Ab initio methods: how/why do they work

D.Svergun

EMBO

EMBL

Major problem for biologists using SAS

- In the past, many biologists did not believe that SAS yields more than the radius of gyration
- Now, an immensely grown number of users are attracted by new possibilities of SAS and they want rapid answers to more and more complicated Questions
- The users often have to perform numerous cumbersome actions during the experiment and data analysis, to become each of the Answers

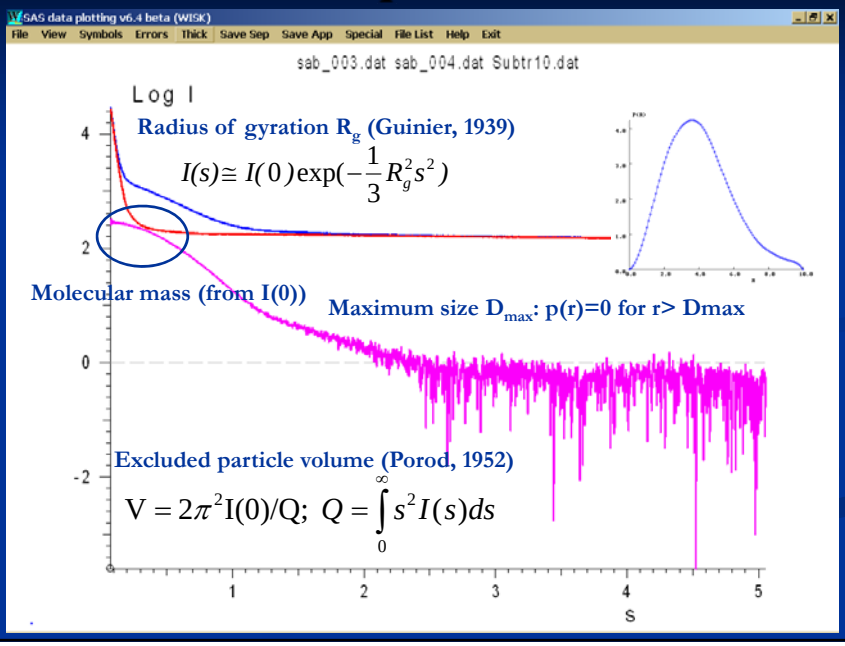
Now we are going through the major steps required on the way

Scattering from dilute macromolecular solutions (monodisperse systems)

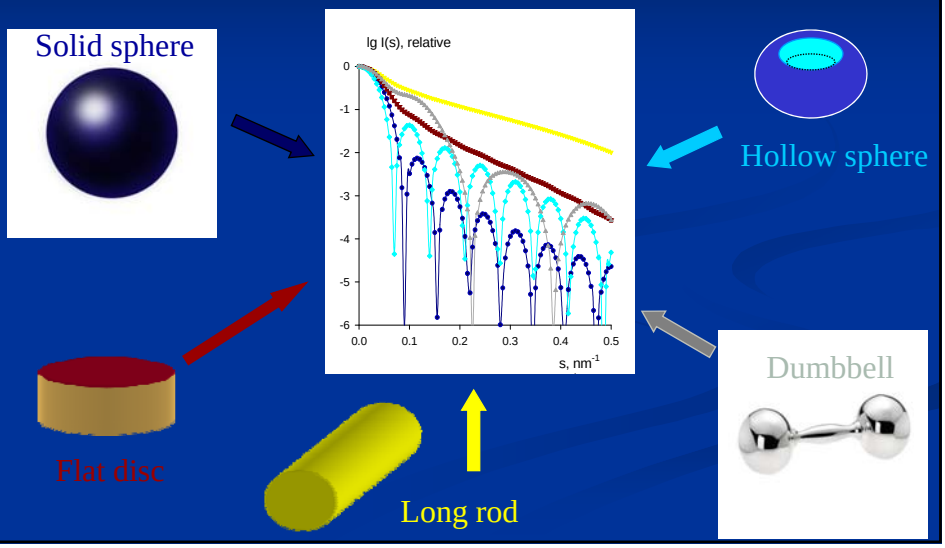
$$I(s) = 4\pi \int_0^D p(r) \frac{\sin sr}{sr} dr$$

The scattering is proportional to that of a single particle averaged over all orientations, which allows one to determine size, shape and internal structure of the particle at low (*1-10 nm*) resolution.

Overall parameters



The scattering is related to the shape (or low resolution structure)

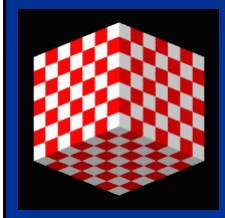


Shape determination: how?

3D search model
M parameters

Trial-and-erro

Non-line
search



Lack of 3D information inevitably leads to ambiguous interpretation, and additional information is always required

Ab initio methods

Advanced methods of SAS data analysis employ spherical harmonics (Stuhrmann, 1970) instead of Fourier transformations

The use of spherical harmonics

SAS intensity is $I(s) = \langle I(\mathbf{s})\rangle_\Omega = \langle\{F[\rho(\mathbf{r})]\}^2\rangle_\Omega$, where F denotes the Fourier transform, $\langle\rangle_\Omega$ stands for the spherical average, and $s=(\mathbf{s}, \Omega)$ is the scattering vector. Expanding $\rho(\mathbf{r})$ in spherical harmonics

$$\rho(\mathbf{r}) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \rho_{lm}(r) Y_{lm}(\omega)$$

the scattering intensity is expressed as

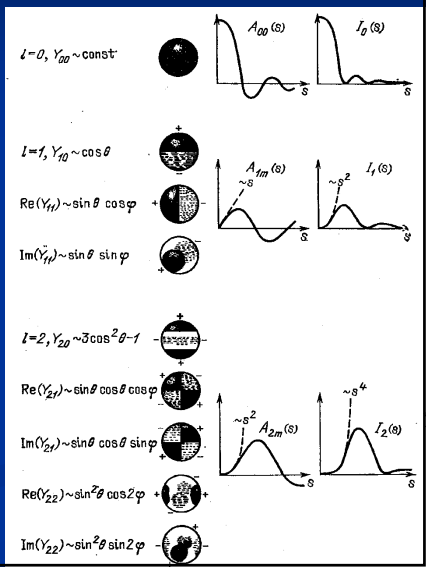
$$I(s) = 2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} |A_{lm}(s)|^2$$

where the partial amplitudes $A_{lm}(s)$ are the Hankel transforms from the radial functions

$$A_{lm}(s) = i^l \sqrt{2/\pi} \int_0^{\infty} \rho_{lm}(r) j_l(sr) r^2 dr$$

and $j_l(sr)$ are the spherical Bessel functions.

Stuhrmann, H.B. Acta Cryst., **A26** (1970) 297.



Structure of bacterial virus T7

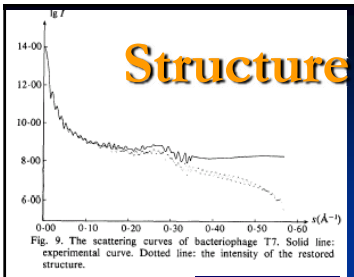


Fig. 9. The scattering curves of bacteriophage T7. Solid line: experimental curve. Dotted line: the intensity of the restored structure.

SAXS, 1982

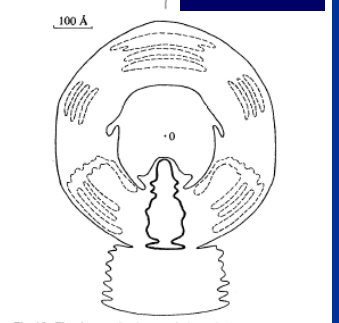
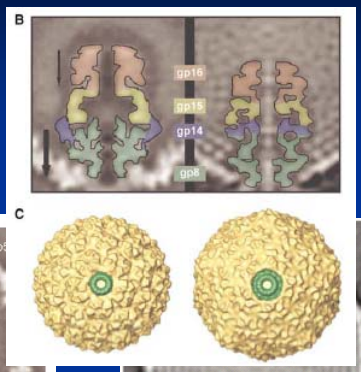
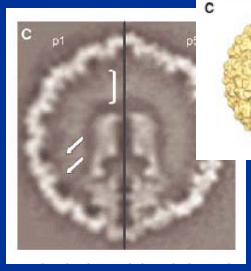


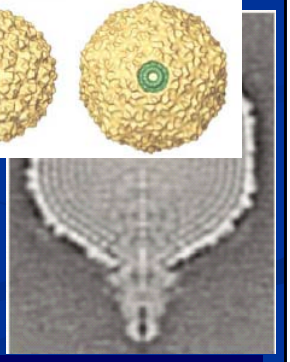
Fig. 10. The electron density map in bacteriophage T7. The cross section containing the axis of rotation (indicated by an arrow) is shown. Solid line: level 0·38 Å⁻³ (hydrated protein); dashed line:

