Journal of
Applied
Crystallography
ISSN 0021-8898

Received 7 March 2005
Accepted 16 August 2005

A global multi-technique approach to study low-resolution solution structures

Marcelo Nöllmann,[a,b]* W. Marshall Stark[a] and Olwyn Byron[b]

[a]Division of Molecular Genetics, IBLS, University of Glasgow, Glasgow G12 8QQ, Scotland, and [b]Division of Infection and Immunity, IBLS, University of Glasgow, Glasgow G12 8QQ, Scotland. Correspondence e-mail: marcnol@berkeley.edu

Finding the conformation of large macromolecular complexes has become an important problem in structural biology, which is not always soluble by high-resolution techniques such as X-ray crystallography and NMR spectroscopy. Solution biophysical properties can provide direct or indirect structural information on these large complexes. A general systematic approach to the construction of a structural model of the macromolecule consistent with all the experimental solution properties is currently lacking. In this paper, such an approach is presented, where generalized rigid-body modelling is combined with a Monte Carlo/simulated-annealing optimization method, to search over a large range of possible conformations for the structure that best fits solution experimental properties derived from small-angle scattering, fluorescence resonance energy transfer and analytical ultracentrifugation.

1. Introduction

Finding the conformation of large multi-protein complexes, DNA–protein complexes and multi-domain proteins is of fundamental importance in understanding a large number of biological problems. High-resolution techniques, such as X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR), have produced structures for many of these macromolecular complexes. In some cases, however, these techniques fall short. Crystals of large macromolecular complexes are difficult to obtain, and it may be possible to determine structures only for segments of the complexes. Crystal packing effects, especially in large systems, might substantially affect the architecture of the complex with respect to that in solution (DePristo *et al.*, 2004). Crystals of a protein in all its relevant conformations, particularly when one or more states in the functional pathway exhibit flexibility, are very difficult to obtain. Similarly, NMR-derived distance information for large complexes, specifically distances between subunits, is demanding to generate and consequently scarce. In addition, flexible or disordered regions may appear to be absent from the final structure. Such regions may involve receptor binding motifs, loops involved in the active site or antigenic epitopes, and might be key to understanding the biological function of the macromolecule or complex.

Solution techniques, such as fluorescence resonance energy transfer (FRET) (Stühmeier, Clegg *et al.*, 2000), small-angle X-ray and neutron scattering (SAXS/SANS) (Feigin & Svergun, 1987) and analytical ultracentrifugation (AUC) sedimentation velocity (SV) (Lebowitz *et al.*, 2002) provide independent direct or indirect structural information on particulate systems in solution. These techniques are well established and have been successful in retrieving structural information on large macromolecular complexes. The structure of the ribosome was studied using small-angle scattering (SAS) (Svergun & Nierhaus, 2000), and more recently using cryo-electron microscopy (c-EM) (Gilbert *et al.*, 2004). Reconstruction of biologically significant DNA structures, such as Holliday junctions and DNA bulges, has been achieved from FRET (Lilley & Clegg, 1993) and SAXS (Nöllmann, Stark & Byron, 2004) data. Hydrodynamic methods and computational modelling have been employed to study the low-resolution conformations of human IgG subclasses by investigating the relative spatial orientations of their Fab′ and Fc domains (Carrasco *et al.*, 2001).

Ab initio methods can be used to deduce the low-resolution conformation of a particle (Stuhrmann, 1970; Chacón *et al.*, 1998; Svergun, 1999; Svergun *et al.*, 2001; Walther *et al.*, 2000; Heller *et al.*, 2003) from the SAS profile. Recently, a computational method was developed to add missing loops and domains to protein models (Petoukhov *et al.*, 2002). These restoration methods have been successfully applied to a large variety of problems (Koch *et al.*, 2003). When an *ab initio* reconstruction algorithm is run several times using the same initial conditions, the outcome is, at best, a family of reconstructed models with minor/moderate conformational differences. The reconstructions can be overlapped and averaged, and the averaged model can be used as a seed for a new reconstruction (Svergun *et al.*, 2001; Kozin & Svergun, 2001). In another approach, the family of reconstructed models can be used to find a consensus model that captures the essential features of the individual models (Heller *et al.*, 2003). Alternatively, SV data have been employed in order to filter out *ab initio* reconstructed models that fit the SAXS data but fail to

doi:10.1107/S0021889805026191

agree with the experimental sedimentation coefficient (Nöllmann, Stark & Byron, 2004) or to gain a greater level of confidence in the retrieved *ab initio* models (Scott *et al.*, 2002; Ackerman *et al.*, 2003; Solovyova *et al.*, 2004). Nonetheless, when applied to macromolecules with anisometric or hollow shapes, *ab initio* reconstruction methods produce a variety of considerably different reconstructions that fit the SAS data equally well (false positive reconstructions) (Heller *et al.*, 2003; Volkov & Svergun, 2003; Rosenzweig *et al.*, 1993).

Provided that the high- or low-resolution structures of the subunits comprising the quaternary complex are known or can be modelled, their arrangement can be found by searching for the quaternary conformations that best fit the SAS experimental data. Variations of this rigid-body modelling approach have been successfully applied to model SAS data (Aslam *et al.*, 2003; Feil *et al.*, 2001; Konarev *et al.*, 2001; Vigil *et al.*, 2001). The use of available structural information has a major advantage, with respect to *ab initio* reconstruction methods, in that the number of false positive reconstructions can be greatly reduced. NMR residual dipolar coupling data have been combined with SAXS data to determine the relative positions of calmodulin (CaM) and trifluoperazine (TFP) in their complex (Mattinen *et al.*, 2002). The NMR data defined the relative rotational positioning of the domains within CaM, whilst the SAXS data defined the molecular envelope of the system. Docking algorithms were used to generate geometrically compatible quaternary structures of purine nucleoside phosphorylase based on available crystallographic information (de Azevedo *et al.*, 2003). The best structural models were selected by finding the highest correlation between modelled and experimental SAXS data.

Rigid-body refinement was employed to improve models for the *Azospirillum brasilense* glutamate synthase holoenzyme based on *ab initio* reconstructions (Petoukhov *et al.*, 2003), and to characterize the nonameric assembly of an Archaeal α-*l*-fucosidase (Rosano *et al.*, 2004). SAXS and SANS were used by Aslam and co-workers (Aslam & Perkins, 2001; Aslam *et al.*, 2003) to study the solution structure of Factor H, composed of 20 short consensus repeat domains, for which there are high-resolution data available. Molecular dynamics (MD) simulations were employed to predict the possible conformations of the linkers in solution, and these were subsequently used to produce models for the whole particle. A large number of structures were chosen by fitting, in the Guinier region, the simulated scattering profiles of the models to the experimental data, and discarding structures whose predicted sedimentation coefficients were considerably different from that determined experimentally. This procedure, however, produced conformations for the whole macromolecule that were strongly restricted by the structures of the linkers. The computer program *MASSHA* (Konarev *et al.*, 2001) implements an algorithm where quaternary structures can be graphically generated by the user from high-resolution models and their SAXS intensity profiles calculated. In the computer program *SASMODEL* (Vigil *et al.*, 2001) the conformation of the macromolecular complex is modelled by a chain of ellipsoids. In this case, the conformation that best fits the scattering data is found by using a Monte Carlo-based method that searches for the values of the Euler angles between the centres of subsequent ellipsoids.

In the method presented in this paper, a rigid-body approach is used to generate a large number of possible macromolecular conformations. Several solution properties are then predicted *in silico* for each macromolecular conformation and compared to the experimental data sets. A general systematic method that searches for the parameters producing the model whose solution properties best fit the available experimental data sets is described.

As a model system to validate this methodology, a series of DNA molecules was chosen comprising three double-stranded DNA (dsDNA) helical fragments (H_{16}, H_n and H_{17}, with n being either 9 or 14) with a single-stranded loop of five deoxyadenosines (A_5 bulge) between each dsDNA fragment, namely $H_{16}A_5H_9A_5H_{17}$ and $H_{16}A_5H_{14}A_5H_{17}$ (Fig. 1). DNA bulges may arise in natural DNA from recombination between imperfectly homologous DNA sequences or from errors in DNA replication. They play an important role in frame-shift mutagenesis (Stassinopoulos *et al.*, 1996) and in specific interactions with RNA-binding proteins (Weeks & Crothers, 1991). Amongst other techniques, FRET (Gohlke *et al.*, 1994) and NMR spectroscopy (Dornberger *et al.*, 1999) have established that a single A_5 bulge introduces a defined kink into the DNA helical axis of about $90 \pm 15^\circ$. FRET was employed on DNA structures containing two A_5 bulges ($H_{16}A_5H_nA_5H_{17}$, with $6 < n < 11$) (Stühmeier, Hillisch *et al.*, 2000), similar to those used in this study, and it was shown that the distance between DNA ends is shortest in the sample with n = 9, for which the dsDNA fragments H_{16} and H_{17} were proposed to be coplanar. In this paper, the new modelling method will be validated by restoring, from simulated SAXS, SV and FRET data sets, the modelled low-resolution structures of two bulged DNA samples with very distinct conformations.

The paper is organized as follows. §2 describes in some detail the computational algorithms and approximations employed. The results section (§4) is devoted to the evaluation of the algorithm in restoring low-resolution structures from simulated data sets.

2. Computational methodology

The macromolecule or macromolecular complex is firstly divided into a number of structural domains. These domains are then combined by using a small number of parameters, such as interdomain distances and angles, in order to generate a large number of macromolecular conformations. The solution properties, such as SAS profiles, FRET distances and sedimentation coefficient, are then predicted *in silico* for each macromolecular conformation and compared with the experimental data sets. A function that measures the discrepancy between simulated and experimental data sets is calculated for each set of values of the parameters defining each macromolecular conformation. A general systematic method that searches for the parameters producing the model whose solution properties best globally fit the available experimental data sets is used.

