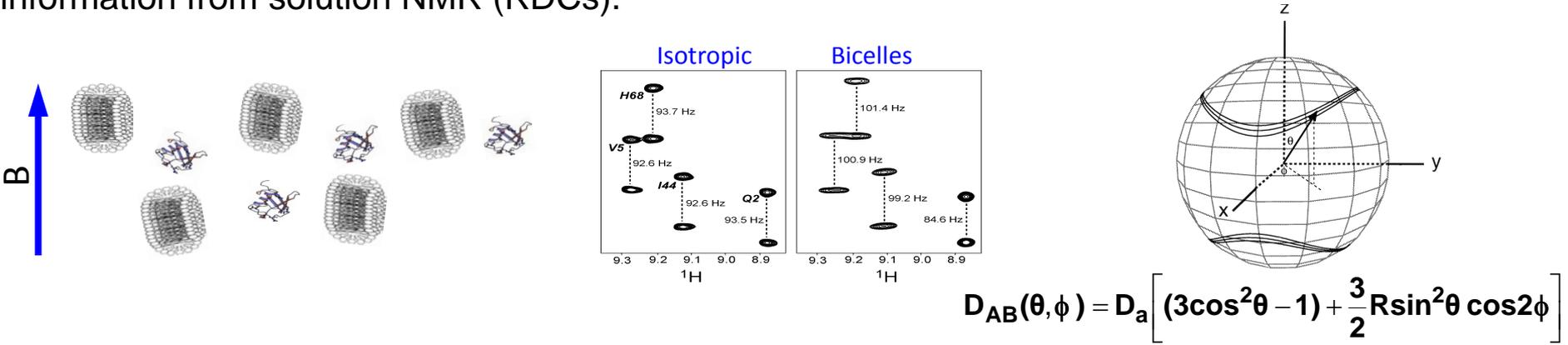
# Using SAXS data for high-res structure determination

## Challenge

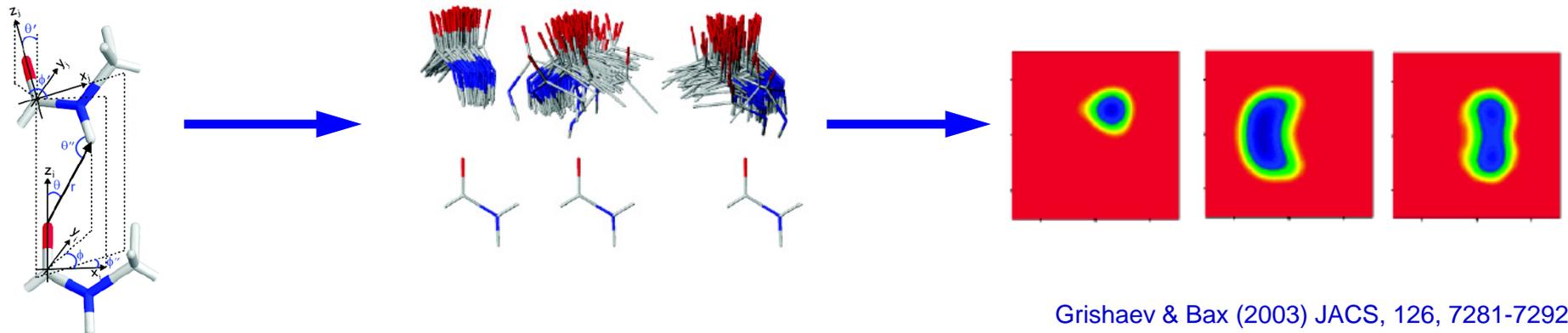
- SAXS data have low information content; scattering curves can be difficult to interpret directly; degenerate solutions are common.

## Solution: hybrid approaches

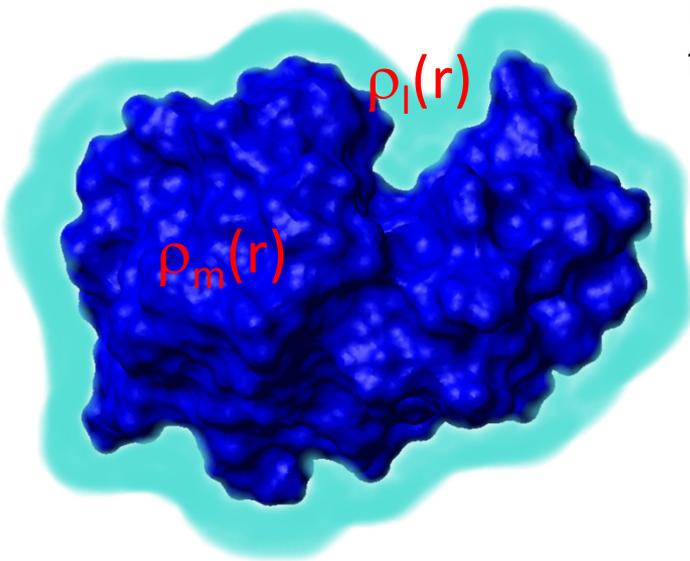
- Translational / shape information from SAS is complementary to site-specific orientational information from solution NMR (RDCs).



- When refining against SAXS data, the best strategy is to keep structure **locally rigid but globally flexible**. Distance and torsion angle NMR restraints are useful as well as H-bonding PMF terms.



Prediction of SAXS data  
from atomic coordinates



$$I(q) = \left\langle \left| A_m(\mathbf{q}) - \rho_s A_s(\mathbf{q}) + \delta\rho A_l(\mathbf{q}) \right|^2 \right\rangle_{\Omega} W.$$

Scattering intensity - average over all orientations

$$A_m(\mathbf{q}) = \sum_{j=1}^N f_j(q) \exp(i\mathbf{q}\mathbf{r}_j)$$

Scattering amplitude *in vacuo*

$$\rho_s A_s(\mathbf{q}) = \sum_{j=1}^N g_j(q) \exp(i\mathbf{q}\mathbf{r}_j)$$

$$g_j(q) = G(q)V_j \exp\left(-\frac{q^2 V_j^{2/3}}{4\pi}\right)$$

Scattering amplitude from the excluded volume  
using *dummy solvent* approximation

Scattering intensity is calculated via *Debye formula*.

$$I(q) = \sum_{i=1}^N \sum_{j=1}^N f_i^s(q) f_j^s(q) \frac{\sin(qr_{ij})}{qr_{ij}}$$

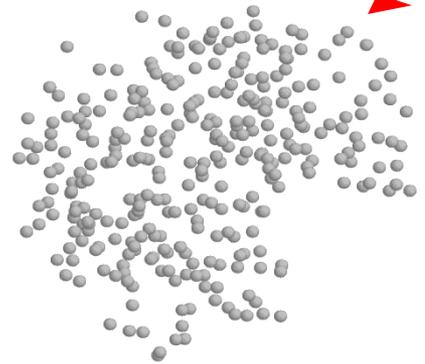
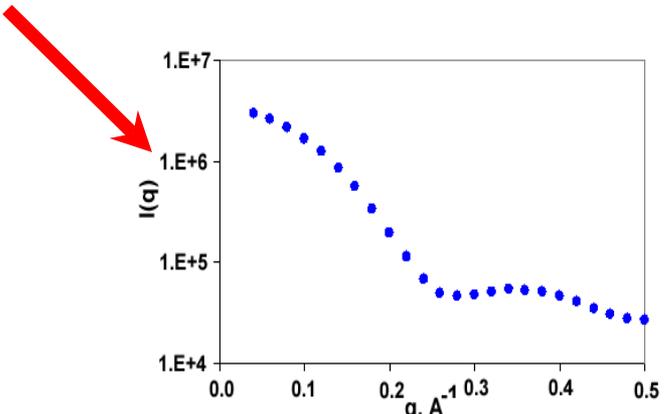
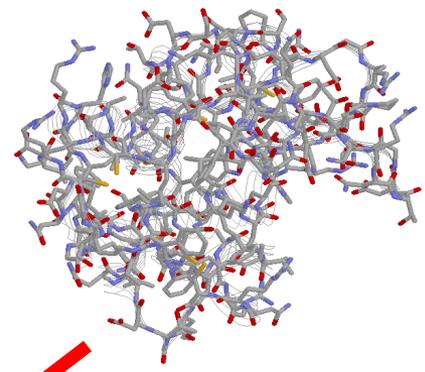
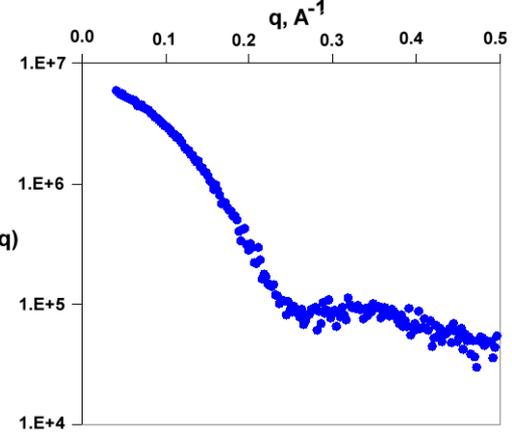
# MD refinement with SAXS data

*Main idea: as long as you can accurately predict SAXS data from the structure, you should be able to refine the structure against it.*

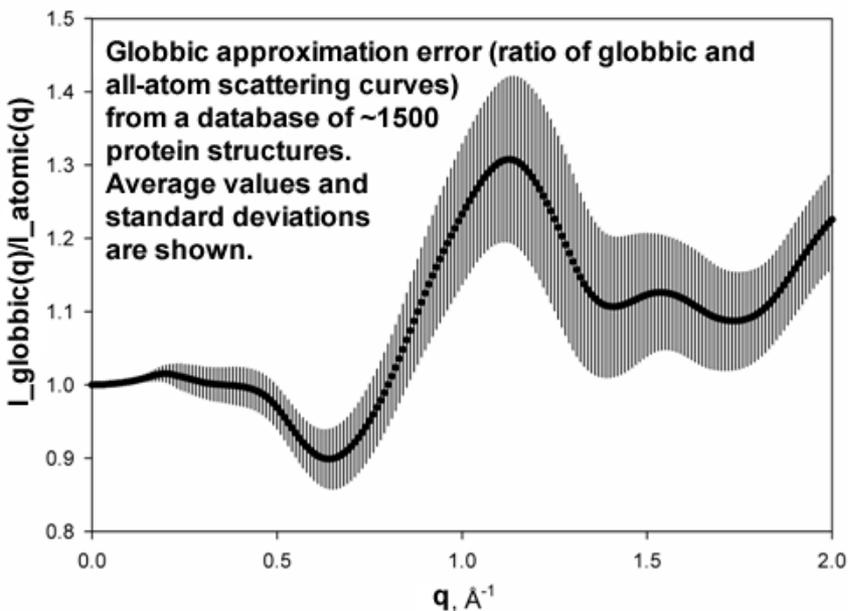
Cost function: 
$$\chi^2 = \frac{1}{N_q - 1} \sum_{k=1}^{N_q} \left[ \frac{c_k I_{calc}(q_k) - I_{expt}(q_k)}{\sigma(q_k)} \right]^2$$
 coded into CNS and Xplor-NIH

Its gradient (force): 
$$\nabla_{r_j} [\chi^2] \approx \sum_{k=1}^{N_q} c_k \frac{c_k I_{calc}(q_k) - I_{expt}(q_k)}{\sigma_k^2} \sum_{i \neq j}^N f_i^s(q_k) f_j^s(q_k) \left[ \cos(q_k r_{ij}) - \frac{\sin(q_k r_{ij})}{q_k r_{ij}} \right] \frac{\mathbf{r}_{ij}}{r_{ij}^2}$$

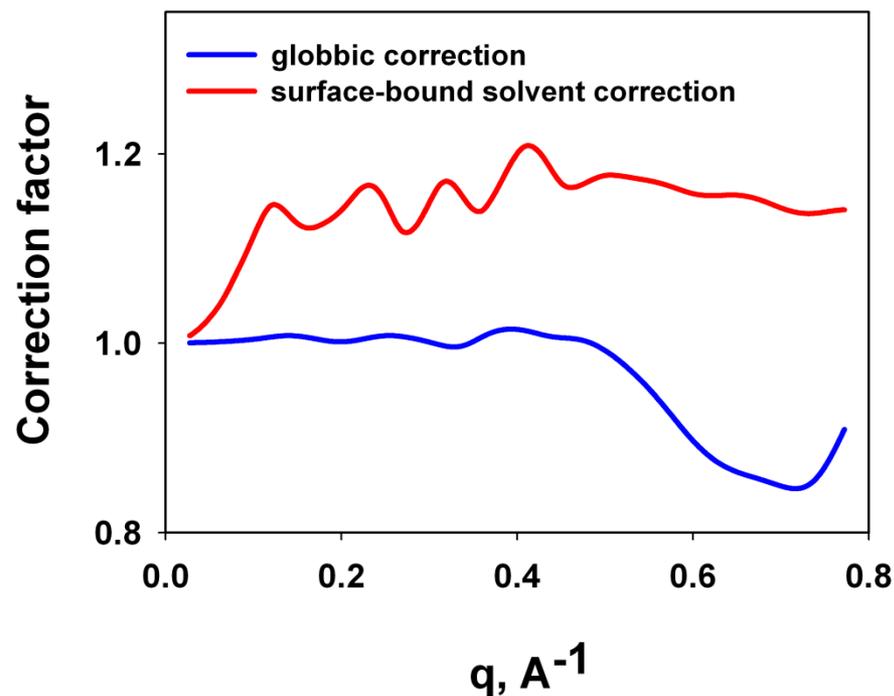
The problem : number of operations  $\sim N^2 * N_q$



# Corrections for the systematic errors in this approach



Globbic correction is a multiplicative factor.



Surface H<sub>2</sub>O scattering is modeled by an additional multiplicative correction factor.

