Hybrid methods and analysis of mixtures

D.Svergun, EMBL-Hamburg

EMBL

Shapes from recent projects at EMBL-HH

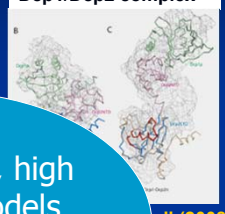
Complexes and assemblies **Domain and quaternary structure**

S-layer proteins

Fagan *et al* Mol. Microbiol (2009)

α-synuclein oligomers

Dcp1/Dcp2 complex



ell (2008)

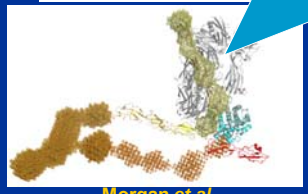
Toxin B

Albesa-Jové *et al* JMB (2010)

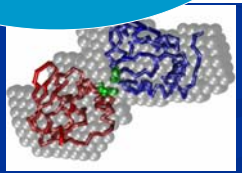
In most cases, high resolution models are drawn inside the shapes

Flexible/trans **ural transitions**

Complement factor H

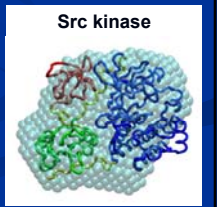


Morgan *et al* NSMB (2011)



Xu et al JACS (2008)

Src kinase



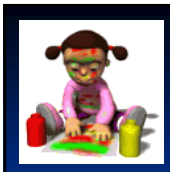
Bernado et al JMB (2008)

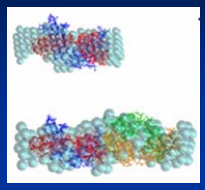
Modern life sciences widely employ hybrid methods

The most known and popular tool is, of course, Photoshop

SAXS also allows for a very effective hybrid model building where high resolution portions are positioned to fit the low resolution scattering data

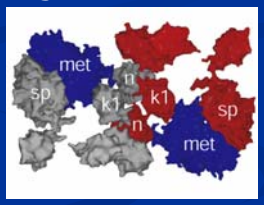
Monodisperse systems



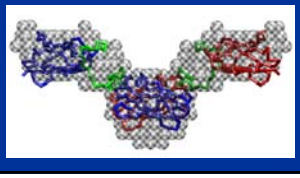


Shape and conformational changes of macromolecules and complexes

Validation of high resolution models and oligomeric organization

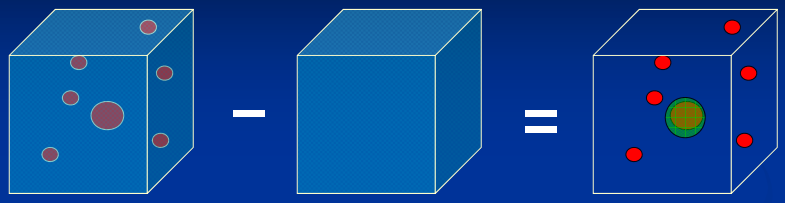


Rigid body models of complexes using high resolution structures



Addition of missing fragments to high resolution models

How to compute SAS from atomic model

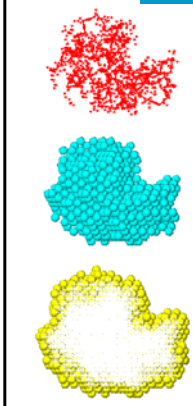


$I_{solution}(s)$ $I_{solvent}(s)$ $I_{particle}(s)$

- To obtain scattering from the particles, solvent scattering must be subtracted to yield effective density distribution $\Delta\rho = \langle\rho(\mathbf{r}) - \rho_s\rangle$, where ρ_s is the scattering density of the solvent
- Further, the bound solvent density may differ from that of the bulk

Scattering from a macromolecule in solution

$$I(s) = \left\langle |A(\mathbf{s})|^2 \right\rangle_\Omega = \left\langle |A_a(\mathbf{s}) - \rho_s A_s(\mathbf{s}) + \delta\rho_b A_b(\mathbf{s})|^2 \right\rangle_\Omega$$



- $A_a(\mathbf{s})$: atomic scattering in vacuum
- $A_s(\mathbf{s})$: scattering from the excluded volume
- $A_b(\mathbf{s})$: scattering from the hydration shell

***CRYSOL* (X-rays):** Svergun et al. (1995). *J. Appl. Cryst.* **28**, 768
***CRYSON* (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, **95**, 2267

The use of multipole expansion

$$I(s) = \left\langle \left| A(\mathbf{s}) \right|^2 \right\rangle_\Omega = \left\langle \left| A_a(\mathbf{s}) - \rho_s E(\mathbf{s}) + \delta\rho_b B(\mathbf{s}) \right|^2 \right\rangle_\Omega$$

If the intensity of each contribution is represented using spherical harmonics

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

the average is performed analytically:

$$I(s) = 2\pi^2 \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| A_{lm}(s) - \rho_0 E_{lm}(s) + \delta\rho B_{lm}(s) \right|^2$$

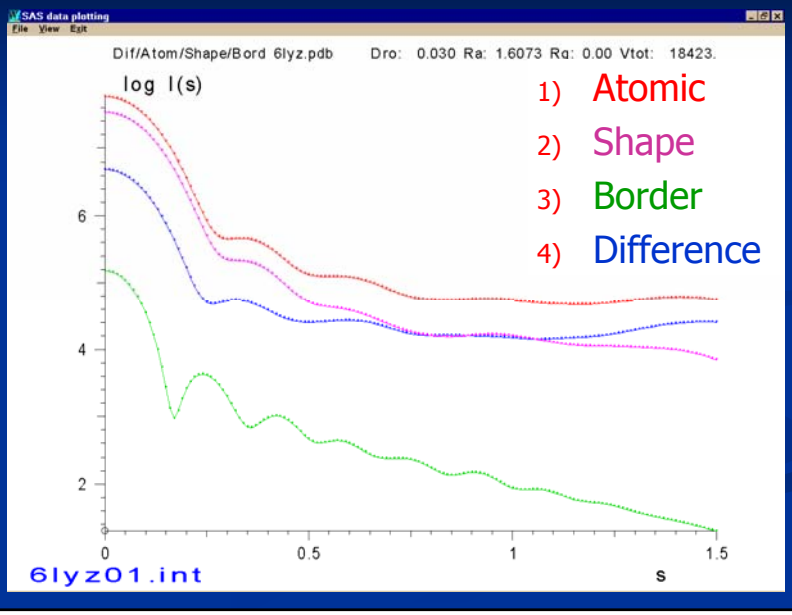
This approach permits to further use rapid algorithms for rigid body refinement