2.1. Construction of the generalized rigid-body model

A general method for generating macromolecular conformations from the structures of individual domains, applicable to a large number of problems, was devised. The conformation of the macromolecule is generated from the structural data available, experimentally derived or computationally modelled atomic or low-resolution structures, for each of the N_D individual domains comprising the macromolecule. The possible conformations that the macromolecule can take are constrained by the definition of the movements of each domain. The allowed movements of each domain have two components: rotations with respect to its centre of mass (CM), and translations and rotations of its CM.

The transformation from any given Cartesian coordinate system to another can be carried out by three successive rotations in a specific sequence, defined by the Euler angles, the choice of which is, within limits, arbitrary. The convention employed here is that used in celestial and applied mechanics, and molecular and solid-state physics (Goldstein, 1980). The sequence of rotations begins with a rotation of the initial coordinate system xyz by an angle φ counterclockwise about the z axis. Secondly, a counterclockwise rotation, by an angle ϕ, about the intermediate x axis is performed. The transformation ends with a counterclockwise rotation by an angle θ about the z' axis, resulting in the new system of coordinates $x'y'z'$. The three Euler angles φ, ϕ and θ represent the three required generalized coordinates specifying the orientation of the $x'y'z'$ system relative to xyz.

The rotations of a domain i around its CM in a system of coordinates $x_i'y_i'z_i'$ are defined by using the Euler angles α_i, β_i and γ_i. The movement of the CM of domain i is specified in a similar manner. A new coordinate system $x_i''y_i''z_i''$ is defined by the rotations specified by a given set of Euler angles. The CM is then translated in the direction of y_i'' a distance r_i, and

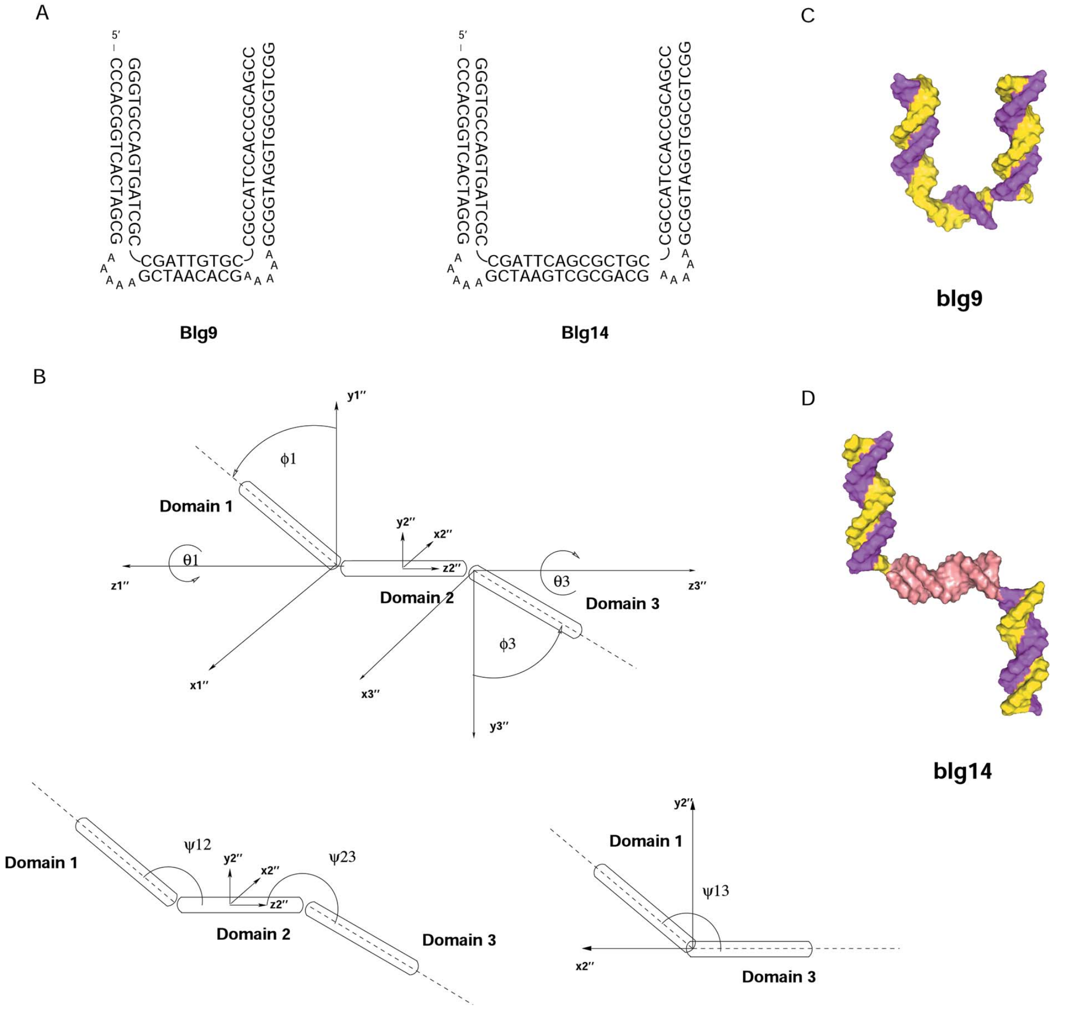

Figure 1
(*A*) Primary structure of the bulged DNA samples used in this study (blg9 and blg14). (*B*) Schematic representation of the parameters employed in the simulations. The helix axes are used as reference lines to define the angles between domains. (*C*), (*D*) Surface representations of the blg9 and blg14 structures generated *in silico*.

arbitrarily rotated in the new coordinate system by using the spherical angles ϕ_i and θ_i (rotations around x_i'' and z_i''). The assignment of the ranges of variation of φ_i^r, ϕ_i^r, θ_i^r, r_i, ϕ_i and θ_i complete the definition of the allowed movements for domain i.

This process is repeated for each domain $i = 1 \ldots N_{\rm D}$. Therefore, this set of $6 \times (N_{\rm D} - 1)$ parameters $\mathbf{r} \equiv (\alpha_1, \beta_1, \gamma_1, \ldots)$ defines a particular configuration of the macromolecule, for which a point in configurational space can be assigned.

2.2. Computation of SAXS/SANS intensities

SAXS intensities were calculated from the atomic coordinate files of the structural model of a particular configuration by using the computer program *CRYSOL* (Svergun *et al.*, 1995), which evaluates the solution scattering $[I(s)]$ of the given structure taking into account the scattering of the particle *in vacuo*, and additionally, the scattering from the excluded volume and the hydration layer around the particle. *CRYSOL* also fits the experimental scattering $[I_{\rm e}(s_i)]$ curve to the simulated one $[I(s)]$ using the average displaced solvent volume and the contrast of the hydration layer as variable parameters. The function $\chi_{\rm X}$ provides a measure of the discrepancy between simulated and experimental scattering curves, and is defined as

$$\chi_{\rm X}^2 = \frac{1}{N_{\rm p}} \sum_{i=1}^{N_{\rm p}} \left[\frac{I_{\rm e}(s_i) - cI(s_i)}{\delta(s_i)} \right]^2, \qquad (1)$$

where $N_{\rm p}$ is the number of experimental points in the scattering curve, $\delta(s_i)$ are the experimental errors and c is a scale factor (Svergun *et al.*, 1995). SANS intensities can be calculated in a similar manner, by employing the computer program *CRYSON* (Svergun *et al.*, 1998). In this case, a function $\chi_{\rm N}$ measuring the discrepancy between simulated and experimental neutron scattering curves can be defined as for $\chi_{\rm X}$.

2.3. Computation of FRET distances

In a FRET experiment, the positions of donor and acceptor molecules are usually not unique. At least one dye (or sometimes both) shows a distribution of positions with respect to its attachment point. Molecular dynamics (MD) simulations can be employed to find these distributions (Stühmeier, Clegg *et al.*, 2000), and to estimate from them an average dye position, referred to as the singular position. The experimentally determined FRET efficiency is employed to calculate the apparent singular distance between donor and acceptor, which, when combined with the MD simulations, can be used to estimate the distance between the attachment points of both dye molecules (referred to as d). For each given configuration of the macromolecule, d is calculated and compared with the experimental value. FRET distances correspond to a particular case of distance constraints, which could actually be derived from other sources, such as SAS or biochemical methods. Hereafter, we will refer throughout this paper to FRET distances, assuming that distance constraints from other methods could be employed similarly.

2.4. Computation of the sedimentation coefficient

The sedimentation coefficient of the macromolecule was calculated by using the computer program *HYDRO* (García de la Torre *et al.*, 1994). Firstly, a low-resolution bead model of the structure of each domain comprising the macromolecule was produced by using a modified version of the *AtoB* algorithm (Byron, 1997), implemented in the computer program *newAtoB* (available from the correspondence author upon request). The *AtoB* program (Byron, 1997) can be used to grid the structure of a macromolecule into a cubic lattice, thus producing a so-called bead model. *newAtoB* was specifically developed to reduce the resolution of atomic-resolution models when performing hydrodynamic calculations. In this implementation, a cubic lattice of grid spacing dx was generated, and the centre of mass (CM) of the structure was placed at the origin of coordinates. The coordinates of each atom t ($\mathbf{r}_t$) were used to determine which voxel (i, j, k) in the lattice it occupied. The resolution of the model was reduced by placing a single bead for each occupied voxel (i, j, k). The coordinates of the bead were calculated from the CM of the atoms that it replaced. Similarly, each bead radius was estimated so that its volume was identical to the total volume of the atoms in the voxel. In order to take into account hydration effects, the coordinates and radii of the beads can be modified using two approaches. One approach is to expand the structure by using the rule $\mathbf{r} = \mathbf{A} \cdot \mathbf{r}_t$, where $\mathbf{A}$ is the outward translation matrix, a matrix with real eigenvalues $\lambda_i > 1$, and to modify the radii of the beads in order to eliminate bead overlaps and voids in an asynchronous manner. In the case of isotropic expansions, $\mathbf{A}$ reduces to $\mathbf{A} = \beta\mathbf{I}$, where $\mathbf{I}$ is the identity matrix and β is the outward translation coefficient ($\beta > 1$). In the case of elongated shapes, such as rods or cylinders, the hydration can be modelled by a matrix $\mathbf{A}$ with nonequal eigenvalues representing an anisotropic expansion. An alternative approach is used by the program *SOMO* (Rai *et al.*, 2005), which is particularly well suited for the hydration of asymmetric biomacromolecules. The *Trans* algorithm within *SOMO* reduces the resolution of an atomic model by placing beads only at centres of main and side chains, and takes into account the hydration by differentially expanding the surface bead radii by an amount that depends on the bead charge. In the present study, all the molecules were isotropically expanded using the first approach.