The final calculated scattering profile:

$$I_{calc}(q|conformation) = I_{globbic}(q|conformation) * \left[ \frac{I_{all-atom}(q)}{I_{glob}(q)} \right] * \left[ \frac{I_{surf-H_2O}(q)}{I_{no surf-H_2O}(q)} \right]$$

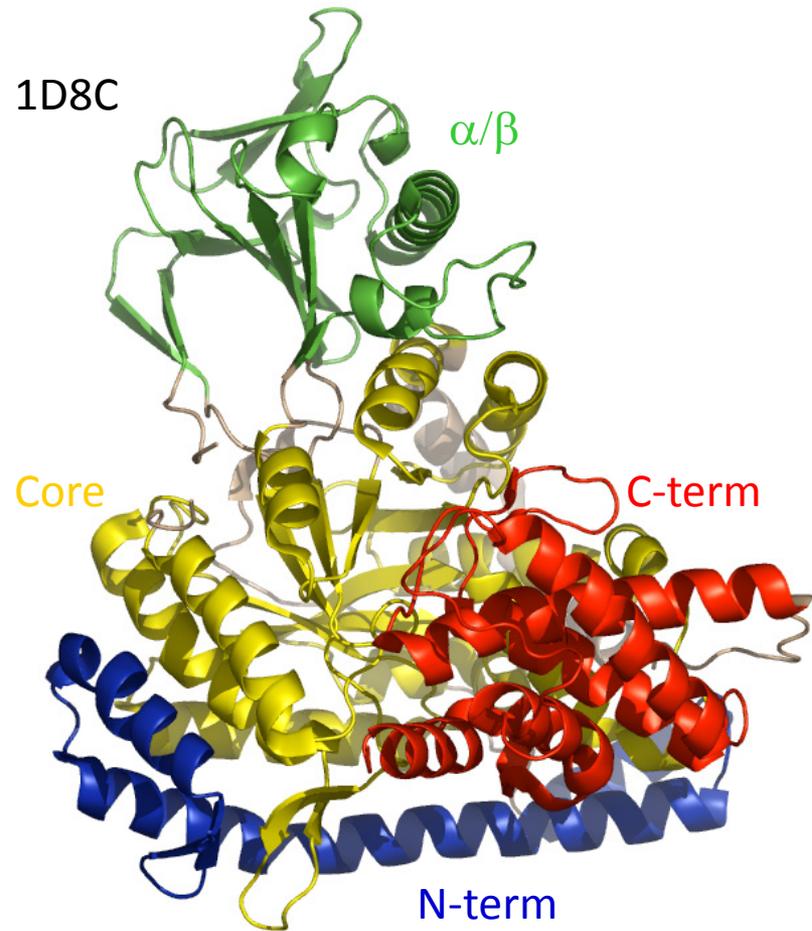
## Summary for Debye formula-based SAXS refinement

1. Decide on your own glob compositions or take the standard ones (2-3 linked heavy atoms + Hs)
2. Define the range of experimental data to be fitted and sparsen it ( $\sim 0.01 \text{ \AA}^{-1}$  step)
3. Calculate globbic form factors (fsglob\_prot code)
4. Based on the current set of models, calculate bound solvent and globbic corrections (crysol and isglob\_prot code)
5. Run the refinement program (Xplor-NIH or CNS) and obtain an updated set of structures
6. Loop steps 4-5 till convergence of the correction profiles.

Detailed instructions and setup/example files available from

[http://spin.niddk.nih.gov/bax/software/saxs\\_distr\\_100308.zip](http://spin.niddk.nih.gov/bax/software/saxs_distr_100308.zip)

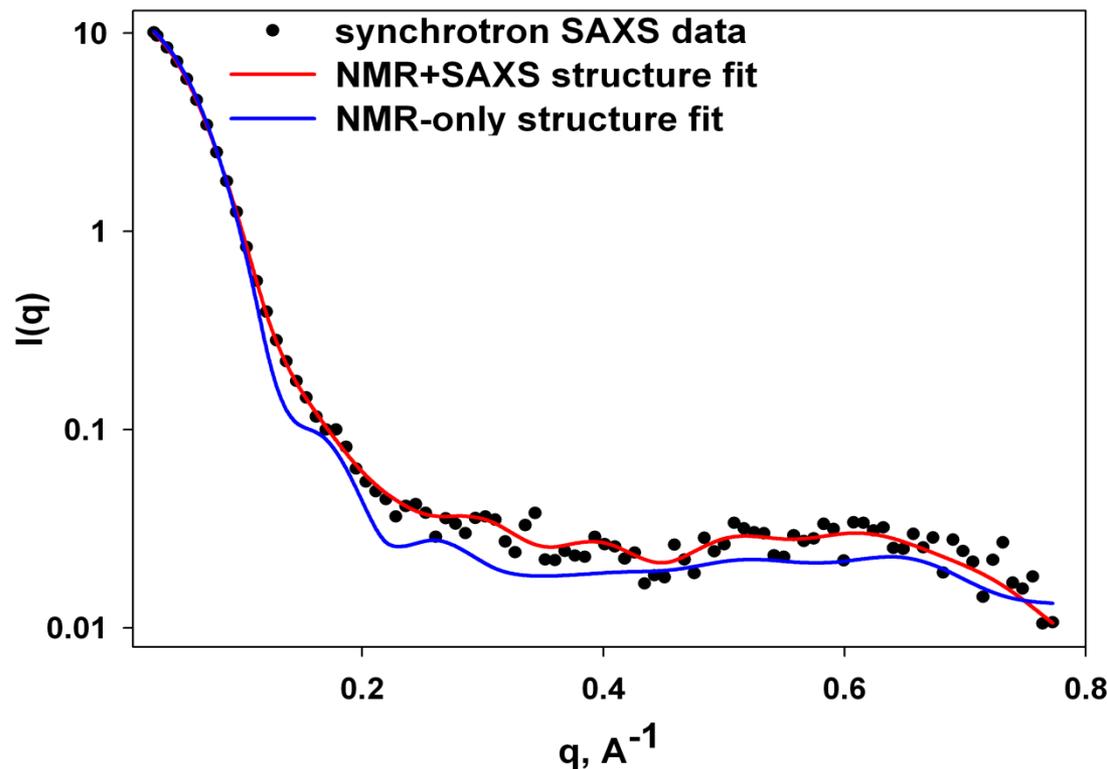
# Example 1: Malate Synthase G



MW = 82 kDa

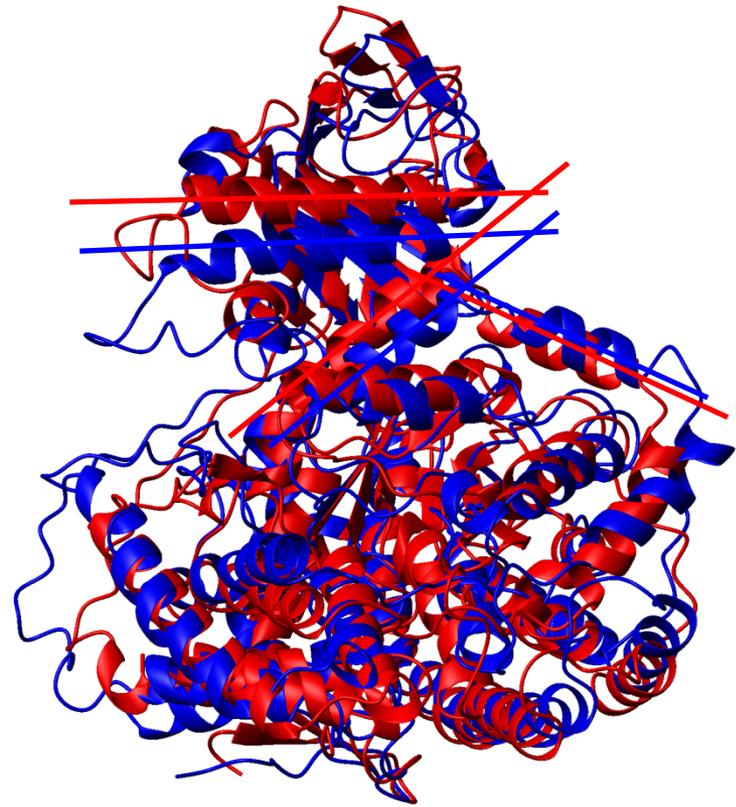
$R_g \sim 26 \text{ \AA}$

NMR data (1Y8B): 1531  $H^N-H^N$ ,  $H^N-CH_3$  and  $CH_3-CH_3$  NOEs, 533 (f,y) dihedral angles, 415  $H^N-N$  RDCs and 300  $^{13}C$ O anisotropic shifts.

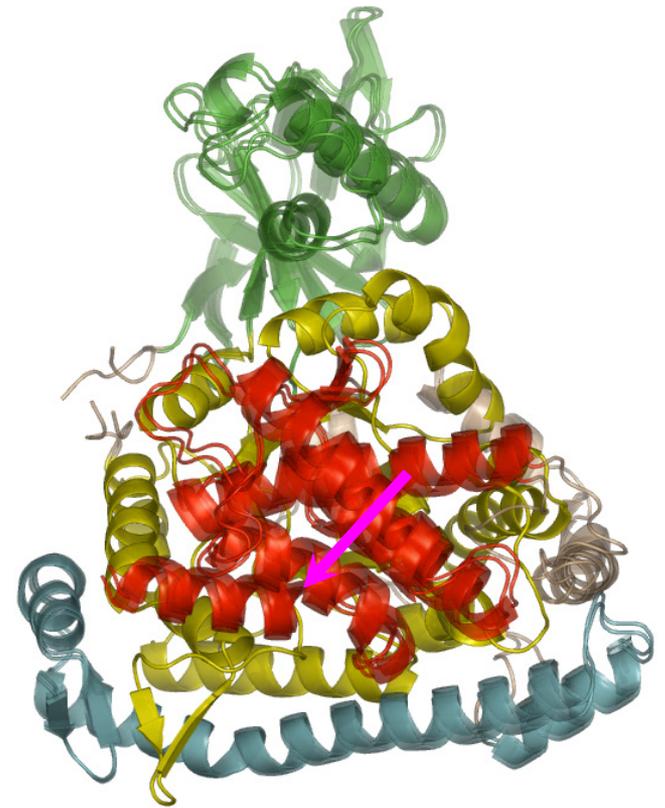


SSRL data: 220-8  $\text{\AA}$  resolution

# Impact of SAXS data on structure of MSG



Fully flexible refinement:  
Rmsd to 1D8C decreases  
from  $\sim 4.6 \text{ \AA}$  to  $\sim 3.2 \text{ \AA}$ .



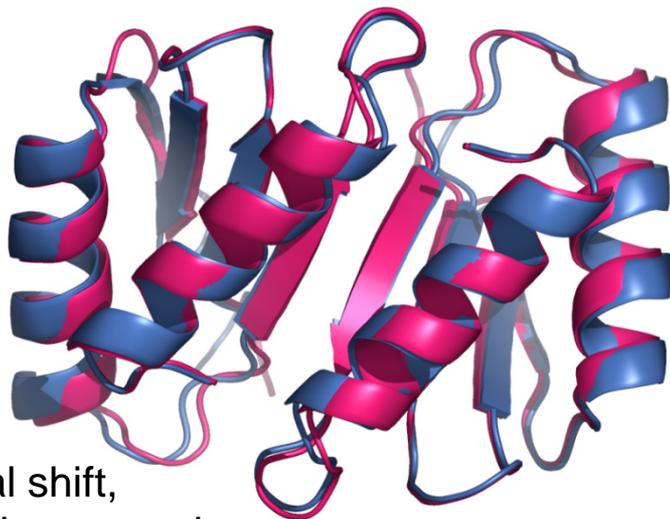
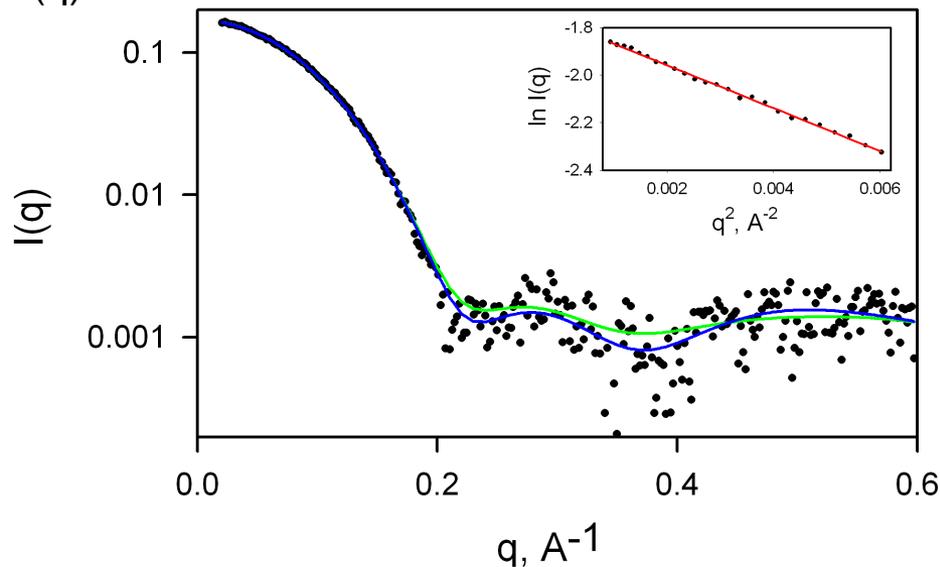
Rigid domain refinement:  
Rmsd to the X-ray (1D8C)  
reaches  $1.4 \text{ \AA}$ .