CRYSOL and CRYSON: X-ray and neutron scattering from macromolecules

$$I(s) = 2\pi^2 \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| A_{lm}(s) - \rho_0 E_{lm}(s) + \delta\rho B_{lm}(s) \right|^2$$

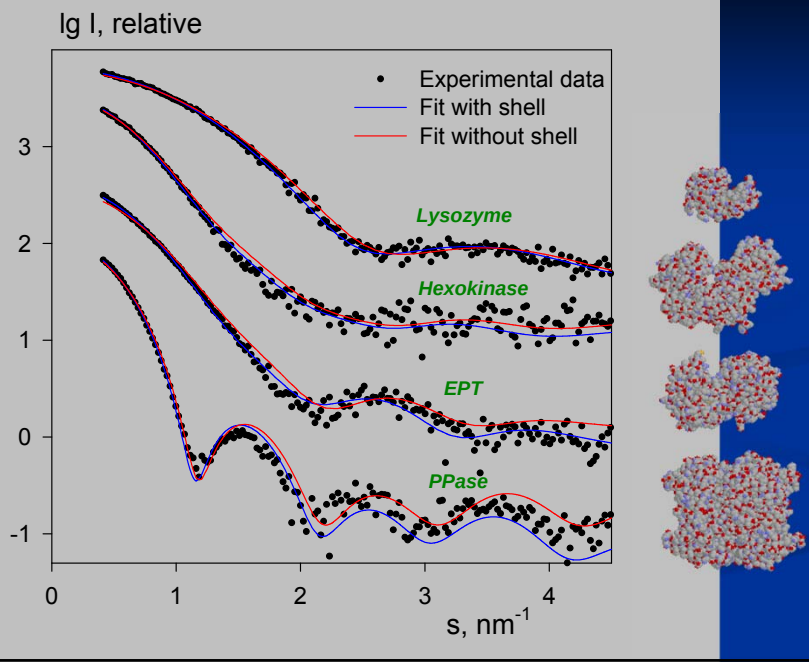
- **The programs:**
  - either fit the experimental data by varying the density of the hydration layer $\delta\rho$ (affects the third term) and the total excluded volume (affects the second term)
  - or predict the scattering from the atomic structure using default parameters (theoretical excluded volume and bound solvent density of 1.1 g/cm^3)
  - provide output files (scattering amplitudes) for rigid body refinement routines
  - compute particle envelope function $F(\omega)$

Scattering components (lysozyme)



1) Atomic
2) Shape
3) Border
4) Difference

Effect of the hydration shell, X-rays



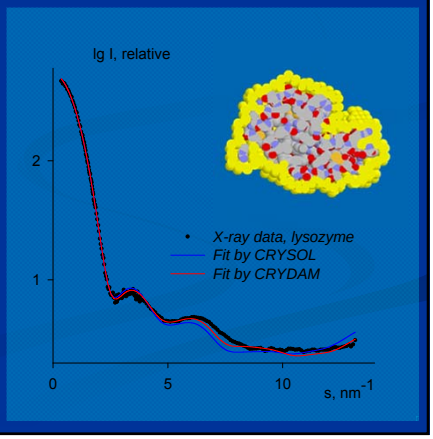
Other approaches/programs I

- **The 'cube method'** (Luzzati et al, 1972; Fedorov and Pavlov, 1983; Müller, 1983) ensures uniform filling of the excluded volume. Could/should/must be superior over the effective atomic factors method at higher angles.

- **CRYDAM** (unpublished)
  - Represents hydration shell by dummy water atoms
  - Handles proteins, carbohydrates, nucleic acids and their complexes
  - Is applicable for wide angle scattering range

  Malfois, M. & Svergun, D.I. (2001), to be submitted

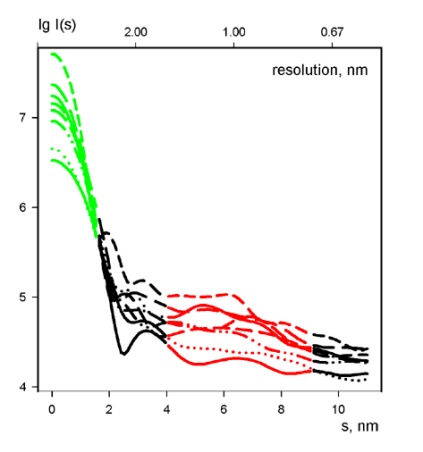
- **CRYSOL 3.0** (is coming)



Other approaches/programs II

- J. Bardhan, S. Park and L. Makowski (2009) SoftWAXS: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules *J. Appl. Cryst.* **42**, 932-943 - *A program to compute WAXS*
- Schneidman-Duhovny D, Hammel M, Sali A. (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. **38** Suppl:W540-4. - *Debye-like computations, Web server*
- Grishaev A, Guo L, Irving T, Bax A. (2010) Improved Fitting of Solution X-ray Scattering Data to Macromolecular Structures and Structural Ensembles by Explicit Water Modeling. *J Am Chem Soc.* **132,** 15484-6. - *Generate bulk and bound waters with MD, do fit the data to the model*
- Poitevin F, Orland H, Doniach S, Koehl P, Delarue M (2011). AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. Nucleic Acids. Res. 39, W184-W189 - *Generate waters around proteins using MD (AquaSol program)*
- Virtanen JJ, Makowski L, Sosnick TR, Freed KF. (2011) Modeling the hydration layer around proteins: applications to small- and wide-angle x-ray scattering. *Biophys J.* **101**, 2061-9. - *Use a "HyPred solvation" model to generate the shell, geared towards WAXS.*