Svergun, D.I., Feigin, L.A. & Schedrin, B.M. (1982) *Acta Cryst.* **A38**, 827

Cryo-EM, 2005







Pro-head

Mature virus

Agirrezabala, J. M. *et al.* & Carrascosa J.L. (2005) *EMBO J.* **24**, 3820

Shape parameterization by spherical harmonics

Homogeneous particle

F(ω) is an envelope function

Scattering density in spherical coordinates (r,ω) = (r,θ,φ) may be described by the envelope function:

$$\rho(\mathbf{r}) = \begin{cases} 1, & 0 \le r \le F(\omega) \\ 0, & r > F(\omega) \end{cases}$$

Shape parameterization by a limited series of spherical harmonics:

$$F(\omega) \cong F_L(\omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm} \cdot Y_{lm}(\omega)$$

$Y_{lm}(\omega)$ – orthogonal spherical harmonics,

f_{lm} – parametrization coefficients,

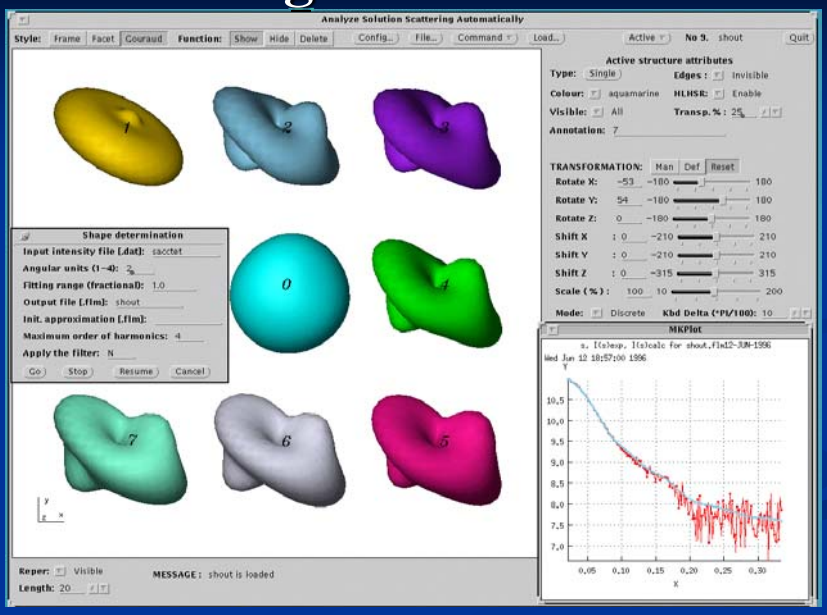
Small-angle scattering intensity from the entire particle is calculated as the sum of scattering from partial harmonics:

$$I_{theor}(s) = \sum_{l=0}^{L} \sum_{m=-l}^{l} 2\pi^2 \left| A_{lm}(s) \right|^2$$

Stuhrmann, H. B. (1970) *Z. Physik. Chem. Neue Folge* 72, 177-198.

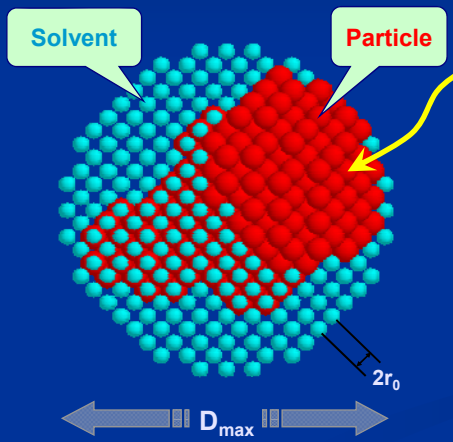
Svergun, D.I. *et al.* (1996) *Acta Crystallogr.* A52, 419-426.

Program SASHA



Bead (dummy atoms) model

A sphere of radius D_{max} is filled by densely packed beads of radius $r_0 << D_{max}$



Vector of model parameters:

$$Position(j) = X(j) = \begin{cases} 1 & \text{if particle} \\ 0 & \text{if solvent} \end{cases}$$

(phase assignments)

Number of model parameters $M \approx (D_{max} / r_0)^3 \approx 10^3$ is too big for conventional minimization methods – Monte-Carlo like approaches are to be used

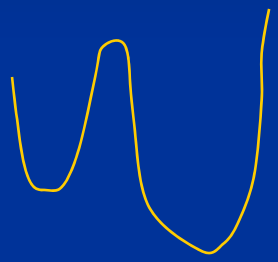
But: This model is able to describe rather complex shapes

Chacón, P. *et al.* (1998) *Biophys. J.* 74, 2760-2775.

Svergun, D.I. (1999) *Biophys. J.* 76, 2879-2886

Finding a global minimum

Pure Monte Carlo runs in a danger to be trapped into a local minimum



Solution: use a global minimization method like simulated annealing or genetic algorithm

Local and global search on the Great Wall

Local search always goes to a better point and can thus be trapped in a local minimum

Pure Monte-Carlo search always goes to the closest local minimum (nature: rapid quenching and vitreous ice formation)

To get out of local minima, global search must be able to (sometimes) go to a worse point

Slower annealing allows to search for a global minimum (nature: normal, e.g. slow freezing of water and ice formation)

Simulated annealing

Aim: find a vector of *M* variables *{x}* minimizing a function *f(x)*

1. Start from a random configuration *x* at a "high" temperature *T*.
2. Make a small step (random modification of the configuration) $x \rightarrow x'$ and compute the difference $\Delta = f(x') - f(x)$.
3. *If* $\Delta < 0$, accept the step; if $\Delta > 0$, accept it with a probability $e^{-\Delta/T}$
4. *Make another step from the old (if the previous step has been rejected) or from the new (if the step has been accepted) configuration.*
5. *Anneal the system at this temperature, i.e.* repeat steps *2-4 "many" (say, 100M* tries or *10M* successful tries, whichever comes first) times, then decrease the temperature *(T' = cT, c<1).*
6. *Continue cooling the system until no improvement in f(x)* is observed.

Shape determination: M≈ 10^3 variables (e.g. 0 or 1 bead assignments in DAMMIN

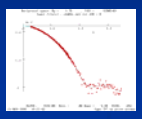
Rigid body methods: M≈ 10^1 variables (positional and rotational parameters of the subunits)

f(x) is always (Discrepancy + Penalty)

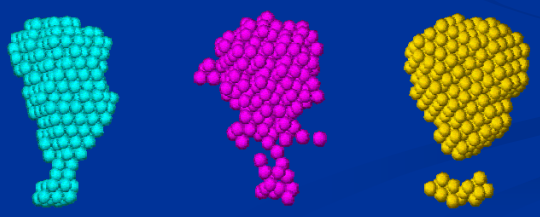
Ab initio program DAMMIN

Using simulated annealing, finds a compact dummy atoms configuration X that fits the scattering data by minimizing

$$f(X) = \chi^2[I_{\exp}(s), I(s, X)] + \alpha P(X)$$



where χ is the discrepancy between the experimental and calculated curves, *P(X)* is the penalty to ensure compactness and connectivity, $\alpha > 0$ its weight.



compact

loose

disconnected

Why/how do *ab initio* methods work

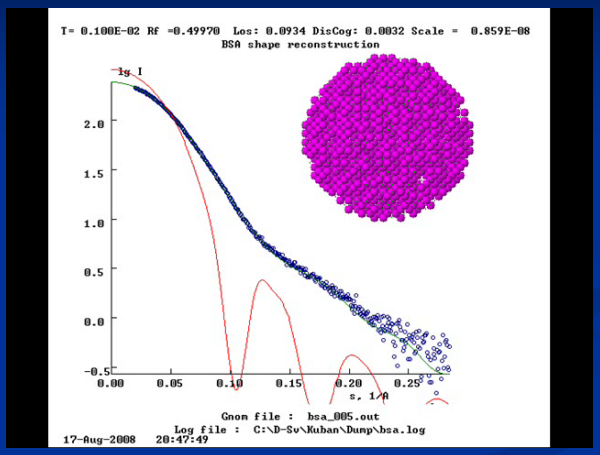
The 3D model is required not only to fit the data but also to fulfill (often stringent) physical and/or biochemical constrains

Why/how do *ab initio* methods work

The 3D model is required not only to fit the data but also to fulfill (often stringent) physical and/or biochemical constrains

A test *ab initio* shape determination run

Program DAMMIN

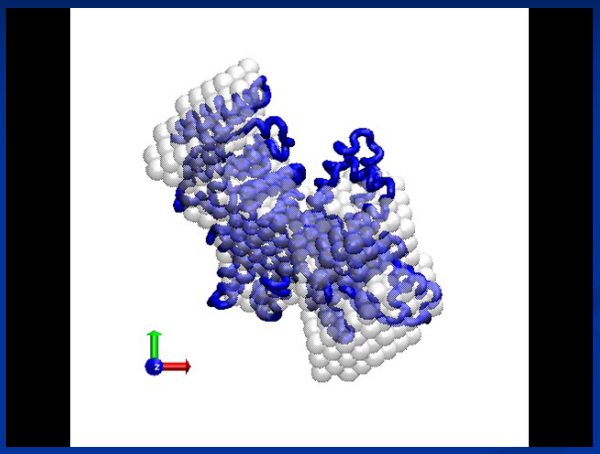


Slow mode

Bovine serum albumin,
molecular mass 66 kDa, no symmetry imposed

A test *ab initio* shape determination run

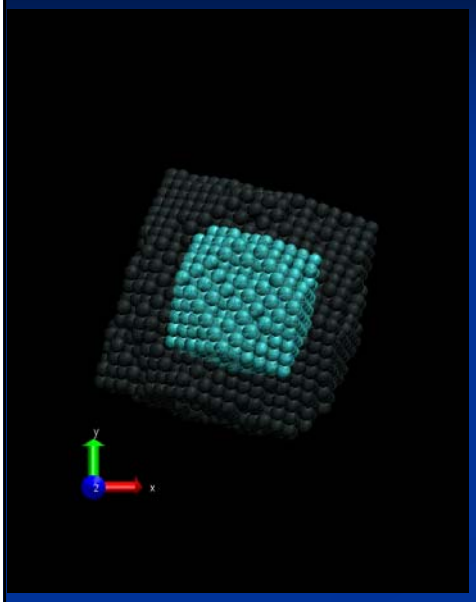
Program DAMMIN



Slow mode

Bovine serum albumin: comparison of the *ab initio* model with the crystal structure of human serum albumin