Adding  $\sim 100$  extra torsion angle  
restrains via TALOS+  
decreases rmsd from  $3.2$  to  $2.6 \text{ \AA}$

## Example 2: ToIR

### Fitted Data:

2441 NOEs,  
121 dihedral angles,  
263 RDCs,  
 $I(q)$  from 0.03 to 0.35  $\text{\AA}^{-1}$

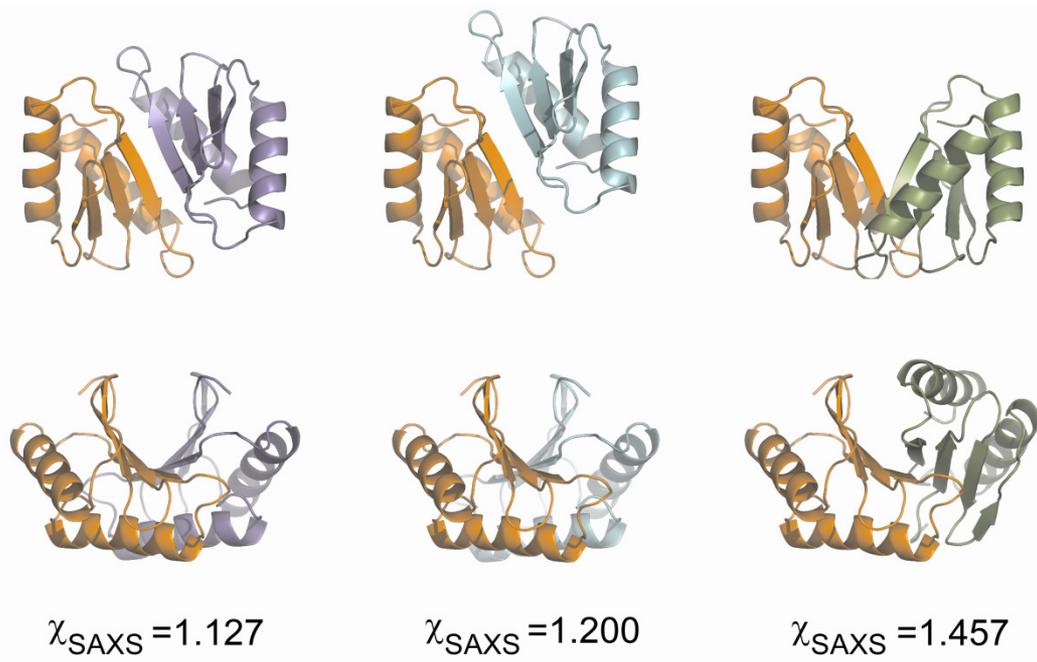
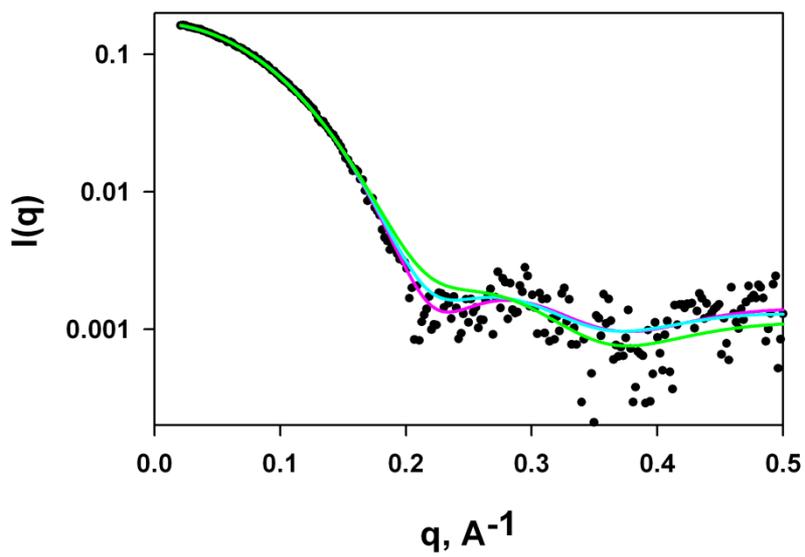
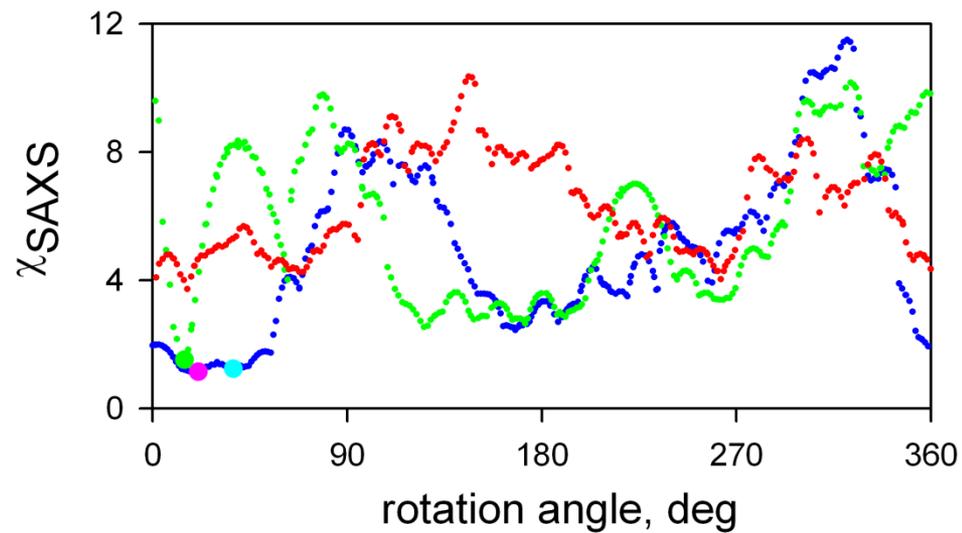
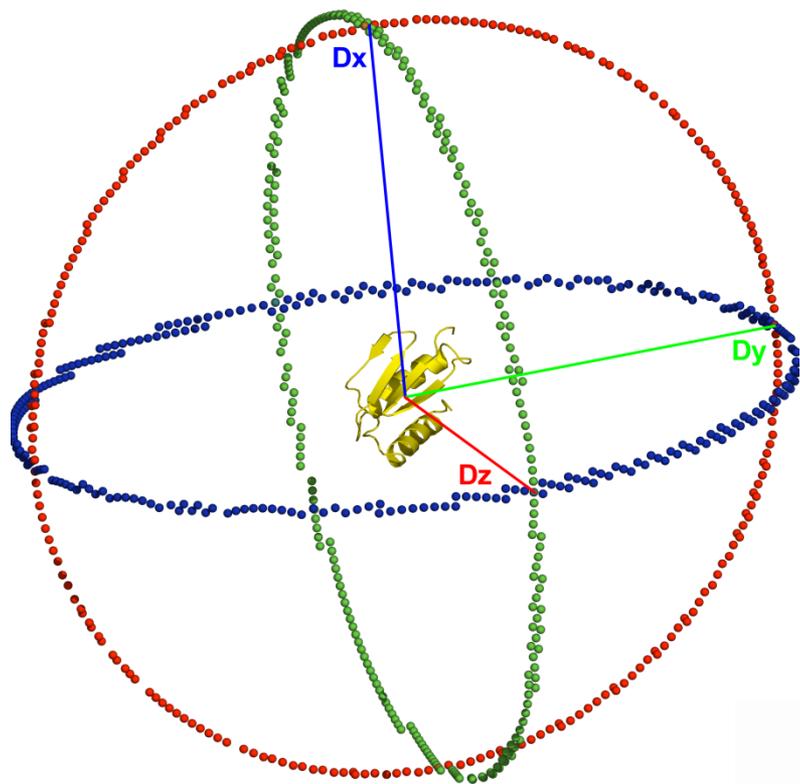


Translational shift,  
 $\sim 0.8 \text{ \AA}$  backbone rmsd

Table 1: NMR and Refinement Statistics

	no SAXS	with SAXS
rmsd from exptl restraints (mean and SD)		
$Q$ -factor (%)	$20.0 \pm 0.2$	$19.1 \pm 0.2$
distances ( $\text{\AA}$ )	$0.015 \pm 0.0005$	$0.015 \pm 0.0005$
torsion angles (deg)	$0.061 \pm 0.044$	$0.009 \pm 0.007$
deviations from idealized geometry		
bond lengths ( $\text{\AA}$ )	$0.0038 \pm 0.0001$	$0.004 \pm 1.0e^{-6}$
bond angles (deg)	$0.68 \pm 0.01$	$0.69 \pm 0.004$
impropers (deg)	$0.56 \pm 0.01$	$0.59 \pm 0.007$
HBDB energies (kT)		
$E(r, \theta, \varphi)$	$-4.58 \pm 0.05$	$-4.67 \pm 0.03$
$E(\theta'' r)$	$0.43 \pm 0.04$	$0.46 \pm 0.02$
Procheck Ramachandran distribution		
core	$96.5 \pm 1.1$	$94.2 \pm 1.0$
allowed	$3.5 \pm 1.1$	$5.8 \pm 1.0$
generous and disallowed	0.0	0.0
other Procheck statistics		
bad contacts per monomer	0.0	$1.9 \pm 0.9$
$G$ -factor	$0.21 \pm 0.02$	$0.16 \pm 0.01$
MolProbity clash score	$7.83 \pm 1.57$	$9.14 \pm 1.41$
average pairwise rmsd ( $\text{\AA}$ ) <sup>b</sup>		
all	$0.66 \pm 0.28$	$0.42 \pm 0.16$
backbone (N, C $^{\alpha}$ , C $^{\prime}$ )	$0.27 \pm 0.24$	$0.17 \pm 0.09$

# ToIR C<sub>2</sub> dimer can be assembled from RDC and SAXS data only



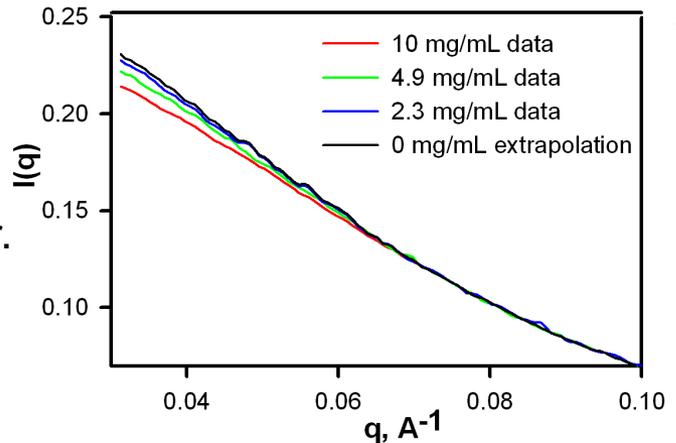
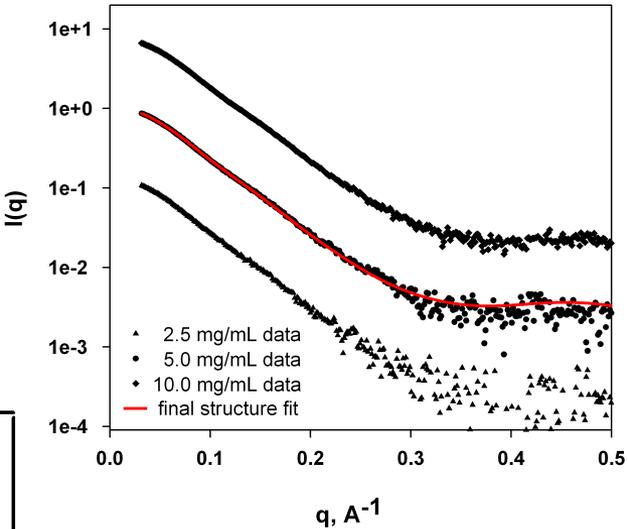
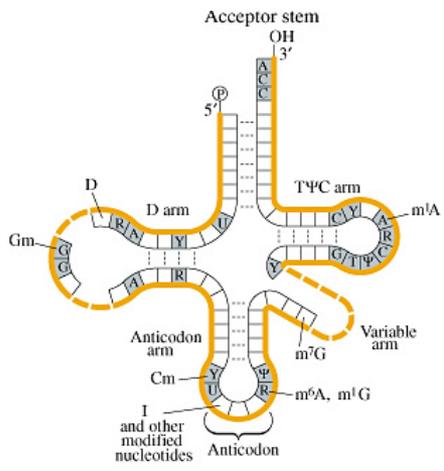
# RNA SAXS data: comparison to proteins

## Advantages:

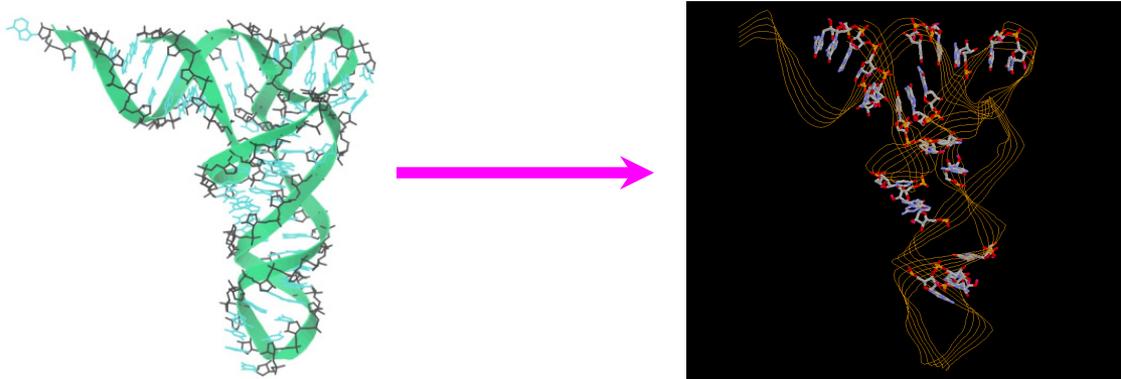
- much higher signal/noise
- much smaller chance of aggregation
- simple secondary structure (A-form helix)

## Challenges:

- more pronounced structure factor.



- RDC and inter-helical NOE restraints are often very scarce.



# RNA structure refinement with SAXS data: problems of the globbic approximation

- Becomes very expensive for large molecules
- Large errors for RNA data

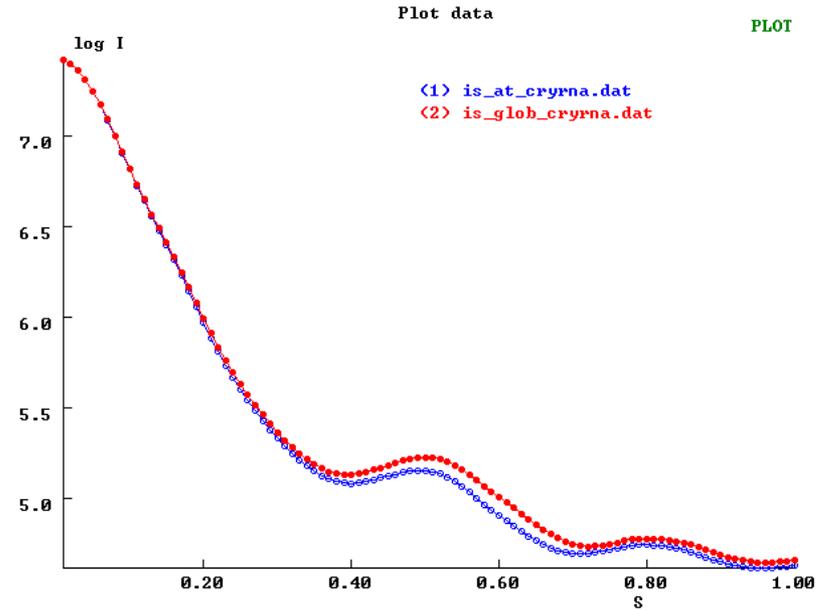
Solution – use approximate angular averaging of the complex scattering amplitude instead

$$I(q) = \left\langle \left| \mathbf{A}_a(\mathbf{q}) - \rho_o \mathbf{A}_s(\mathbf{q}) \right|^2 \right\rangle_{\Omega}$$

$$I(q) = \left\langle \left| \mathbf{A}(\mathbf{q}) \right|^2 \right\rangle_{\Omega} = \left\langle (\text{Re}[\mathbf{A}(q)])^2 + (\text{Im}[\mathbf{A}(q)])^2 \right\rangle_{\Omega}$$

The final expression for the gradient becomes

$$\nabla_m \chi^2 = \frac{4}{N_{dat} N_{grid}} \sum_{j=1}^{N_{dat}} c_j \frac{c_j I(\mathbf{q}_j) - I^o(\mathbf{q}_j)}{\sigma_j^2} g_m(\mathbf{q}_j) \sum_{k=1}^{N_{grid}} \mathbf{q}_{jk} \left\{ \cos(\mathbf{q}_{jk} \cdot \mathbf{r}_m) \text{Im}[A(\mathbf{q}_{jk})] - \sin(\mathbf{q}_{jk} \cdot \mathbf{r}_m) \text{Re}[A(\mathbf{q}_{jk})] \right\}$$