DARA, a database for rapid characterization of proteins

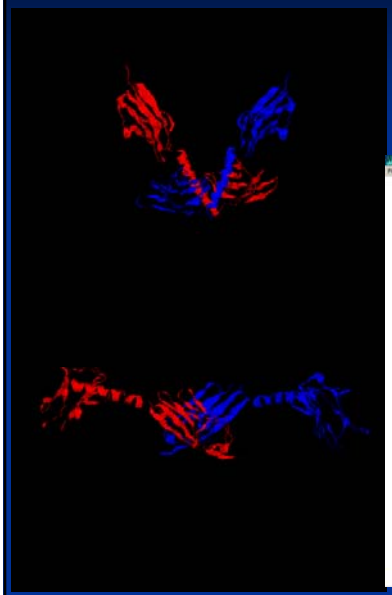


http://dara.embl-hamburg.de/

- About 20000 atomic models of biologically active molecules are generated from the entries in Protein Data Bank and the scattering patterns computed by CRYSOL
- Rapidly identifies proteins with similar shape (from low resolution data) and neighbors in structural organization (from higher resolution data)
- Recent developments: recalculation of the curves, new interface, new search (A.Kikhney, A.Panjkovich)

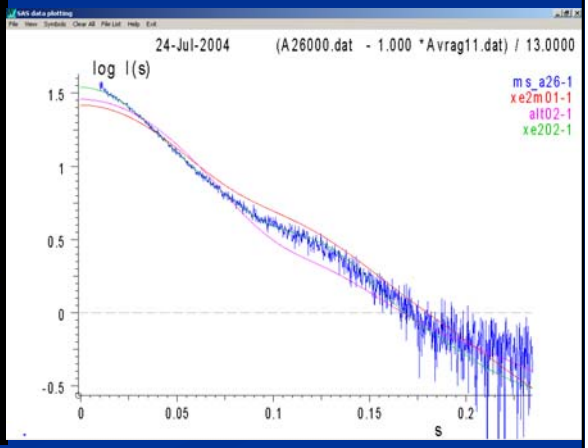
Sokolova, A.V., Volkov, V.V. & Svergun, D.I. (2003) *J. Appl. Crystallogr.* **36**, 865-868

Identification of biologically active oligomers



Biologically active dimer of myomesin-1

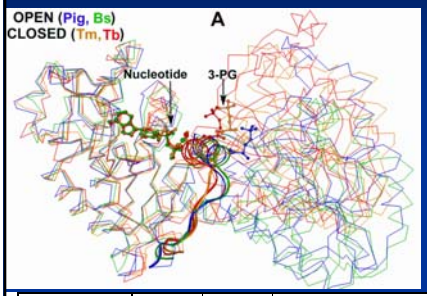
Experiment started: 24-07-2004 at 21:09
Final result obtained: 24-07-2004 at 21:48



Pinotsis, N., Lange, S., Perriard, J.-C., Svergun, D.I. & Wilmanns, M. (2008) *EMBO J*. **27,** 253-264

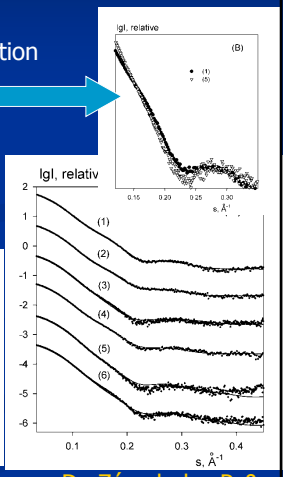
Domain Closure in 3-Phosphoglycerate Kinase

Closure of the two domains of PGK upon substrate binding is essential for the enzyme function. Numerous crystal structures do not yield conclusive answer, which conditions are required for the closure



A SAXS fingerprint of open/closed conformation (human PGK)

SAXS proves that binding of both substrates induces the closure

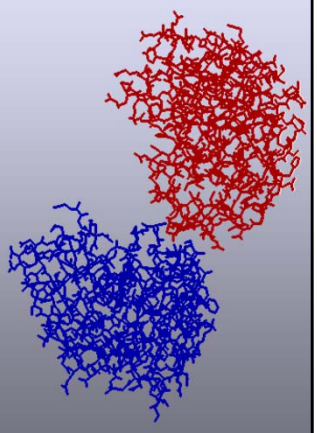


Ligands/ Parameters	Pig PGK	Bs PGK	Pig PGK				Tm PGK	Tb PGK
	Substr. free	MgADP binary	MgATP binary	3-PG binary	^{a}tern1	^{a}tern2	^{a}tern1	^{a}tern2
No	2.746	4.332	3.524	3.158	3.664	4.767	9.135	9.560
3-PG	2.678	5.329	3.297	1.958	3.655	4.234	6.052	6.125
MgATP	3.855	2.848	2.409	3.389	7.827	7.766	3.179	3.910
MgADP	1.486	3.235	1.627	1.140	1.780	2.463	5.151	6.193
MgATP*3-PG	6.140	6.044	4.656	5.307	5.146	4.805	2.247	1.611
MgADP*3-PG	2.303	3.522	2.795	2.049	2.712	2.810	2.018	2.922
R_g (theor), A	24.25	24.34	24.02	23.97	24.24	24.16	23.26	22.64

Varga, A., Flachner, B., Konarev, P., Gráczer, E., Szabó, J., Svergun, D., Závodszky, P. & Vas, M. (2006) *FEBS Lett.* **580**, 2698-2706.

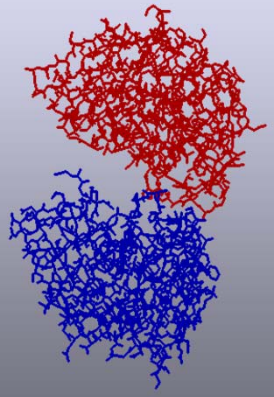
The idea of rigid body modeling

- The structures of two subunits in reference positions are known.
- Arbitrary complex can be constructed by moving and rotating the second subunit.
- This operation depends on three Euler rotation angles and three Cartesian shifts.

