electronic reprint

Journal of
Applied Crystallography
ISSN 0021-8898

Prototype of a database for rapid protein classification based on solution scattering data

Anna V. Sokolova, Vladimir V. Volkov and Dmitri I. Svergun

Prototype of a database for rapid protein classification based on solution scattering data

Anna V. Sokolova,[ab] Vladimir V. Volkov[a] and Dmitri I. Svergun[ba*]

[a]Institute of Crystallography RAS, Leninsky pr., 59, 117333, Moscow, Russia, and [b]European Molecular Biology Laboratory c/o DESY, Notkestrasse 85, D-22603, Hamburg, Germany. E-mail: svergun@embl-hamburg.de

Finding similarities in sequences and/or structures is one of the fastest ways of characterizing proteins. Similarity in atomic structure of proteins is generally more recognizable than similarity in sequence and may be more closely related to similarity in function. If the atomic structure is not available, useful information can be obtained by X-ray solution scattering. *Ab initio* methods to analyse the scattering data require longer computation time and yield low resolution models only. We present an approach to rapidly characterize proteins with unknown structure based on comparison of experimental scattering profiles with a database of scattering patterns calculated from known structures.

Keywords: **Small angle scattering; protein classification; database**

1. Introduction

The relationship between function and structure for biological macromolecules is one of the fundamental paradigms of molecular biology. Similarity in structure is generally more recognizable than similarity in sequence and the structure-based tools for the analysis of structure–function relationship such as databases SCOP (Murzin *et al*., 1995), and CATH (Pearl *et al*., 2000) allow more informative transfer of functional description than sequence alone. The current databases utilize similarities in high resolution three-dimensional (3D) structures and can thus not be used for proteins with unknown atomic structure. This paper describes the development of a method for rapid characterization of proteins in solution based on small-angle X-ray scattering (SAXS) data.

Small-angle scattering intensity I(s) from a dilute solution of macromolecules is an isotropic function of momentum transfer $s = 4\pi \sin\theta / \lambda$, where 2θ is the scattering angle, and λ is the X-ray wavelength. A low-angle part of a SAXS pattern in the range of s from 0 to about 1.5 nm^{-1} (corresponding to resolution up to about 4 nm) provides information about the overall shape of the protein. A medium angle part of the curve (range of s from about 4 to 9 nm^{-1}; resolution from 1.5 to 0.7 nm) contains information about the internal (tertiary and secondary) structure (Fig. 1). It was therefore decided to separately analyse the low and medium angle parts of the curves to develop comparison criteria for the scattering patterns. At this stage, all computations were made using theoretical scattering curves predicted from atomic models using the program CRYSOL (Svergun *et al*., 1995).

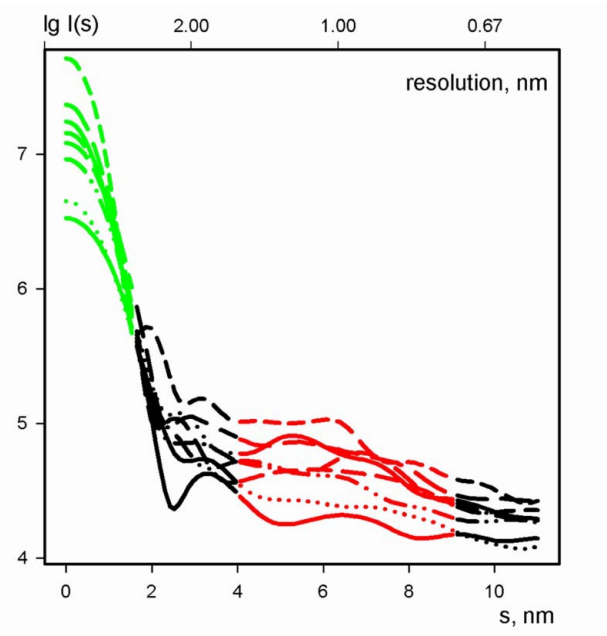

Figure 1

A set of SAXS patterns from 8 proteins with molecular mass from 100 to 300 kDa. The internal parts of these curves are marked by green and the medium parts are coloured red.