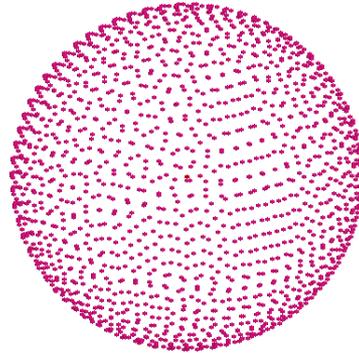
06-Nov-2007 01:59:32

C:\Alex\NIH\SAXS\tRNA

# Quasi-uniform vector grids for approximate angular averaging

## Fibonacci number-based algorithm

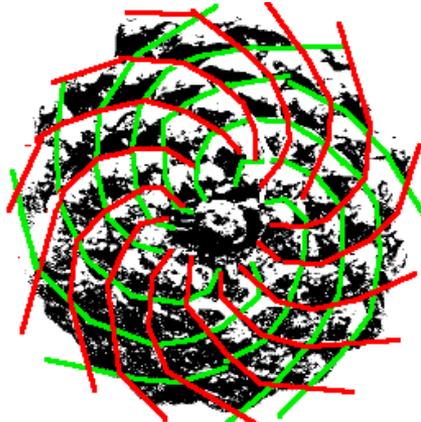
$F(1) \dots F(n)$   
 $F(i) = F(i-1) + F(i-2)$   
0, 1, 1, 2, 3, 5, 8, 13, 21, ...  
 $N_{\text{grid}} = F(n) + 1$   
 $q(j) = \arccos(1 - 2(j-1)/F(n))$   
 $f(j) = 2 * p * \text{mod}(j * F(n-1)/F(n))$



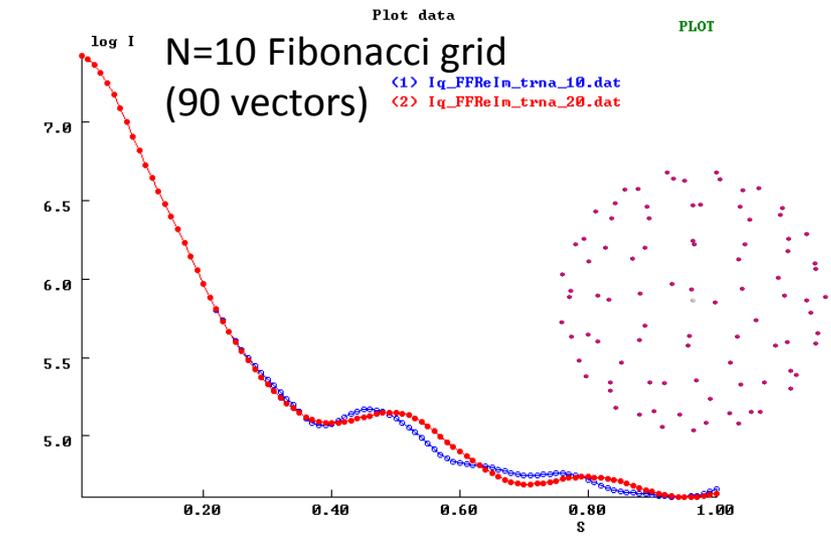
For a large grid ( $> \sim 10^3$  vectors)  
angular averaging is nearly exact

## spiral algorithm

1.....N  
 $h(j) = -1 + 2(j-1)/(n-1)$   
 $q(j) = \arccos(h(j))$   
 $f(0) = f(N) = 0$   
 $f(j) = f(j-1) + 3.6/\sqrt{n * (1 - h^2)}$

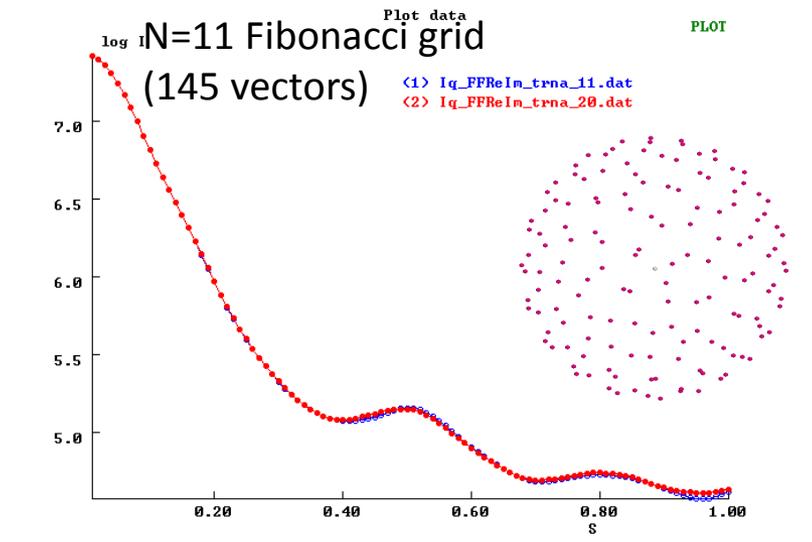


# Accuracy of data representation with small grid sizes



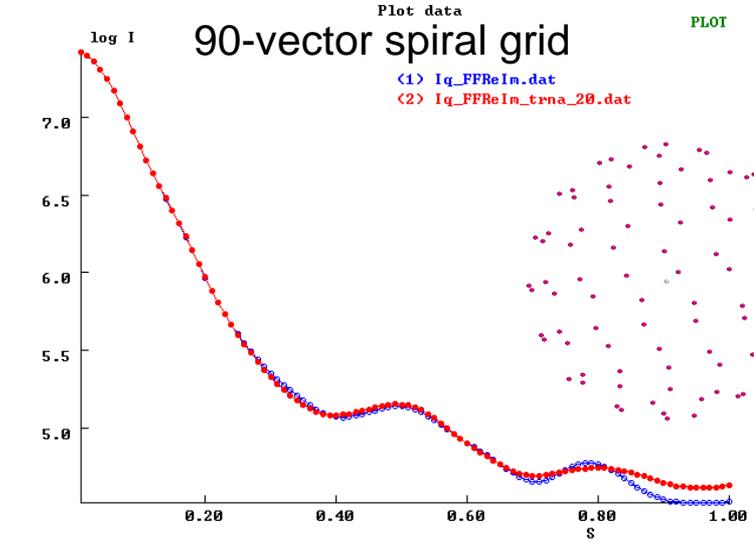
06-Nov-2007 00:59:36

C:\Alex\NIH\SAXS\trna



06-Nov-2007 01:00:53

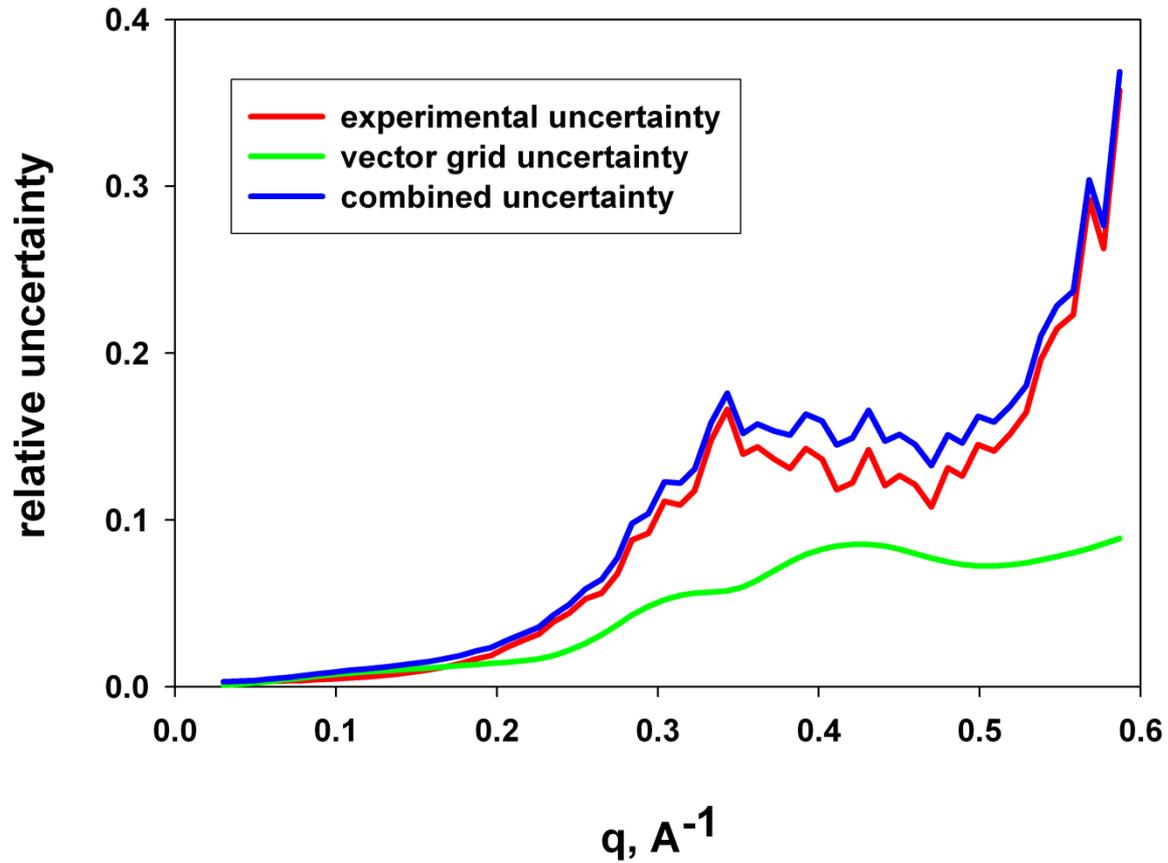
C:\Alex\NIH\SAXS\trna



06-Nov-2007 01:09:36

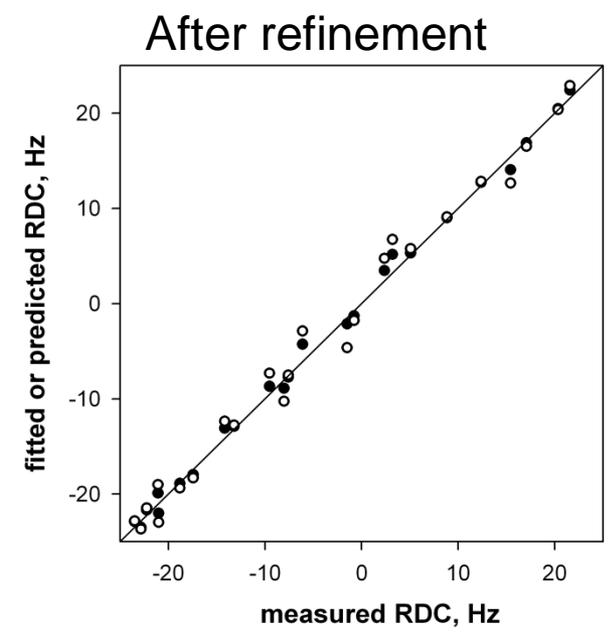
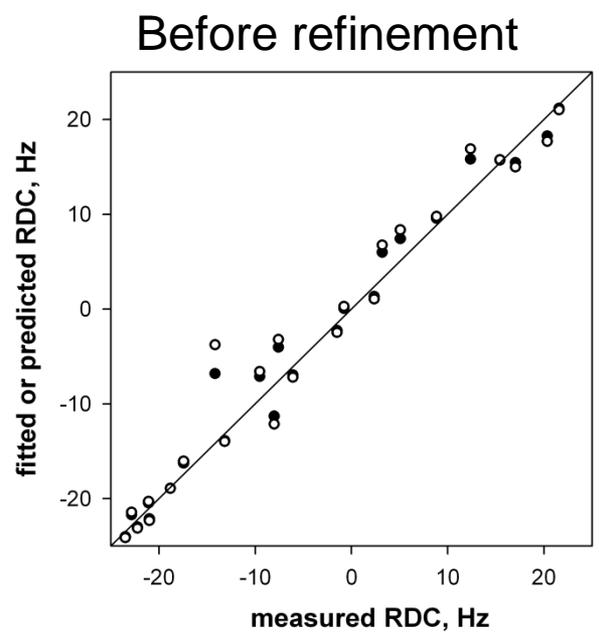
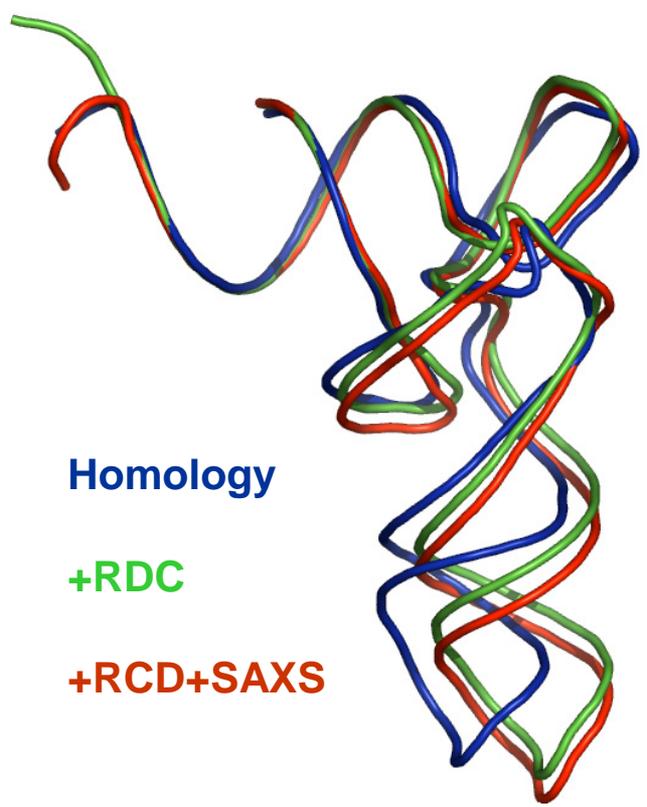
C:\Alex\NIH\SAXS\trna

The price of a 90-versor spiral grid...