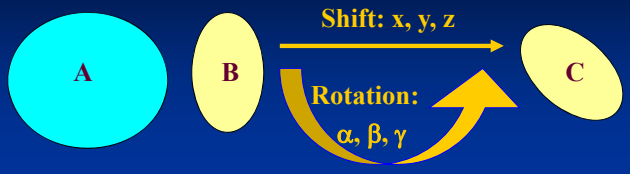


The idea of rigid body modeling

- The structures of two subunits in reference positions are known.
- Arbitrary complex can be constructed by moving and rotating the second subunit.
- This operation depends on three Euler rotation angles and three Cartesian shifts.



Equation for rigid body modeling



Using spherical harmonics, the amplitude(s) of arbitrarily rotated and displaced subunit(s) are analytically expressed *via* the initial amplitude and the six positional parameters: $C_{lm}(s) = C_{lm}(B_{lm}, \alpha, \beta, \gamma, x, y, z)$.

The scattering from the complex is then rapidly calculated as

$$I(s) = I_A(s) + I_B(s) + 4\pi^2 \sum_{0}^{\infty} \sum_{-l}^{l} \mathrm{Re}\left[A_{lm}(s) C_{lm}^{*}(s)\right]$$

Svergun, D.I. (1991). *J. Appl. Cryst.* **24**, 485-492

Constraints for rigid body modelling

- *Interconnectivity*
- *Absence of steric clashes*
- Symmetry
- Intersubunit contacts (from chemical shifts by NMR or mutagenesis)
- Distances between residues (FRET or mutagenesis)
- Relative orientation of subunits (RDC by NMR)
- Scattering data from subcomplexes

Petoukhov & Svergun (2005) *Biophys J.* **89**, 1237; (2006) *Eur. Biophys. J.* **35**, 567.

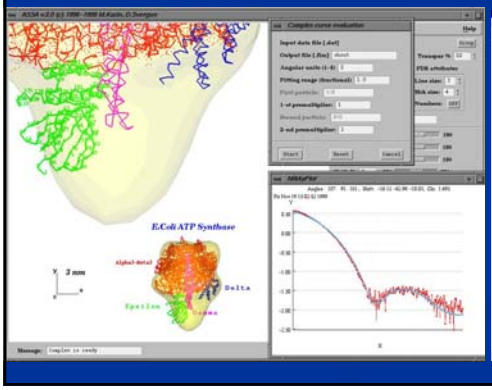
Interactive and local refinement

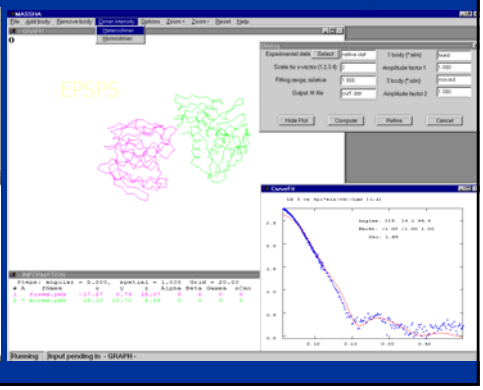
♦ ***ASSA*** (SUN/SGI/DEC)

Kozin & Svergun (2000). *J. Appl. Cryst.* **33**, 775-777

♦ ***MASSHA*** (Win9x/NT/2000)

Konarev, Petoukhov & Svergun (2001). *J. Appl. Cryst.* **34**, 527-532

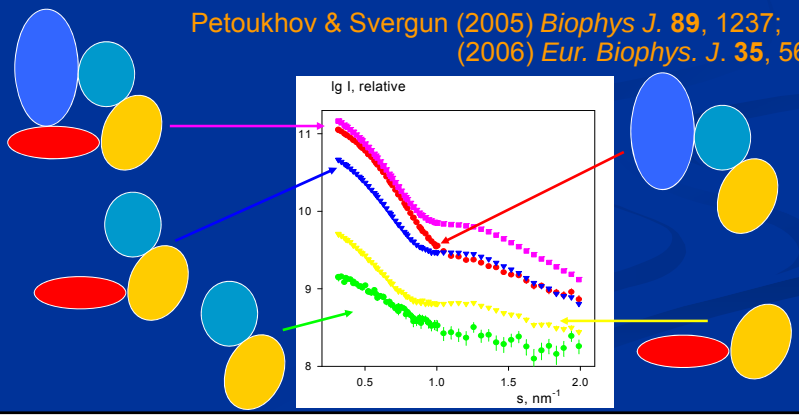




Global rigid body modelling (SASREF)

- Fits (multiple X-ray and neutron) scattering curve(s) from partial constructs or contrast variation using simulated annealing
- Requires models of subunits, builds interconnected models without steric clashes
- Uses constraints: symmetry, distance (FRET or mutagenesis) relative orientation (RDC from NMR), if applicable

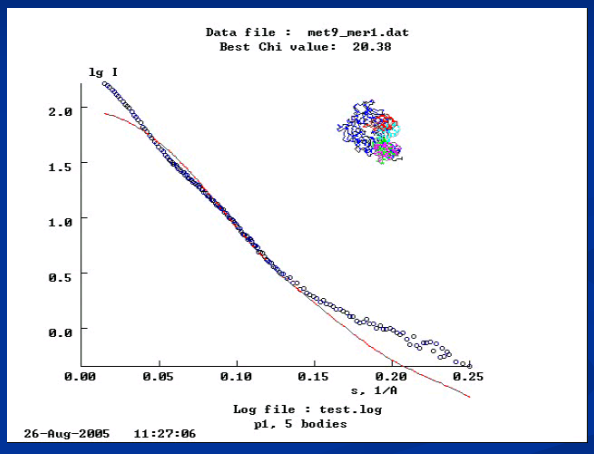
Petoukhov & Svergun (2005) *Biophys J.* **89**, 1237;
(2006) *Eur. Biophys. J.* **35**, 567.