2. Realisation of a database of protein structures

To generate a representative database of protein structures, high resolution models of proteins ranging from 50 to 3000 aminoacids were downloaded from the Brookhaven Protein Data Bank (PDB; Bernstein *et al*., 1977). Only the structures obtained by X-ray diffraction were taken (excluding theoretical and NMR models), and the structural homologues as defined in the PDB were removed. The PDB files do not explicitly provide biologically active forms of the molecules but rather contain independent parts of the crystallographic unit cell. Programs were written to automatically analyze the PDB files, read proper transformation matrices and construct the biologically active molecules. To exclude the influence of the molecular mass (MM), all structures were divided in 26 sets with an increment of 50 aminoacids residues, and all subsequent comparisons were made within the individual sets. The database of high resolution models of biologically active protein forms currently contains about 1500 structures.

3. Selection of comparison criteria between scattering patterns

Adequate comparison of small-angle scattering patterns is not a trivial task, as the curves are rapidly decaying functions of the scattering vector. A criterion to compare two one-dimensional scattering curves $I_1(s)$ and $I_2(s)$ (R-factor) can be generally written as:

$$Rf_{I_1(s),I_2(s)} = \frac{\sum_{i=1}^{N}(Weight_i \cdot (ScaFac_i \cdot I_2(s_i) - I_1(s_i)))^2}{\sum_{i=1}^{N} I_1(s_i) \cdot ScaFac_i \cdot I_2(s_i) \cdot (Weight_i^2)} \quad (1)$$

where the scaling multiplier *ScaFac* yielding the best least-squares fit is calculated as:

$$ScaFac = \frac{\sum_{i=1}^{N} I_1(s_i) \cdot I_2(s_i) \cdot Weight_i^2}{\sum I_2(s_i) \cdot Weight_i^2} \quad (2)$$

Several types of the weighting function "$Weight_i$" were analyzed ($Weight_i = s_i$, $Weight_i = s_i^2$, as well as comparison on a logarithmic scale). It was found on simulated examples that $Weight_i = s_i$ (which corresponds to the weight used in the Shannon sampling theorem (Shannon & Weaver, 1949)) provides the most sensitive criterion and can be used for the low and medium angle parts of the scattering data. Below, the R-factors for the two parts will be called *RfI* and *RfM*, respectively.

4. Analysis of internal parts of SAXS curves and shape analogs

Analyzing the low resolution parts of the SAXS patterns the main question to answer was, what is the limiting value of the *RfI*, below which the overall shapes could be considered virtually identical. In all the molecular mass sets, the values of *RfI* were calculated for every pair of proteins, and the differences in shape between the two proteins were simultaneously characterized using the program SUPCOMB (Kozin & Svergun, 2001). The latter program automatically aligns two arbitrary 3D structures represented by ensembles of points and provides a measure of dissimilarity between the aligned structures. This measure called a normalized spatial discrepancy (NSD) is computed as follows. For every point in the first model, the minimum value among the distances between this point and all points in the second model is found, and the same is done for the points in the second model. These distances are added and normalized against the average distances between the neighboring points for the two models so that the value NSD < 1 indicates that the two models are similar. To transform the atomic models to low resolution, a program was written for converting an all-atom representation of a protein into a densely packed bead model with a desired bead radius. This procedure allowed to significantly reduce computation time using SUPCOMB and to compute the NSD values adequately representing discrepancy between the low resolution shapes.

Analysis of correlation between *RfI* and NSD (an example is presented in Fig.2 for 54 proteins with the polypeptide chain length between 400 and 450 residues) allows one to make the following conclusions:

(i) The minimum value of *RfI* corresponding to virtually equal low resolution structures (NSD < 0.8) is 2 %;

(ii) If the value of *RfI* is larger than 5 %, the shapes of the two proteins are significantly different (NSD > 1.3).