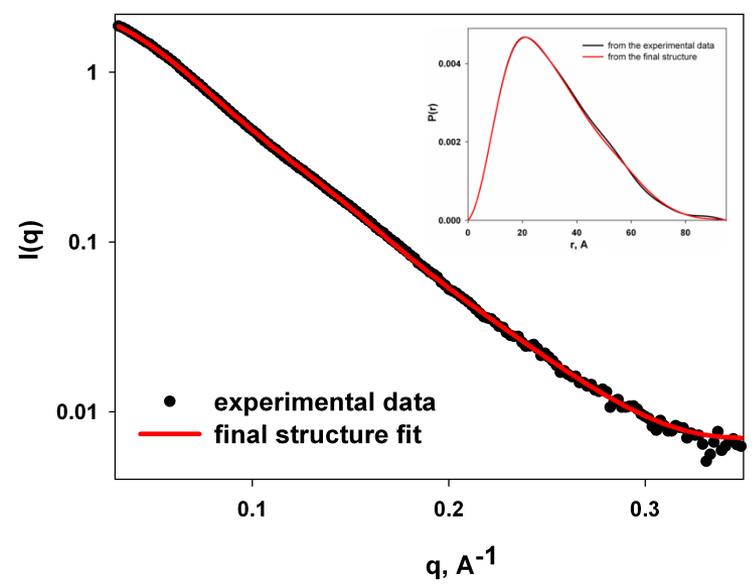
calculation time <1 sec/step

# Example: tRNA<sup>Val</sup>



$$Q = \frac{\text{rms}(D_{\text{pred}} - D_{\text{obs}})}{\sqrt{[D_a^2(4 + 3R^2)]/5}}$$

	Q <sup>free</sup> (Pf1)	RMSD to stage 1 all heavy atoms, nt. 1-72
Homology model	0.247	
+SAXS	0.211	2.61±0.26
+SAXS + RDC <sup>Pf1/MSA</sup>	0.142	2.81±0.21



## Summary for scattering amplitude-based SAXS refinement

1. Decide on the type of equidistant vector grid (Fibonacci vs Spiral) and the number of vectors.  
Compromise between calculation speed and data predictions accuracy. With spiral grid,  $N \sim 100$  works well up to  $q = 0.3 - 0.4 \text{ \AA}^{-1}$ . Decide on glob usage and composition.
2. Define the range of experimental data to be fitted and sparsen it ( $\sim 0.01 \text{ \AA}^{-1}$  step)
3. Calculate globbic form factors (fsglob\_ffreim code)
4. Evaluate  $q$ -dependent uncertainty due to the finite grid size (iq\_ffreim\_gridunc code) and add it to the experimental uncertainty as  $\sqrt{\text{expt\_unc}^2 + \text{grid\_unc}^2}$
5. Based on the current set of models, calculate bound solvent and globbic corrections (crysol and iq\_ffreim code)
6. Run the refinement program (Xplor-NIH or CNS) and obtain an updated set of structures
6. Loop steps 5-6 till convergence of the correction profiles.

Setup/example files available from <http://spin.niddk.nih.gov/bax/software/FastSAXS.zip>

# Structure refinement against SAXS data: summary

**Xplor-NIH** and **CNS** code available

Two models of calculation:

**Debye formula / globbic approximation**

$$I(q) = \sum_{i=1}^N \sum_{j=1}^N f_i^s(q) f_j^s(q) \frac{\sin(qr_{ij})}{qr_{ij}}$$

$$\nabla_{r_j} [\chi^2] \approx \sum_{k=1}^{N_q} c_k \frac{c_k I_{calc}(q_k) - I_{expt}(q_k)}{\sigma_k^2} \sum_{i \neq j}^N f_i^s(q_k) f_j^s(q_k) \left[ \cos(q_k r_{ij}) - \frac{\sin(q_k r_{ij})}{q_k r_{ij}} \right] \frac{\mathbf{r}_{ij}}{r_{ij}^2}$$

**Complex scattering amplitude/ pseudo-uniform angular averaging**

$$I(q) = \left\langle \left| \mathbf{F}_a(\mathbf{q}) - \rho_o \mathbf{F}_s(\mathbf{q}) \right|^2 \right\rangle_{\Omega} = \left\langle |\mathbf{F}(\mathbf{q})|^2 \right\rangle_{\Omega} = \left\langle (\text{Re}[\mathbf{F}(q)])^2 + (\text{Im}[\mathbf{F}(q)])^2 \right\rangle_{\Omega}$$

$$\nabla_m \chi^2 = \frac{4}{N_{dat} N_{grid}} \sum_{j=1}^{N_{dat}} c_j \frac{c_j I(\mathbf{q}_j) - I^o(\mathbf{q}_j)}{\sigma_j^2} g_m(\mathbf{q}_j) \sum_{k=1}^{N_{grid}} \mathbf{q}_{jk} \left\{ \cos(\mathbf{q}_{jk} \cdot \mathbf{r}_m) \text{Im}[A(\mathbf{q}_{jk})] - \sin(\mathbf{q}_{jk} \cdot \mathbf{r}_m) \text{Re}[A(\mathbf{q}_{jk})] \right\}$$

The choice depends on the nature and size of the system.

Debye formula is faster and accurate for smaller proteins.

Complex scattering amplitude is preferable for large proteins (>30-40 kDa) and RNA/DNA.

## **3b. Using SAXS to define global structures of complex**

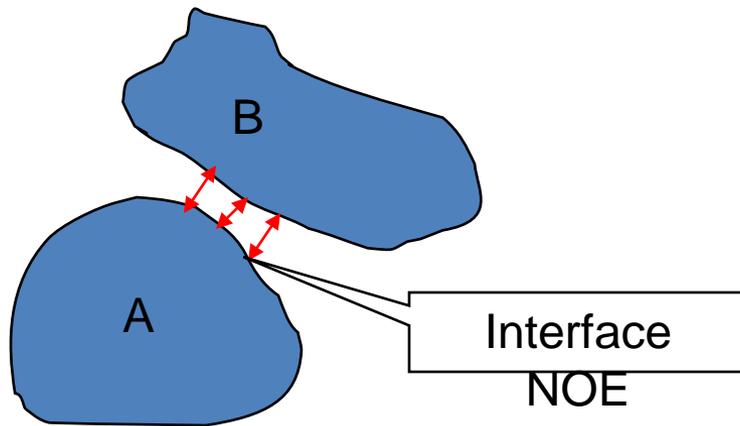
**GASR program**(Global Architecture derived from SAXS and RDC) determine the global architectures of proteins, RNAs or complexes in solution using:

1. Residual dipolar couplings (RDCs).
2. Small-angle X-ray scattering (SAXS).

**J. AM. CHEM. SOC. 2009, 131, 10507–10515**

# Structure Determination of Complexes by Solution NMR

Structures of complex are generally determined using distance, RDC restraints and subunit structures.



Interface NOE  $\longrightarrow$  Translation

RDC  $\longrightarrow$  Orientation

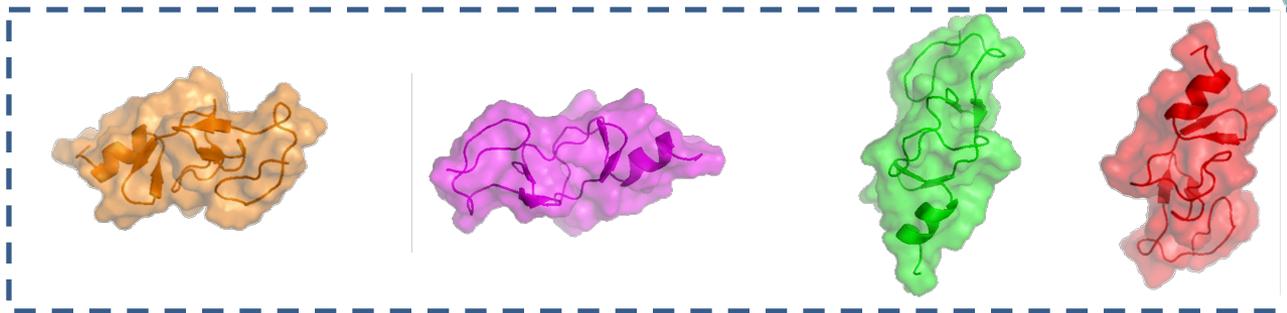
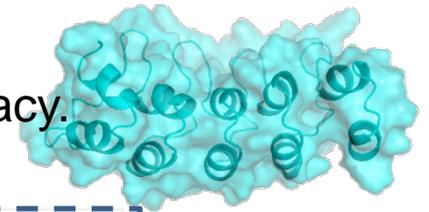
Interface NOE:

- various sophisticated
- labor-intensive
- costly isotope-labeled samples
- Less distance restraints

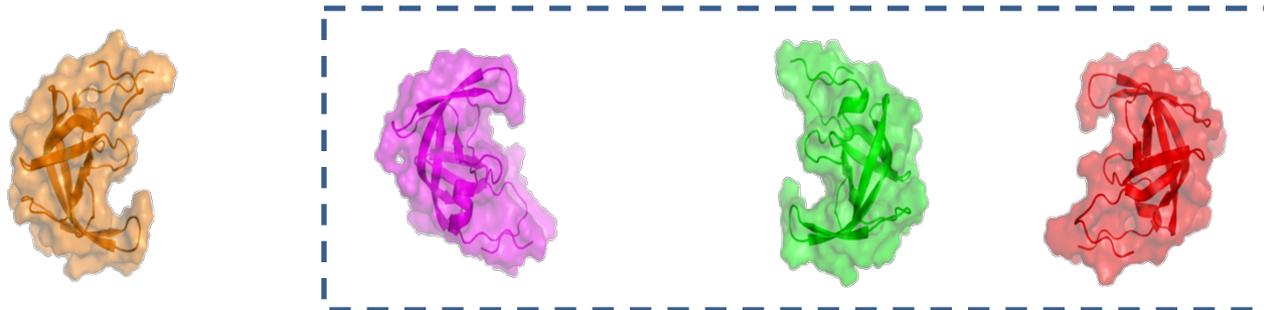
# RDC Degeneracy

Four-fold degeneracy is inherent to the orientation of any 3D structure relative to a molecular alignment tensor.

Subunit of heterodimeric complex has 4-fold degeneracy.

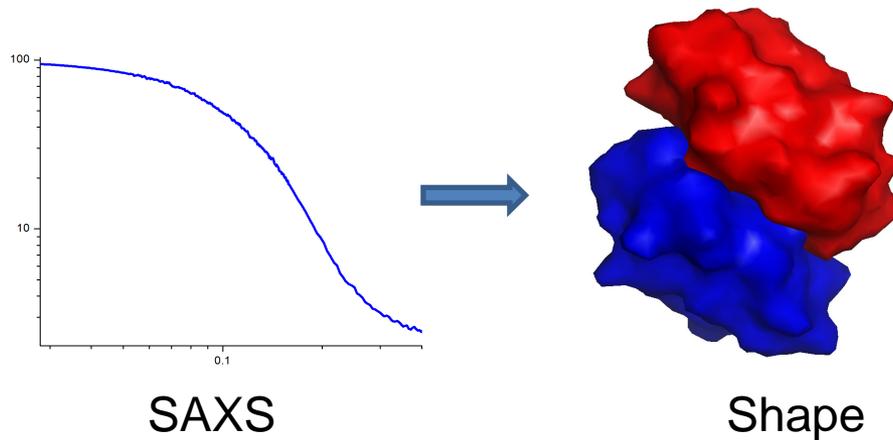


Subunit of homodimeric complex has 3-fold degeneracy.



# Small Angle X-ray Scattering(SAXS)

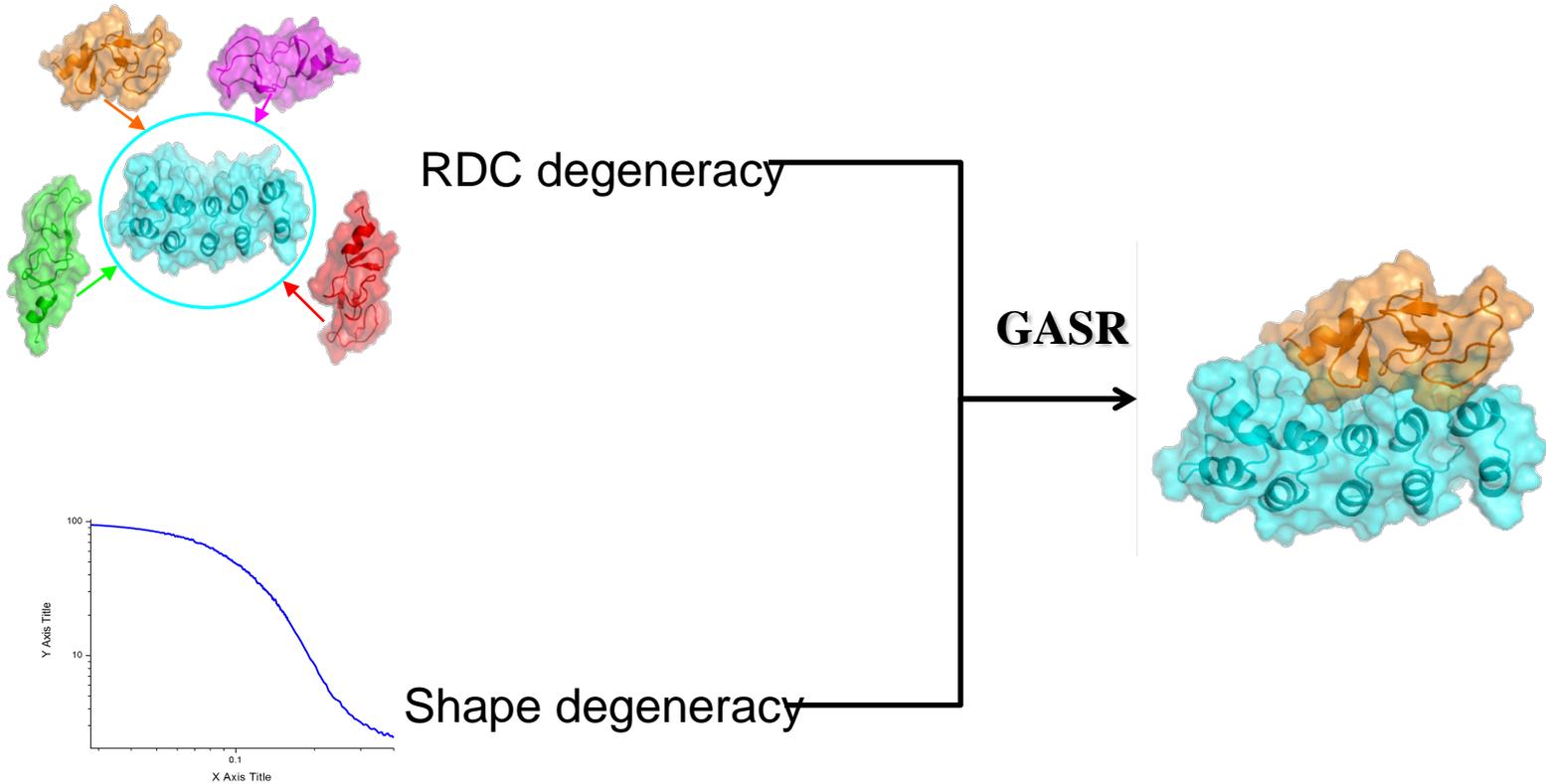
SAXS data provide global shape information, which helps determine the relative positions of subunits.



