

Rapid automated superposition of shapes and macromolecular models using spherical harmonics

Petr V. Konarev,^a Maxim V. Petoukhov^b and Dmitri I. Svergun^{b*}

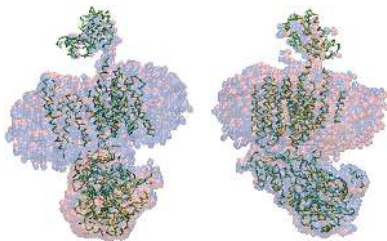
^aLaboratory of Reflectometry and Small-Angle Scattering, A.V. Shubnikov Institute of Crystallography of Federal Scientific Research Centre 'Crystallography and Photonics', Russian Academy of Sciences, Leninsky prospekt 59, Moscow, 119333, Russian Federation, and ^bHamburg Outstation, European Molecular Biology Laboratory, Notkestrasse 85, Hamburg, 22607, Germany. *Correspondence e-mail: svergun@embl-hamburg.de

A rapid algorithm to superimpose macromolecular models in Fourier space is proposed and implemented (*SUPALM*). The method uses a normalized integrated cross-term of the scattering amplitudes as a proximity measure between two three-dimensional objects. The reciprocal-space algorithm allows for direct matching of heterogeneous objects including high- and low-resolution models represented by atomic coordinates, beads or dummy residue chains as well as electron microscopy density maps and inhomogeneous multi-phase models (*e.g.* of protein–nucleic acid complexes). Using spherical harmonics for the computation of the amplitudes, the method is up to an order of magnitude faster than the real-space algorithm implemented in *SUPCOMB* by Kozin & Svergun [*J. Appl. Cryst.* (2001), **34**, 33–41]. The utility of the new method is demonstrated in a number of test cases and compared with the results of *SUPCOMB*. The spherical harmonics algorithm is best suited for low-resolution shape models, *e.g.* those provided by solution scattering experiments, but also facilitates a rapid cross-validation against structural models obtained by other methods.

1. Introduction

Small-angle scattering of X-rays and neutrons (SAXS and SANS) is more and more actively employed for structural studies of biomacromolecular solutions, including hybrid modelling in combination with other structural methods (Svergun *et al.*, 2013). The SAXS/SANS analysis methods provide three-dimensional models of different nature and resolution, and comparisons between such heterogeneous models are often required for cross-validation of structural results obtained by different techniques. The comparisons usually require automated best-matching superposition of three-dimensional structures. Superposition of heterogeneous models, which may not only be different in resolution and representations (atomic models, bead models, dummy residue chains, density maps, surfaces *etc.*) but also contain parts with significantly different density (*e.g.* nucleoprotein complexes), is not a trivial task.

We previously developed a program *SUPCOMB* (Kozin & Svergun, 2001) for matching high- and low-resolution three-dimensional structures, which uses a normalized spatial discrepancy (NSD) as a proximity measure between the objects. For every point (bead or atom) in the first model, the closest neighbouring point in the second model is found, and the same is done for all points in the second model. The squared proximal distances are added and normalized against the squared average distances between neighbouring points for the two models, yielding the NSD value for the given



relative positions of the models. Starting from an inertia-axes alignment, the algorithm minimizes the NSD and finds the best-matching alignment of the structures. However, the CPU time used by *SUPCOMB* is proportional to the product of the number of points (beads or atoms) representing the two objects and for large macromolecular complexes the program becomes computationally expensive.

There are a number of fitting algorithms and programs in electron microscopy (EM) to fit high-resolution models into lower-resolution electron-density maps, such as *SITUS* (Chacón & Wriggers, 2002), *FRM* (Kovacs *et al.*, 2003), *FOLDHUNTER* (Jiang *et al.*, 2001), *COAN* (Volkman & Hanein, 2003), *NORMA* (Suhre *et al.*, 2006) and *ADP_EM* (Garzón *et al.*, 2007). All these methods use a real-space cross-correlation coefficient as a proximity measure between the models, some of them (*e.g.* *ADP_EM*) utilizing a spherical harmonics expansion of the density function. Most algorithms employ an exhaustive search of the six-dimensional parameter space (three rotations and three translations) using different mathematical transformations to speed up the process (fast Fourier transforms and/or spherical harmonics), while some of them perform rapid local optimization algorithms making use of the score gradient (*e.g.* steepest ascent). The above methods are similar in performance and the accuracy of the results largely depends on the software implementation.

Here, we propose an algorithm for a fast matching of macromolecular models based on the spherical harmonics representation of the scattering amplitudes in Fourier space. The method uses a normalized integrated cross-term of the scattering amplitudes as a proximity measure between two low-resolution three-dimensional objects. The method is implemented in a program which is significantly faster than *SUPCOMB* and is directly applicable for comparisons of low-resolution shapes and heterogeneous models of different nature (atoms, beads, EM maps *etc.*).

2. A normalized correlation coefficient as a proximity measure between three-dimensional objects

For an arbitrary three-dimensional object with scattering density $\rho_A(\mathbf{r})$ in real space, the scattering amplitude $A(\mathbf{s})$ can be represented using spherical harmonics as (Stuhrmann, 1970)

$$A(\mathbf{s}) = \sum_{l=0}^L \sum_{m=-l}^l A_{lm}(s) Y_{lm}(\Omega), \quad (1)$$

where the momentum transfer $s = 4\pi \sin(\theta/2)/\lambda$ with θ the scattering angle and λ the wavelength, Ω is the solid angle in reciprocal space, $\mathbf{s} = (s, \Omega)$, and the truncation value L defines the resolution. For an object represented by N points (atoms, beads or density values) the partial amplitudes $A_{lm}(s)$ are calculated as

$$A_{lm}(s) = 4\pi i^l \sum_{j=1}^N f_j(s) j_l(sr_j) Y_{lm}^*(\omega_j), \quad (2)$$

where the sum runs over all elements with coordinates $\mathbf{r}_j = (r_j, \omega_j) = (r_j, \theta_j, \varphi_j)$ and $f_j(s)$ are the corresponding form factors.

