Journal of
Applied Crystallography
ISSN 0021-8898

Received 10 October 2000
Accepted 31 July 2001

Using *Situs* for the registration of protein structures with low-resolution bead models from X-ray solution scattering

Willy Wriggers* and Pablo Chacón

Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Rd, La Jolla, CA 9203, USA. Correspondence e-mail: wriggers@scripps.edu

Three-dimensional bead models of proteins in solution are routinely determined from one-dimensional small-angle X-ray scattering (SAXS) data. The *Situs* software provides a novel set of visualization and registration procedures to facilitate the localization of protein structures in low-resolution SAXS bead models. The docking algorithm takes advantage of a reduced representation of the input data sets by means of topology-representing neural networks to expedite the rigid-body search. The precision of the docking was tested on ten different simulated bead models: for >100 beads typically arising in SAXS models, a docking precision of the order of an ångström can be achieved. The shape-matching score captured the correct solutions in all ten trial cases and was sufficiently stringent to yield unique matches in seven systems. A size-invariant shape descriptor of 'sphericity' is proposed to assess the onset of ambiguity in the matching of globular molecules. The software, a tutorial and supplementary data are available at http://situs.scripps.edu/saxs.

1. Introduction

Shape reconstruction from one-dimensional small-angle X-ray scattering (SAXS) data is emerging as a powerful tool to characterize the gross structural features of biopolymers in solution. In particular, bead modeling is now routinely applied to refine three-dimensional shapes against SAXS data (Chacón *et al.*, 1998; Svergun, 1999; Walther *et al.*, 2000), and several recent experimental applications demonstrate the value of this methodology (Chacón *et al.*, 2000; Bada *et al.*, 2000; Svergun *et al.*, 2000). Typically, the resulting low-resolution models exhibit a small variability in the distribution and the number of beads, caused by noise and the intrinsic degeneracy of the inverse scattering problem. Hence, the superposition and comparison of bead models with high-resolution structures is a nontrivial problem (Kozin & Svergun, 2001). *Situs* takes advantage of a reduced representation of the three-dimensional data sets. By aligning gross features, the method is insensitive to local perturbations.

Vector quantization with topology-representing networks (Martinetz & Schulten, 1994) offers a flexible way to develop a discrete representation of three-dimensional data (Wriggers *et al.*, 1998). As described in detail elsewhere (Wriggers *et al.*, 1999), vector quantization places a number of so-called 'codebook vectors' into three-dimensional data at characteristic positions. The vectors form a set of point landmarks that are robust under changes in resolution, identify gross features, and thereby provide information about the shape and density distribution of the biological object. Codebook vectors are therefore suitable for the registration of corresponding features and have been used successfully in the fitting of atomic structures to low-resolution data from electron microscopy (Wriggers *et al.*, 1998, 1999). Here, we present routines that were specifically adapted for the registration of atomic structures with SAXS bead models, and routines for the visualization of the results. Also, for the first time, rigid-body registration with *Situs* has been tested on simulated low-resolution data. The results should be of interest in all areas of biophysics where multi-resolution three-dimensional data sets are aligned.

2. Design of the *Situs* package and examples of use

The software consists of modular stand-alone C programs for visualization, vector quantization, and docking of three-dimensional bead models and atomic structures in PDB format (Protein Data Bank, http://www.rcsb.org/pdb). Each program is self-explanatory as the user is asked to enter all relevant information at the shell prompt during interactive use on a Unix workstation. The series of steps and the programs that are required to dock an atomic resolution structure into low-resolution SAXS bead models are shown schematically in Fig. 1.

The bead data are first mapped to a three-dimensional lattice with a hard-sphere kernel-convolution of each bead center that transforms the PDB-formatted data into a three-dimensional density. The *pdblur* convolution program (Fig. 1) uses a radial hard-sphere density distribution of the functional form $\max\{0, A[1-\frac{1}{2}(r/R)^{60}]\}$. Input parameters include the kernel amplitude A and the half-max kernel radius R, which should correspond to the user-defined bead radius. The package contains a variety of additional routines (Wriggers & Birmanns, 2001) for the inspection and manipulation of volumetric data (*e.g.* cross sections of the resulting three-dimensional density data can be inspected with the visualization program *volslice*, and the voxel histogram of the density values is computed with the *histovox* program).