A global refinement run with distance constraints

A tyrosine kinase MET (118 kDa) consisting of five domains

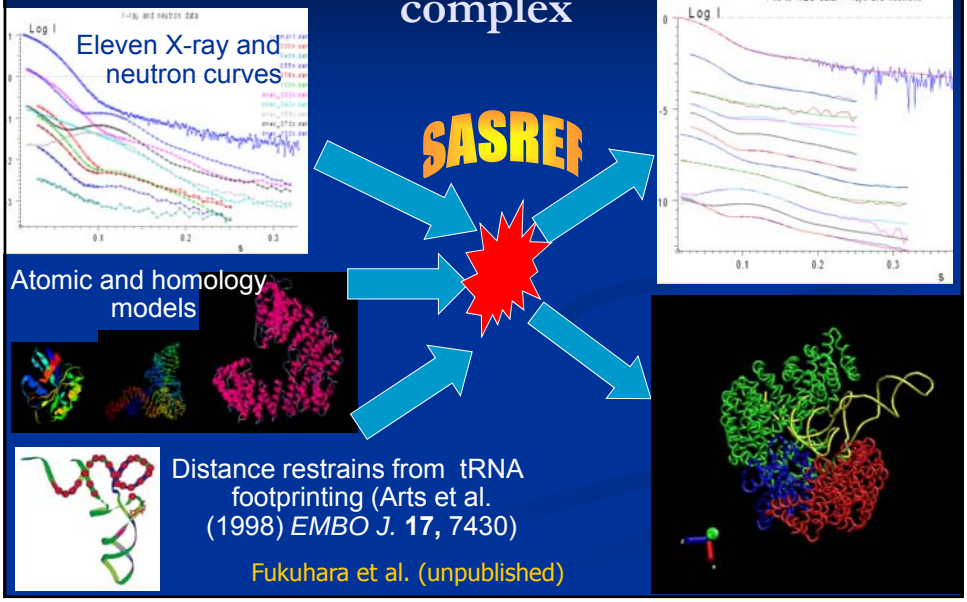
Program SASREF



Single curve fitting with distance constraints: C to N termini contacts

Gherardi, E., Sandin, S., Petoukhov, M.V., Finch, J., Youles, M.E., Ofverstedt, L.G., Miguel, R.N., Blundell, T.L., Vande Woude, G.F., Skoglund, U. & Svergun, D.I. (2006) *PNAS USA,* **103**, 4046.

Rigid body modelling of the Xpot ternary complex

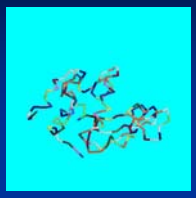


Addition of missing fragments

- Flexible loops or domains are often not resolved in high resolution models or genetically removed to facilitate crystallization
- Tentative configuration of such fragments are reconstructed by fixing the known portion and adding the missing parts to fit the scattering from the full-length macromolecule.

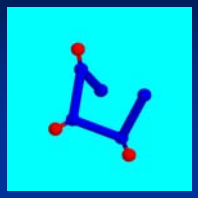
Building native-like folds of missing fragments

- Using DR-type models and protein-specific penalty functions

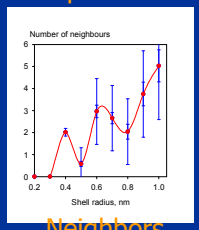


Primary sequence

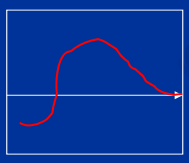
Secondary structure



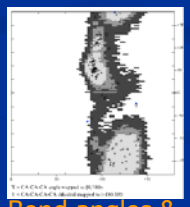
Excluded volume



Neighbors distribution



Knowledge-based potentials

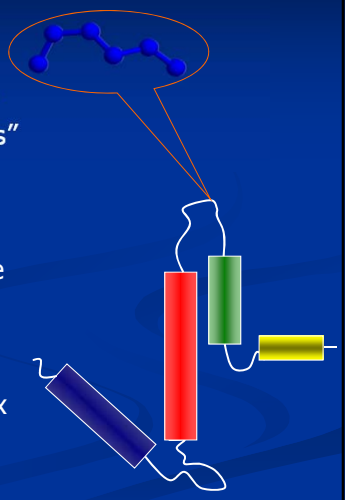


Bond angles & dihedrals distribution

Petoukhov, M.V., Eady, N.A.J., Brown, K.A. & Svergun, D.I. (2002) *Biophys. J.* **83**, 3113

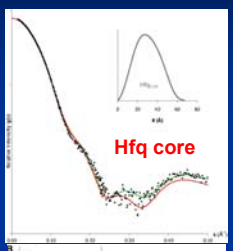
Addition of missing fragments: BUNCH

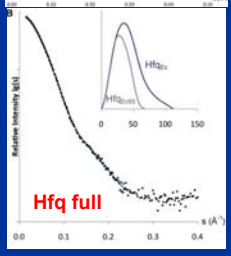
- BUNCH combines rigid body and *ab initio* modelling to find the positions and orientations of rigid domains and probable conformations of flexible linkers represented as "dummy residues" chains
- Multiple experimental scattering data sets from partial constructs (e.g. deletion mutants) can be fitted simultaneously with the data of the full-length protein.
- BUNCH accounts for symmetry, allows one to fix some domains and to restrain the model by contacts between specific residues