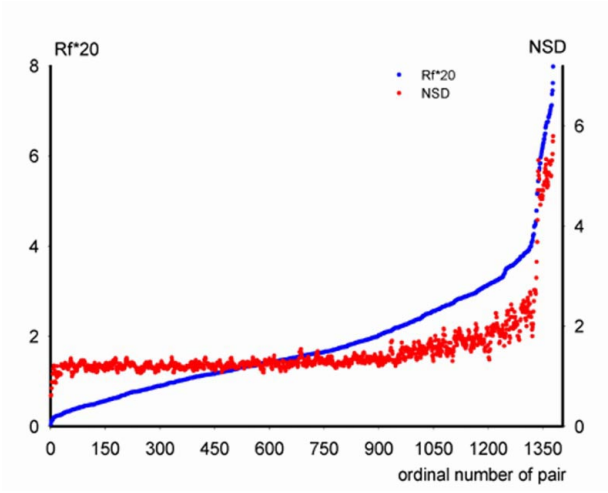

Figure 2

An example of correlation between *RfI* and NSD.

Fig.3 presents examples of the structures with similar overall shapes and two corresponding scattering patterns calculated by CRYSOL (top and bottoms panels, respectively).

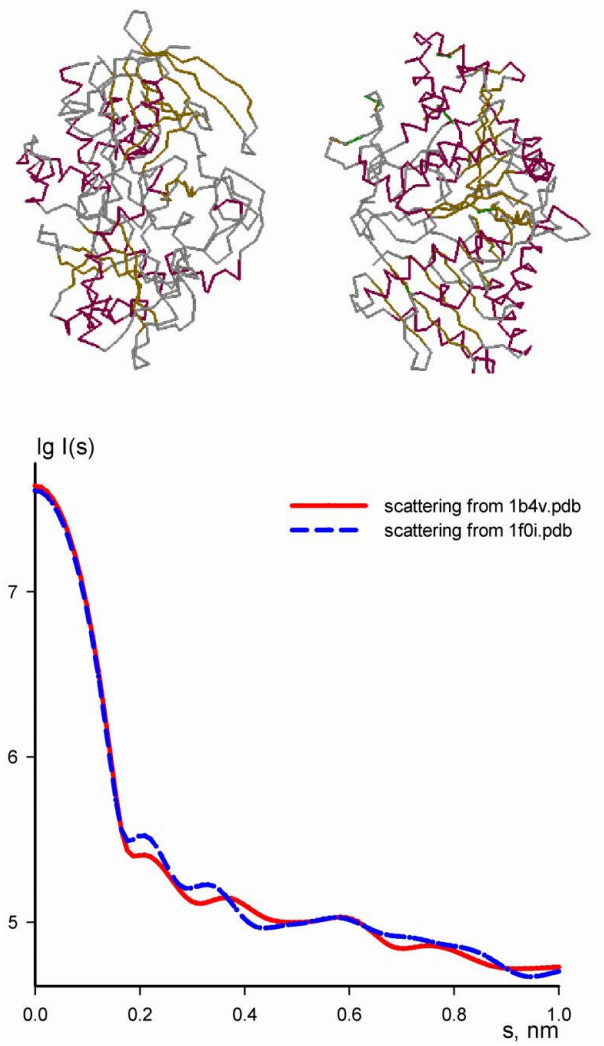

Figure 3

Proteins with similar overall shape (left: 1b4v.pdb, right: 1f0i.pdb) and the corresponding SAXS patterns yielding *RfI* = 0.2%.