Two vector quantization routines are provided by the *Situs* package: *qpdb*, for the quantization of atomic resolution data, and *qvol*, for the quantization of bead-model densities. Let us assume that the data sets each are represented by k codebook vectors $\mathbf{x}_i$, corre-

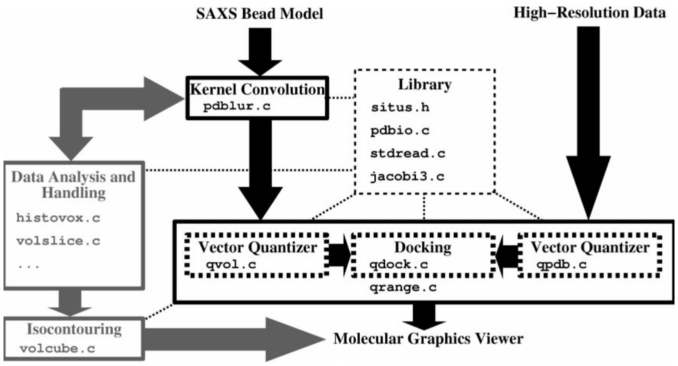

Figure 1
Schematic diagram of SAXS-related routines of the *Situs* package (version 1.4). Individual C program components are classified by their functionality. The main data flow is indicated using black arrows. The visualization routines (gray) for the rendering of the bead models (see also Fig. 2) are optional. The main procedures are discussed in the text. The programs are supported by a header file (`situs.h`), and by auxiliary library programs that handle input and output of atomic coordinates (`pdbio.c`), input of data at the command prompt (`stdread.c`) and eigenvector computation for real symmetric 3 × 3 matrices (`jacobi3.c`). Additional documentation is available at http://situs.scripps.edu/saxs.

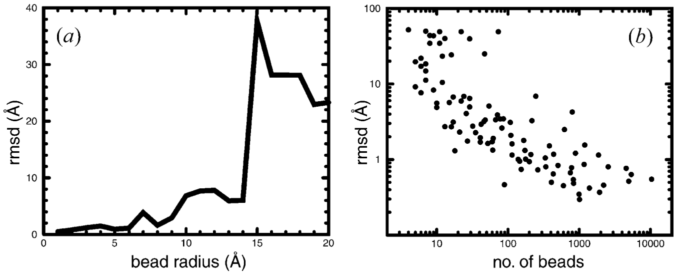

Figure 3
Docking precision tested on simulated bead models. (*a*) Nitrito-reductase (see Table 1): r.m.s. deviation of the *Situs*-docked structure from the initial structure as a function of bead radius. Supplementary data, available at http://situs.scripps.edu/saxs, show the size-dependent r.m.s. deviation for all models listed in Table 1. (*b*) Scatter plot of the ten trial systems (see Table 1): r.m.s. deviation of the docked structures from the initial structures as a function of the number of beads in the models. All validation runs were performed with *Situs* (version 1.3) utilities. The optimum vector number k (see text) was automatically selected by minimizing the k-dependent average of the vector r.m.s. deviation and statistical vector variability (Wriggers & Birmanns, 2001) for $3 \le k \le 9$.

