Biophysical Journal, Volume *114*

Supplemental Information

Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions

Daniel Franke, Cy M. Jeffries, and Dmitri I. Svergun

Supplementary Materials and Methods

Preparation of native ribonuclease A (RNAse) and carboxyamidomethylated ribonuclease A (cam-RNAse).

Lyophilised bovine pancreatic ribonuclease A (Sigma) was resuspended in phosphate buffered saline, pH 7.0, (PBS) and dialysed overnight at 4 °C against the same buffer to obtain a sample of natively-folded RNAse. The final sample concentration was 5.24 mg/ml (determined at $Abs_{280\ nm}$ using an $E_{0.1\%}$ = 0.71 ml/mg calculated from the amino acid sequence (1). The post dialysis buffer was used as an exact solvent blank for the SAXS measurements.

The preparation of disulfide-reduced and carboyxamidomethylated RNAse (*cam*-RNAse) followed the procedure as described by Wang, Trewhella, & Goldenberg (2008). Briefly, lyophilised RNASe powder (approximately 8–10 mg) was dissolved in 1 ml of 6 M guanidine hydrochloride (Gdn.HCl), 10 mM ethylenediaminetetraacetic acid (EDTA), 10 mM dithiothreitol (DTT) and 100 mM Tris, with a final (combined) pH of 8.0. The solution was incubated for 1.5 hr at room temperature with gentle mixing to effect protein unfolding and disulphide reduction. At the completion of the high-pH denaturation step, fresh iodoacetamide (180 mM stock in H_2O) was added to a final concentration of 30 mM and the system left for approximately 45 min to effect sulfhydryl and histidine alkylation. Concentrated HCl (1 M in H_2O) was then added to the RNAse with rapid mixing to a final concentration of 100 mM. The protein solution was dialysed overnight at 4 °C against 10 mM HCl in water. To two individual aliquots of post-dialysis *cam*-RNAse were removed and a stock solution of 8 M urea in 10 mM glycine (combined pH = 2.5) was added to final concentrations of 1 or 2 M, respectively. The protein concentrations of the *cam*-RNAse samples were: 5.97 mg/ml (10 mM HCl), 5.27 mg/ml (10 mM HCl, 1 M urea) and 4.58 mg/ml (10 mM HCl, 2 M urea). For the SAXS measurements, the post-dialysis 10 mM HCl solution was used as an exact solvent blank, with the addition of an equivalent mass of 8 M urea solution (+/- 1 mg) as used for the 1 M and 2 M urea *cam*-RNAse samples. The corresponding SAXS data of folded and unfolded RNAse can be located in the SASBDB entries SASDDL3 and SASDDM3, respectively.

Preparation of Candida antarctica lipase B.

A solution of *Candida antarctica* lipase B (Hampton Research) was dialysed overnight at room temperature against 100 mM NaCl, 20 mM Na_2HPO_4, containing 0, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5 or 6 M Gdn.HCl (combined pH = 6.0, adjusted using HCl/NaOH). In all instances, the respective post-dialysis buffer was used as the solvent blank for the SAXS measurements. A set of lipase B samples were also prepared under reducing conditions. DTT (1 M stock in H_2O) was added to each lipase B sample/solvent blank to a final concentration of 10 mM immediately prior to SAXS. The final concentrations of the lipase B samples were assessed using an $Abs_{280\ nm}$ $E_{0.1\%}$ = 1.239 ml/mg (1) and are summarised in Table 2. The SAXS data of both Lipase B both with and without DTT and the Gdn.HCl unfolding series can be located in the SASBDB entries SASDDJ3 and SASDDK3.

Circular dichroism spectropolarimetry: native RNAse and unfolded *cam*-RNAse.

Circular dichroism (CD) measurements were performed at room temperature using a Chirascan (Applied Photophysics) spectropolarimeter with a quartz cell pathlength (l) of 1 mm. The RNAse samples used for SAXS (described above) were diluted 50-fold in their respective supporting solvents. The approximate protein concentrations, C, in mg/ml used for the CD measurements are reported in Table 2.