The output of *newAtoB* is in both *BEAMS* (Spotorno *et al.*, 1997) and *HYDRO* (García de la Torre *et al.*, 1994) formats. In this study, only *HYDRO* was employed to calculate the hydrodynamic parameters of reduced models. The value of β was estimated as follows: (i) the sedimentation coefficient of the original domain structure was calculated by using the computer program *HYDROPRO* (García de la Torre *et al.*, 2000), which can be used to model macromolecular hydration; (ii) the computer program *HYDRO* was then used to calculate the sedimentation coefficient of the bead model produced with *newAtoB*; (iii) the value of β was then modified and the process from (ii) restarted, until the sedimentation coefficient calculated from the bead model equalled that of the original atomic-resolution structure as calculated with *HYDROPRO*. Only values of β that produced typical values for hydration (0.3–0.4 g water per g macromolecule) were accepted (Tanford, 1961; García de la Torre, 2001). When available, experimental sedimentation coefficients can be used instead of those obtained from *HYDROPRO*.

2.5. Estimation of steric clashes

Configurations causing steric clashes between the comprising domains were discarded. Domain volumes were calculated either from bead models or atomic-resolution structures by the routine *PDB2volume*, based on an algorithm similar to that used to estimate the Gaussian electron density in a cubic lattice and then estimate the macromolecular volume by adding the volumes of the occupied voxels (Lee & Richards, 1971; Gerstein, 1992). The maximum total volume ($V_{\max}$) of the macromolecule was estimated as the sum of the volumes of each domain. The volume of a particular configuration (V) was calculated using the same algorithm, and was accepted if $V \geq \gamma V_{\max}$, where γ is the overlap threshold coefficient, usually set at a value of 0.95.

2.6. Search for the best configuration

The Monte Carlo simulated-annealing method (MC/SA) (Kirkpatrick *et al.*, 1983) has been widely used in statistical physics (Landau & Binder, 2000) to locate a global minimum in a rugged landscape containing many local minima. The method can be summarized as follows: (i) a random point $\mathbf{r}_0$ in configurational space (space defined by the parameters) is chosen, and a quantity $E(\mathbf{r}_0)$ (energy or scoring function) is calculated from the values of the parameters; (ii) a random modification in one random parameter is introduced and the new energy $E(\mathbf{r}_1)$ is calculated; (iii) the change is accepted with a Boltzmann probability factor $\exp\{-[E(\mathbf{r}_1) - E(\mathbf{r}_0)]/T\}$, where T is the 'temperature'. The temperature T represents the 'thermal energy'. T has energy units, which do not correspond to real but rather to arbitrary units. Note that the change is always accepted if $E(\mathbf{r}_1) < E(\mathbf{r}_0)$, but it might still be accepted even if $E_1 > E_0$, depending on the value of the temperature T; (iv) the process is repeated by restarting from step (iii) for a large number of steps $N_{\rm steps}$, after which the temperature of the system is decreased by a factor α, so that $T_{\rm new} = \alpha T_{\rm old}$ (with $0 < \alpha < 1$), and the whole process is restarted from (i). The first configuration at this new temperature $T_{\rm new}$ is taken from the best configuration found at the previous temperature $T_{\rm old}$.

The computation is started at an initial temperature T_0 and stopped when the system reaches a predefined minimum temperature $T_{\rm F}$, where no further decrease in energy is registered. It is worth noting that T_0, $T_{\rm F}$ and α are related by the relation $\alpha^{N_a} = T_{\rm F}/T_0$, where $N_{\rm a}$ is the total number of temperature updates in a given computation.

The energy E is a function of all the parameters characteristic of a given configuration and decreases as the fit to the available experimental solution properties improves. The functions $\chi_{\rm X}$ and $\chi_{\rm N}$ [equation (1)] provide a measure of discrepancy between simulated and experimental SAXS/SANS curves. For the sedimentation coefficient $s^{\rm o}_{\rm w,20}$, an experimental variable with only one value, the measure of discrepancy between experimental ($s^{\rm E}_{\rm w,20}$) and simulated ($s^{\rm S}_{\rm w,20}$) values can be defined as

$$\chi_s = \left| \frac{s^{\rm E}_{\rm w,20} - s^{\rm S}_{\rm w,20}}{\delta_s} \right|, \tag{2}$$

where δ_s is the experimental error in $s^{\rm E}_{\rm w,20}$. An identical expression can be used for the definition of the measure of discrepancy between experimental and simulated FRET d values. These measures of discrepancy between experimental and simulated data can be simultaneously combined in a single expression defining the total energy, as follows:

$$E(\mathbf{r}) = \eta_{\rm X}\chi_{\rm X}(\mathbf{r}) + \eta_{\rm N}\chi_{\rm N}(\mathbf{r}) + \eta_{\rm F}\chi_d + \eta_{\rm SV}\chi_s, \tag{3}$$

where $\eta_{\rm X}$, $\eta_{\rm N}$, $\eta_{\rm F}$, $\eta_{\rm SV}$ are user-defined penalties for each technique (SAXS, SANS, FRET and SV, respectively), the determination of which is described below. In this way, the configuration that globally satisfies all the available experimental data sets can be found by minimizing the total energy $E(\mathbf{r})$ as a function of the parameters $\mathbf{r}$.

At high temperatures, the configurational space is effectively explored in the ranges available to each variable. Even if they do not produce acceptable fits to the experimental data, a large number of different configurations for the quaternary structure are investigated in this regime. At intermediate temperatures T, the system will still be able to climb energy barriers smaller than T, but will tend to be localized in regions of low energy. Only configurations with reasonable fits to the experimental data are allowed, but the system is still exploring all the accessible regions of low energy. At the lowest temperatures, only moves that reduce the energy are accepted, and so the system can only descend on the energy landscape. Provided that $T_{\rm F}$ is low enough so that no change in $E(\mathbf{r})$ is observed, and the annealing is sufficiently slow, the configuration with the best fit to the experimental data is found.

Ideally, each different technique should be equally important in determining the final configuration. In other words, each separate term contributing to the total energy in equation (3) should take, on average, similar numerical values when the algorithm explores the phase space at the highest temperature T_0. In practical terms, this implies that the η_i can be determined so that, for all i, $\eta_i \times \max(|\chi_i|) \simeq 1$, where max() is a function that calculates the maximum value taken by $|\chi_i|$ at temperature T_0. The values of the penalties η_i are chosen so that, when simultaneously employing several data sets, the fluctuations in the values of each variable (see below) have a similar temperature evolution. This thwarted the domination of a single data set in driving the convergence to a minimum that satisfied only itself, and not the other data sets. The information content of a scattering experiment is larger (10–15 times) than that of a FRET or a SV experiment. This was not taken into account when normalizing the penalties in the energy function, as it would have favoured one technique over the others to the point at which the search in configuration space would have not been a global search, but rather a search for the configuration that best fitted the scattering experiment. These normalizations avoided the domination of one technique over the others in approaching the global minimum, and thus optimized the global search procedure.

2.7. Parameter and configuration likelihood estimators

For any individual simulation i, the optimum value of any given parameter r_j is determined by its value $r_{j,i}(T_{\rm F})$ at the lowest temperature $T_{\rm F}$. A family of simulations is a set of N_r simulations performed under the same conditions but with random starting values for the $N_{\rm p}$ parameters. By generating a histogram of the final parameters r_j obtained in each family of simulations, it is possible to infer whether the reconstruction process produces a unique solution or there are different families of solutions that simultaneously fit all data sets. In the first case, the average value of each parameter r_j was calculated from the final values of each individual run as

$$\langle r_j^{\rm F} \rangle = \frac{1}{N_r} \sum_{i=1}^{N_r} r_{j,i}(T_{\rm F}). \tag{4}$$

In addition, for each family of simulations, the uncertainty σ_{r_j} in the value of a parameter r_j was calculated as

$$\sigma_{r_j} = \left\{ \frac{1}{N_r} \sum_{i=1}^{N_r} \left[r_{j,i}(T_{\rm F}) - \langle r_j^{\rm F} \rangle \right]^2 \right\}. \tag{5}$$

A family of simulations produces a set of N_r final configurations $\mathbf{r}_k^{\rm F}$ ($k = 1 \ldots N_r$). In order to measure the dispersion of a group of configurations around a given fixed configuration $\mathbf{R}$, a function F was defined as

$$F(\mathbf{r}^{\rm F}) = \frac{1}{N_r N_{\rm p}} \sum_{i=1}^{N_r} \sum_{j=1}^{N_{\rm p}} \left(r_{i,j}^{\rm F} - R_j \right)^2, \tag{6}$$

where $N_{\rm p}$ is the total number of parameters, $r_{i,j}^{\rm F}$ is the final value of parameter j in run i, and R_j is the value of parameter j in configuration $\mathbf{R}$. The lower the F value, the closer the N_r final configurations are to configuration $\mathbf{R}$. If the data sets used to run the simulations were generated from a given configuration $\mathbf{R}$, the success of a family of simulations in retrieving that configuration can be measured by calculating its F value.