An example of structures with different overall shapes (NSD=1.832) and corresponding scattering curves is shown on Fig.4 (top and bottom panels, respectively). It is interesting to note that not a single case was found where two structures with significantly different shapes gave a low (< 2%) *RfI*, in other words, it appears that the low resolution portion of a SAXS curve uniquely defines the particle shape. This result is not trivial as the scattering patterns are one-dimensional functions and a lot of structural information is lost because of chaotic orientation of particles in solution (Feigin & Svergun, 1987). This finding can be correlated with recent progress in *ab initio* shape determination methods (Svergun *et al*., 2001; Svergun, 1999). On the other hand, some of the geometrical parameters (e.g. the maximum size of the particle) were found to be poorly correlated with *RfI*.

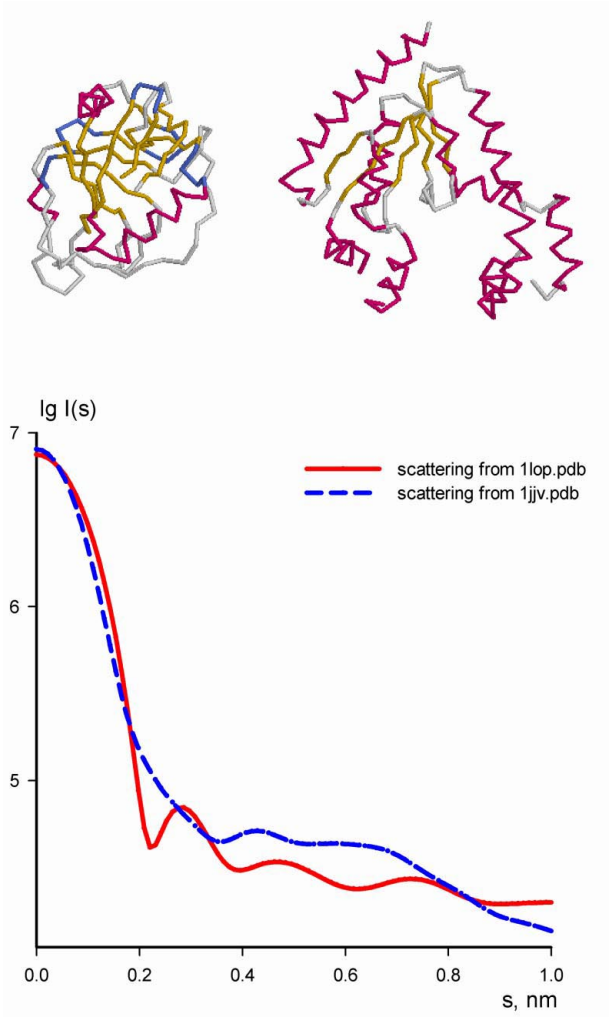

Figure 4

Structures with different overall shapes (left: 1lop.pdb, right: 1jjv.pdb) and the corresponding SAXS curves yielding *RfI* = 14.9%.

5. Analysis of the external parts of SAS curves and analogs in internal structure

The first question to answer was to which extent could the *RfM* value be correlated with the level of homology between two proteins. At the first step, the values of *RfM* were pairwise calculated as described above for *RfI*, and the pairs with *RfM* exceeding 15%, were discarded as corresponding to particles with different internal structure.

All structures were separated into three sets according to the CATH classification, which is a hierarchical classification of 18577 domains into evolutionary families and structural groups. A class (first level of classification) is assigned automatically by considering the percentage of **α**-helix and **β**-strands and the secondary structure packing. The three major classes are mainly-**α**, mainly-**β** and **α+β** proteins. The average scattering curves of the three classes $I_{\alpha}(s)$, $I_{\beta}(s)$ and $I_{\alpha+\beta}(s)$ were calculated for each set and the program OLIGOMER (Svergun D.I., Volkov V.V. and Sokolova A.V., unpublished information) was implemented to decompose the scattering patterns of proteins into a linear superposition:

$$I(s) = w_{\alpha} I(s)_{\alpha} + w_{\beta} I(s)_{\beta} + w_{\alpha+\beta} I(s)_{\alpha+\beta} \qquad (3)$$

The OLIGOMER yields the contributions w_{α}, w_{β} and $w_{\alpha+\beta}$ of the average "class" curve into the given scattering profile. In most cases the class of the protein was correctly recognized, i.e. the scattering from a protein belonging to a given class was represented by 100% the average scattering from this class. However, several cases were found when this was not the case (see example in Fig.5). Further subclasses in the CATH classification could not be distinguished based in the scattering data.