Owing to the orthogonal properties of the spherical harmonics, a spherically averaged scattering intensity $I_A(s)$ (*e.g.* as measured in a small-angle experiment from an ensemble of randomly oriented particles) is expressed as the sum of individual multipole contributions:

$$I_A(s) = 2\pi^2 \sum_{l=0}^L \sum_{m=-l}^l |A_{lm}(s)|^2. \quad (3)$$

Let us now consider two three-dimensional objects with densities $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$. The best superposition of these objects should maximize the integral correlation expressed as $\int [\rho_A(\mathbf{r}) + \rho_B(\mathbf{r})]^2 d\mathbf{r}$. Following Parseval's theorem (Parseval des Chênes, 1799) this integral is equal to the total intensity in reciprocal space:

$$\begin{aligned} T(A, B) &= \int [A(\mathbf{s}) + B(\mathbf{s})]^2 d\mathbf{s} \\ &= \int [I_A(\mathbf{s}) + I_B(\mathbf{s}) + 2A(\mathbf{s})B^*(\mathbf{s})] d\mathbf{s} \\ &= 4\pi \int [I_A(s) + I_B(s) + 2I_{AB}(s)] s^2 ds \end{aligned} \quad (4)$$

where $I_A(s)$ and $I_B(s)$ are the averaged scattering intensities of the objects A and B, respectively, and $I_{AB}(s)$ denotes the cross-term. Note that the individual integrated intensities do not depend on the position and only the integral from the cross-term changes when the particles are moved/rotated. Using the spherical harmonics expansion [equations (1)–(3)], the three terms are readily calculated in terms of the partial amplitudes and a quantitative measure of the agreement between the two real-space three-dimensional objects can therefore be expressed in reciprocal space as a normalized correlation coefficient (NCC):

$$\text{NCC}(A, B) = \frac{\int_0^{s_m} [A_{lm}(s)B_{lm}^*(s) + A_{lm}^*(s)B_{lm}(s)] s^2 ds}{2 \left[\int_0^{s_m} I_A(s) s^2 ds \right]^{1/2} \left[\int_0^{s_m} I_B(s) s^2 ds \right]^{1/2}}, \quad (5)$$

where $A_{lm}(s)$ and $B_{lm}(s)$ are the scattering amplitudes of the objects A and B, respectively. Here, all integrals are evaluated in a restricted angular range up to $s = s_m$ defining the resolution of the objects. Theoretically (for an infinite upper limit of integration), NCC varies between 0 and 1, the latter value corresponding to an ideal overlap of two identical structures. Thus, NCC can serve as a convenient proximity measure and its maximization should allow one to obtain the best overlap of two objects.

3. The superposition algorithm for the alignment of two heterogeneous objects

Numerical minimization of a proximity measure between three-dimensional heterogeneous objects with respect to the positional and rotational parameters is a non-trivial task. The target function may display multiple local minima, where the search algorithms may be trapped. Global minimization algorithms (*e.g.* simulated annealing) can overcome local minima but they are not computationally efficient. A practical approach to solving this problem is the use of a local mini-

Table 1

Test objects superimposed using *SUPALM* and *SUPCOMB*.

The template objects represent high-resolution structures (their scattering amplitudes were calculated with *CRY SOL*) and the matched objects are low-resolution *ab initio* shapes (with the scattering amplitudes calculated by *DAM2ALM*) obtained by *DAMMIN* (Svergun, 1999), *DAMMIF* (Franke & Svergun, 2009), *GASBOR* (Svergun *et al.*, 2001) and *MONSA* (Svergun & Nierhaus, 2000). The NSD values were computed between the template and the matched objects obtained by *SUPCOMB* and *SUPALM*. The NCC values were evaluated by *SUPALM*. The computing gain factor is obtained as the ratio of CPU time of *SUPCOMB* to that of *SUPALM*. The default parameters of *SUPALM* ($L_{\max} = 5$ and $N_{\text{sh}} = 7$) were used for all tests.

No.	Template object	Matched object	MW (kDa)	NSD, <i>SUPCOMB</i>	NCC, <i>SUPCOMB</i>	NSD, <i>SUPALM</i>	NCC, <i>SUPALM</i>	CPU gain factor <i>SUPALM</i> versus <i>SUPCOMB</i>
1	2a38 (PDB)	z1z2 (<i>DAMMIF</i>)	21	1.70	0.87	1.79	0.96	1.6
2	1got (PDB)	G protein (<i>GASBOR</i>)	40	1.12	0.72	1.14	0.76	2.8
3	1o6v (PDB)	Internalin (<i>DAMMIN</i>)	50	1.69	0.74	1.94	0.95	3.2
4	4uis (PDB)	EMD-2974 (EMD map)	140	3.65	0.86	3.75	0.93	70.0†
5	1hqk (PDB)	Lumazine synthase (<i>DAMMIN</i>)	960	1.22	0.83	1.24	0.87	10.1
6	4v4w (PDB)	70S ribosome (<i>MONSA</i>)	2150	1.10	0.81	1.27	0.94	25.2

† For the EMD map file (EMD-2974) the comparison with *SUPCOMB* was made in the following way. For *SUPCOMB* the electron-density map was first transformed into a bead PDB file (containing 29 374 atoms) using the density threshold 0.04 by the program *EM2DAM* from the *ATSAS* package and then it was superimposed with the PDB structure (4uis). *SUPALM* performed the direct superposition of PDB and EMD files. The RMSD between the deposited fitted EMDDataBank model and the *SUPALM* fit is 5.5 Å.

mization starting from pre-aligned positions. These can be obtained from the inertia-axes matching procedure as implemented in *SUPCOMB* (Kozin & Svergun, 2001). Here, the principal axes of inertia are found for both objects as the eigenvectors of the inertia tensor matrix representing linear combinations of the second central moments of distribution around the mass centres. The two objects are set in canonical positions where they are origin centred and rotated so that their principal inertia axes, taken in ascending order of eigenvalues, are aligned along the *X*, *Y* and *Z* axes, respectively. Depending on whether enantiomorph structures are allowed or not, there are four or eight sign combinations of the eigenvalues.