In the case of a family of simulations producing different solutions that fit the data sets but that do not represent the same shape, the employment of an objective clustering technique, such as self-organizing neural networks (Kohonen, 1989), in order to sort the solutions into different clusters is needed. For each family, the approach suggested in equations (4)–(6) can be used to estimate the accuracy of the parameters and the F value. This same approach could also be employed to identify the characteristics of the families of solutions that often arise from a set of *ab initio* restorations, and its implementation is beyond the scope of the present study.

The mean value $[\langle r_{j,i}(T) \rangle]$ and standard deviation $[\sigma_{r_{j,i}}(T)]$ of the set of values taken by each parameter r_j in accepted configurations during a particular run i were calculated as a function of temperature. The temperature evolution of the standard deviation of the parameters was used to analyse the transition from configurational space exploration to localization (see below).

3. Availability and running times

The routines described in this paper were implemented in the computer program *rayuela*. All described computer programs were written in the C programming language and are available as precompiled binaries for Intel Linux from the correspondence author upon request. A run of the program *rayuela* on the presented examples, using simultaneously SV, FRET and SAXS data sets, with $N_r = 1500$ and $N_{\rm a} = 10$, on a 2 GHz Intel Linux PC, takes approximately 4 h. This running time increases for larger macromolecules, mainly as a result of the increased simulation time for calculating the hydrodynamic and scattering properties. This time can be reduced, however, by lowering the resolution of the *HYDRO* model (by increasing the grid size in *newAtoB*), and by decreasing the $s_{\rm max}$ for *CRYSOL/CRYSON* simulations. Reasonable running times can be thus assured even for large macromolecular complexes.

3.1. Construction of *in silico* models

For the construction of blg9 (see §4 and Fig. 1*C*), the helical axis of domain 2 was set parallel to z_2'', and its centre of mass (CM) placed at the origin of the $x_2''y_2''z_2''$ coordinate system (Fig. 1*B*). Initially, the helical axis of domain 1 was set parallel to z_1'', and its CM placed at the origin of coordinates of $x_2''y_2''z_2''$. Domain 1 was then rotated by an angle −90° in the x_2'' axis, and its CM translated to (0, 33 Å, −22 Å) in the $x_2''y_2''z_2''$ coordinate system. The helical axis of domain 3 was initially set parallel to z_1'', and its CM placed at the origin of coordinates of $x_2''y_2''z_2''$. Domain 3 was then rotated by an angle 90° in the x_2'' axis, and its CM translated to (0, 35 Å, 25 Å) in the $x_2''y_2''z_2''$ coordinate system.

For the construction of blg14 (see §4 and Fig. 1*D*), domain 2 and the initial positions and orientations of domains 1 and 3 were as for blg9. Domain 1 was placed at its final position by a rotation by an angle −90° in the x_2'' axis, and a translation of its CM to (0, 33 Å, −33 Å) in the $x_2''y_2''z_2''$ coordinate system. Domain 3 was placed at its final position by a rotation by an angle 90° in the x_2'' axis, and a translation of its CM to (0, 33 Å, 32 Å) in the $x_2''y_2''z_2''$ coordinate system.

4. Results

4.1. Simulated data sets

The methodology developed in this paper was tested on two bulged DNA structures, $H_{16}A_5H_9A_5H_{17}$ (referred to as blg9) and $H_{16}A_5H_{14}A_5H_{17}$ (also referred to as blg14) (Fig. 1*A*), where H_x refers to dsDNA with x base-pairs, A_5 is a single stranded loop of five nucleotides, and n is the number of base-pairs in the central dsDNA fragment between the bulges. The sequence of blg9 was identical to that used in a previous study

(Stühmeier, Hillisch *et al.*, 2000). The possible conformations of the bulges were modelled by combining three dsDNA fragments of appropriate sizes (16, 9 and 17 bp for blg9, and 16, 14 and 17 bp for blg14) as shown in Fig. 1. These fragments are hereafter referred to as domains 1, 2 and 3. The Euler angles ϕ and θ for domains 1 and 3 were employed to model all the possible conformations of the bulges, while domain 2 was kept fixed aligned with the z_1'' axis (see Fig. 1*B*). Additional variables were not necessary for modelling the possible conformations of these molecules. The *in silico* structures of blg9 and blg14 were generated by using the angles (ϕ_1, θ_1, ϕ_3, θ_3) = (0°, 0°, 0°, 0°) and (ϕ_1, θ_1, ϕ_3, θ_3) = (0°, 0°, 180°, 0°), respectively (see §2 and Figs. 1*C* and 1*D*). The use of these parameters was based on the NMR spectroscopy and FRET data available for DNA fragments containing one (Gohlke *et al.*, 1994; Dornberger *et al.*, 1999) or two (Stühmeier, Hillisch *et al.*, 2000) A_5 bulges. The SV, SAXS and FRET data were simulated from the structures of blg9 and blg14 DNA produced *in silico*.

Sedimentation coefficients were simulated from the *in silico* structures by using the computer program *HYDROPRO*. The simulated sedimentation coefficients for the blg9 and blg14 structures were 3.36 and 3.32 S (1 S = 10^{-13} s) (data sets SV9 and SV14, respectively).

The SAXS intensity profiles were similarly predicted from the *in silico* structures of blg9 and blg14 by using the computer program *CRYSOL* (data sets SAXS9 and SAXS14, Fig. 2*A*). Different levels of white noise (noise whose frequency spectrum is constant) were added to the original simulated curve for blg9 (SAXS9), in order to evaluate the robustness of the method (data sets $\mathrm{SAXS9}_{20}$ and $\mathrm{SAXS9}_{40}$, Fig. 2*B*).

Finally, FRET distances were calculated from *in silico* structures of blg9 and blg14 by measuring the distance between atom 480 in domain 1 and atom 558 in domain 3, both situated at the ends of the DNA fragments near the axis of the double helix. The average distances measured were 48 (5) Å for blg9 (data set FRET9) and 135 (10) Å for blg14 (data set FRET14). The average distance for blg14 (135 Å) is outside the present FRET experimental range of accessibility, but other means, such as SAS (*e.g.* D_{MAX}) or microscopical (atomic force or electron microscopy) methods could be used for its experimental determination. In the case of blg9, the proposed distance between the ends of domains 1 and 3 agrees with that reported in a previous study (53.7 Å) (Stühmeier, Hillisch *et al.*, 2000).

4.2. Method validation

The structures of both blg9 and blg14 were reconstructed by using different combinations of the data sets and the methodology described above. In all simulations, the ranges of variations of the angular parameters were as follows: $0 < \phi_1 < 180°$, $0 < \theta_1 < 360°$, $180 < \phi_3 < 360°$ and $0 < \theta_3 < 360°$. Different combinations of the parameters give rise, however, to effectively the same low-resolution structure [for instance (ϕ_1, θ_1, ϕ_3, θ_3) = (0°, 0°, 0°, 0°) and (0°, 90°, 0°, 90°)]. For this reason, in order to compare results from different runs, the four aforementioned parameters were reduced to only three parameters: $\psi_{12} = -\sin(\phi_1)$ representing the angle between domains 1 and 2, $\psi_{23} = -\sin(\phi_3)$ the angle between domains 2 and 3, and finally

$$\begin{aligned}\psi_{13} = a\cos[&\cos(\phi_1)\cos(\theta_1)\cos(\phi_3)\cos(\theta_3)\\ &+\cos(\phi_1)\sin(\theta_1)\cos(\phi_3)\sin(\theta_3)\\ &+\sin(\phi_1)\sin(\phi_3)],\end{aligned}$$

which is the angle between domains 1 and 3, in a plane containing their helical axes (Fig. 1*B*). Using this convention, blg9 is defined by the angles (ψ_{12}^{9}, ψ_{23}^{9}, ψ_{13}^{9}) = (90°, 90°, 0°), whereas blg14 is defined by the angles (ψ_{12}^{14}, ψ_{23}^{14}, ψ_{13}^{14}) = (90°, 90°, 180°) (see Figs. 1*C* and 1*D*). It is worth noting that two enantiomorphic conformations would have the same values of ψ_{12}, ψ_{23} and ψ_{13}.

Families of simulations using only one of the available data sets were performed. In addition, other families of simulations employing both SAXS and SV, SAXS and FRET, or SV, FRET and SAXS data were made. The settings utilized for the simulations are shown in Table 1. Comparison of these simu-

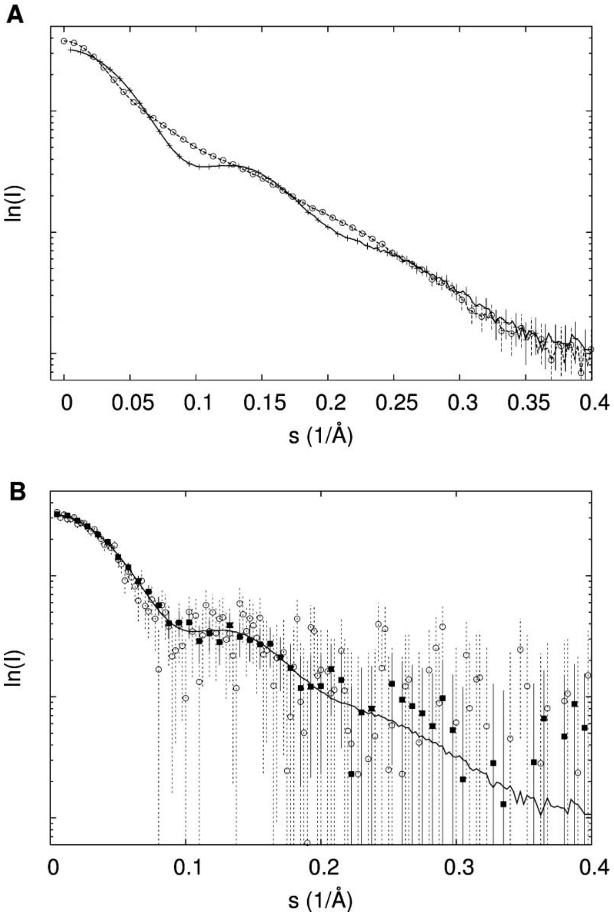

Figure 2
(*A*) Simulated SAXS data for blg9 (data set SAXS9, crosses, solid line) and blg14 (data set SAXS14, open circles, dashed line). (*B*) Simulated SAXS data for blg9 with 0 (data set SAXS9, solid line), 2000% (data set $\mathrm{SAXS9}_{20}$, filled boxes) and 4000% (data set $\mathrm{SAXS9}_{40}$, open circles) added noise. Solid/dashed vertical lines represent error bars.