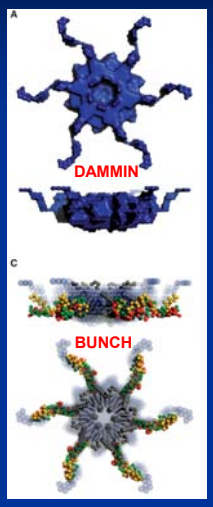


Petoukhov, M. V. & Svergun, D. I. (2005). *Biophys. J.* **89**, 1237-1250

Dynamics and function of the C-terminus of the *E. coli* RNA chaperone Hfq







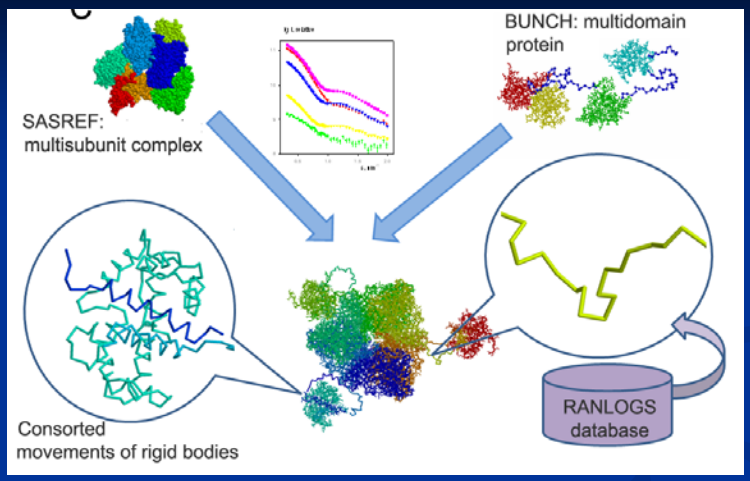
The hexameric Hfq (HfqEc) is involved in riboregulation of target mRNAs by small trans-encoded RNAs. Hfq proteins of different bacteria comprise an evolutionarily conserved core, whereas the C-terminus is variable in length.

By bioinfomatics, NMR, synchrotron CD and SAXS the C-termini are demonstrated to be flexible and to extend laterally away from the hexameric core. The flexible C-terminal moiety is capable of tethering long and structurally diverse RNA molecules.

Beich-Frandsen M, Vecerek B, Konarev PV, Sjöblom B, Kloiber K, Hämmerle H, Rajkowitsch L, Miles AJ, Kontaxis G, Wallace BA, Svergun DI, Konrat R, Bläsi U and Djinovic-Carugo K. (2011) Nucleic Acids Res. **39**, 4900-15

27

Addition of missing fragments: CORAL



- A merger of SASREF and BUNCH: advanced methods to account for missing loops in multi-subunit protein structures (RANLOGS, CORAL)

M.V. Petoukhov, D. Franke, A. Shkumatov, G. Tria, A.G. Kikhney, M. Gajda, C. Gorba, H.D.T. Mertens, P.V. Konarev, D.I. Svergun (2012). *J. Appl. Cryst.* **45**, 342-350.

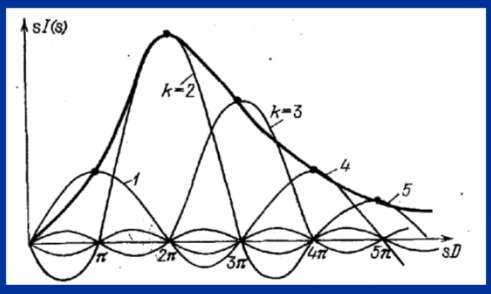
Some words of caution

Or Always remember about ambiguity!

Sampling formalism

Shannon sampling theorem: the scattering intensity from a particle with the maximum size D is defined by its values on a grid $s_k = k\pi/D$ (Shannon channels):

$$sI(s) = \sum_{k=1}^{\infty} s_k a_k \left[\frac{\sin D(s - s_k)}{D(s - s_k)} - \frac{\sin D(s + s_k)}{D(s + s_k)} \right]$$



Shannon sampling was utilized by many authors (e.g. Moore, 1980). An estimate of the number of channels in the experimental data range ($N_s = s_{max}D/\pi$) is often used to assess the information content in the measured data.

Simple explanations do not work in SAS

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)

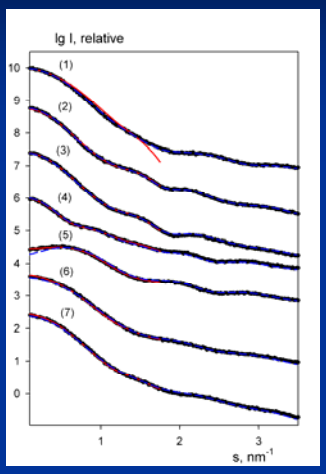
From the informational point of view, rigid body modeling should provide unique or at least much less ambiguous models than shape determination

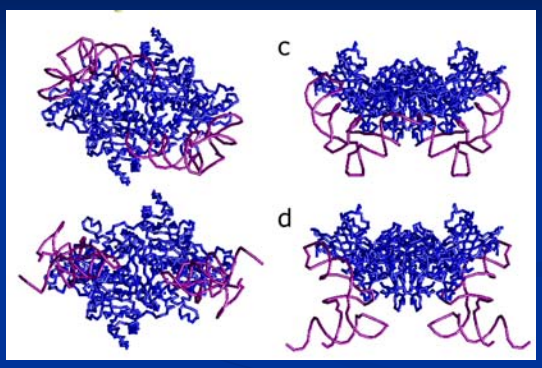
NO WAY

As all the problems are non-linear, the number of Shannon channels does not give you exact number of parameters, which is possible to extract from the scattering data (depending on accuracy, this number varies between zero and infinity).