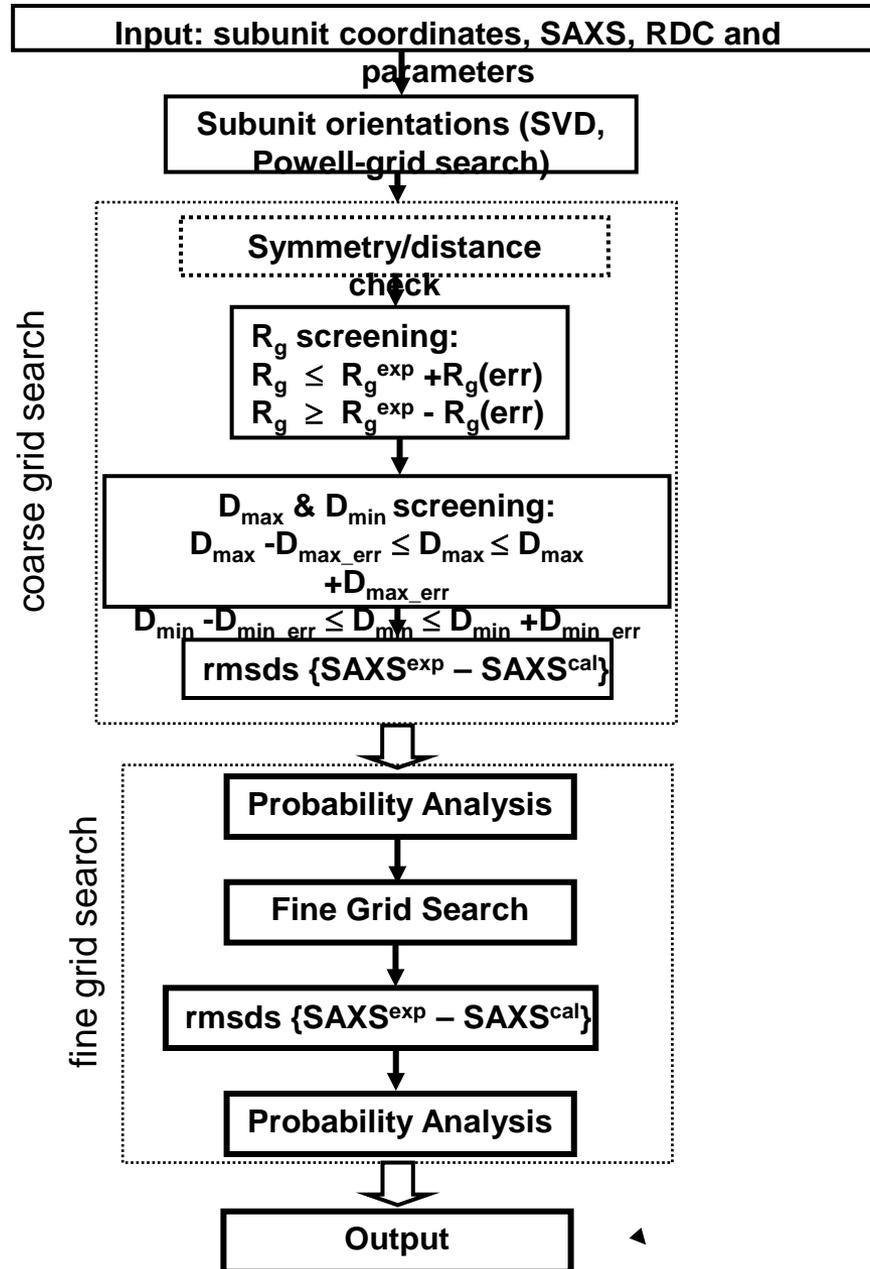
## **SAXS shape degeneracy:**

SAXS data may provide a possible list of different combinations of subunits. They all satisfy the SAXS data within experimental error range.

# GASR Concept

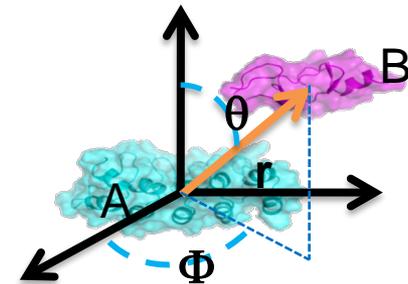
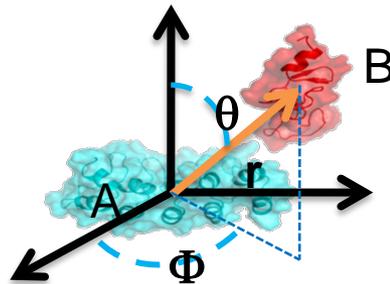
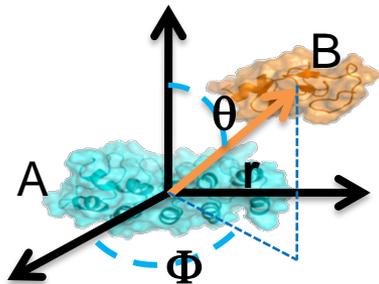
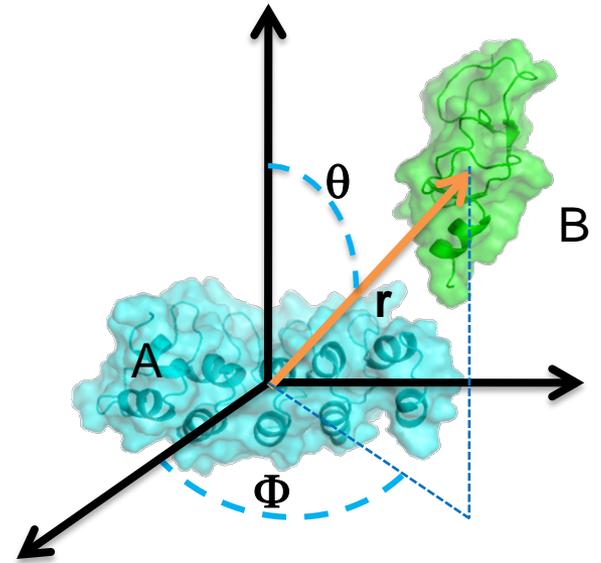


# GASR Program Flowchart



# Grid Search

- Assume Rigid-body.
- Fix the position and orientation of subunit A at the center.
- Search subunit B with one of four possible orientations in polar space  $r$ ,  $\theta$  and  $\phi$ .
- Filter each grid by  $R_g$ ,  $D_{\max}$ ,  $D_{\min}$  and rmsd.



# SAXS Data Analysis

The simplified Debye Formula is used in GASR method.

$$I(q) = \sum_{i=1}^N I_i(q) + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^N F_i(q) F_j(q) \frac{\sin qr_{ij}}{qr_{ij}}$$

$$I(q) = I_a(q) + I_b(q) + I_{ab}(q)$$

$$I_a(q) = \sum_{i=1}^m I_i(q) + 2 \sum_{i=1}^{m-1} \sum_{j=i+1}^m A_i(q) A_j(q) \frac{\sin qr_{ij}}{qr_{ij}}$$

$$I_b(q) = \sum_{i=m+1}^N I_i(q) + 2 \sum_{i=m+1}^{N-1} \sum_{j=i+1}^N A_i(q) A_j(q) \frac{\sin qr_{ij}}{qr_{ij}}$$

$$I_{ab}(q) = 2 \sum_{i=1}^m \sum_{j=m+1}^N A_i(q) A_j(q) \frac{\sin qr_{ij}}{qr_{ij}}$$

**Score function:**

$$rmsd\% = \frac{1}{N-1} \sum_q \left[ \frac{I_{\text{exp}}(q) - cI_{\text{cal}}(q)}{I_{\text{exp}}(q)} \right]^2 \times 100\%$$

# Boundary Effects on Grid Search

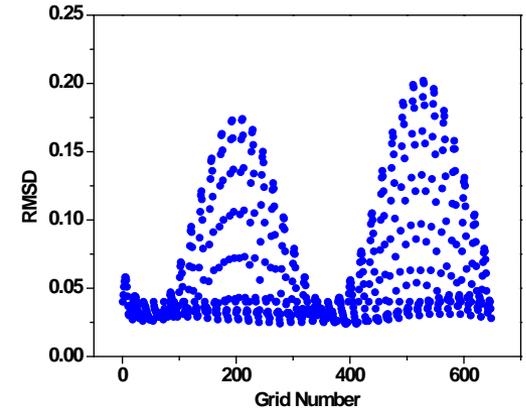
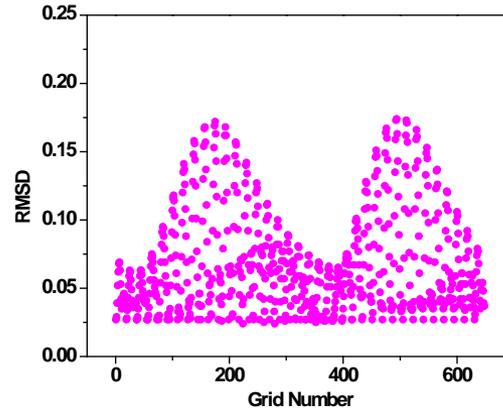
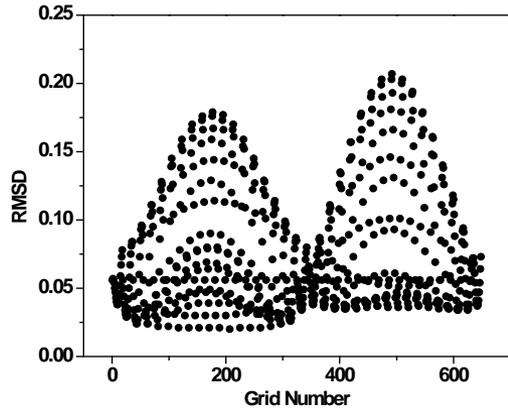
GB1 homodimer coarse search  
 Grid size:  $r=18\text{\AA}$ ,  $\Delta\theta, \Delta\phi=10^\circ$   
 Number of grids =  $18(\theta) \times 36(\phi) \times 3(\text{orientation}) = 378$

Orientation 0

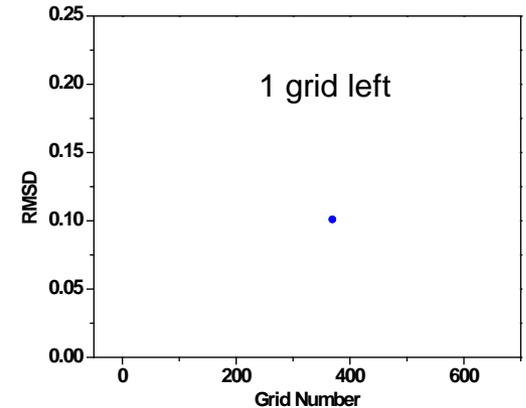
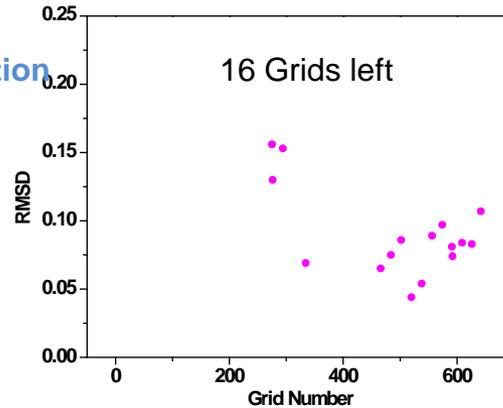
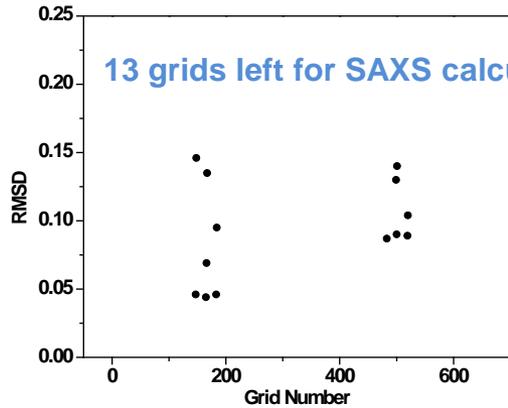
Orientation 2

Orientation 3

Without boundary restraints



With boundary restraints



GASR program is fast and efficient. GB1 (2x14KD) 10mins in a laptop.

# GASR Solved Complex Structures

## Homodimeric proteins:

HIV-1 protease complex with simulated RDCs.

## Two-domain proteins with a linker between two domains:

L11 and  $\gamma$ D-Crystallin.

## Heterodimeric complex(tightly bound):

ILK ARD-PINCH LIM1 complex.

## Homodimeric Protein Complex(weakly associating proteins) :

GB1side-by-side dimer.

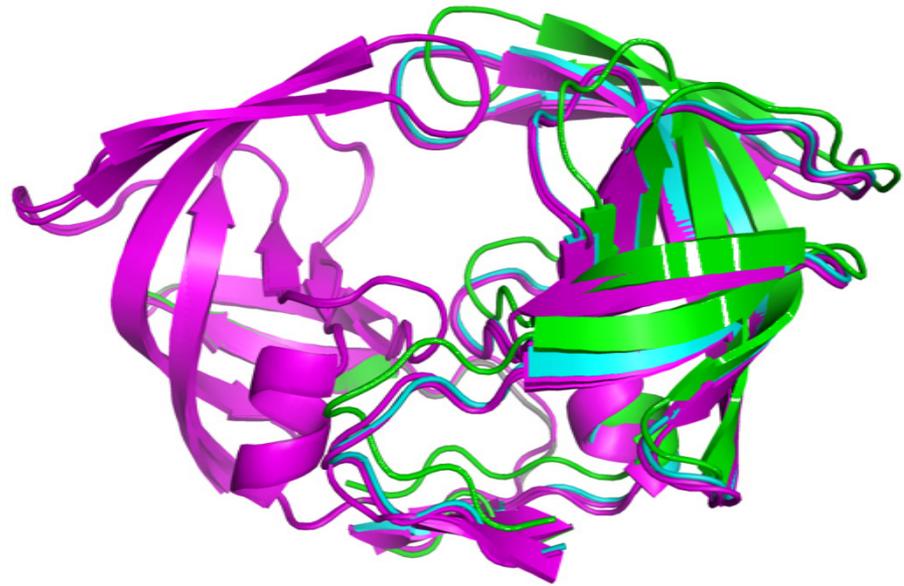
## RNA complex:

Homodimeric tetraloop-receptor RNA complex.