Table 1
Settings employed for the different simulation families (see text for a full explanation of the column headings).

Identifier	Data set	N_r	T_0	$T_{\rm F}$	$\eta_{\rm SV}$	$\eta_{\rm F}$	$\eta_{\rm X}$
S9-H	SV9	12	2.0	0.004	1	0	0
S9-F	FRET9	10	2.0	0.004	0	1	0
S9-X	SAXS9	10	2.0	0.004	0	0	1
S9-X-20	$\rm SAXS9_{20}$	10	2.0	0.004	0	0	1
S9-X-40	$\rm SAXS_{40}$	10	2.0	0.004	0	0	1
S9-HX-0	SV9, SAXS9	10	6.0	0.0001	1	0	6
S9-HXF-0	SV9, SAXS9, FRET9	10	6.0	0.01	1	0.1	10
S9-HXF-20	SV9, $\rm SAXS9_{20}$, FRET9	10	2.0	0.004	1	1	6
S14-H	SV14	10	2.0	0.004	1	0	0
S14-F	FRET14	10	2.0	0.004	0	1	0
S14-X	SAXS14	10	2.0	0.004	0	0	1
S14-HXF	SV14, SAXS14, FRET14	10	6.0	0.004	1	0.1	10

lations was used to evaluate the advantages of using this multi-technique global modelling approach.

In all simulations, the annealing was performed by using at least ten temperature update cycles ($N_{\rm a}$ = 10). The corresponding temperature update factor α was calculated from T_0, $T_{\rm F}$ and $N_{\rm a}$ as described above (§2). At each temperature, $N_{\rm steps}$ = 1500 configurations were evaluated.

The rejection probability factor was defined as the number of rejected configurations over the total number of configurations evaluated at each temperature. As the annealing process evolves, and the temperature decreases, the rejection probability increased from 0, at the highest temperature, to 1 at the lowest (data not shown). This evolution reflects the fact that, as the temperature decreases, the chance of a configuration with high energy being accepted decreases. The evolution of the values taken by the parameters ϕ_1, θ_1, ϕ_3 and θ_3 also exhibited this trend.[1]

The values of the variables, such as $\chi_{\rm X}$, d and $s_{\rm w,20}$, also evolve with temperature. At high temperatures, the variables explore a wide range of values, limited only by the allowed configurations. For instance, at T = 2.0, the sedimentation coefficient $s_{\rm w,20}$ varies between 2.9 S and 3.35 S, and thus takes all accessible values given the ranges of variation of the different parameters defining a configuration (Fig. 3*B*). The fluctuations in the variables, and thus their standard deviations, diminish with decreasing temperatures, and take their final values once the minimum energy configuration has been found. This behaviour in the temperature evolution is observed for all variables (Figs. 3*A*–3*C*).

The effectiveness of each individual technique to retrieve the original parameters giving rise to the two bulged DNA structures, blg9 and blg14, was first evaluated. The F value (defined in §2, the **R** configuration being here the configuration defined by the parameters generating the *in silico* structures of blg9 and blg14) provides a measure of the effectiveness of a family of simulations in retrieving the original configuration. In all cases, the structure of blg9 was properly reconstructed (Table 2). Simulations employing only

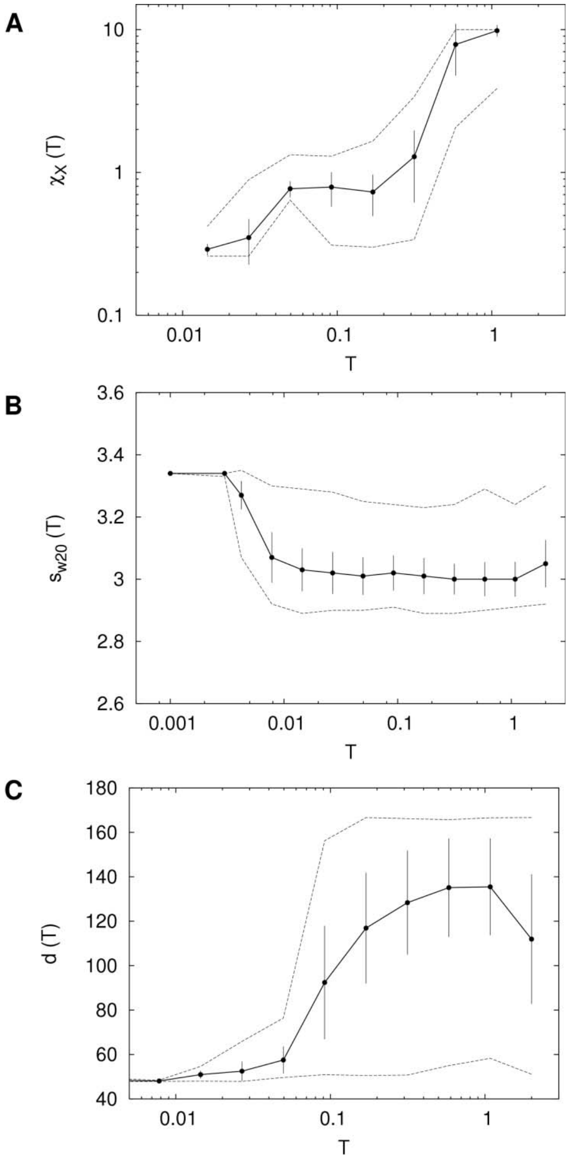

Figure 3
Values of (*A*) χ_X, (*B*) $s_{\rm w,20}$, and (*C*) d as a function of the annealing temperature for a single run in the (*A*) S9-X, (*B*) S9-H and (*C*) S9-F family of simulations. Vertical bars represent the standard deviation of the values taken by each variable at a fixed annealing temperature, whereas dashed lines indicate their limits of variation.

[1] Supplementary data are available from the IUCr electronic archives (Reference: HI5568). Services for accessing these data are described at the back of the journal.

Table 2
Final parameter statistics for S9 runs using the blg9 structure.

Identifier	ψ_{12}^F (°)	ψ_{23}^F (°)	ψ_{13}^F (°)	F
In silico	90	90	0	
S9-H	91.4 (18)	88.8 (18)	7.0 (7)	5.8
S9-F	90.2 (3)	89.7 (45)	4.3 (32)	2.7
S9-X	90.2 (14)	89.7 (23)	2.5 (19)	1.85
S9-HX	90.1 (1)	89.8 (15)	1.4 (8)	0.93
S9-HXF	90.09 (9)	89.8 (15)	0.89 (5)	0.63

Table 3
Final parameter statistics for simulations of blg9 with different noise levels.

Identifier	ψ_{12}^F (°)	ψ_{23}^F (°)	ψ_{13}^F (°)	F
In silico	90	90	0	
S9-X	90.2 (14)	89.7 (23)	2.5 (19)	1.85
S9-X-20	90.9 (8)	89.4 (6)	5.5 (39)	3.76
S9-X-40	91.2 (13)	89.1 (15)	7.9 (42)	5.3
S9-HXF-20	90.3 (2)	89.9 (14)	2.6 (16)	1.74

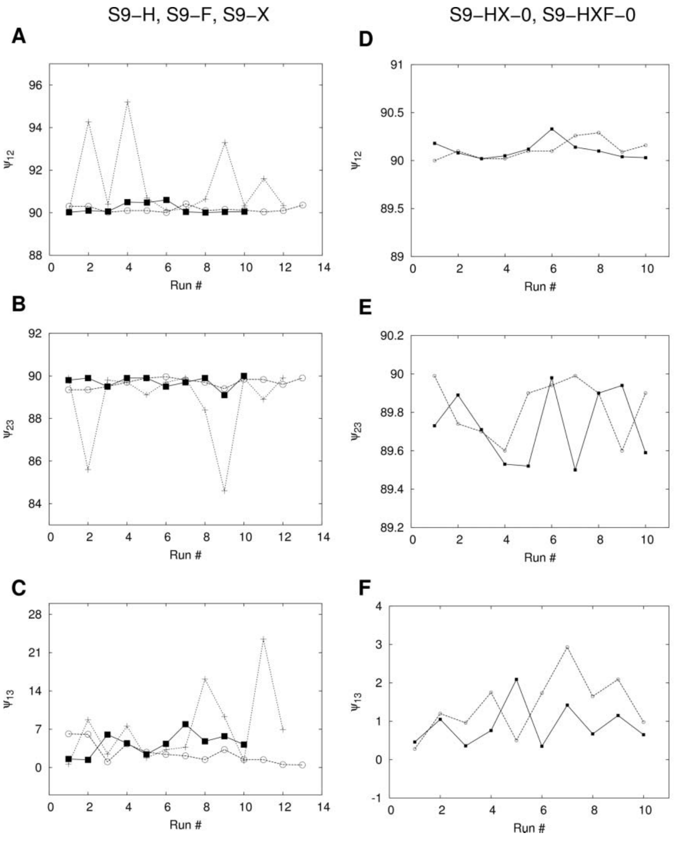

Figure 4
Final values of the parameters in different runs of simulations for blg9. (*A*)–(*C*) Angles ψ_{12}, ψ_{23} andψ_{13} for S9-H (dotted line, crosses), S9-F (solid line, filled boxes) and S9-X (dashed line, open circles). (*D*)–(*F*) Angles ψ_{12}, ψ_{23} and ψ_{13} for S9-HX-0 (dotted line, crosses), S9-HXF-0 (solid line, filled boxes). Note that the *y*-axis ranges in (*D*)–(*F*) are smaller than those in (*A*)–(*C*).