DAMMIF, a fast DAMMIN



DAMMIF is a completely reimplemented DAMMIN written in object-oriented code

- **About 25-40 times faster than DAMMIN (in fast mode, takes about 1-2 min on a PC)**
- **Employs adaptive search volume**
- **Makes use of multiple CPUs**

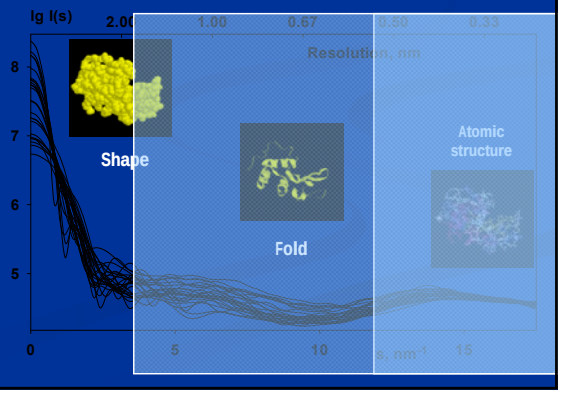
Franke, D. & Svergun, D. I. (2009)
J. Appl. Cryst. **42**, 342–346

Limitations of shape determination

- Very low resolution
- Ambiguity of the models

Accounts for a restricted portion of the data

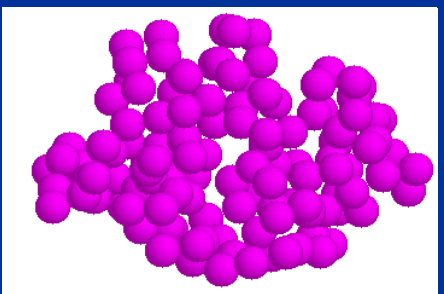
How to construct *ab initio* models accounting for higher resolution data?



Ab initio dummy residues model

- Proteins typically consist of folded polypeptide chains composed of amino acid residues

At a resolution of 0.5 nm a protein can be represented by an ensemble of K dummy residues centered at the Cα positions with coordinates $\{\boldsymbol{r}_i\}$

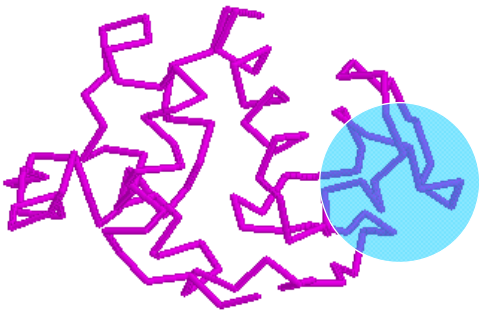


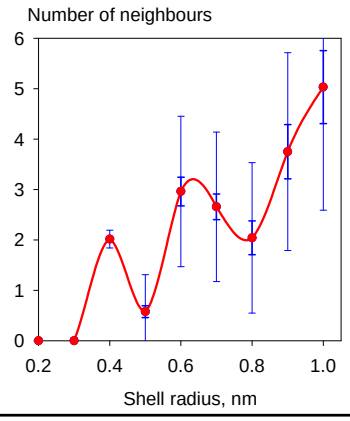
Scattering from such a model is computed using the Debye (1915) formula.

Starting from a random model, simulated annealing is employed similar to DAMMIN

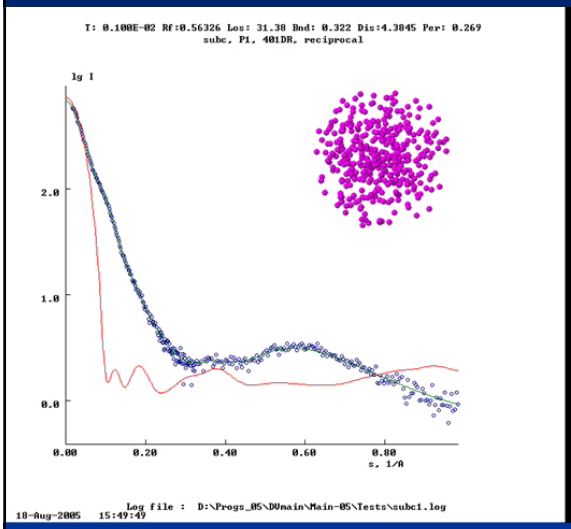
Distribution of neighbors

Excluded volume effects and local interactions lead to a characteristic distribution of nearest neighbors around a given residue in a polypeptide chain



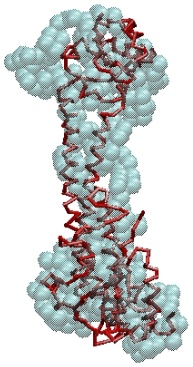


GASBOR run on C subunit of V-ATPase



Starting from a random "gas" of 401 dummy residues, fits the data by a locally chain-compatible model

GASBOR run on C subunit of V-ATPase

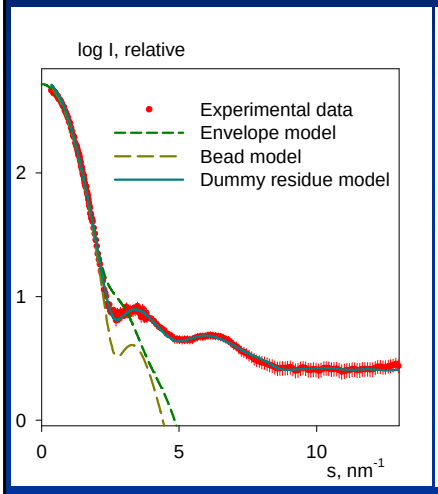


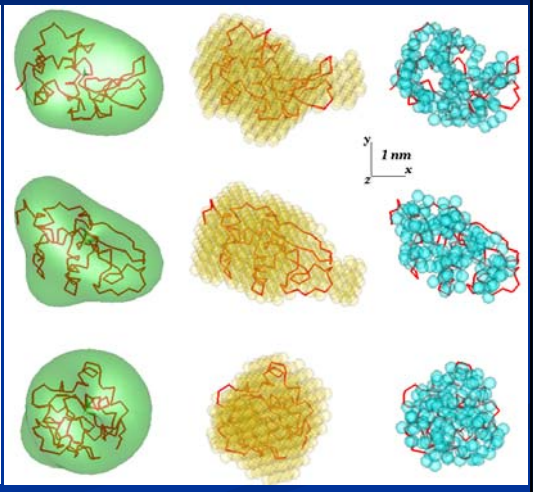
Beads: Ambruster *et al.* (2004, June) *FEBS Lett.* **570,** 119

C_α trace: Drory *et al.* (2004, November), *EMBO reports,* **5**, 1148

Benchmarking *ab initio* methods

Envelope **Bead model** **Dummy residues**

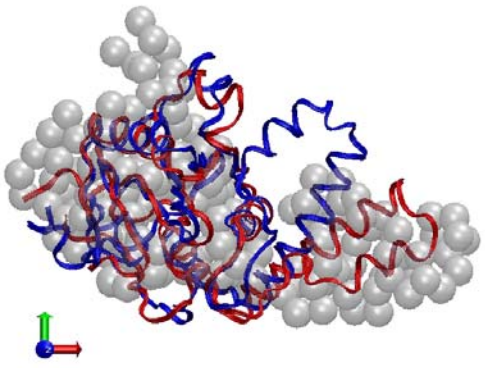


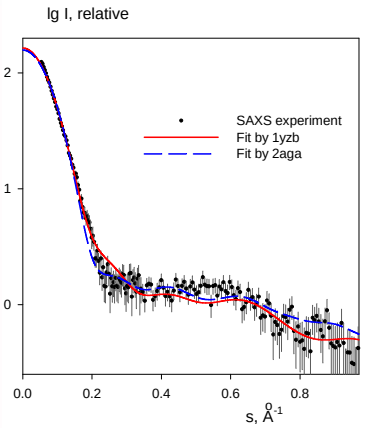


Comparison with the crystal structure of lysozyme

SASHA 1996 **DAMMIN 1999** **GASBOR 2001**

Validating NMR models by SAXS

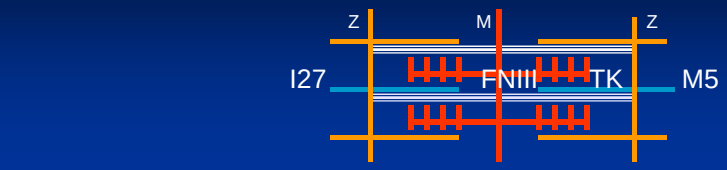




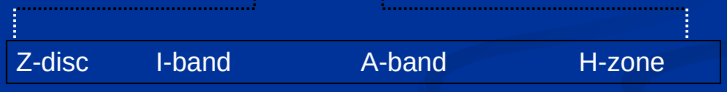
NMR models of the Josephin domain of ataxin-3: red curve and chain: **1yzb**, Nicastro et al. (2005) *PNAS USA* **102**, 10493; blue curve and chain: **2aga**, Mao et al. (2005) *PNAS USA* **102**, 12700.