Here, we employ the following algorithm for the superposition of two three-dimensional objects (model A and model B):

(i) Inertia tensors and their eigenvectors are computed for both model A and model B.

(ii) Model B is rotated by $\mathbf{M}_A \mathbf{M}_B^T$ and shifted by $\mathbf{T}_B - \mathbf{T}_A$ to align its principal axes with those of model A (here, \mathbf{M}_A and \mathbf{M}_B are the rotation matrices of model A and model B, respectively, and \mathbf{T}_A and \mathbf{T}_B are the corresponding shifts of the centres of mass from the origin of the coordinate system). The signs of the columns of the rotation matrix \mathbf{M}_B are selected out of four sign combinations (or eight, if enantiomorphs are allowed).

(iii) The scattering amplitudes of the two objects $A_{lm}(s)$ and $B_{lm}(s)$ are evaluated. Here, one can use *CRY SOL* (Svergun *et al.*, 1995) for the high-resolution atomic models or *DAM2ALM* [available in the *ATSAS* package (Petoukhov *et al.*, 2012)] for low-resolution *ab initio* shapes. Options for the direct reading of electron-density maps (EMD files, EMDDataBank; <http://emdatbank.org/index.html>) or multi-phase *ab initio* bead models [obtained by the program *MONSA* (Svergun & Nierhaus, 2000)] and calculation of their scattering amplitudes are also provided.

(iv) The position and orientation of model B are refined by minimizing the value of $1/\text{NCC}$ using a local minimization by the nonlinear least-squares package *NL2SOL* (Dennis *et al.*,

1981). The obtained best NCC value is reported as an estimate of proximity measure between the objects.

Steps (i) and (ii) are similar to the ones used in *SUPCOMB*; however, steps (iii) and (iv) are different and offer several advantages. First, the use of the pre-computed amplitudes allows one to easily match heterogeneous objects and models of different nature. Second, the changes in $B_{lm}(s)$ upon rotations and displacements of object B are rapidly calculated using the finite rotation elements matrix (Edmonds, 1957; Svergun, Volkov *et al.*, 1997). Importantly, the scattering amplitudes are calculated only once and the computational cost of the algorithm does not depend on the complexity of the structures to be superimposed (in contrast to *SUPCOMB*, where the calculation time is proportional to the product of the numbers of elements in the two objects, $N_A N_B$).

4. Optimal resolution parameters for the NCC calculation

The method was implemented in a computer program called *SUPALM*, and its performance was tested in a number of test cases on various high- and low-resolution models of biological macromolecules and compared with *SUPCOMB*. Most of the low-resolution models in Tables 1–3 are experimentally determined *ab initio* protein shapes reconstructed from SAXS patterns measured at the beamlines X33 and P12 (EMBL c/o DESY) using *DAMMIN* (Svergun, 1999), *DAMMIF* (Franke & Svergun, 2009) and *GASBOR* (Svergun *et al.*, 2001). All these shapes are superimposed with the available crystal structures of these proteins taken from the appropriate Protein Data Bank (PDB) entries. To illustrate the performance of *SUPALM* on EM maps, a bead model generated directly from the experimental EMD map is superposed with the high-resolution cryoEM-derived model of the human gamma-secretase (Sun *et al.*, 2015). As an example of a heterogeneous assembly the model of the protein–RNA distribution in the 70S ribosome of *Escherichia coli* derived from X-ray and neutron scattering data (Svergun, Burkhardt

Table 2

Test objects superimposed by *SUPALM* with different maximum number of harmonics L_{\max} used in the calculations of the scattering amplitudes.

The objects are the same as in Table 1. The angular data range corresponding to seven Shannon channels ($N_{\text{sh}} = 7$) was used for the tests. Values in bold correspond to the default setting for the maximum number of harmonics L_{\max} .

No.	Template object	Matched object	MW (kDa)	NSD/NCC ($L_{\max} = 3$)	NSD/NCC ($L_{\max} = 4$)	NSD/NCC ($L_{\max} = 5$)	NSD/NCC ($L_{\max} = 6$)	NSD/NCC ($L_{\max} = 7$)
1	2a38 (PDB)	z1z2 (<i>DAMMIF</i>)	21	1.82/0.97	1.84/0.96	1.79/0.96	1.80/0.95	1.79/0.95
2	1got (PDB)	G protein (<i>GASBOR</i>)	40	1.28/0.77	1.16/0.76	1.14/0.76	1.14/0.75	1.13/0.74
3	1o6v (PDB)	Internalin (<i>DAMMIN</i>)	50	2.08/0.97	1.96/0.96	1.94/0.95	1.95/0.94	1.93/0.93
4	4uis (PDB)	EMD-2974 (EMD map)	140	4.10/0.95	3.89/0.94	3.75/0.93	3.75/0.92	3.74/0.91
5	1hqk (PDB)	Lumazine synthase (<i>DAMMIN</i>)	960	1.27/0.88	1.26/0.87	1.24/0.87	1.25/0.87	1.24/0.86
6	4v4w (PDB)	70S ribosome (<i>MONSA</i>)	2150	1.30/0.96	1.28/0.95	1.27/0.94	1.28/0.94	1.26/0.93

Table 3

Test objects superimposed by *SUPALM* using different angular data ranges N_{sh} for the calculations of the scattering amplitudes.