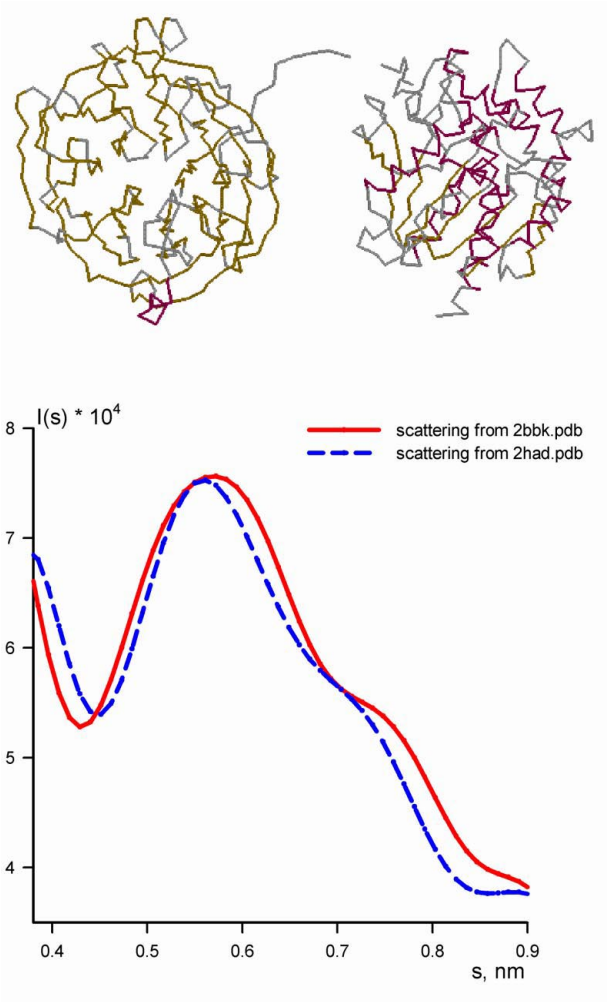

Figure 5

Two structures belonging to different structural classes: 2bbk.pdb (left, class mainly-**β**) and 2had.pdb (right, class (**α+β**)) and the calculated medium angle SAXS patterns yielding *RfM* = 3.8%.

For several target proteins, the homology with the neighbors in *RfM* was assessed using the database DALI (Holm & Sander, 1993). This algorithm allows one to optimally align two protein structures and to numerically estimate their structural homology by a statistical significance of the similarity criterion called **Z** (the higher **Z,** the larger homology). A weak correlation was observed between the *RfM* and **Z**: the average *RfM* value for similar structures (**Z** > 2) was (6±3)%, whereas for different structures (**Z** < 2) it was (12±6)%. In most cases, homologous structures (**Z** > 10) yield similar scattering curves (*RfM* < 6%, see example in Fig.6), but in some cases similar proteins have different scattering patterns (*RfM* > 10%).