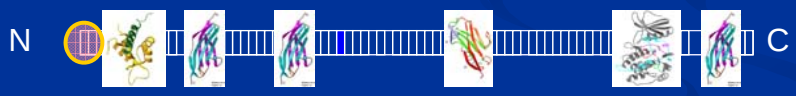
Nicastro, G., Habeck, M., Masino, L., Svergun, D.I. & Pastore, A. (2006) *J. Biomol. NMR*, **36**, 267.

Modular structure of a giant mucsle protein titin



26 926 aa
1.2 µm





	Z1Z2	*Z7*	*I1*	*I27*	*Ax*	*TK*	*M5*
fold	IG	EF	IG	IG	FN-III	kinase	IG
method	X	NMR	X	NMR	NMR	X	NMR

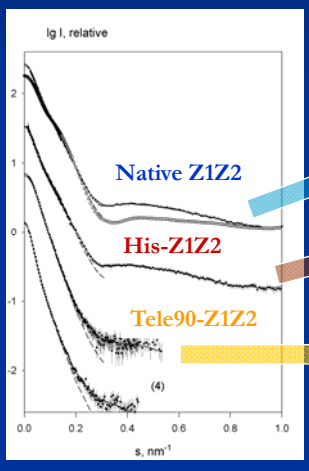
NMR data: Pastore lab; X-ray data: Wilmanns lab

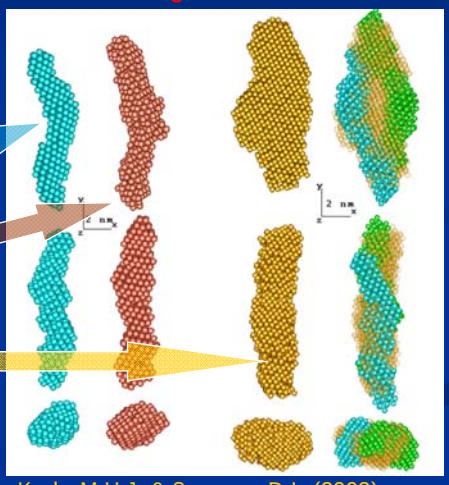
Solution structure of Z1Z2-telethonin complex

- Z1Z2 includes two modules at the N-terminal of the Z-disc of titin and interacts with telethonin

Shape of Z1Z2 and localization of the his-tag

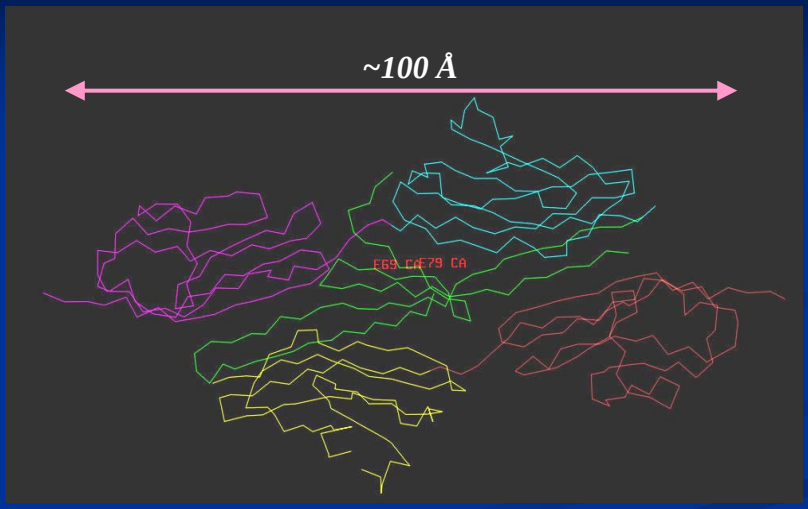
Cross-linking function of telethonin





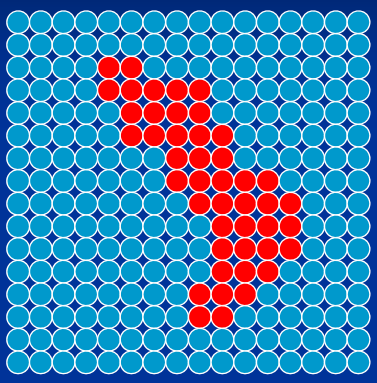
Zou, P., Gautel, M., Geerlof, M., Wilmanns, M., Koch, M.H.J. & Svergun, D.I. (2003) *J. Biol. Chem.* **278**, 2636

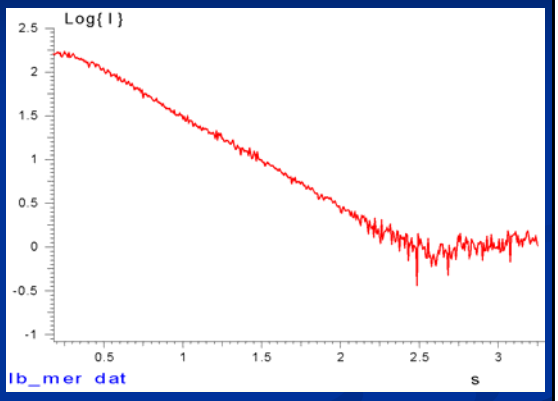
Crystal structure of Z1Z2-telethonin complex



Zou P., Pinotsis N., Lange S., Song Y.H., Popov A., Mavridis I., Mayans O.M., Gautel M. & Wilmanns M. (2006) *Nature* **439**, 229-33.

Shape analysis for multi-component systems: principle

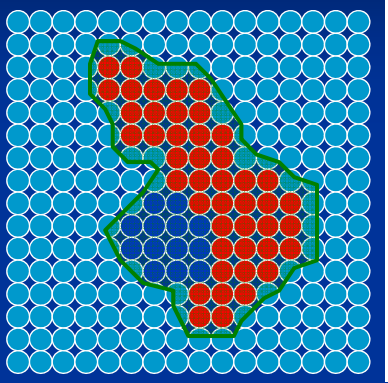


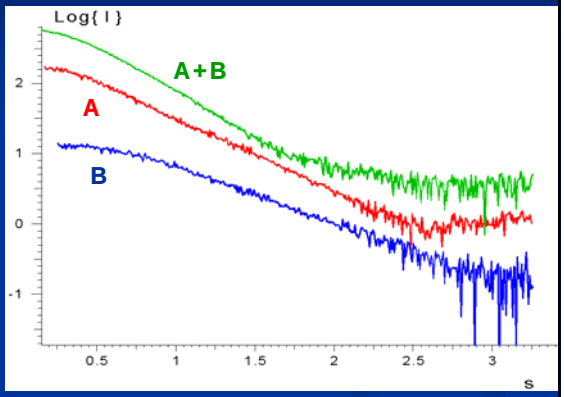


One component, one scattering pattern: "normal" shape determination

Chacón, P. et al. (1998) Biophys. J. 74, 2760-2775
Svergun, D.I. (1999) Biophys. J. **76**, 2879-2886

Shape analysis for multi-component systems: principle

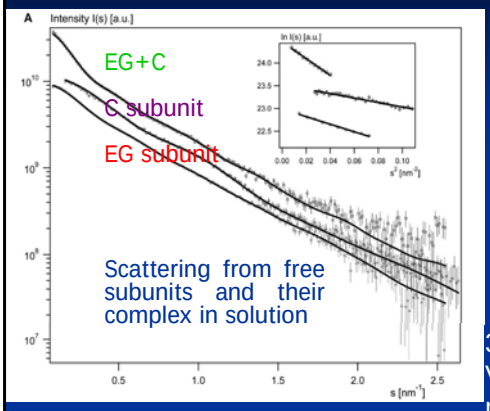


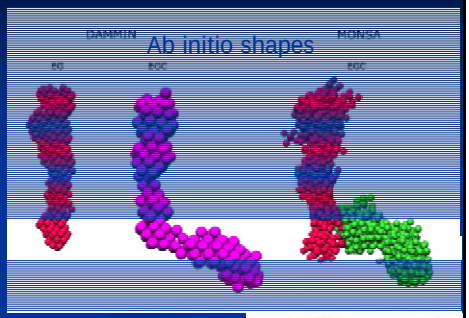


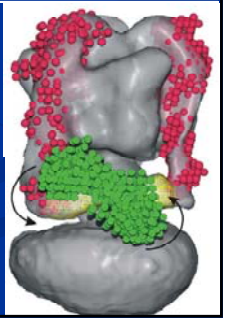
Many components, many scattering patterns: shape and internal structure