The objects are the same as in Table 1. The default number of spherical harmonics ($L_{\max} = 5$) was used for the tests. Values in bold correspond to the default setting for the angular data range.

No.	Template object	Matched object	MW (kDa)	NSD/NCC ($N_{\text{sh}} = 5$)	NSD/NCC ($N_{\text{sh}} = 6$)	NSD/NCC ($N_{\text{sh}} = 7$)	NSD/NCC ($N_{\text{sh}} = 8$)	NSD/NCC ($N_{\text{sh}} = 9$)
1	2a38 (PDB)	z1z2 (<i>DAMMIF</i>)	21	1.95/0.97	1.87/0.96	1.79/0.96	1.80/0.95	1.79/0.94
2	1got (PDB)	G protein (<i>GASBOR</i>)	40	1.17/0.77	1.16/0.76	1.14/0.76	1.15/0.76	1.14/0.75
3	1o6v (PDB)	Internalin (<i>DAMMIN</i>)	50	2.13/0.96	2.02/0.95	1.94/0.95	1.96/0.94	1.95/0.94
4	4uis (PDB)	EMD-2974 (EMD map)	140	5.33/0.95	3.91/0.94	3.75/0.93	3.74/0.93	3.73/0.92
5	1hqk (PDB)	Lumazine synthase (<i>DAMMIN</i>)	960	1.28/0.88	1.25/0.87	1.24/0.87	1.24/0.87	1.23/0.86
6	4v4w (PDB)	70S ribosome (<i>MONSA</i>)	2150	1.30/0.97	1.29/0.95	1.27/0.94	1.28/0.94	1.27/0.93

et al., 1997) is compared with the recent cryo-EM structure of the ribosome (Mitra *et al.*, 2006). As seen from Tables 1–3, for all these different objects the NSD values yielded by *SUPCOMB* agree well with the NSDs computed from the models provided by *SUPALM*. In Figs. 1–3 the template models and those superimposed by *SUPCOMB/SUPALM* are presented to illustrate the similarity of the results of the two programs. Note that, for the case of human gamma-secretase where the largest NSD values are observed, the atomic resolution fragments of the cryo-EM model were manually fitted inside the EMD map using *Coot* (Emsley & Cowtan, 2004) in the original publication (Sun *et al.*, 2015).

The results of Table 2 indicate that, in all test cases, a maximum number of spherical harmonics of $L_{\max} = 5$ was sufficient to ensure a good-quality superposition of the two objects. If a lower number of harmonics are used, the matching quality becomes significantly worse (as can be judged from the higher NSD values), whereas higher L_{\max} values increase the computation times but do not significantly improve the overlap.

The angular data range employed to compute the integrals in equation (5) can also be optimally selected. As illustrated in Table 3, s_m values corresponding to seven Shannon channels N_{sh} (where $N_{\text{sh}} = D_{\max} s_m / \pi$ and D_{\max} is the maximum size of the particle) are sufficient for a reliable positioning. The use of wider angular intervals does not significantly improve the NSD values but, again, slows down the calculations. On the basis of these results the default values for *SUPALM* are selected to be $L = 5$ and $s_m = 7\pi / D_{\max}$ (if needed, these values can be changed by the user from the command-line arguments).

5. Comparison of the performance of *SUPCOMB* and *SUPALM*

SUPALM with the default parameters is about 1.5–2 times faster than *SUPCOMB* for small macromolecules with molecular weights lower than 100 kDa (represented by about 10^3 atoms). For large macromolecular complexes (*e.g.* 1 MDa, about 10^5 atoms) *SUPALM* performs more than ten times faster compared to *SUPCOMB*. For superpositions of crystal structures with EM density maps the speed of *SUPALM* is of the same order of magnitude as that of the high-performance EM docking programs like *SITUS* or *ADP_EM*.

The superimposed models obtained by *SUPCOMB* and *SUPALM* are, as expected, not identical as the quantities to be minimized (NSD and $1/\text{NCC}$, respectively) are different. Still, the overlaps in Figs. 1–3 and comparison of the NSD values in Table 1 indicate that, although the solutions by *SUPALM* reveal slightly higher NSD values, their positions essentially coincide with those given by *SUPCOMB*. Fig. 4 displays an example of NSD and $1/\text{NCC}$ contour profiles in the vicinity of a *SUPALM* solution. The plots of both functions display well defined minima, and both behave as analytical functions smoothly approaching the minimum values. Overall, in all the test examples, *SUPALM* provides practically the same results as *SUPCOMB* but runs much faster, especially for large macromolecular complexes.