Further, uniqueness of reconstruction depends largely on the complexity of the function f(x) to be minimized

Ambiguity of rigid body analysis





- A synthetic example: two different orientations of tRNA in a dimeric complex with aspartyl-tRNA synthetase obtained by rigid body modelling and compatible with X-ray and contrast variation neutron scattering data

Petoukhov, M.V. & Svergun, D. I. (2006) *Eur. Biophys. J.* **35**, 567-576

Constraints and restrains used in global modelling procedures

- Information about contacting residues from other experiments (spin labelling, site-directed mutagenesis, FRET, chemical shifts etc)
- Information about symmetry
- Avoiding steric clashes
- For missing loops and linkers: contiguous chain, excluded volume, Ramachandran plot for Cα, knowledge-based potentials etc

AND STILL, one must always cross-validate SAS models against all available biochemical/biophysical information

Use of high resolution models

CRYSOL/CRYSON: computation of X-ray and neutron scattering from atomic models

Interactive and brute force programs

MASSHA/ASSA: interactive manipulations

DIMFOM: homo/heterodimers

GLOBSYMM: symmetric oligomers

Heuristic methods

SASREF: a universal rigid body program

BUNCH, CORAL: allow addition of missing fragments

Recent EMBL SAXS collaborative projects

Transcription factors

De *et al* PNAS (2014)

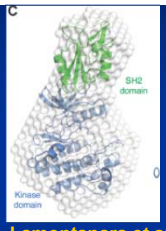
PDH complex E2 core

Marrott *et al* FEBS J (2012)

Human chromatin remodeler CHD4

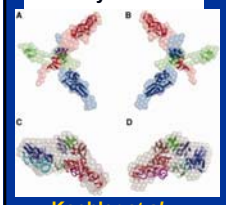
Watson *et al* JMB (2012)

SH2 domain of ABL kinase



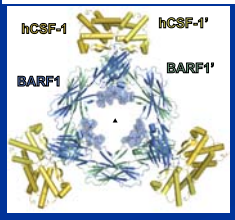
Lamontanara *et al* Nat Commun. (2014)

Arc1p-aminoacyl-tRNA synthetases



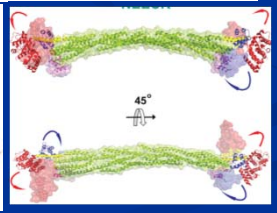
Koehler *et al* NAR (2013)

BARF1/hCSF-1 complex



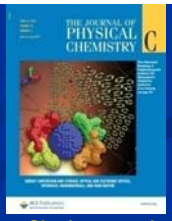
Elegheert *et al* NSMB (2012)

Human Muscle α-Actinin



de Almeida Ribeiro *et al* Cell (2014)

Nanocomposites

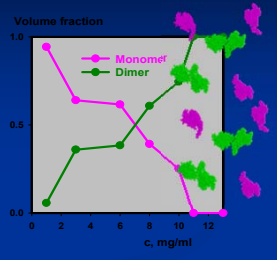


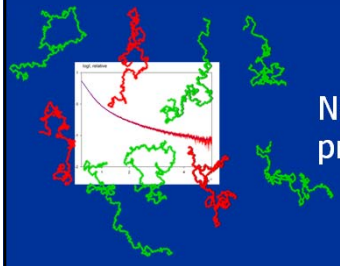
Shtykova *et al* JPC (2012)

Life is more complicated: mixtures and processes

Equilibrium oligomeric mixtures

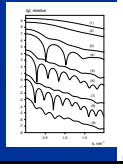
Stoichiometry and complex formation

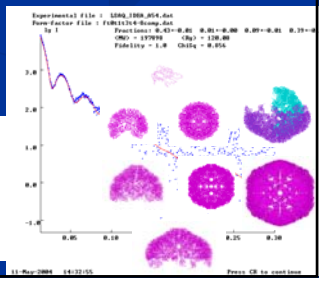




Natively unfolded proteins and multidomain proteins with flexible linkers

Protein folding/unfolding kinetics





Assembly/disassembly processes

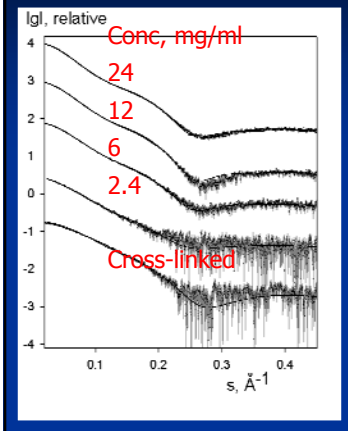
Scattering from mixtures

$$I(s)=\sum_k \nu_k I_k(s)$$

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.

For equilibrium and non-equilibrium mixtures, solution scattering permits to determine the number of components and given their scattering intensities $I_k(s)$, also the volume fractions

An encounter adrenodoxin/cytochrome C complex



Volume fractions

Mon:Dim:Tri:Tet

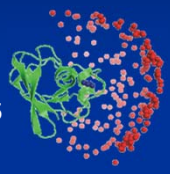
0 : 0 : 50 : 50

0 : 8 : 47 : 45

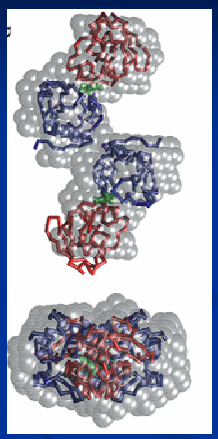
5 : 25 : 55 : 15

25 : 25 : 50 : 0

0 : 100 : 0 : 0 NMR data



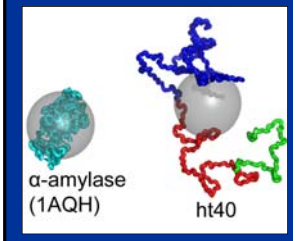
A model of heterotetrameric native complex at high salt and high solute concentration

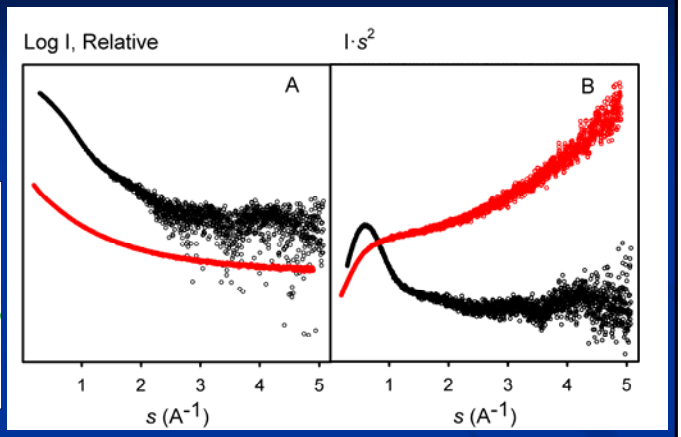


The cross-linked complex Ad/CC is always a heterodimer independent on conditions
The native complex strongly depends on the sample concentration and on the amount of NaCl in the buffer.