Svergun, D.I. (1999) Biophys. J. **76**, 2879-2886
Svergun, D.I. & Nierhaus, K.H. (2000) J. Biol. Chem. 275, 14432-14439

EGC stator sub-complex of V-ATPase





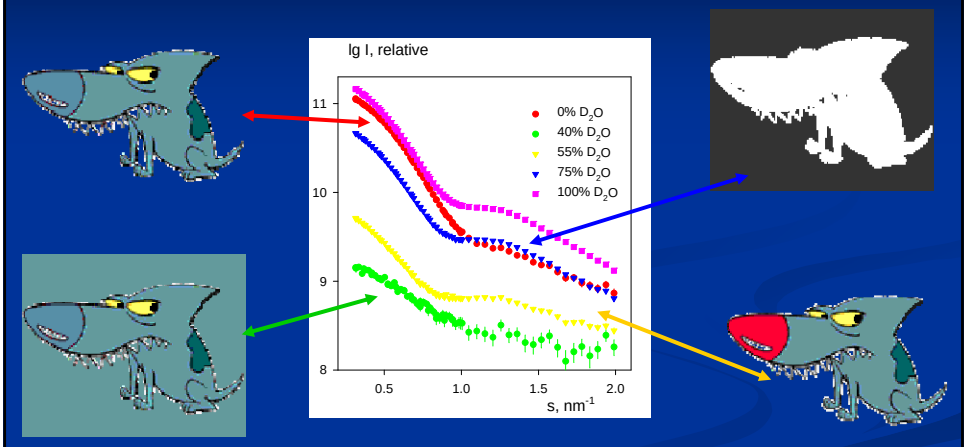


3D map of the yeast V-ATPase by electron microscopy.

In solution, EG makes an L-shaped assembly with subunit-C. This model is supported by the EM showing three copies of EG, two of them linked by C. The data further indicate a conformational change of EGC during regulatory assembly/disassembly.

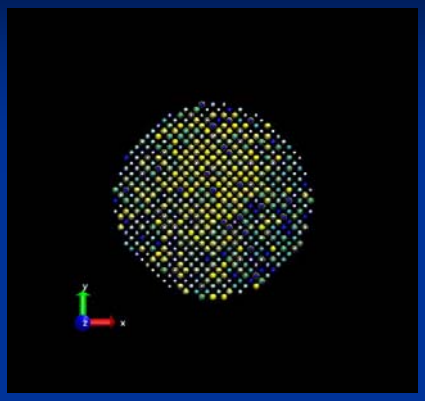
Diepholz, M. *et al.* (2008) Structure **16**, 1789-1798

Scattering from a multiphase particle

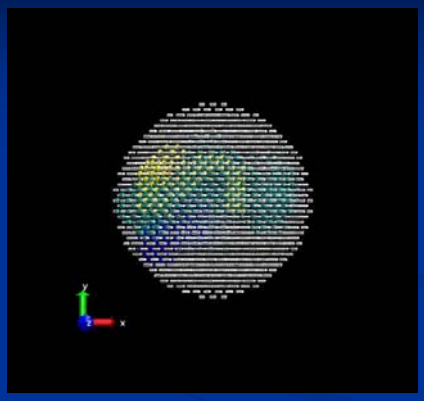


$$I_m(s) = \sum_j (\Delta\rho_j^m)^2 I_j(s) + 2\sum_{j>k} \Delta\rho_j^m \Delta\rho_i^m I_{jk}(s)$$

Ab initio multiphase modelling



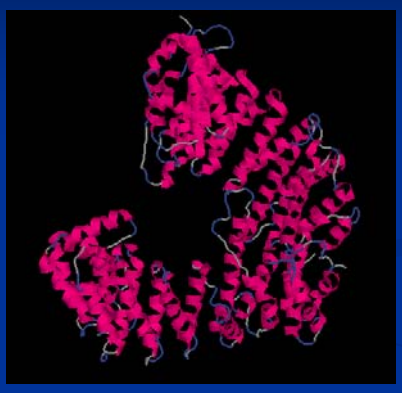
Start: random phase assignments within the search volume, no fit to the experimental data

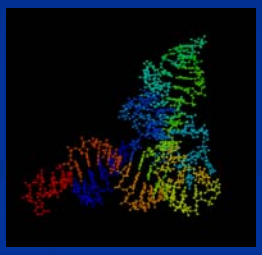


Finish: condensed multiphase model with minimum interfacial area fitting multiple data sets

Program MONSA, Svergun, D.I. (1999) *Biophys. J.* **76**, 2879;
Petoukhov, M.V. & Svergun, D. I. (2006) *Eur. Biophys. J.* **35**, 567.

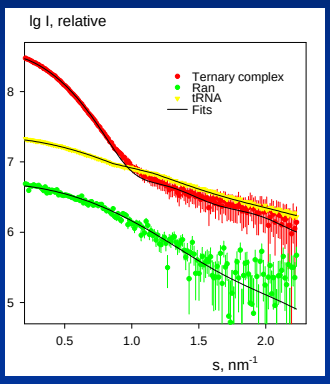
Ternary complex: Exportin-t/Ran/tRNA

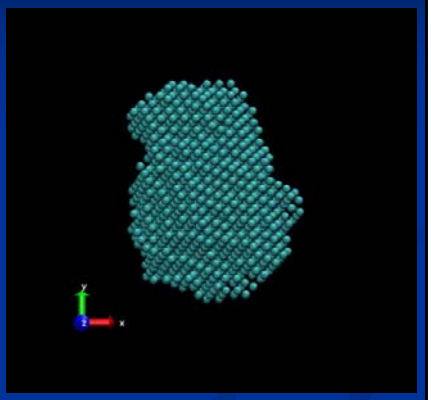




Ran (structure known) Exportin-t (tentative homology model) t-RNA (structure known)

X-rays: *ab initio* overall shape





One X-ray scattering pattern from the ternary complex fitted by DAMMIN

Fukuhara, N., Fernandez, E., Ebert, J., Conti, E. & Svergun, D. I. (2004) *J. Biol. Chem.* **279**, 2176

Scattering data from Exportin-t/Ran/tRNA

X-ray scattering

- From Exportin-t, Ran, tRNA 3 curves

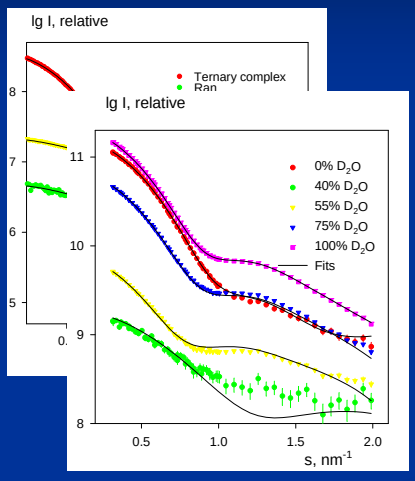
Neutron scattering

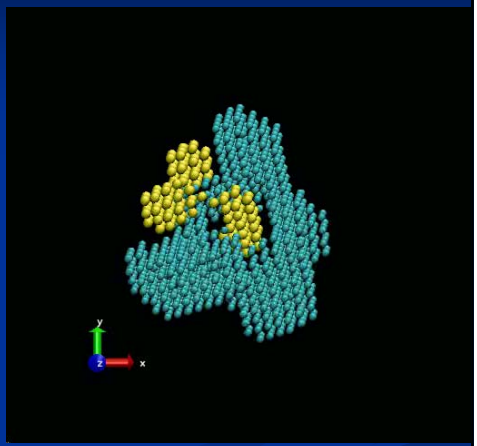
- Ternary complex with protonated Ran
  in 0, 40, 55, 75, 100% D_2O 5 curves

- Ternary complex with deuterated Ran
  in 0, 40, 55, 70, 100% D_2O 5 curves

TOTAL ***13 curves***

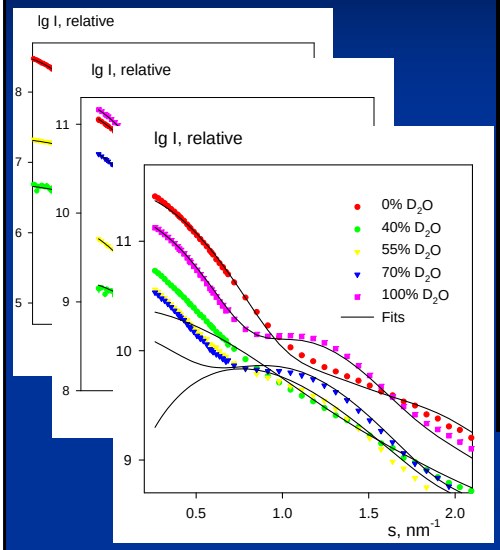
Contrast variation: localization of tRNA