# HIV-1 Protease Structures

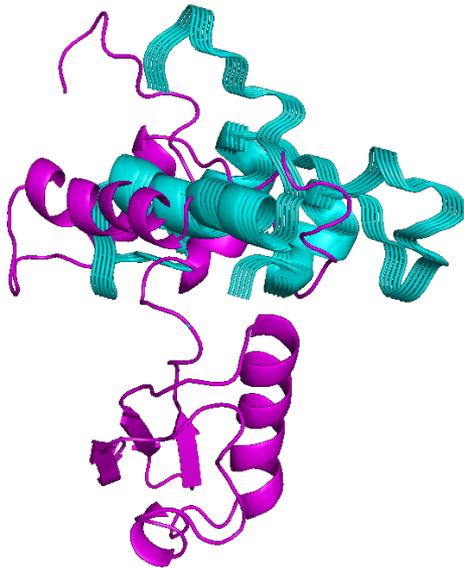
The HIV-1 protease is a homodimeric globular protein, with each subunit comprising 99 residues.

Original NMR structure (1BVE, magenta, **218** inter-NOE)  
GASR structure (simulated RDC, cyan, backbone rmsd=0.21Å)  
GASR structure (simulated RDC with added noise, green, backbone rmsd=0.87Å)



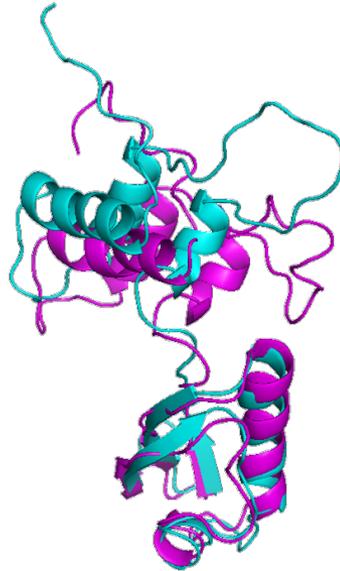
# Ribosomal Protein L11 Structures

A



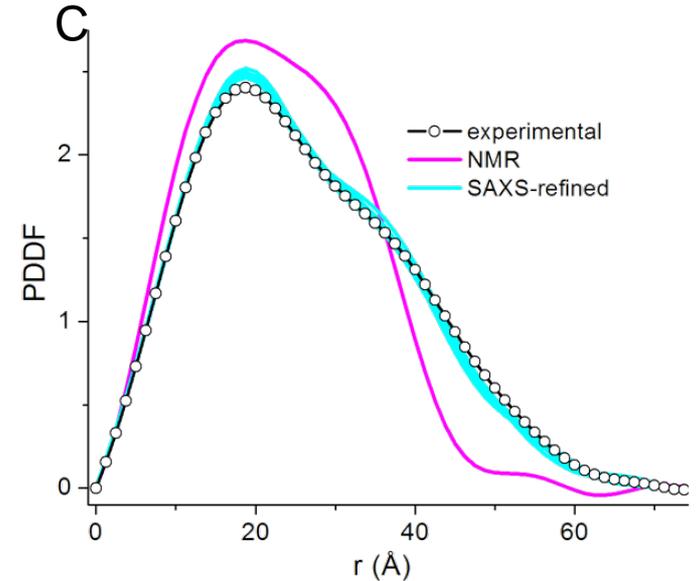
Top 10% GASR structures (cyan)

B



SAXS-refined GASR structure (cyan)

C



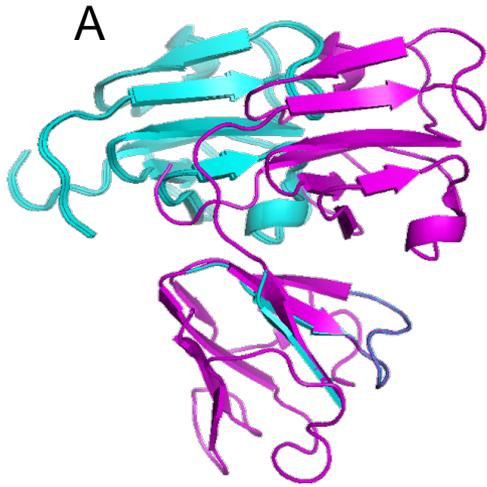
PDDF curves

Non-SAXS refined NMR structure  
(magenta, few inter-NOE, PDB  
code: 2e35)

147-aa ribosomal protein L11 was previously determined using NOE distance-, RDC- and the small-angle neutron scattering (SANS). Two domains, 5-residue linker.

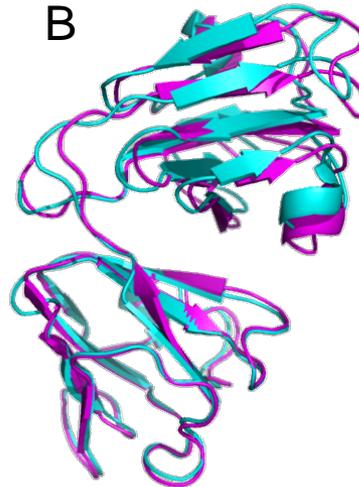
GASR treats L11 as a heterodimeric protein by breaking the covalent bond Pro72 and 73 in the linker and each individual domain was used as component input. low to medium quality starting structure.

# $\gamma$ D-Crystallin (P23T Mutant) Structures



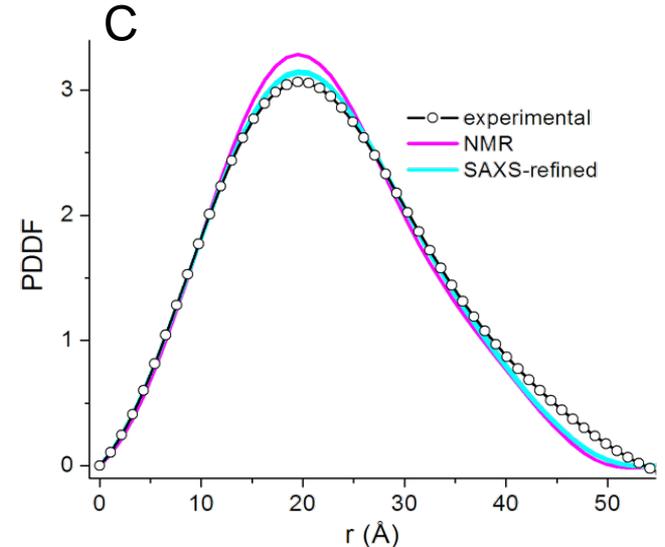
GASR structure (cyan)

Non-SAXS refined  
NMR structure  
(magenta)



SAXS-refined GASR structure

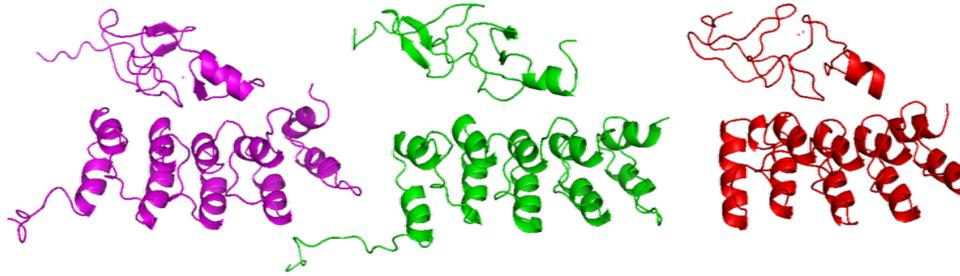
(cyan)  
RMSD=1.1Å



PDDF curves

human  $\gamma$ D-Crystallin protein (177aa) consists of two domains that are linked by a short and nonflexible linker. Both domains are highly globular

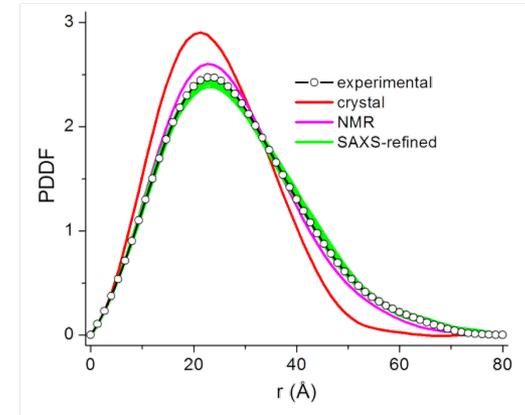
# ILK ARD/PINCH LIM1 Complex Structures



NMR structure (magenta) (PDB ID: 2kbx)

SAXS-refined GASR structure (green)

X-ray crystal structure (red) (PDB ID:6f6q)

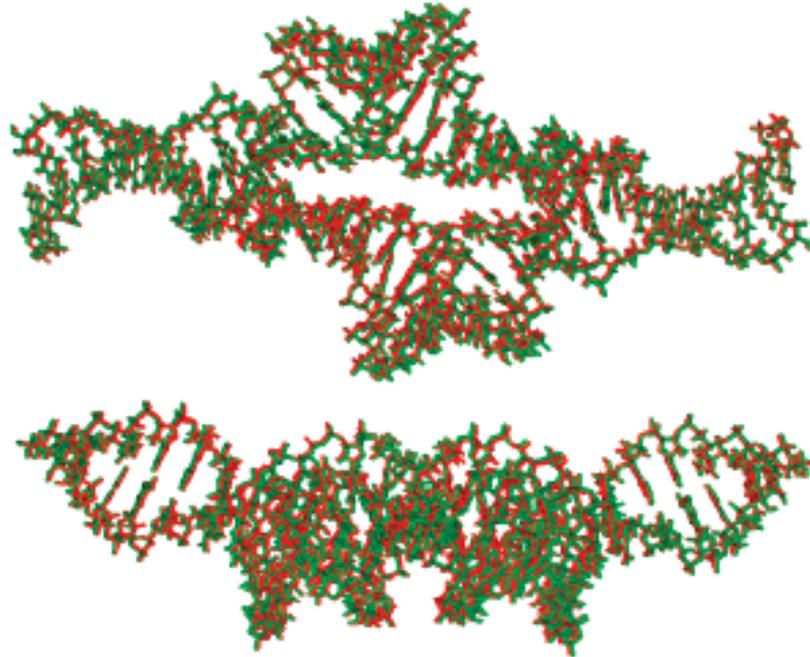


PDDF curves

The large difference in PDDF between the back-calculated curve from the X-ray structure and the experimental one suggest that a significant difference between the structure in solution and the crystal is present.

- (i) the solution NMR structures of the ILK ARD and the PINCH LIM1 are of a low (LIM1) and medium (ILK) quality;
- (ii) both proteins contain relatively large numbers of nonstructured regions, complicating the interpretation of the SAXS and RDC data;
- (iii) relatively a low quality of RDC data, especially for PINCH LIM1, resulted in uncertainty in the determination of the alignment tensor and translated into errors in the four discrete orientations of each subunit in the complex.

# homodimeric tetraloop-receptor RNA comp



Side and top views of the tetraloop receptor homodimeric structures, refined **with 36 2 distance and hydrogen bond restraints** (red) (accession code: 2jyj) and defined by **SAXS data** using a rigid-body calculation (green) (accession code: 2jyh).

The rmsd between the two structures is 0.4 Å.

Rg comparison: 25.1 Å (NMR), 23.0 Å (GASR) and 23.2 Å (experiment)

# GASR Input

Input data file names →

```
# monomer A
fn_ma = gbl_a.pdb

# monomer B
fn_mb = gbl_b.pdb

# SAXS experimental data
fn_saxs = gbl_saxs.dat

# Euler angles
fn_euler = gbl_euler.dat
```

Boundary Parameters →

```
# Rg
Rg = 14.60
Rg_err = 1.0

# Dmax
Dmax = 45.00
Dmax_err = 10.0

# Dmin
Dmin_min_cutoff = 2.00
Dmin_max_cutoff = 3.00
Dmin_fsmax_cutoff = 3.00

# Search Regions search regions
Search_region = [0, 1, 2, 3]
```

Vicinity & Symmetry →

```
# Vicinity restraint
# Vicinity restraints identify parts of components that neighbor each
# other in form of DISTANCE
# DISTANCE filter

DISTANCE = OFF

# if DISTANCE is turned on, you have define the following:
# resid atom resid atom distance lower_error upper_error
# distance_max = distance - lower_error
# distance_max = distance + upper_error

# C2 Symmetry (only for homodimer)
SYMMETRY = On
# Symmetry check conditions
sym_cutoff = 0.1
```

Other Parameters →

```
# SAXS Parameters
rmsd_cutoff = 0.2500
#rom is the Expansion factor in SAXS back-calculation
#1.000 is recommended for calculations of proteins
#q_max is the maximum q range that is used for the GASR calculation
rom = 1.000
q_max = 0.250
```

# GASR Input--Input Files

# input file for a GASR calculation (a sample input file actually used # for GB1 calculation)

# monomer A  
fn\_ma = gb1\_a.pdb

# monomer B  
fn\_mb = gb1\_b.pdb

# SAXS experimental data  
fn\_saxs = gb1\_saxs.dat

# Euler angles  
fn\_euler = gb1\_euler.dat

```
# monomer A
fn_ma = gb1_a.pdb

# monomer B
fn_mb = gb1_b.pdb

# SAXS experimental data
fn_saxs = gb1_saxs.dat

# Euler angles
fn_euler = gb1_euler.dat
```

```
# Rg
Rg = 14.60
Rg_err = 1.0

# Dmax
Dmax = 45.00
Dmax_err = 10.0

# Dmin
Dmin_min_cutoff = 2.00
Dmin_max_cutoff = 3.00
Dmin_fsmax_cutoff = 3.00

# Search Regions search regions
Search_region = [0, 1, 2, 3]

# Vicinity restraint
# Vicinity restraints identify parts of components that neighbor each
# other in form of DISTANCE
# DISTANCE filter

DISTANCE = OFF

# if DISTANCE is turned on, you have define the following:
# resid atom resid atom distance lower_error upper_error
# distance_max = distance - lower_error
# distance_max = distance + upper_error

# C2 Symmetry (only for homodimer)
SYMMETRY = On
# Symmetry check conditions
sym_cutoff = 0.1

# SAXS Parameters
rmed_cutoff = 0.2500
#rcm is the Expansion factor in SAXS back-calculation
#1.000 is recommended for calculations of proteins
#q_max is the maximum q range that is used for the GASR calculation
rcm = 1.000
q_max = 0.250
```

# GASR Input—Boundary and Other Parameters

# Rg

Rg = 14.60

Rg\_err = 1.0

# Dmax

Dmax = 45.00

Dmax\_err = 10.0

# Dmin

Dmin\_min\_cutoff = 2.00

Dmin\_max\_cutoff = 3.00

Dmin\_fsmax\_cutoff = 3.00

# Search orientations

Search\_region = [0, 2, 3]