X. Xu, W. Reinle, F. Hannemann, P. V. Konarev, D. I. Svergun, R. Bernhardt & M. Ubbink (2008) *JACS*, **130**, 6395-6403

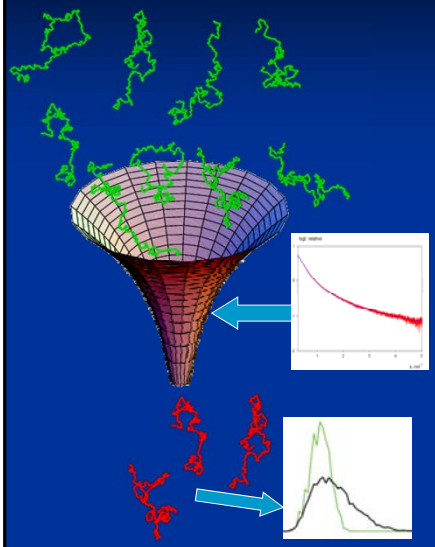
SAXS from folded *vs* disordered protein





Folded: relatively small R_g and D_{max}, bell-shaped Kratky plot
(e.g. for folded α-amylase (448 AAs) R_g=2.4 nm)
Disordered: large R_g and D_{max}, increasing Kratky plot
(e.g. for IUP tau (441 AAs) Rg=6.5 nm)

Quantitative assessment of flexibility



- Automated classification (folded, partially or completely unfolded) is available

D.Franke DATCLASS

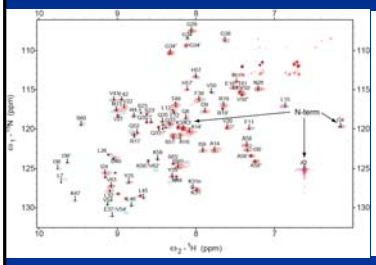
- More quantitative estimates are provided by ensemble methods
- One generates a large pool covering the conformational space and selects sub-ensemble(s) such that their mixture fits the available experimental data
- The structural properties of the selected ensemble(s) are compared to those of the pool

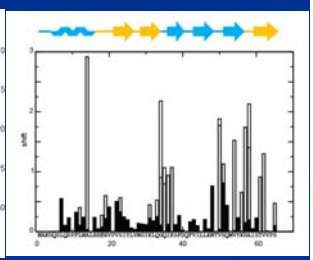
EOM, Bernadó et al. (2007)
J. Am. Chem. Soc. **129**, 5656.

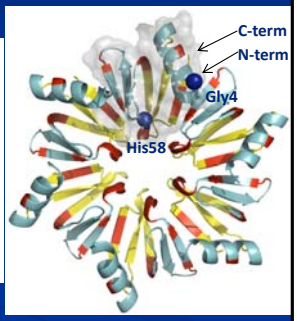
40

DsrA domain II bound to the RNA chaperone Hfq

A small regulatory RNA (DsrA) associates with the RNA chaperone Hfq and requires this protein for regulation of target *E.coli rpoS* mRNA, encoding the stationary phase sigma-factor. Previous studies revealed that the hexameric E. coli Hfq (HfqEc) binds sRNAs on the proximal site.





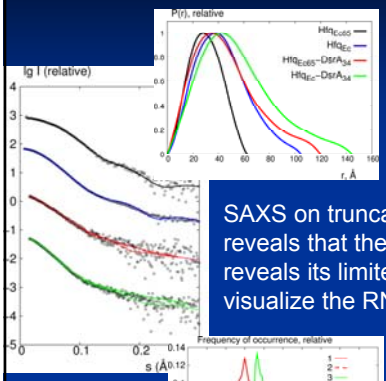


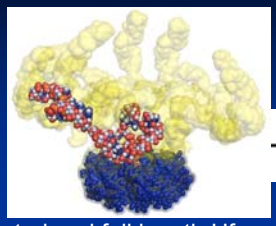
NMR data: superposition of the ^{1}H-^{15}N HSQC spectra of Hfq_{Ec65} RNA free form (blue) and in complex with $DsrA_{34}$ (red) and chemical shift differences. The residues with differences above the threshold are coded red in the Hfq_{Ec65} model

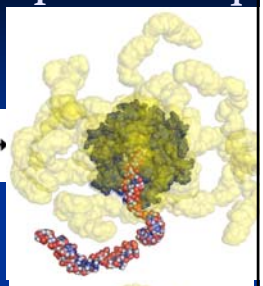
Almeida Ribeiro, E., Beich-Frandsen, M., Konarev, P. V., Shang, W., Vecerek, B., Kontaxis, G., Hammerle, H., Peterlik,H., Svergun, D. I., Blasi, U. & Djinovic-Carugo, K. (2012) *Nucleic Acids Res.* **40**, 8072-8.

41

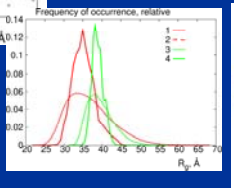
DsrA domain II bound to the RNA chaperone Hfq

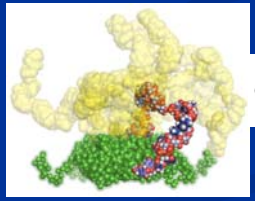






SAXS on truncated and full length Hfq reveals that the RNA is largely unfolded, reveals its limited flexibility and allows one to visualize the RNA conformational space





Almeida Ribeiro, E., Beich-Frandsen, M., Konarev, P. V., Shang, W., Vecerek, B., Kontaxis, G., Hammerle, H., Peterlik,H., Svergun, D. I., Blasi, U. & Djinovic-Carugo, K. (2012) *Nucleic Acids Res.* **40**, 8072-8.

42

Analysis of mixtures

SVDPOLT: singular value decomposition
OLIGOMER: mixtures of components with known structure
EOM: flexible systems and intrinsically unfolded proteins

And now let us awake for the LUNCH

After lunch we shall start with

ATSAS tutorials