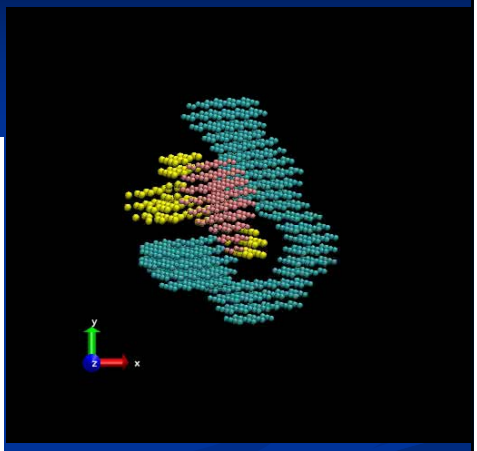




Three X-ray and five neutron data sets fitted by MONSA

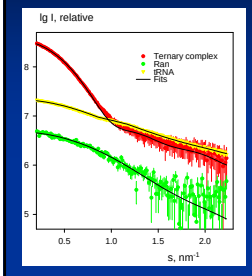
Specific deuteration: highlighting d-Ran

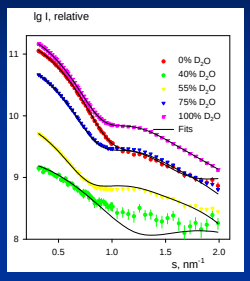


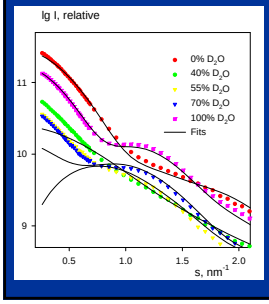


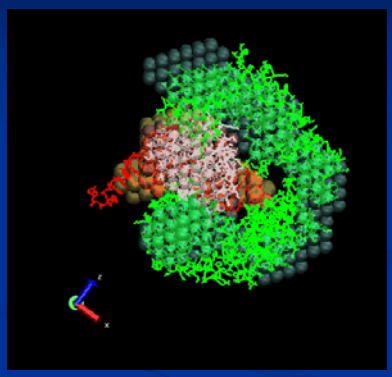
Three X-ray and ten neutron data sets fitted by MONSA

Ternary complex: Exportin-t/Ran/tRNA







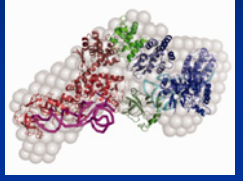


High resolution models of the components docked into the three-phase ab initio model of the complex based on X-ray and neutron scattering from selectively deuterated particles

Shapes from recent projects at EMBL-HH

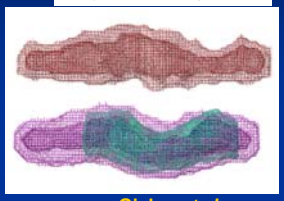
Complexes and assemblies

Aminoacyl-tRNA synthetases complex



Koehler *et al* NAR (2013)

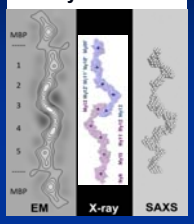
α-synuclein oligomers



Giehm *et al* PNAS USA (2011)

Domain and quaternary structure

Myomesin



Pinotsis *et al* Plos Biol (2012)

Toxin B

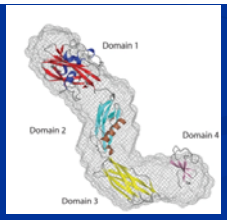
Albesa-Jové *et al* JMB (2010)

Flexible/transient systems

Complement factor H

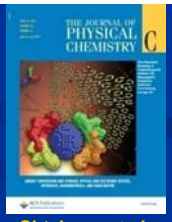
Morgan *et al* NSMB (2011)

Colonization factor GbpA



Wong *et al* , Plos Pathog (2012)

Nanocomposites



Shtykova *et al* JPC (2012)

Ab initio programs for SAS

- Genetic algorithm DALAI_GA (Chacon et al., 1998, 2000)
- 'Give-n-take' procedure SAXS3D (Bada et al., 2000)
- Spheres modeling program GA_STRUCT (Heller et al., 2002)
- Envelope models: SASHA(1) (Svergun et al., 1996)
- Dummy atoms: DAMMIN(1,4) & MONSA(1,2) (Svergun, 1999)
- Dummy residues: GASBOR(1,3) (Petoukhov et al., 2001)

(1) Able to impose symmetry and anisometry constrains

(2) Multiphase inhomogeneous models

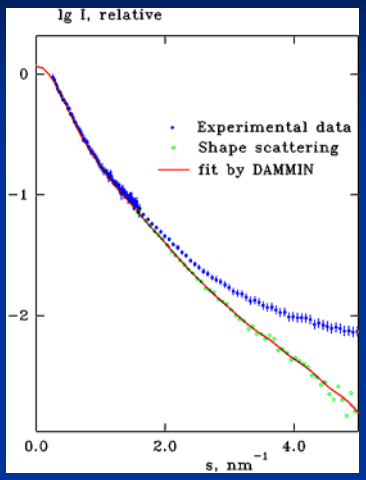
(3) Accounts for higher resolution data

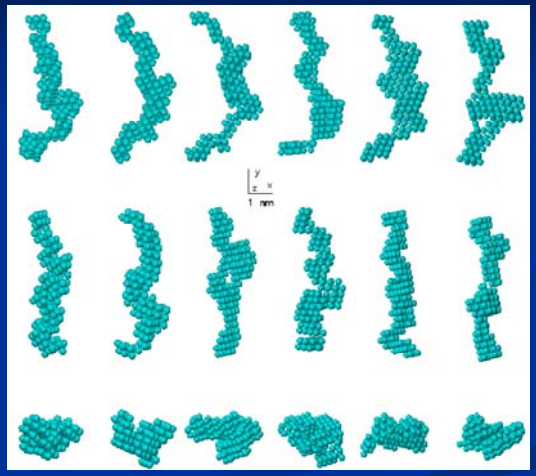
(4) DAMMIF is 30 times faster (D.Franke & D.Svergun, 2009)

Some words of caution

Or Always remember about ambiguity!

Shape determination of 5S RNA: a variety of DAMMIN models yielding identical fits

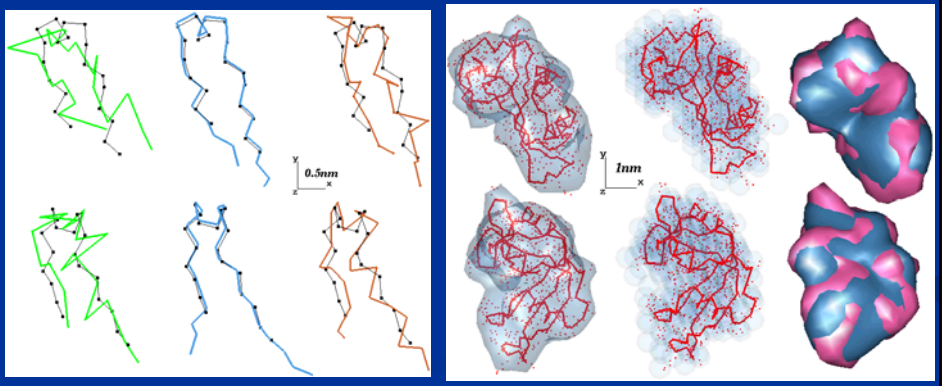




Funari, S., Rapp, G., Perbandt, M., Dierks, K., Vallazza, M., Betzel, Ch., Erdmann, V. A. & Svergun, D. I. (2000) *J. Biol. Chem.* **275**, 31283-31288.

Program SUPCOMB – a tool to align and conquer

- Aligns heterogeneous high- and low-resolution models and provides a dissimilarity measure (NSD)
- For shape determination, allows one to find common features in a series of independent reconstructions



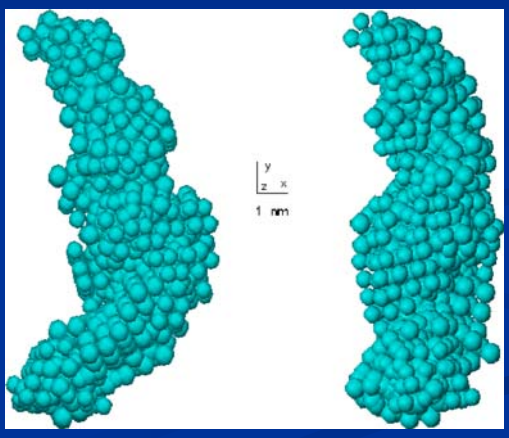
Kozin, M.B. & Svergun, D.I. (2001) *J. Appl. Crystallogr.* **34,** 33-41

Automated analysis of multiple models

1. Find a set of solutions starting from random initial models and superimpose all pairs of models with SUPCOMB.
2. Find the most probable model (which is on average least different from all the others) and align all the other models with this reference one.
3. Remap all models onto a common grid to obtain the solution spread region and compute the spatial occupancy density of the grid points.
4. Reduce the spread region by rejecting knots with lowest occupancy to find the most populated volume
5. These steps are automatically done by a package called DAMAVER if you just put all multiple solutions in one directory