the SV9 data set (S9-H) were sufficient to obtain structures that resembled, at low-resolution, the structure of blg9 generated *in silico* (Figs. 4*A*–4*C*, dotted lines, and S9-H in Table 2). This occurred also when only the FRET9 or the SAXS9 data sets (S9-F, S9-X) were employed separately. In the case of the FRET9 data set, the final parameters also agreed with the parameters of blg9, but with a lower *F* value and lower uncertainties (Figs. 4*A*–4*C*, solid lines, and S9-F in Table 2). Finally, the final parameters obtained by using the SAXS9 data set had the lowest *F* value and uncertainties (Figs. 4*A*–4*C*, dashed lines, and S9-X in Table 2).

The combination of two techniques not only reduced the uncertainties of the final parameters but also decreased the *F* values (Figs. 4*D*–4*F*, dashed and dotted lines, and S9-HX in Table 2). This trend was even more pronounced when three data sets, SV9, SAXS9 and FRET9, were employed in the reconstruction process (S9-HXF in Table 2, and Figs. 4*D*–4*F*, solid lines, and Fig. 6*A*). With combination of techniques, the configuration with minimum energy was found in a smaller number of annealing cycles (data not shown). The user-defined penalties (Table 1) were chosen in order to optimize the search process by avoiding the domination of any single technique in the process leading to finding the global minimum in parameter space.

The robustness of the reconstruction method in retrieving the original parameters was evaluated by applying different levels of noise to the simulated scattering data. As expected, not only the difference between retrieved and original parameters (resulting in higher *F* values) but also the parameter uncertainties augmented with increasing noise levels and the inclusion of more data sets improved the reconstruction process (see Table 3). Even with large noise levels, the reconstruction process was successful in retrieving the *in silico* structures of blg9 and blg14 (see superpositions of *in silico* and reconstructed structures in Figs. 6*B* and 6*C*). This demonstrates the robustness of the method with respect to the introduction of high levels of noise in the data sets.

Several families of simulations were performed by using different combinations of the blg14 simulated and data sets. The reconstructions performed using only the FRET14 data set could not retrieve the original parameters (S14-F, Table 4, Fig. 5, dashed lines). A similar failure was observed when the SV14 data set was used alone (S14-H, Table 4, Fig. 5, dotted lines). These failures are reflected in large *F* values. Without any other constraint, there were far too many conformations for blg14 having an end-to-end distance of 135 Å, or a sedimentation coefficient of 3.32 S, resulting in different runs producing very different final values for the parameters. This represented a typical case where the energy space defined by the data sets is degenerate (contains several minima), resulting in many conformations that fit the data sets equally well.

On the contrary, when using only the SAXS14 data set, the retrieved parameters converged towards the original ones. This is reflected by a dramatic reduction in the *F* value (S14-X, Table 4, Fig. 5). When simultaneously using the SAXS14, FRET14 and SV14 data sets, there was only a marginal improvement in the *F* value and in the parameter uncertainties

Table 4
Final parameter statistics for simulations using blg14.

Identifier	ψ^F_{12} (°)	ψ^F_{23} (°)	ψ^F_{13} (°)	F
In silico	90	90	180	
S14-H	70 (40)	96 (6)	81 (7)	66
S14-F	123 (29)	56 (23)	85 (9)	65
S14-X	91.7 (6)	88.8 (7)	158.9 (13)	12.4
S14-HXF	91.6 (7)	88.5 (13)	159.7 (27)	11

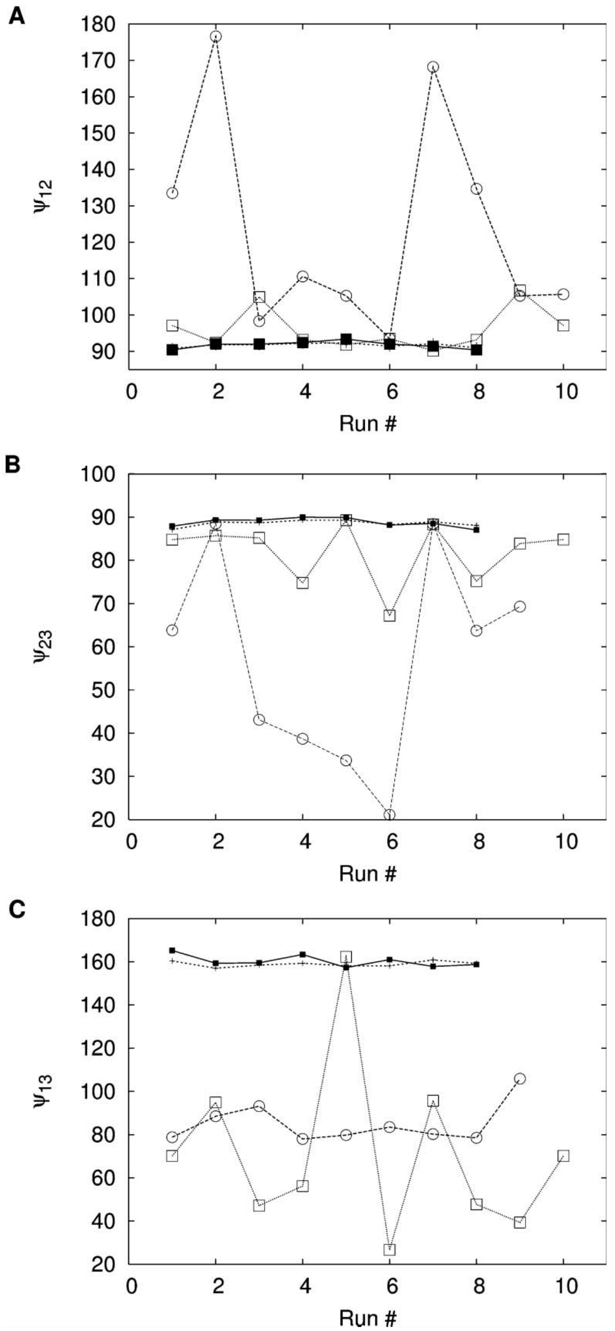

Figure 5
Final results for simulations of blg14. (*A*)–(*C*) Angles ψ_{12}, ψ_{23} and ψ_{13} for S14-F (dashed line, open circles), S14-H (dotted line, open boxes), S14-X (solid line, filled boxes), and S14-HXF (dashed line, crosses).

(S14-HXF, Table 4, Fig. 5). This, again, confirmed the existence of many conformations of blg14 that equally fit the FRET14 and SV14 data sets. Apart from a systematic deviation in the final value of ψ_{13} with respect to that in the *in silico* structure, the reconstruction process of blg14 was successful (Fig. 6*D*).

5. Discussion

In this paper, a general methodology that can be used to reconstruct the low-resolution solution structure of a macromolecular complex from several sources of experimental data has been presented. The macromolecule is first divided into domains for which structural data are available. The domains are combined by using a small number of parameters to produce a hypothetical conformation of the macromolecule. The algorithm reconstructs the low-resolution shape of the macromolecule by finding the relative positioning of each domain so that a number of solution properties (SAXS/SANS profiles, SV and FRET data) are simultaneously satisfied. The assembly of the structure of the macromolecule in terms of its domains has been implemented in a generalized manner, so that the methodology can be applied to a large variety of problems. The MC/SA algorithm employed to search for the best conformation is easily scalable to problems with large numbers of domains. The procedure was validated against two bulged DNA samples with very different overall conformations. The conformations of blg9 and blg14 were generated *in silico* from previous data (Gohlke *et al.*, 1994; Dornberger *et al.*, 1999; Stühmeier, Hillisch *et al.*, 2000) by using two sets of values for the parameters (see §3). SAXS, SV and FRET data were simulated from these *in silico* structures. The methodology was tested by using different combinations of the SAXS, FRET and SV data sets to restore the original *in silico* structures of the macromolecules.

The use of data sets simulated from *in silico* structures allowed a direct quantification of the effectiveness of the reconstruction process. The reconstructed models were directly compared with the *in silico* structures, and any discrepancy was unambiguously attributed to the reconstruction process rather than to other causes that ultimately lead to differences between experimental and simulated data sets, such as differences between solution and crystallography structures, systematic errors in the experimental data sets, or problems in the prediction of experimental properties.