sponding to high-resolution data, and by k codebook vectors $\mathbf{y}_j$, corresponding to low-resolution data ($i, j = 1, \ldots, k$). Furthermore, let the index map $I: j \to i$ define the k pairs of corresponding vectors. In practical situations, I is not known *a priori*, and all $k! = k(k-1)\ldots(3 \times 2)$ possible permutations $[I(1), \ldots, I(k)]$ have to be explored. The program *qdock* carries out an exhaustive search of the permutations and returns a list of best fits, ranked by the remaining r.m.s. (root mean square) deviation after least-squares fitting (Kabsch, 1976, 1978) of the vectors $\mathbf{x}_{I(j)}$ to the $\mathbf{y}_j$. If a user wishes to find the optimum number k of vectors automatically, the program *qrange* consolidates the functionality of the vector quantization and docking routines into a single program and carries out searches for a range of $k, 3 \le k \le 9$. A maximum of nine vectors is suitable for more complex shapes, although in most cases a small number ($\ge$3) is sufficient for the rigid-body docking. Fig. 2 presents examples of optimally superimposed data sets using codebook vectors.

The work of our collaborators (Chacón *et al.*, 2000) required the development of SAXS bead-model visualization routines. The rendering of the bead model by solid or transparent spheres (Fig. 2*a*) occludes much of the docked protein and becomes unpractical for a large number of beads. Therefore, software was developed that solely renders an outer contour surface of the bead model.

In *Situs*, the model surface may be rendered starting with a kernel-convolution of the original PDB-formatted bead data that transforms the pseudo-atomic model into a three-dimensional density. The *pdblur* program (Fig. 1) provides, in addition to the hard-sphere kernel, a choice of four 'softer' density kernels: Epanechnikov, $\max\{0, A[1-\frac{1}{2}(r/R)^2]\}$; 'semi-Epanechnikov', $\max\{0, A[1-\frac{1}{2}(r/R)^{3/2}]\}$; triangular, $\max\{0, A[1-\frac{1}{2}(r/R)]\}$; Gaussian, $A\exp[-\frac{3}{2}(r/\sigma)^2]$. Input parameters include the kernel amplitude A and the half-max kernel (bead) radius R, or the R-dependent standard deviation σ.

Using the volumetric density, the isocontour program *volcube* (Fig. 1) generates wireframe meshes or solid surfaces of bead-model contours that can be displayed with atomic structures using the free molecular graphics package *VMD* (Humphrey *et al.*, 1996), available at http://www.ks.uiuc.edu/Research/vmd. Input parameters of *volcube* include the rendering style (wireframe or solid) and the mesh size for the rendering of the contours (the input grid is automatically interpolated). An isocontour threshold of $A/2$ (A is the kernel amplitude chosen in *pdblur*) must be entered in *volcube* to draw a contour that is consistent with the desired bead radius.

Figs. 2(*b*) and (*c*) (see also Chacón *et al.*, 2000) show wireframe examples of *Situs* contour rendering of bead models on a hexagonal

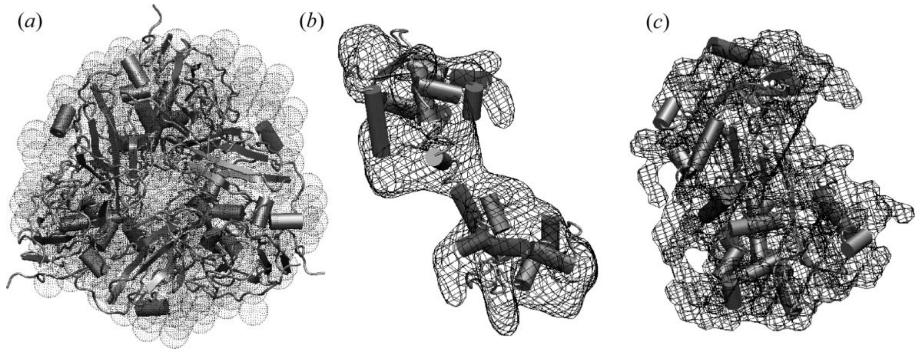

Figure 2
Docking and visualization of experimental SAXS bead models. The bead models were generated from SAXS data in the Andreu laboratory in Madrid, as described elsewhere (Chacón *et al.*, 2000). (*a*) Nitrito-reductase (PDB entry 2nrd). (*b*) Troponin C (1top). (*c*) Ovalbumin (1ova). The *Situs*-docked structures are shown in cartoon representation with *VMD* (Humphrey *et al.*, 1996). Three different *Situs* rendering styles are shown (see text): (*a*) h.c.p. lattice (dotted spheres); (*b*) triangular kernel smoothing (isocontour wireframe); (*c*), (*b*) Epanechnikov kernel smoothing (isocontour wireframe). Supplementary color figures, available at http://situs.scripps.edu/saxs, demonstrate the docking for all ten experimental bead models (Chacón *et al.*, 2000) corresponding to the proteins listed in Table 1.