We have also implemented an alternative procedure for speeding up the model superposition of *SUPCOMB*, using a different real-space metric for the proximity measure of three-dimensional objects, the normalized overlapped volume (NOV). Here, both objects are voxelized on a cubic grid in real space and the maximization of NOV provides the best

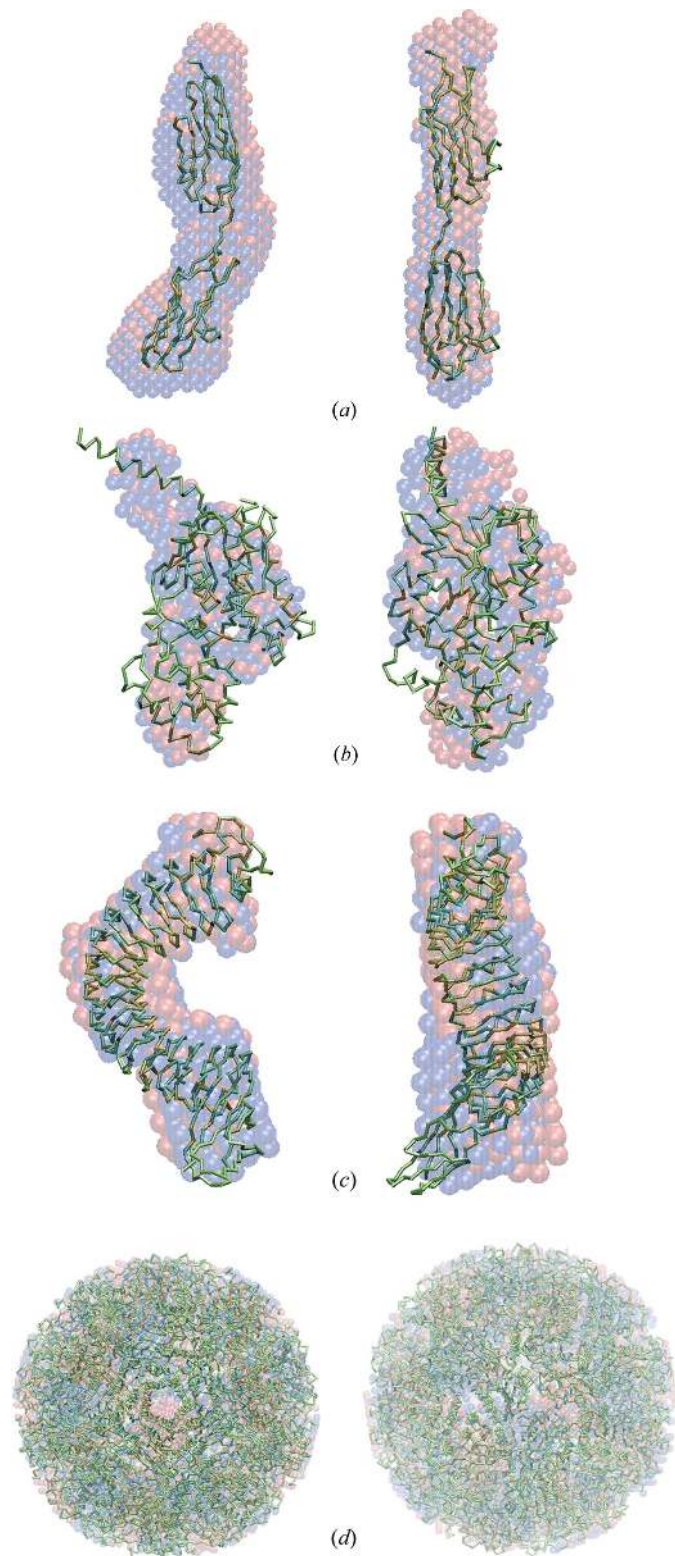


Figure 1
The crystal structures of several proteins (green α traces) with superimposed *ab initio* shapes obtained by SUPALM (red spheres) and by SUPCOMB (blue spheres). (a) Z1Z2 protein (PDB code 2a38; Marino *et al.*, 2006), (b) G protein (PDB code 1got; Lambright *et al.*, 1996), (c) internalin (PDB code 1o6v; Schubert *et al.*, 2002), (d) lumazine synthase (PDB code 1hqk; Zhang *et al.*, 2001). The right view is rotated counterclockwise by 90° around the vertical axis.

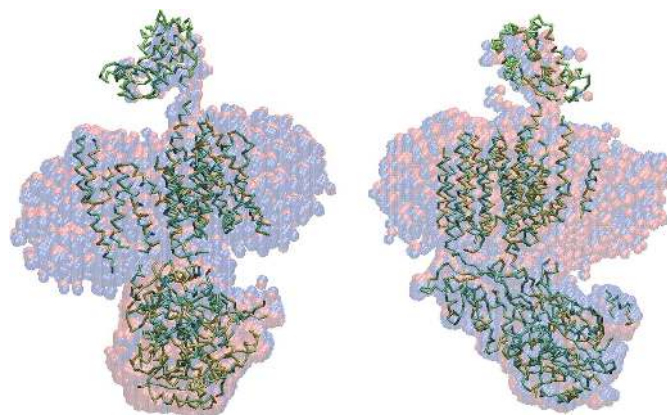


Figure 2
A hybrid high-resolution model of the human gamma-secretase derived from cryoEM (PDB code 4uis; Sun *et al.*, 2015) (green α traces) and the bead models from the electron-density map (ID EMD-2974.map) superimposed onto it by SUPALM (red spheres) and by SUPCOMB (blue spheres). The right view is rotated counterclockwise by 90° around the vertical axis.