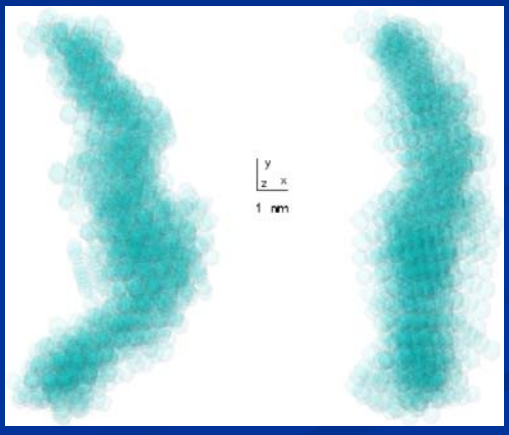
Program DAMAVER, Volkov & Svergun (2003) *J. Appl. Crystallogr*. 36, 860

5S RNA: ten shapes superimposed



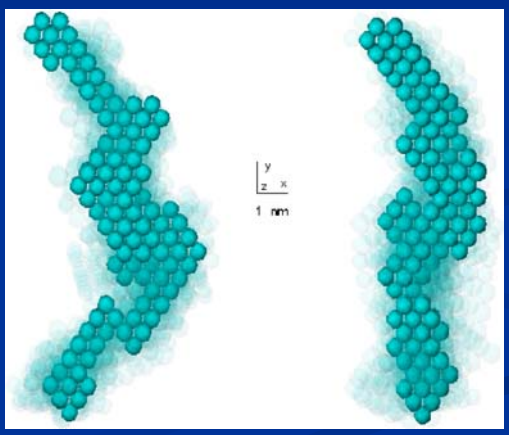
Solution spread region

5S RNA: ten shapes superimposed



Most populated volume

5S RNA: final solution



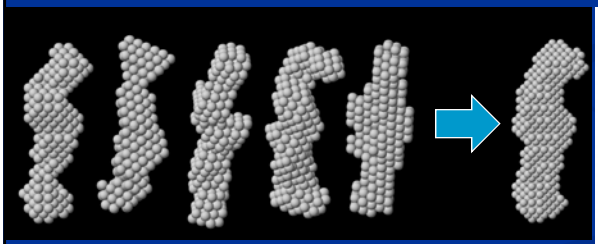
The final model obtained within the solution spread region

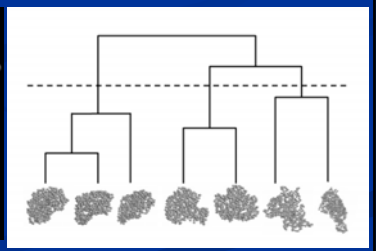
Damaver and Damclust

Damaver superposes multiple *ab initio* models, computes deviations between them using a normalized spatial discrepancy (NSD), finds the most probable and an averaged model.

Outliers, if any, are discarded

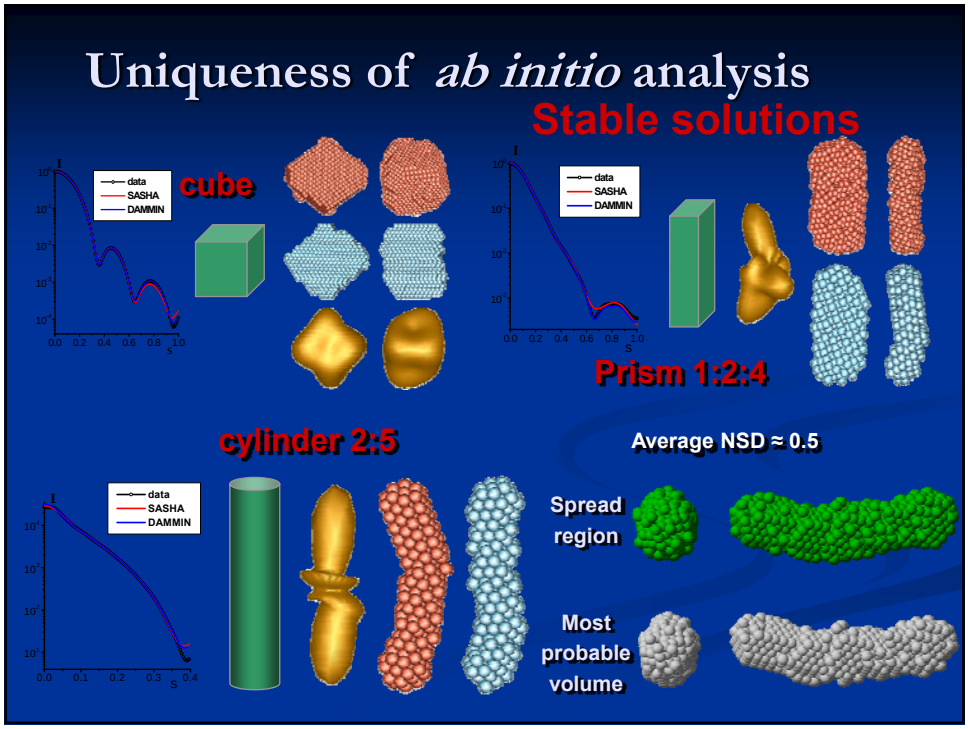
Damclust superposes multiple *ab initio* or rigid body models, computes deviations between them using NSD or RMSD, and attributes the models to distinct clusters