Table 1
Characteristics of simulated SAXS bead models.

Protein used for generating bead model	PDB entry	Critical bead size (Å)†	Docking precision (Å)‡	Oligomeric symmetry	Degeneracy of best fit§	Sphericity ς
Catalase	7cat	16	0.8	4×	4×	0.57
β-4-Integrin	1qg3	4	0.8	1×	1×	0.16
Chymotrypsinogen A	2cga	5	2.4	1×	2×	0.66
Myoglobin	1mbn	8	0.8	1×	2×	0.56
Nitrito-reductase	2nrd	15	0.8	3×	6×	0.61
Ovalbumin	1ova	7	0.6	1×	1×	0.48
Spermadhesin	1spp	8	0.8	1×	1×	0.48
Superoxide dismutase	1xso	9	2.7	2×	2×	0.46
Troponin C	1top	9	0.7	1×	1×	0.18
$\alpha\beta$-Tubulin	1tub	11	1.8	1×	1×	0.48

† The value given is the smallest bead radius for which the r.m.s. deviation of the docked structure with respect to the target structure exceeded 10 Å. For the r.m.s. deviation evaluation, the fit with the lowest r.m.s. deviation among any degenerate fits was selected. ‡ The stated value is the r.m.s. deviation of the docked structure with respect to the target structure averaged for bead radii of 1, 2 and 3 Å. For the r.m.s. deviation evaluation, the fit with the lowest r.m.s. deviation among any degenerate fits was selected. § The degeneracy is the number of optimum fits (at sub-critical bead size) that were empirically found to cluster within a narrow numeric range of the optimum score, as a result of symmetry effects or ambiguity of matching.

close-packed (h.c.p.) lattice. Note how the choice of kernel affects the rendering. The triangular kernel (Fig. 2*b*) yields a smooth surface. In contrast, the Epanechnikov kernel (Fig. 2*c*) yields a segmented surface that closely follows the embedded beads.

3. Precision of the docking

The atomic structures of ten trial proteins (Table 1) were projected onto an h.c.p. lattice to measure the precision and reliability of *Situs*-based registration. A bead of radius q was placed on the lattice if it contained at least $0.1q^3$ Å^{-3} atoms (the threshold value was adjusted empirically to match closely the volume of the atomic structure). H.c.p. lattices with $q = 1, 2, \ldots, 20$ Å were created, unless the number of resulting beads was <3. For each system, the size-dependent r.m.s. deviation of the docked structure with respect to the initial structure was evaluated. Fig. 3(*a*) shows the r.m.s. deviation of nitrito-reductase as a function of radius. The docking error typically increases with bead size but remains smaller than the bead size up to a critical size value, at which the docking breaks down as the result of a catastrophic misalignment. Table 1 shows that this critical bead size is system-dependent because of the different sizes of the proteins studied. Fig. 3(*b*) shows the r.m.s. deviation values of all trial systems as a function of the number of beads. This number was found to be a more system-independent measure of the coarseness introduced by the bead representation in the cases studied here. The major result is that for >100 beads, typically arising in SAXS models, a docking precision of the order of an ångström can be achieved (see also Table 1).