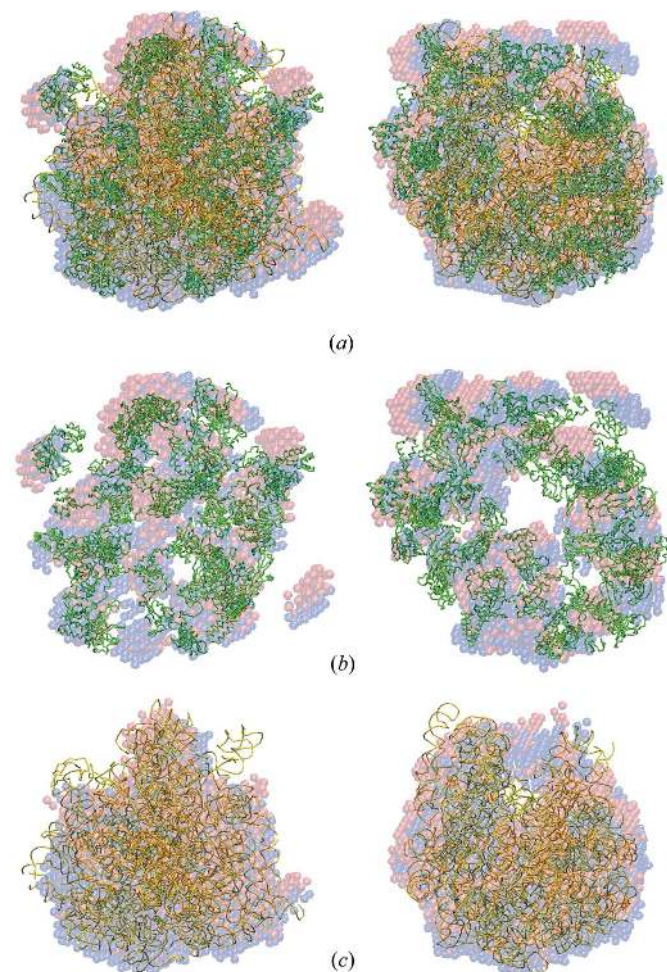


Figure 3
The crystal structure of the 70S ribosome (PDB code 4v4w; Mitra *et al.*, 2006) (green α traces correspond to the protein parts, the RNA parts are shown in yellow) with superimposed two-phase *ab initio* shapes obtained by SUPALM (red spheres) and by SUPCOMB (blue spheres). (a) The complete 70S ribosome models, (b) the protein parts, (c) the RNA parts. The right view is rotated counterclockwise by 90° around the vertical axis.

matching of two objects. We found that the performance of the accelerated *SUPCOMB* using this metric is comparable to that of *SUPALM*; however, the NOV-based superpositions are not

sufficiently sensitive to the finer details of the objects. For complicated shapes, the use of NOV may yield worse NSD values compared to *SUPALM* and to the standard (NSD-driven) mode of *SUPCOMB*. The rapid NOV mode may still be useful for high-throughput superpositions of simple shapes and will therefore be offered to *SUPCOMB* users.

An important feature of *SUPALM* (not available in *SUPCOMB*) is the ability to work directly with electron-density maps (EMD files) and with multi-phase *ab initio* MONSA models of inhomogeneous particles (e.g. protein–RNA and/or protein–DNA complexes). In the case of EMD models, the three-dimensional grid of voxels and the corresponding electron densities are directly used to calculate the scattering amplitudes; in the case of inhomogeneous bead models additional input files with the scattering density values for each bead phase (e.g. protein or DNA/RNA part of the complex) must be provided by the user. After the appropriate computation of the scattering amplitudes, the superposition runs automatically as in the plain bead or atomic model case.

6. Additional test cases for *SUPALM* alignments

The examples considered in Table 1 utilized *ab initio* shapes restored from the experimental data to illustrate similarity between *SUPALM* and *SUPCOMB* results; therefore no ‘true’ positions of the structures were available, with *SUPCOMB* solutions used as reference. To test the ability of *SUPALM* to reconstruct ‘true’ alignments, an additional series of model calculations were performed. First, low-resolution bead envelopes were generated from the high-resolution structures in their reference positions. Then, the bead models were arbitrarily moved/rotated as rigid bodies and *SUPALM* was employed to reconstruct their initial position. In all cases presented in Table 1, the initial position was restored, with a r.m.s.d. of about 1 Å, demonstrating the ability of *SUPALM* to adequately find the best overlap. The small differences between the ‘true’ and matched structures are within the resolution limit used in the calculations.

Another test was conducted to check the stability of *SUPALM* when superposing variable conformations. For this, an NMR ensemble of protein structures (PDB code 2ma0 containing ten conformers of the small protein UVI31+; Chary, Rout & Rao, unpublished work) was selected. The first model in the PDB file was taken as a reference and all conformers were superimposed onto it by *SUPALM*. Fig. 5 displays the original NMR ensemble (where the rigid parts overlap with each other but the flexible portions display great variability) and the *SUPALM* superpositions, where the differences are, as expected, more uniformly distributed between the structures, but still the common parts of the ensemble structures are well superposed.

As a rule of thumb, for *SUPCOMB* alignments, NSD values around unity indicate good similarity between the two objects, whereas significantly larger values point to poor similarity. For NCC, the cut-off value for a good overlap depends on the molecular size, shape complexity and model resolution used in the calculations. From our test examples, the lower limit for a