In the majority of cases, each individual data set, when used separately, was able to restore the original parameters. The combination of more than one data set was shown to produce better restorations, in that the final restored parameters were more similar to those used for the *in silico* simulation of the original structures of the bulges (lower F values). The uncertainties in the values of the parameters in the final reconstructions were shown to diminish as more data sets were used for the restoration process. The introduction of noise in the original data sets produced similar restored parameters, but with higher F values and uncertainties. All in all, the method was able to restore the original *in silico* structures even

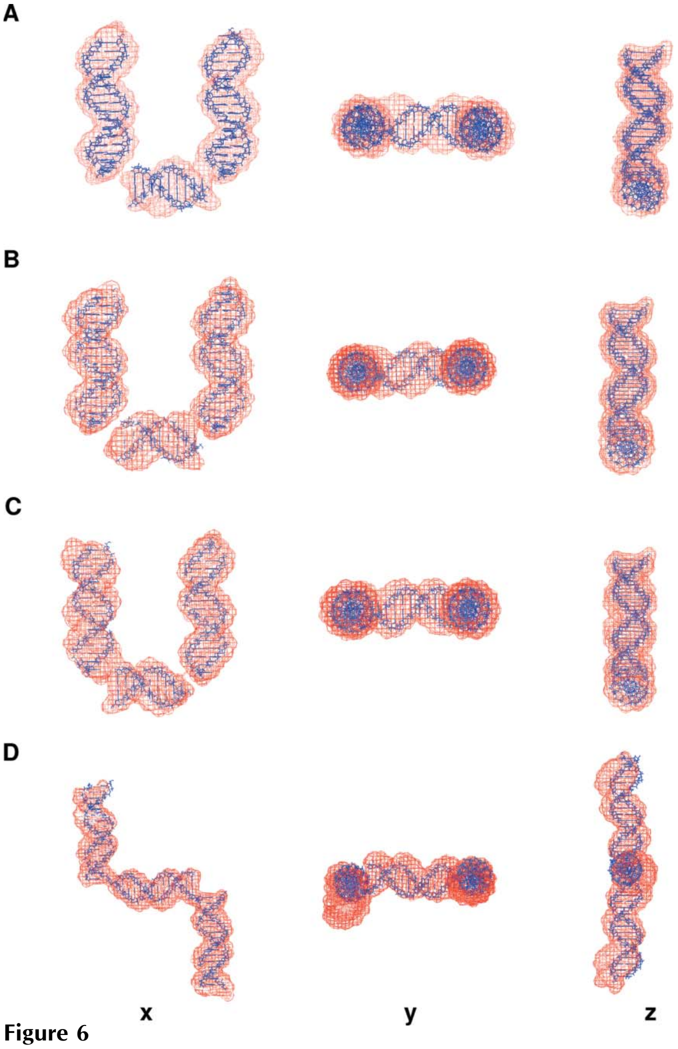

Figure 6
Final reconstructions of blg9 and blg14. Each panel contains the atomic-resolution *in silico* model of blg9 (panels *A*–*C*) or blg14 (panel *D*) as blue sticks and a reconstruction with the final average parameters from simulations (*A*) S9-HFX, (*B*) S9-HXF-20, (*C*) S9-X-40 and (*D*) S14-HXF (in red).

when a considerable amount of noise was introduced into the data sets.

In most, but not all cases, the use of multiple data sets considerably improved the restored parameters. The ability simultaneously to fit multiple data sets generally resulted in a reduction of the parameter uncertainties and a decrease in the F values (improved restored parameters). The implemented methodology provides a single integrated framework for finding solution conformations that would potentially fit any individual data set. The user-defined penalties in the energy function were chosen in order to thwart the domination of any single technique over the search process. This improved the result as it avoided energy minima that satisfied only one technique, and thus allowed the simulated annealing search procedure to find the minimum that simultaneously fitted all the given data sets. In cases where the available data sets, when fitted individually, predict several possible models for the conformation of the macromolecule, the global modelling approach implemented in this algorithm allows one to find the model(s) that simultaneously satisfies all the data sets. A single model that fits, at the same time, a range of data sets is more likely to represent the real conformation of the macromolecule in solution. In some circumstances, the values for the parameters predicted by different data sets might be contradictory. In such cases, it is very important to be able to fit the data sets individually and to manually compare the results provided by each data set.

In this study, a reduced number of parameters was employed to describe the possible conformations of the two test cases. However, the algorithm presented was designed so that it can be applied, without substantial changes, to larger macromolecular complexes where more variables are necessary in order to describe their possible conformations. In such cases, the computing time in the simulation of hydrodynamic parameters can be reduced by lowering the resolution of the *HYDRO* model, by increasing the grid size in *newAtoB*. Similarly, the scattering curve of a large macromolecule will decay more abruptly (increased R_g) than that of a smaller one, the important structural information needed to reconstruct its shape being shifted to smaller s values. In this case, the settings in *CRYSOL* can be modified by using a shorter $s_{\max}$. The computation of distance constraints would not be affected by the size of the macromolecule under study. These proposed changes would reduce the computation time per iteration and thus allow for a reconstruction to be carried out in a reasonable time, at the expense of the resolution of the final restored model. This reduction in resolution with macromolecular size also occurs with *ab initio* shape reconstruction methods.

The approach proposed in this manuscript has been successfully applied to the reconstruction of a protein–DNA complex and a protein macromolecular complex. The low-resolution structure of the complex formed by four Tn3 resolvase subunits and two DNA fragments was previously deduced from experimental small-angle neutron and X-ray scattering data (Nöllmann, He *et al.*, 2004). The same low-resolution structure was reconstructed using the algorithm proposed in this manuscript on the available SAXS/SANS experimental data set. Recently, the algorithm has also been used successfully to reconstruct the low-resolution conformation of proteins that are part of the human pyruvate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) and E3 binding protein (Smolle, Prior, Brown, Cooper, Byron & Lindsay, manuscript in preparation).

5.1. Comparison with other reconstruction algorithms

Other rigid-body modelling approaches have previously been implemented for reconstructing single SAXS data sets. *MASSHA* (Konarev *et al.*, 2001) allows for the manual construction of the low-resolution conformation of a macromolecular complex from the high-resolution atomic structures or low-resolution models comprising it. This user-constructed model can be automatically refined to improve the fit to the experimental scattering data by allowing small domain movements. This algorithm is thus optimal to refine a user-generated model in terms of its fit to the scattering data, but not to perform long-range conformational searches as is the objective of the program presented in this paper.

The computer program *SASMODEL* (Vigil *et al.*, 2001) models the macromolecular complex as a chain of ellipsoids, and is thus ideal for long-range exploration of the conformational space. This simplification, however, does not make use of all the structural data often available on the individual domains comprising the macromolecule, thereby producing a model of lower resolution. Both algorithms search for the macromolecular conformation that best fits a single SAXS data set. The algorithm shown here directly uses high-resolution structures to model the conformation of the macromolecule and presents the possibility of using multiple data sets, which improves the search process and increases the likelihood of the final reconstruction.

SAXS *ab initio* restoration methods are uniquely suited to producing low-resolution reconstructions of macromolecules of unknown structure. For simple shapes, different *ab initio* restorations usually differ only in minor details. In these cases, the reconstructed models can be superimposed and averaged in order to find the consensus reconstruction (Heller *et al.*, 2003; Kozin & Svergun, 2001). In some cases, however, the restoration process produces different reconstructions that fit the SAS data equally well (Heller *et al.*, 2003; Rosenzweig *et al.*, 1993; Volkov & Svergun, 2003; Walther *et al.*, 2000). Some of these models would in fact be false positives. It was recently shown that *ab initio* reconstruction methods are able to reconstruct shapes with small anisometries and, sometimes, with voids (Volkov & Svergun, 2003). Shapes with larger anisometries or smaller voids often cannot be reconstructed at all, even if the reconstruction process is stable (*i.e.* always reproduces a similar shape). For this situation, there is no method currently available that would systematically sort the reconstructed models into families of similar models, or that would provide a reliable measure of the probability of certain models/families of models being false positive reconstructions.

DAMMIN (Svergun, 1999), arguably the most recognized *ab initio* reconstruction method, was applied to the SAXS9, SAXS9$_{20}$, SAXS9$_{40}$ and SAXS14 data sets in order to test its ability to reconstruct the original *in silico* models. *DAMMIN* was able to reconstruct the blg9 model when using the noiseless SAXS9 data set (Fig. 7*A*). However, the reconstructed models differed greatly from the structure of blg9 when the noisy data sets were used (Figs. 7*B* and 7*C*), reflecting the inability of this *ab initio* reconstruction method to deal with data of considerable noise levels. The method presented in this paper was successful in reconstructing the topology of blg14 and its overall conformation, apart from a small (10%) systematic deviation in the final value of ψ_{13} with respect to that in the *in silico* structure. *DAMMIN* was unable, however, to reconstruct the conformation of blg14 (see Fig. 7*D*). This test case ultimately reflects the inherent limitations of solution techniques in dealing with the conformations of molecules of complex topology, but suggests that rigid-body reconstruction algorithms are better suited than *ab initio* methods in those cases.

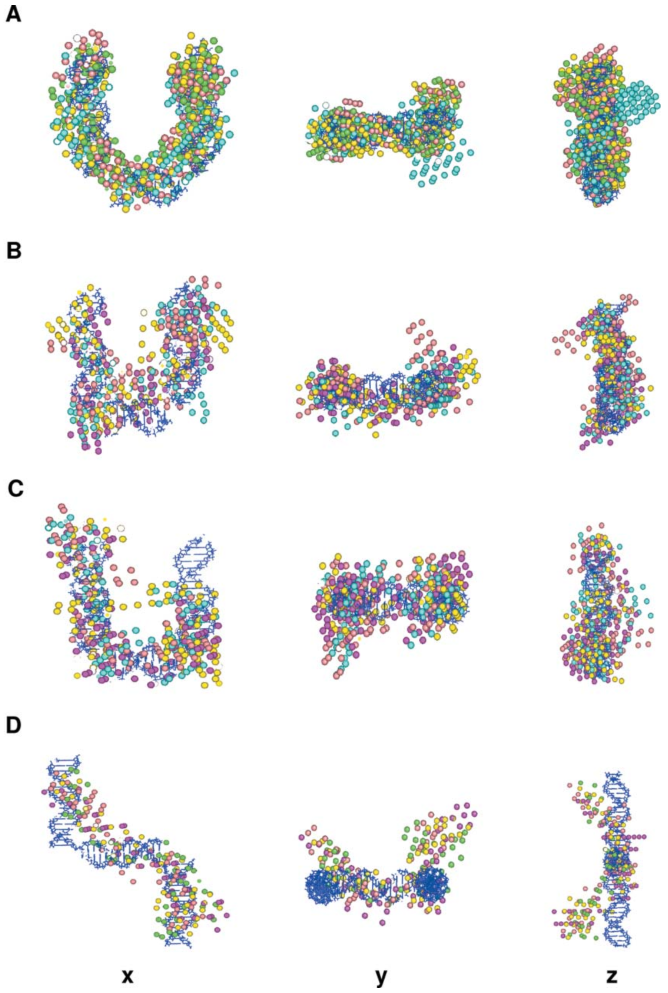

Figure 7
Three views of various superimposed *DAMMIN* reconstructions of blg9 and blg14. Each panel contains the atomic-resolution *in silico* model of blg9 (panels *A*–*C*, blue lines) or blg14 (panel *D*, blue lines) and four *DAMMIN* reconstructions (represented as yellow, cyan, orange and green beads) from the data sets (*A*) SAXS9, (*B*) SAXS9$_{20}$, (*C*) SAXS9$_{40}$ and (*D*) SAXS14 (see §4.1).

