Ab initio methods: how/why do they work

D.Svergun

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

- Radius of gyration \( R_g \) (Guinier, 1939)
  \[ I(s) \approx I(0) \exp\left(-\frac{1}{3}R_g^2s^2\right) \]

- Molecular mass (from \( I(0) \))

- Maximum size \( D_{max} \): \( p(r) = 0 \) for \( r > D_{max} \)

- Excluded particle volume (Porod, 1952)
  \[ V = 2\pi^2 I(0)/Q; \quad Q = \int_0^\infty s^2 I(s) ds \]
The scattering is related to the shape (or low resolution structure)

Solid sphere
Hollow sphere
Flat disc
Long rod

Shape determination: how?

3D search model
M parameters

Trial-and-error
Non-linear 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 \( I(s) = \langle |I(s)|^2 \rangle = \langle |F[\rho(r)]|^2 \rangle \), where \( F \) denotes the Fourier transform, \( \langle \cdot \rangle \) stands for the spherical average, and \( s=(s, \Omega) \) is the scattering vector. Expanding \( \rho(r) \) in spherical harmonics

\[
\rho(r) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \rho_{lm}(r) Y_{lm}(\Omega)
\]

the scattering intensity is expressed as

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

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

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

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

Structure of bacterial virus T7


SAXS, 1982

Cryo-EM, 2005

Pro-head
Mature virus

Shape parameterization by spherical harmonics

Homogeneous particle

Scattering density in spherical coordinates $(r, \omega) = (r, \theta, \phi)$ may be described by the envelope function:

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

Shape parameterization by a limited series of spherical harmonics:

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

$Y_{lm}(\omega)$ – 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:

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


Shape parameterization by spherical harmonics

Homogeneous particle

\[ F(\omega) = f_{00} A_{00}(s) + f_{11} A_{11}(s) + f_{20} A_{20}(s) + f_{22} A_{22}(s) + \ldots \]

Spatial resolution: \[ \delta = \frac{\pi R}{(L+1)} \]

Number of model parameters \( f_{lm} \) is \((L+1)^2\).

One can easily impose symmetry by selecting appropriate harmonics in the sum. This significantly reduces the number of parameters describing \( F(\omega) \) for a given \( L \).

Program SASHA
Vector of model parameters:

\[ \text{Position} (j) = x(j) = \begin{cases} 1 & \text{if particle} \\ 0 & \text{if solvent} \end{cases} \]

Number of model parameters

\[ M \approx \left( \frac{D_{\text{max}}}{r_0} \right)^3 \approx 10^3 \]

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.


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: $M \approx 10^3$ variables (e.g. 0 or 1 bead assignments in DAMMIN)

Rigid body methods: $M \approx 10^4$ 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

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

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

---

**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

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 $\{r_i\}$

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.
GASBORB run on C subunit of V-ATPase

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

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

---

Benchmarking *ab initio* methods

Comparison with the crystal structure of lysozyme

<table>
<thead>
<tr>
<th>SASHA</th>
<th>DAMMIN</th>
<th>GASBORB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996</td>
<td>1999</td>
<td>2001</td>
</tr>
</tbody>
</table>
Modular structure of a giant muscle protein titin

- **Z-disc**, I-band, A-band, H-zone
- 26,926 aa
- 1.2 μm
- Z1Z2 includes two modules at the N-terminal of the Z-disc of titin and interacts with telethonin

Solution structure of Z1Z2-telethonin complex

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

Crystal structure of Z1Z2-telethonin complex


Validating NMR models by SAXS

Shape analysis for multi-component systems: principle

One component, one scattering pattern: “normal” shape determination


Shape analysis for multi-component systems: principle

Many components, many scattering patterns: shape and internal structure

EGC stator sub-complex of V-ATPase

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.

**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


---

**Ternary complex: Exportin-t/Ran/tRNA**

Ran (structure known)  Exportin-t  t-RNA (structure known)
(tentative homology model)
X-rays: *ab initio* overall shape

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


---

### 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$_2$O 5 curves
- Ternary complex with deuterated Ran in 0, 40, 55, 70, 100% D$_2$O 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

- α-synuclein oligomers: Giehm et al, PNAS USA (2011)

Structural transitions

- Complement factor H: Morgan et al, NSMB (2011)
- Cytochrome/adrenodoxin: Xu et al, JACS (2008)
- Src kinase: Bernado et al, JMB (2008)
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


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

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.


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
- Cylinder 2:5
- Prism 1:2:4

Average NSD ≈ 0.5

**Spread region**

**Most probable volume**
Fair stability

Ring 1:3:1

Average NSD = 0.9

Disk 5:1

Poor stability

Disk 10:1

This structure can not be restored without use of additional information

Use of symmetry

However: symmetry biases the results and must also be used with caution. Always run in P1 first!

Shape determination of $V_1$ ATPase

Progress in *ab initio* methods

And now let us awake for lunch
Then – yet more exciting topics

M. Petoukhov

*Rigid body analysis*

P. Konarev,
M. Petoukhov,
C. Blanchet:

*Practical demonstration*