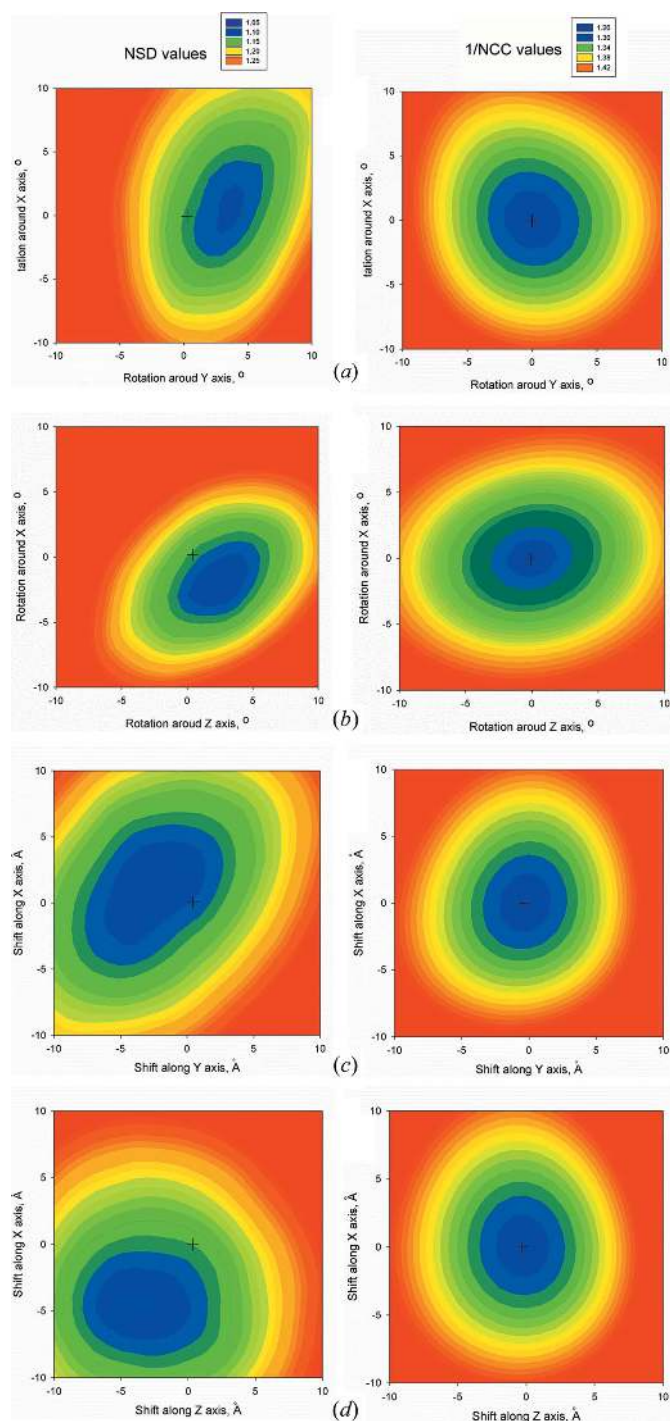


Figure 4 Two-dimensional contour plots of NSD (left column) and 1/NCC (right column) in the vicinity of the *SUPALM* solution for G protein (see Fig. 1b). The *SUPALM* solution corresponds to the origin of the coordinates for all plots (marked with a cross symbol). (a), (b) The contour plots versus the rotations around *x/y* and *x/z* axes, respectively. (c), (d) The contour plots versus the shifts along *x/y* and *x/z* axes, respectively. The ‘true’ *SUPCOMB* solution (the minimum of the two-dimensional contour plot) is positioned close to the *SUPALM* solution (RMSD = 3.3 Å).

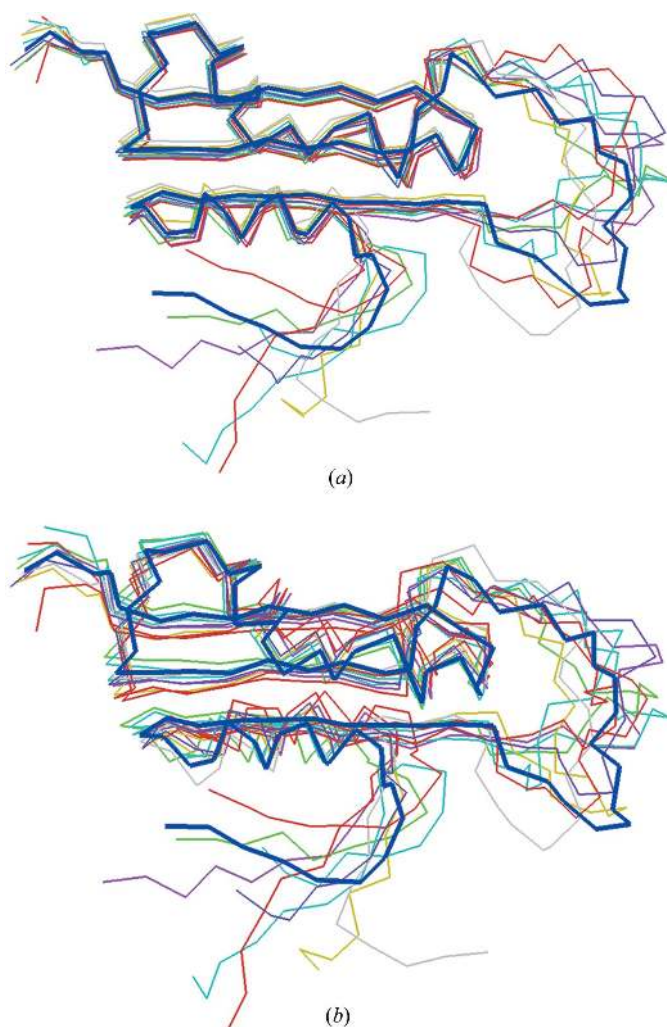


Figure 5
Different conformers of UVI31+ protein as determined by NMR (PDB code 2ma0) (a) and *SUPALM* alignments of these conformations onto the reference model (b). The reference NMR model is shown with a bold blue $C\alpha$ trace in both panels.

good similarity for NCC is around 0.7–0.8 at the resolution level used ($L = 5$ and seven Shannon channels).

7. Discussion and conclusions

An algorithm for rapid automated superposition of high- and low-resolution models was proposed and tested on a number of macromolecular structures to demonstrate the reliability of the method. Contrary to *SUPCOMB* (Kozin & Svergun, 2001) where the CPU time is proportional to the product of the number of points in the two objects, the performance of the reciprocal-space superposition depends solely on the maximum number of spherical harmonics L and on the angular data range $[0, s_m]$ used to calculate the correlation coefficient. The default values of these parameters ($L = 5$ and seven Shannon channels in the range of integration) ensure a good quality of the superposition and at the same time provide a significant gain in the computing time against *SUPCOMB*. The gain is especially noticeable for large macromolecular

structures where the Fourier space positioning becomes over ten times faster than in *SUPCOMB*.