In a previous study (Nöllmann, Stark & Byron, 2004), it was shown that *ab initio* shape restoration methods were able to reconstruct the shape of a Holliday junction. However, in that study, the conformations and topologies of the macromolecules employed were very different to those used in the present study. In addition to these differences, in that study, the *ab initio* restorations were performed using symmetry constraints, which reduce the size of the search space and the number of false positives, thus considerably simplifying the reconstruction process. Based on the cases presented in this paper, and on other studies (Volkov & Svergun, 2003; Walther *et al.*, 2000), we expect similar problems with other *ab initio* approaches.

The methodology proposed in this paper is restricted to cases where previous structural data for the subunits comprising the macromolecule are available or can be modelled on the basis of available experimental data. However, it makes full use of these structural constraints and of several sources of solution properties to reduce the number of false positives in the reconstruction process, and increase its reliability. The method shown here is more suitable than *ab initio* reconstruction approaches for reconstructing macromolecular shapes of complex nature, whereas *ab initio* shape determination procedures are expected to produce more reliable results in situations where the solution conformation of the macromolecule is different from that in the crystal.

Ab initio reconstruction methods can only reconstruct particles with homogeneous electron density (atomic density for SANS) (Koch *et al.*, 2003). Accordingly, they cannot be used for the reconstruction of macromolecular complexes whose parts are inhomogeneous, such as protein–DNA or protein–polysaccharide complexes. A simplified version of the methodology presented here has recently been shown to produce excellent results when applied to the reconstruction of the solution conformation of a protein–DNA complex (Nöllmann, He *et al.*, 2004). The ability to reconstruct inhomogenous macromolecular complexes represents another advantage of this approach with respect to *ab initio* retrieval methods.

We thank Sean Colloms for a critical reading of this manuscript and useful suggestions throughout this project. The authors are grateful to the Wellcome Trust for financial support in the form of a studentship (to MN).

References

Ackerman, C. J., Harnett, M. M., Harnett, W., Kelly, S. M., Svergun, D. I. & Byron, O. (2003). *Biophys. J.* **84**, 489–500.

Aslam, M., Guthridge, J. M., Hack, B. K., Quigg, R. J., Holers, V. M. & Perkins, S. J. (2003). *J. Mol. Biol.* **329**, 525–550.

Aslam, M. & Perkins, S. (2001). *J. Mol. Biol.* **309**, 1117–1138.

Azevedo, W. F. de, dos Santos, G. C., dos Santos, D. M., Olivieri, J. R., Canduri, F., Silva, R. G., Basso, L. A., Renard, G., da Fonseca, I. O., Mendes, M. A. S. P. M. & Santos, D. S. (2003). *Biochem. Biophys. Res. Commun.* **309**, 923–928.

Byron, O. (1997). *Biophys. J.* **72**, 408–415.

Carrasco, B., García de la Torre, J., Davis, K. G., Jones, S., Athwal, D., Walters, C., Burton, D. R. & Harding, S. E. (2001). *Biophys. Chem.* **93**, 181–196.

Chacón, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.

DePristo, M. A., de Bakker, P. I. W. & Blundell, T. L. (2004). *Structure (Camb.)*, **12**, 831–838.

Dornberger, U., Hillisch, A., Gollmick, F. A., Fritzsche, H. & Diekmann, S. (1999). *Biochemistry*, **38**, 12860–12868.

Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering.* New York: Plenum Press.

Feil, I. K., Malfois, M., Hendle, J., van der Zandt, H. & Svergun, D. I. (2001). *J. Biol. Chem.* **276**, 12024–12029.

García de la Torre, J. (2001). *Biophys. Chem.* **93**, 159–170.

García de la Torre, J., Huertas, M. L. & Carrasco, B. (2000). *Biophys. J.* **78**, 719–730.

García de la Torre, J., Navarro, S., López Martínez, M. C., Diaz, F. G. & López Cascales, J. (1994). *Biophys. J.* **67**, 530–531.

Gerstein, M. (1992). *Acta Cryst.* A**48**, 271–276.

Gilbert, R. J. C., Fucini, P., Connell, S., Fuller, S. D., Nierhaus, K. H., Robinson, C. V., Dobson, C. M. & Stuart, D. I. (2004). *Mol. Cell*, **14**, 57–66.

Gohlke, C., Murchie, A. I., Lilley, D. M. & Clegg, R. M. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 11660–11664.

Goldstein, H. (1980). *Classical Mechanics.* New York: Addison-Wesley.

Heller, W. T., Krueger, J. K. & Trewhella, J. (2003). *Biochemistry*, **42**, 10579–10588.

Kirkpatrick, S., Gelatt, C. D. & Vecci, M. P. (1983). *Science*, **220**, 671–680.

Koch, M. H., Vachette, P. & Svergun, D. I. (2003). *Q. Rev. Biophys.* **36**, 147–227.

Kohonen, T. (1989). *Self-Organization and Associative Memory.* Berlin: Springer-Verlag.

Konarev, P., Petoukhov, M. & Svergun, D. (2001). *J. Appl. Cryst.* **34**, 527–532.

Kozin, M. & Svergun, D. (2001). *J. Appl. Cryst.* **34**, 33–41.

Landau, D. P. & Binder, K. (2000). *A Guide to Monte Carlo Simulations in Statistical Physics.* Cambridge University Press.

Lebowitz, J., Lewis, M. S. & Schuck, P. (2002). *Protein Sci.* **11**, 2067–2079.

Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.

Lilley, D. & Clegg, R. (1993). *Q. Rev. Biophys.* **26**, 131–175.

Mattinen, M.-L., Paakkonen, K., Ikonen, T., Craven, J., Drakenberg, T., Serimaa, R., Waltho, J. & Annila, A. (2002). *Biophys. J.* **83**, 1177–1183.

Nöllmann, M., He, J., Byron, O. & Stark, W. M. (2004). *Mol. Cell*, **16**, 127–137.

Nöllmann, M., Stark, W. M. & Byron, O. (2004). *Biophys. J.* **86**, 3060–3069.

Petoukhov, M. V., Eady, N. A. J., Brown, K. A. & Svergun, D. I. (2002). *Biophys. J.* **83**, 3113–3125.

Petoukhov, M. V., Svergun, D. I., Konarev, P. V., Ravasio, S., van den Heuvel, R. H. H., Curti, B. & Vanoni, M. A. (2003). *J. Biol. Chem.* **278**, 29933–29939.

Rai, N., Nöllmann, M., Spotorno, B., Tassara, G., Byron, O. & Rocco, M. (2005). *Structure*, **13**, 723–734.

Rosano, C., Zuccotti, S., Cobucci-Ponzano, B., Mazzone, M., Rossi, M., Moracci, M., Petoukhov, M. V., Svergun, D. I. & Bolognesi, M. (2004). *Biochem. Biophys. Res. Commun.* **320**, 176–182.

Rosenzweig, A. C., Frederick, C. A., Lippard, S. J. & Nordlund, P. (1993). *Nature (London)*, **366**, 537–543.

Scott, D. J., Grossmann, J. G., Tame, J. R. H., Byron, O., Wilson, K. S. & Otto, B. R. (2002). *J. Mol. Biol.* **315**, 1179–1187.

Solovyova, A. S., Nöllmann, M., Mitchell, T. J. & Byron, O. (2004). *Biophys J.* **87**, 540–552.

Spotorno, B., Piccinini, L., Tassara, G., Ruggiero, C., Nardini, M., Molina, F. & Rocco, M. (1997). *Eur. Biophys. J. Biophys. Lett.* **25**, 373–384.

Stassinopoulos, A., Ji, J., Gao, X. & Goldberg, I. H. (1996). *Science*, **272**, 1943–1946.

Stühmeier, F., Clegg, R., Hillisch, A. & Diekmann, S. (2000). *DNA–Protein Interactions*, edited by A. Travers & M. Buckle, ch. 6, pp. 77–94. Oxford University Press.

Stühmeier, F., Hillisch, A., Clegg, R. M. & Diekman, S. (2000). *J. Mol. Biol.* **302**, 1081–1100.

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

Svergun, D., Richard, S., Koch, M., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.

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

Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

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.

Tanford, C. (1961). *Physical Chemistry of Macromolecules.* New York: John Wiley.

Vigil, D., Gallagher, S. C., Trewhella, J. & Garcia, A. E. (2001). *Biophys. J.* **80**, 2082–2092.

Volkov, V. & Svergun, D. (2003). *J. Appl. Cryst.* **36**, 860–864.

Walther, D., Cohen, F. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.

Weeks, K. M. & Crothers, D. M. (1991). *Cell*, **66**, 577–588.