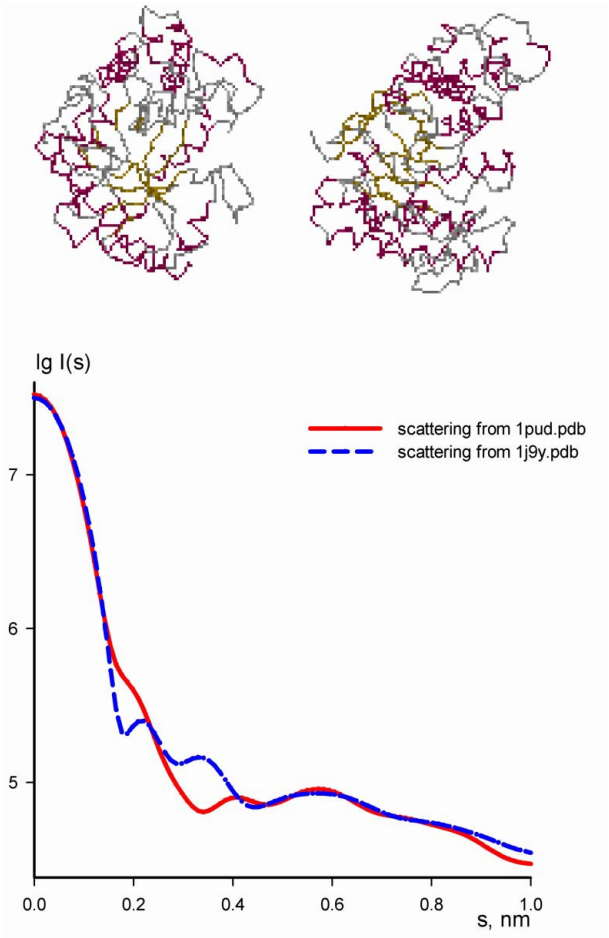

Figure 6

Homologous structures 1j9y.pdb (left) and 1pud.pdb (right) with statistical significance **Z** = 10.5 and their scattering curves yielding *RfH* = 4.6%.

Surprisingly we have found that all the neighbours in *RfM* had the low resolution shapes similar to those of the target proteins, i.e. the shape similarity was a necessary prerequisite for the similarity of the scattering patterns in the range of momentum transfer $0.4 < s < 0.9$ nm^{-1}. This suggests that for similar molecular mass the medium angle scattering pattern may provide a useful signature of the overall structure, additional to the low angle data.

Further analysis of the *RfM* neighbours is now in progress. In particular, a low pass filter in reciprocal space with the frequency of about 2 nm should be able to reduce the influence of the overall structure and thus to improve contrast for the comparison of the internal structure based on the medium angle scattering patterns.

6. Conclusions

A database containing scattering patterns from biologically active protein molecules is being created. Quantitative criteria were formulated to find virtually identical and obviously different shapes based on low resolution scattering data. Some correlations were observed between the medium angle scattering curves and the classifications in the databases CATH and DALI. It was found that the medium angle data still contain significant information about the overall shape of the protein. Further work on internal structure classification is in progress.

Sokolova A.V. acknowledges support from the fellowship by FEBS, DAAD (Grant A/02/24151), by INTAS (Grant YSF 2001/2-133). The work was also supported by the INTAS Grant 00-243.

References

Bernstein, F.C., Koetzle, T.F., Williams, C.J., Meyer, E.E.Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542; http://www.rcsb.org

Feigin, L.A. & Svergun, D.I. (1987). New York: Plenum Press.

Holm, L. & Sander, C. (1993). *J. Mol. Biol.* **233**, 123-138. http://www2.ebi.ac.uk/dali.

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

Muller, J.J., Damaschun, G. & Schrauber, H. (1990). *J.Appl Cryst.* **23**, 26-34.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536-540; http:// scop.berkeley.edu.

Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. & Orengo, C.A. (2000). *Nucleic Acids Research* **28**(1), 77-282; http://www.biochem.ucl.ac.uk/bsm/cath_new.

Shannon, C.E. & Weaver, W. (1949). *University of Illinois Press, Urbana, IL*

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

Svergun, D.I., Barberato, C. & Koch, M.H.J. (1995). *J. Appl. Cryst.* **28**, 768-733.

Svergun, D.I., Petoukhov, M.V. & Koch, M.H.J. (2001). *Biophys. J.* **80**, 2946-2953.