The use of the normalized correlation coefficient based on the spherical harmonics representation employs a similar principle to the crystallographic fast rotation function method (Crowther, 1972), but contrary to the rotation function, we utilize only low-order harmonics at small angles to find the best overlap. Further, the number of harmonics in *SUPALM* can be defined by the user and it is not bound to a power of two as in the packages utilizing the fast Fourier transform technique (like *SITUS*, *FRM* or *ADP_EM*). We stress that the proposed method is not meant as an alternative to the mentioned EM packages, rather that the main aim of *SUPALM* is a rapid and convenient object matching at a level of the overall shape (e.g. using low-resolution models obtained from SAXS). Still, *SUPALM* permits the direct input and overlap of electron-density maps (EMD files) and multi-phase *ab initio* bead models (*MONSA*) and this significantly extends the applicability of the superposition method compared to *SUPCOMB*. *SUPALM* eliminates the need for intermediate steps (e.g. the transformation of the EMD file into a PDB model) and increases the accuracy of the superposition of heterogeneous structures (as the electron densities of each coordinate of the object are taken into account during the calculation of the scattering amplitudes, instead of using the simple approximation of a homogeneous particle). This option should facilitate a faster and more accurate comparison of SAXS models with those from complementary methods (such as EM, SANS, NMR and X-ray crystallography).

The proposed method reliably operates with default parameters, does not require user input and is therefore applicable in automated pipelines. The procedure, implemented in a program module *SUPALM* included in the *ATSAS* package (<http://www.embl-hamburg.de/biosaxs/software.html>), is freely available to academic users together with other programs from the *ATSAS* 2.7 release.

Acknowledgements

The authors acknowledge the support of the Bundesministerium für Bildung und Forschung (BMBF) project BIOSCAT, grant No. 05K20912, and of the European Commission, FP7 Infrastructures program grant BioStruct-X, project No. 283570.

References

- Chacón, P. & Wriggers, W. (2002). *J. Mol. Biol.* **317**, 375–384.
- Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon and Breach.
- Dennis, J., Gay, D. & Walsh, R. E. (1981). *ACM Trans. Math. Softw.* **7**, 348–368.
- Edmonds, A. R. (1957). *Angular Momentum in Quantum Mechanics*. Princeton University Press.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126–2132.
- Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.
- Garzón, J. I., Kovacs, J., Abagyan, R. & Chacón, P. (2007). *Bioinformatics*, **23**, 427–433.
- Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). *J. Mol. Biol.* **308**, 1033–1044.

- Kovacs, J. A., Chacón, P., Cong, Y., Metwally, E. & Wriggers, W. (2003). *Acta Cryst. D* **59**, 1371–1376.
- Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 33–41.
- Lambright, D. G., Sondek, J., Bohm, A., Skiba, N. P., Hamm, H. E. & Sigler, P. B. (1996). *Nature*, **379**, 311–319.
- Marino, M., Zou, P., Svergun, D., Garcia, P., Edlich, C., Simon, B., Wilmanns, M., Muhle-Goll, C. & Mayans, O. (2006). *Structure*, **14**, 1437–1447.
- Mitra, K., Schaffitzel, C., Fabiola, F., Chapman, M. S., Ban, N. & Frank, J. (2006). *Mol. Cell*, **22**, 533–543.
- Parseval des Chênes, M.-A. (1799). *Mémoire sur les séries et sur l'intégration complète d'une équation aux différences partielles linéaire du second ordre, à coefficients constants*. Paris.
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
- Schubert, W.-D., Urbanke, C., Ziehm, T., Beier, V., Machner, M. P., Domann, E., Wehland, J., Chakraborty, T. & Heinz, D. W. (2002). *Cell*, **111**, 825.
- Stuhrmann, H. B. (1970). *Acta Cryst.* **A26**, 297–306.
- Suhre, K., Navaza, J. & Sanejouand, Y.-H. (2006). *Acta Cryst.* **D62**, 1098–1100.
- Sun, L., Zhao, L., Yang, G., Yan, C., Zhou, R., Zhou, X., Xie, T., Zhao, Y., Wu, S., Li, X. & Shi, Y. (2015). *Proc. Natl Acad. Sci. USA*, **112**, 6003–6008.
- Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
- Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Svergun, D. I., Burkhardt, N., Pedersen, J. S., Koch, M. H. J., Volkov, V. V., Kozin, M. B., Meerwink, W., Stuhrmann, H. B., Diedrich, G. & Nierhaus, K. H. (1997). *J. Mol. Biol.* **271**, 602–618.
- Svergun, D. I., Koch, M. H. J., Timmins, P. A. & May, R. P. (2013). *Small Angle X-ray and Neutron Scattering From Solutions of Biological Macromolecules*. Oxford University Press.
- Svergun, D. I. & Nierhaus, K. H. (2000). *J. Biol. Chem.* **275**, 14432–14439.
- Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
- Svergun, D. I., Volkov, V. V., Kozin, M. B., Stuhrmann, H. B., Barberato, C. & Koch, M. H. J. (1997). *J. Appl. Cryst.* **30**, 798–802.
- Volkman, N. & Hanein, D. (2003). *Methods Enzymol.* **374**, 204–225.
- Zhang, X., Meining, W., Fischer, M., Bacher, A. & Ladenstein, R. (2001). *J. Mol. Biol.* **306**, 1099–1114.