The highest-scoring fits are degenerate in certain cases. This degeneracy (Table 1) closely follows the number of symmetry-related oligomeric subunits in the trial systems. Only in three cases the algorithm returned spurious fits that were not symmetry related (chymotrypsinogen A, myoglobin, nitrito-reductase). In these systems, each correct (or symmetry-related) high-scoring fit was paired with an additional incorrect fit that was indistinguishable by means of the codebook vector docking score. This ambiguity arises in cases of globular molecules that do not exhibit significant shape features suitable for the docking. To provide a measure for the onset of ambiguity in the matching of globular structures, a 'sphericity' shape descriptor is implemented in *qrange* and *qdock*. The sphericity is defined as the density $\varsigma = cM/R_g^3$, where M is the mass of the molecule, R_g is the radius of gyration, and the constant c = 0.119 Å^3 amu^{-1} was parametrized to yield a value ς = 1 for spherical protein and polynucleotide excisions from PDB entries of various sizes and origins. Since the packing density of proteins is conserved (Tsai *et al.*, 1999), ς is a size-invariant shape descriptor that is maximal for spherical distributions. Table 1 shows that all three non-trivial degenerate cases exhibit high sphericity ($\varsigma \geq$ 0.56), and the unambiguous fits exhibit low sphericity ($\varsigma \leq 0.48$).

4. Conclusions

We have developed, tested and disseminated a set of procedures for the reproducible docking and the visualization of atomic structures and SAXS bead models. Researchers in the SAXS community who seek to validate reconstruction methods by matching SAXS-based models with known atomic structures will find the tools a welcome addition to their existing routines.

The lack of prior knowledge about the mutual correspondence of features complicates the shape-based registration of three-dimensional data sets from various biophysical sources. By testing *Situs* systematically against simulated low-resolution data, we addressed three questions of interest in applications of structural docking algorithms: (i) whether the docking is correct, at least to within bead-size accuracy (yes), (ii) whether correct solutions are inadvertently missed as a result of the reduced search space (no), and (iii) whether incorrect (spurious) fits give rise to ambiguities (yes, but spurious fits resulted only in the case of globular molecules).

Applications of *Situs* are not limited to rigid-body docking alone. The most intriguing problems may well arise in situations where the solution structure of a protein deviates from its crystal structure. For example, the structure of calmodulin was shown first by SAXS to compact in solution relative to the sole crystallographic conformation known at the time (Heidorn & Trewhella, 1988). We have recently devised routines based on *Situs* and molecular dynamics simulation that bring deviating global features of structures into register, while preserving the atomic structure locally (Wriggers *et al.*, 2000; Wriggers & Birmanns, 2001). The SAXS modules described in this paper are fully compatible with the flexible docking routines.

The present work was funded by National Institutes of Health grants P41RR12255 and 1R01GM62968, and by the La Jolla Interfaces in Science program/Burroughs Wellcome Fund.

References

Bada, M., Walther, D., Arcangioli, B., Doniach, S. & Delarue, M. (2000). *J. Mol. Biol.* **300**, 563–574.
Chacón, P., Díaz, J. F., Morán, F. & Andreu, J. M. (2000). *J. Mol. Biol.* **299**, 1289–1302.
Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Heidorn, D. & Trewhella, J. (1988). *Biochemistry*, **27**, 909–915.
Humphrey, W. F., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kabsch, W. (1978). *Acta Cryst.* A**34**, 827–828.
Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 33–41.

Martinetz, T. & Schulten, K. (1994). *Neural Networks*, **7**, 507–522.

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

Svergun, D. I., Malfois, M., Koch, M. H., Wigneshweraraj, S. R. & Buck, M. (2000). *J. Biol. Chem.* **275**, 4210–4214.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *J. Mol. Biol.* **290**, 253–266.

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

Wriggers, W., Agrawal, R. K., Drew, D. L., McCammon, J. A. & Frank, J. (2000). *Biophys. J.* **79**, 1670–1678.

Wriggers, W. & Birmanns, S. (2001). *J. Struct. Biol.* **133**, 193–202.

Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). *J. Struct. Biol.* **125**, 185–195.

Wriggers, W., Milligan, R. A., Schulten, K. & McCammon, J. A. (1998). *J. Mol. Biol.* **284**, 1247–1254.