# SAXS Parameters

rmsd\_cutoff = 0.2500

#rom is the Expansion factor in SAXS back-calculation

#1.000 is recommended for calculations of proteins

#q\_max is the maximum q range that is used for the GASR calculation

rom = 1.000

q\_max = 0.250

```
# monomer A
fn_ma = gbl_a.pdb

# monomer B
fn_mb = gbl_b.pdb

# SAXS experimental data
fn_saxs = gbl_saxs.dat

# Euler angles
fn_euler = gbl_euler.dat

# Rg
Rg = 14.60
Rg_err = 1.0

# Dmax
Dmax = 45.00
Dmax_err = 10.0

# Dmin
Dmin_min_cutoff = 2.00
Dmin_max_cutoff = 3.00
Dmin_fsmax_cutoff = 3.00

# Search Regions search regions
Search_region = [0, 1, 2, 3]

# Vicinity restraint
# Vicinity restraints identify parts of components that neighbor each
# other in form of DISTANCE
# DISTANCE filter

DISTANCE = OFF

# if DISTANCE is turned on, you have define the following:
# resid atom resid atom distance lower_error upper_error
# distance_max = distance - lower_error
# distance_max = distance + upper_error

# C2 Symmetry (only for homodimer)
SYMMETRY = On
# Symmetry check conditions
sym_cutoff = 0.1

# SAXS Parameters
rmsd_cutoff = 0.2500
#rom is the Expansion factor in SAXS back-calculation
#1.000 is recommended for calculations of proteins
#q_max is the maximum q range that is used for the GASR calculation
rom = 1.000
q_max = 0.250
```

# GASR Input—Vicinity & Symmetry

# Vicinity restraint

# Vicinity restraints identify parts of components that neighbor each

# other in form of DISTANCE

# DISTANCE filter

DISTANCE = OFF

# if DISTANCE is turned on, you have define the following:

# resid atom resid atom distance lower\_error upper\_error

# distance\_max = distance - lower\_error

# distance\_max = distance + upper\_error

# C2 Symmetry (only for homodimer)

SYMMETRY = On

# Symmetry check conditions

sym\_cutoff = 0.1

DISTANCE = On

72 C 73 N 5 4 4

```
# monomer A
fn_ma = gbl_a.pdb

# monomer B
fn_mb = gbl_b.pdb

# SAXS experimental data
fn_saxs = gbl_saxs.dat

# Euler angles
fn_euler = gbl_euler.dat
```

```
# Rg
Rg = 14.60
Rg_err = 1.0

# Dmax
Dmax = 45.00
Dmax_err = 10.0
```

```
# Dmin
Dmin_min_cutoff = 2.00
Dmin_max_cutoff = 3.00
Dmin_fsmax_cutoff = 3.00
```

```
# Search Regions search regions
Search_region = [0, 1, 2, 3]
```

```
# Vicinity restraint
# Vicinity restraints identify parts of components that neighbor each
# other in form of DISTANCE
# DISTANCE filter
```

```
DISTANCE = OFF
```

```
# if DISTANCE is turned on, you have define the following:
# resid atom resid atom distance lower_error upper_error
# distance_max = distance - lower_error
# distance_max = distance + upper_error
```

```
# C2 Symmetry (only for homodimer)
SYMMETRY = On
# Symmetry check conditions
sym_cutoff = 0.1
```

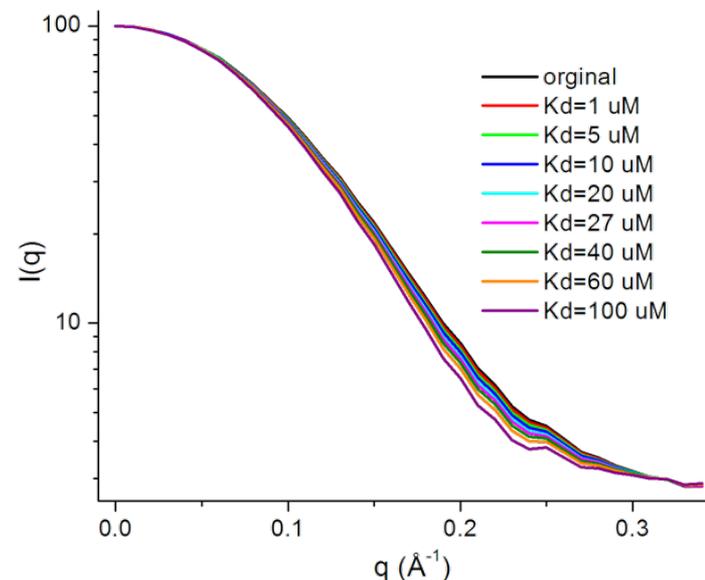
```
# SAXS Parameters
rmed_cutoff = 0.2500
#rcm is the Expansion factor in SAXS back-calculation
#1.000 is recommended for calculations of proteins
#q_max is the maximum q range that is used for the GASR calculation
rcm = 1.000
q_max = 0.250
```

# GB1 SAXS Input Curve after correction

#q	Iq
0	100
0.01	99.30151
0.02	97.23993
0.03	93.89665
0.04	89.40052
0.05	83.94819
0.06	78.08219
0.07	70.69714
0.08	63.29852
0.09	55.62186
0.1	48.94141
0.11	42.22433
...	

Two column data  
q range: 0~0.25

Both monomer and dimer species coexist in equilibrium in solution and their relative proportions were calculated from the dissociation constant that was measured as  $27 \pm 4 \mu\text{M}$  at room temperature. The scattering contribution from dimer was calculated by subtracting the monomer contribution.



Simulated SAXS curves using GB1-A34F coordinates (accession code: 2rmm), at various  $K_d$  values for the protein concentration used in the actual SAXS experiment.

# Derive GB1 Complex Structure

Input:

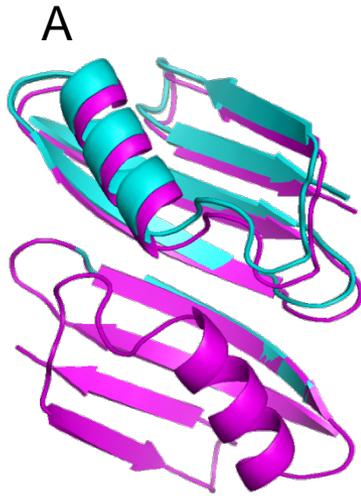
```
C:\ Prompt
```

```
C:\GASR\examples\gb1>..\..\gasr.exe gb1.inp
```

Output:

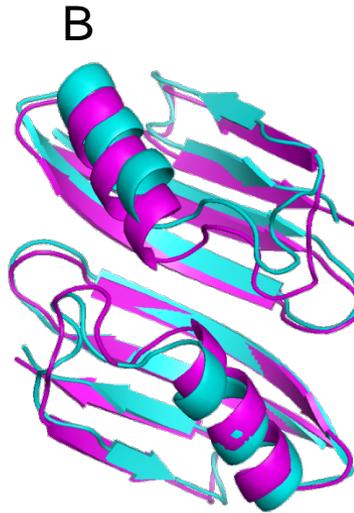
```
# top 10 percent conformers: (20 conformers)
# Rg= 14.48 Dmax= 46.17 Dmin= 2.08 Iq_rmsd= 0.0338 pdb: gb1_r0_1.pdb
# Rg= 14.48 Dmax= 46.15 Dmin= 2.15 Iq_rmsd= 0.0340 pdb: gb1_r0_2.pdb
# Rg= 14.48 Dmax= 46.13 Dmin= 2.21 Iq_rmsd= 0.0341 pdb: gb1_r0_3.pdb
# Rg= 14.48 Dmax= 46.12 Dmin= 2.28 Iq_rmsd= 0.0343 pdb: gb1_r0_4.pdb
# Rg= 14.48 Dmax= 46.10 Dmin= 2.34 Iq_rmsd= 0.0344 pdb: gb1_r0_5.pdb
# Rg= 14.48 Dmax= 46.06 Dmin= 2.16 Iq_rmsd= 0.0344 pdb: gb1_r0_6.pdb
# Rg= 14.48 Dmax= 46.09 Dmin= 2.41 Iq_rmsd= 0.0345 pdb: gb1_r0_7.pdb
# Rg= 14.48 Dmax= 46.05 Dmin= 2.24 Iq_rmsd= 0.0345 pdb: gb1_r0_8.pdb
# Rg= 14.48 Dmax= 46.07 Dmin= 2.46 Iq_rmsd= 0.0346 pdb: gb1_r0_9.pdb
# Rg= 14.48 Dmax= 46.04 Dmin= 2.31 Iq_rmsd= 0.0346 pdb: gb1_r0_10.pdb
# Rg= 14.48 Dmax= 46.06 Dmin= 2.49 Iq_rmsd= 0.0347 pdb: gb1_r0_11.pdb
# Rg= 14.48 Dmax= 46.03 Dmin= 2.34 Iq_rmsd= 0.0347 pdb: gb1_r0_12.pdb
# Rg= 14.48 Dmax= 46.04 Dmin= 2.52 Iq_rmsd= 0.0348 pdb: gb1_r0_13.pdb
# Rg= 14.48 Dmax= 46.02 Dmin= 2.37 Iq_rmsd= 0.0348 pdb: gb1_r0_14.pdb
# Rg= 14.48 Dmax= 46.03 Dmin= 2.55 Iq_rmsd= 0.0349 pdb: gb1_r0_15.pdb
# Rg= 14.48 Dmax= 46.01 Dmin= 2.40 Iq_rmsd= 0.0349 pdb: gb1_r0_16.pdb
# Rg= 14.48 Dmax= 46.01 Dmin= 2.59 Iq_rmsd= 0.0350 pdb: gb1_r0_17.pdb
# Rg= 14.48 Dmax= 46.01 Dmin= 2.43 Iq_rmsd= 0.0350 pdb: gb1_r0_18.pdb
# Rg= 14.48 Dmax= 46.00 Dmin= 2.46 Iq_rmsd= 0.0350 pdb: gb1_r0_19.pdb
# Rg= 14.48 Dmax= 45.99 Dmin= 2.49 Iq_rmsd= 0.0351 pdb: gb1_r0_20.pdb
#
# save 10 percent structures in the directory: top10
# region 0: 100.0(%) in top 10 percent(20 trjs) by rmsd(Iq)
```

# GB1 (A34F Mutant) Structures

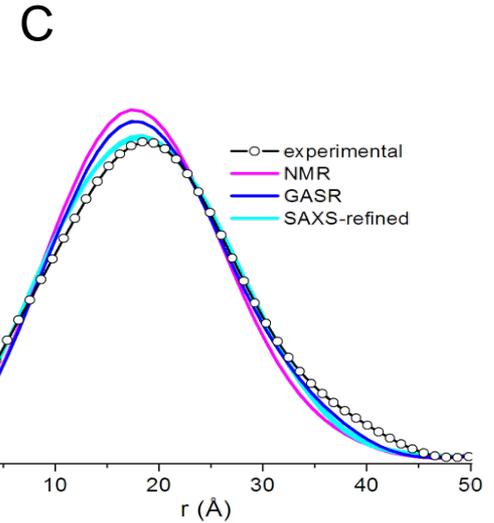


GASR structure (cyan)

Non-SAXS refined NMR  
structure (magenta)



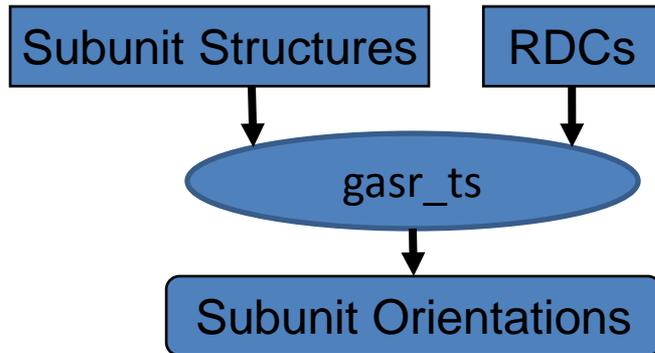
SAXS-refined GASR structure (cyan)  
RMSD=0.7Å



PDDF curves

Solution NMR (PDB code: 2rmm) : 50 inter-NOE distances

# Determination of the relative orientation of domains



Xplor-NIH  
Format

## RDC Input File

```
!RDC_restraints

assign ( resid 500 and name OO )
  ( resid 500 and name Z )
  ( resid 500 and name X )
  ( resid 500 and name Y )
  ( resid 2 and name H )
  ( resid 2 and name N ) -0.9100 1.2000 1.2000

assign ( resid 500 and name OO )
  ( resid 500 and name Z )
  ( resid 500 and name X )
  ( resid 500 and name Y )
  ( resid 3 and name H )
  ( resid 3 and name N ) -21.6100 1.2000 1.2000

.....
```

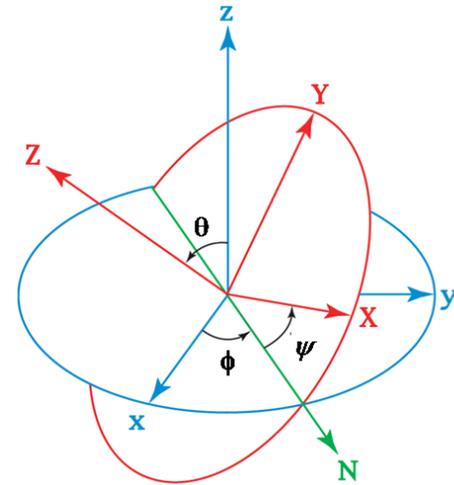
# Calculate Euler Tensor

Using **gasr\_ts.exe**(in windows) or **gasr\_ts.py** in linux:

```
gasr_ts.exe fn_pdb fn_rdc
```

fn\_pdb: the pdb file of subunit.

fn\_rdc: the rdc file in Xplor format.



**Euler Angles**

```
c:\ Prompt
C:\GASR\examples\gb1>..\..\gasr_ts.exe gb1_a.pdb gb1_rdc.tbl_
```

Part of output:

```
REMARK
REMARK VarTensor phi : 353.81 353.81 173.81 173.81
REMARK VarTensor theta : 67.05 247.05 112.95 292.95
REMARK VarTensor psi : 10.60 349.40 349.40 10.60
# Powell result : phi= 353.8 the= 67.0 psi= 10.6 Da= 16.41 R= 0.65 Dc= 0.00
# save lines into file: gb1_a_euler.dat
```

# GASR Program

Data evaluation using GASR is relatively straightforward and less labor intensive than recording and interpreting various types of heteronuclear multidimensional NMR spectra.

GASR program is fast and efficient.

**GASR approach is ideally suited to aid in the structure determination of multicomponent proteins and complexes in solution.**

**Limitations:**

Ambiguous results can ensue when component shapes are highly globular or very symmetrical, or initial structures of components are less well determined.

**One powerful remedy** – “proximity”(distance information) readily be obtained

from chemical cross-linking, heuristic biochemical information, compensatory mutagenesis, chemical shift perturbation or paramagnetic relaxation enhancement.