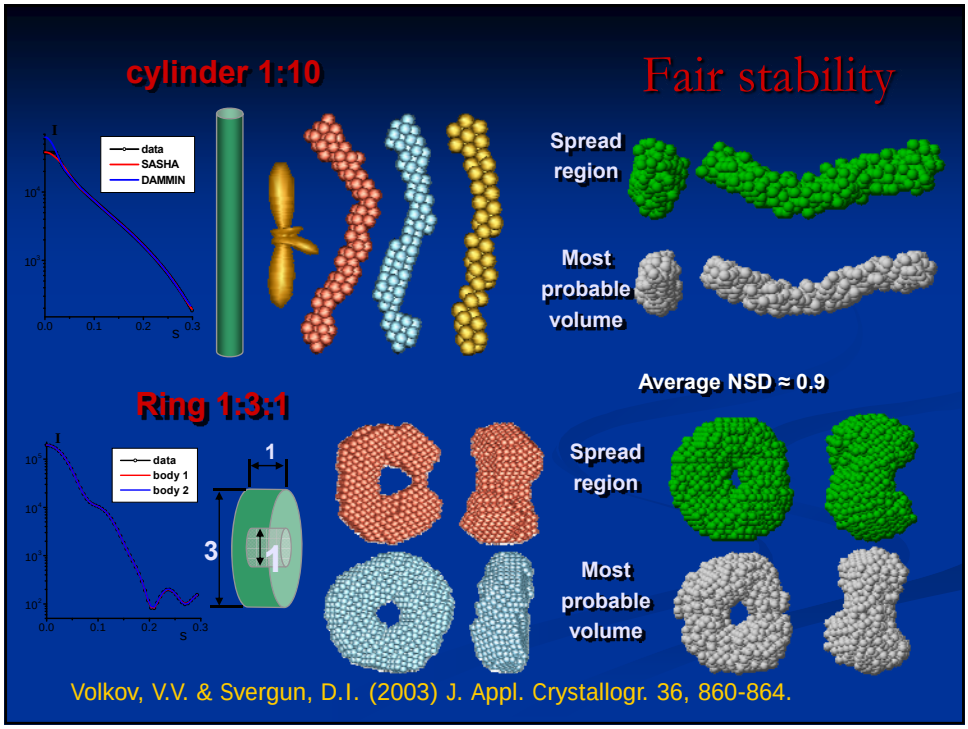


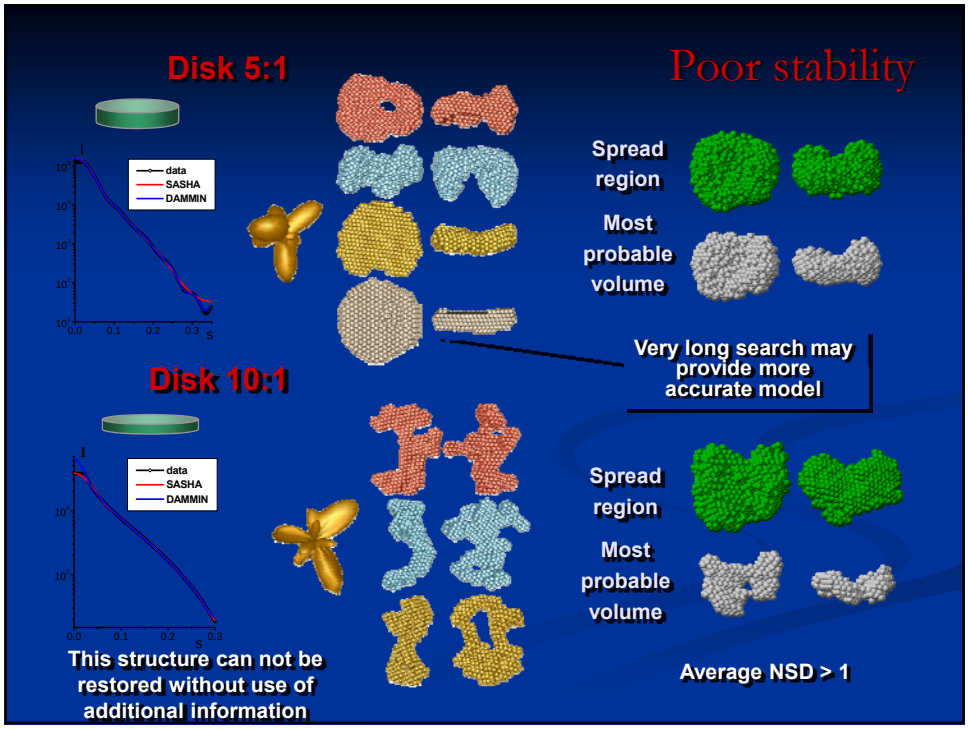


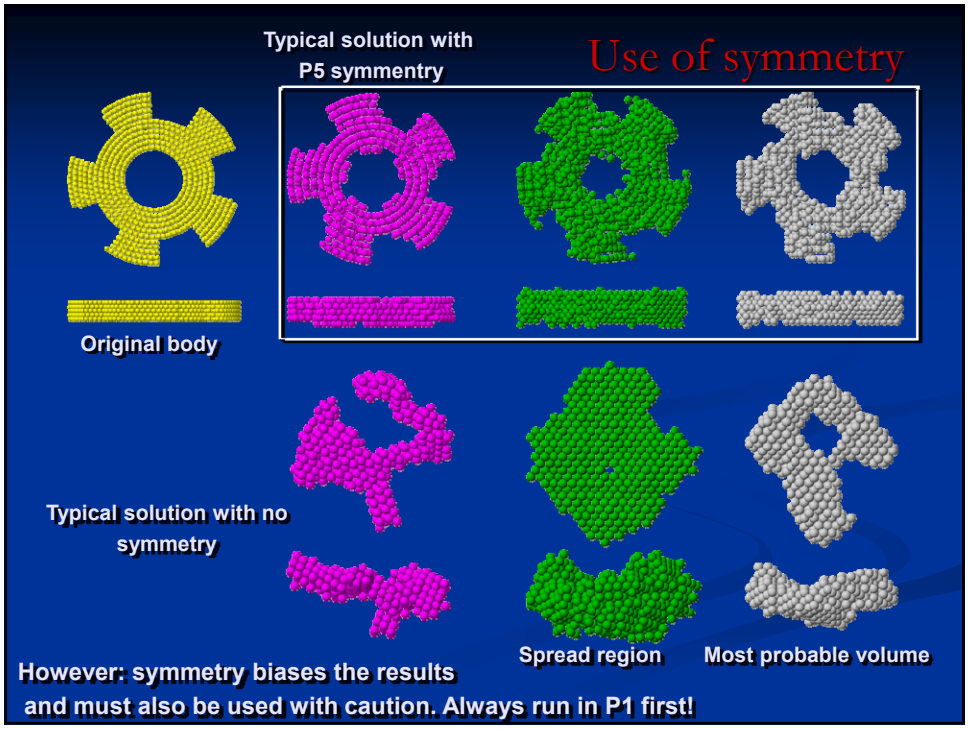
Result: Most probable model
Single averaged model

Representatives of clusters
Distances between clusters

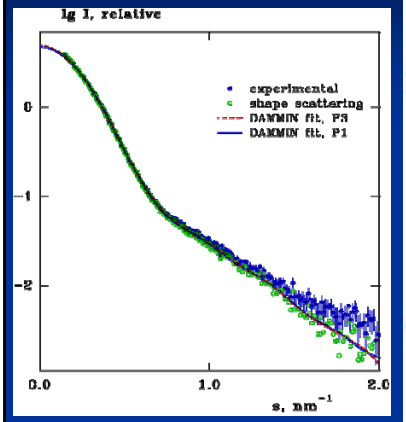

Uniqueness of *ab initio* analysis
Stable solutions
cube
data
SASHA
DAMMIN
Prism 1:2:4
cylinder 2:5
Average NSD ≈ 0.5
Spread region
Most probable volume

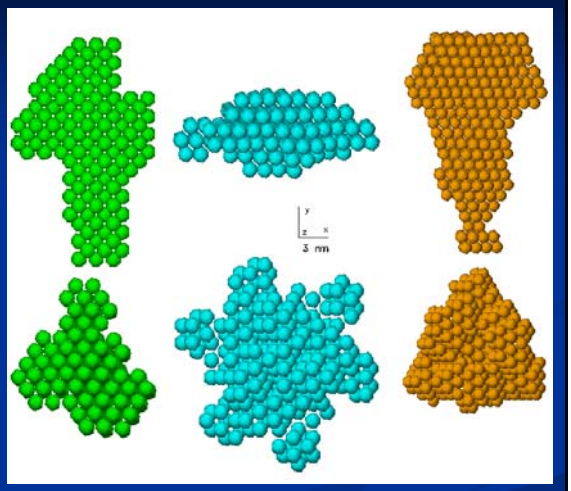

cylinder 1:10
Fair stability
data
SASHA
DAMMIN
Spread region
Most probable volume
Average NSD ≈ 0.9
Ring 1:3:1
data
body 1
body 2
1
3
1
Volkov, V.V. & Svergun, D.I. (2003) J. Appl. Crystallogr. 36, 860-864.


Disk 5:1
Poor stability
data
SASHA
DAMMIN
Spread region
Most probable volume
Very long search may provide more accurate model
Disk 10:1
data
SASHA
DAMMIN
Spread region
Most probable volume
This structure can not be restored without use of additional information
Average NSD > 1


Typical solution with P5 symmentry
Use of symmetry
Original body
Typical solution with no symmetry
Spread region
Most probable volume
However: symmetry biases the results and must also be used with caution. Always run in P1 first!

Shape determination of V_1 ATPase



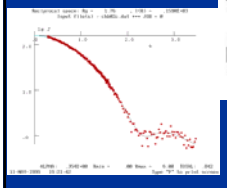


P1 P3 P3, prolate

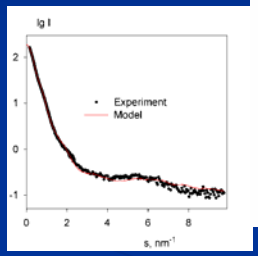
Svergun, D.I., Konrad, S., Huss, M., Koch, M.H.J., Wieczorek, H., Altendorf, K.-H., Volkov, V.V. & Grueber, G. (1998) Biochemistry 37, 17659-17663.

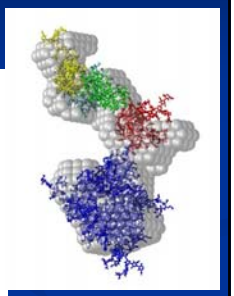
Progress in *ab initio* methods

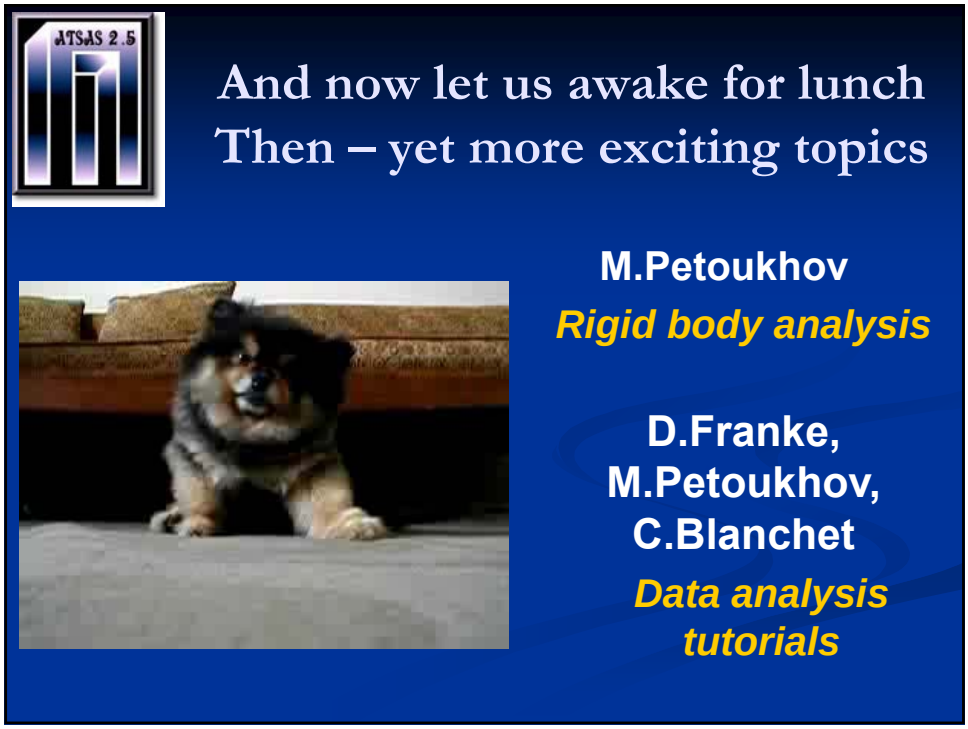
1993



2014






ATSAS 2.5
And now let us awake for lunch
Then – yet more exciting topics
M.Petoukhov
Rigid body analysis
D.Franke,
M.Petoukhov,
C.Blanchet
Data analysis
tutorials