The CD spectra were acquired across 175–280 nm using a time constant of 0.5 s at 1 nm wavelength intervals (1 nm bandwidth). The presented data (Supp. Fig. 1) represent the solvent-subtracted average of these scans for each sample, quoted as mean residue ellipticity, θ in deg.cm^2/dmol versus wavelength, λ in nm (where the molecular weight, MW, of RNAse = 13690 Da, and the number of amino acids, N= 124). The conversion from machine units (mdeg) to θ followed:

$$\theta = (\text{mdeg}*\text{MW})/(Nl\text{C}).$$

Those data with unduly high absorbance at low wavelength were discarded to produce the final spectra for:

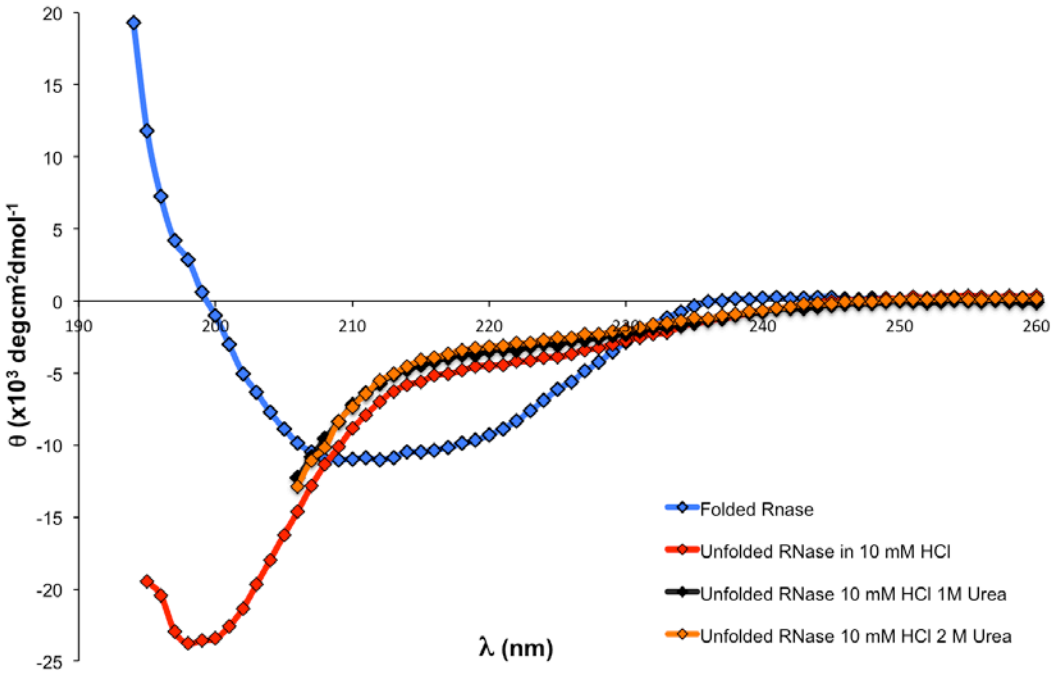
Native RNAse; 197–260 nm,
cam-RNAse in 10 mM HCl; 197–260 nm,
cam-RNAse in 10 mM HCl, 1 M urea; 205–260 nm,
cam-RNAse in 10 mM HCl, 2 M urea; 206–260 nm.

Secondary structure analysis was performed using the online BeStSel single spectrum analysis and fold recognition server, http://bestsel.elte.hu/ (2). Spectra were converted into absorption units, i.e., as the differential molar extinction coefficient, $\Delta\varepsilon$ ($M^{-1}.cm^{-1}$) vs λ, (where $\Delta\varepsilon = \theta/3298.2$) and analysed for secondary structure content using the 200–250 nm option of the BeStSel server. Only those spectra for native RNAse and *cam*-RNAse in 10 mM HCl access sufficiently low wavelengths for secondary structure analysis (Table 1), and consequently the secondary structure content of the remaining *cam*-RNASe samples were not assessed. The experimental results were compared to the secondary structure content extracted from the X-ray crystal structure of RNAse A (PDB: 3MZQ) and that reported for RNAse in the Protein Circular Dichroism Data Bank (http://pcddb.cryst.bbk.ac.uk/home.php, PCDDBID: CD0000063000).

Tryptophan fluorescence spectroscopy: native and unfolded lipase B.

Intrinsic tryptophan fluorescence spectroscopy measurements from lipase B and denatured lipase B in Gdn.HCl or Gdn.HCL plus 10 mM DTT (Table 2) were performed using a Tecan Infinite M1000 spectrometer. Scans were performed at 25 °C using an excitation wavelength of 295 nm, with the emission spectra recorded from 310–600 nm using an emission wavelength step size of 1 nm (flash frequency, 400 Hz; 50 flashes per nm). The fluorescence yields were normalised to protein concentration and the wavelength corresponding to the maximum flourescence yield for each scan was recorded to qualitatively assess red-shifts in the tryptophan emission spectra (Supp. Fig. 2).

Supplementary Figures and Tables



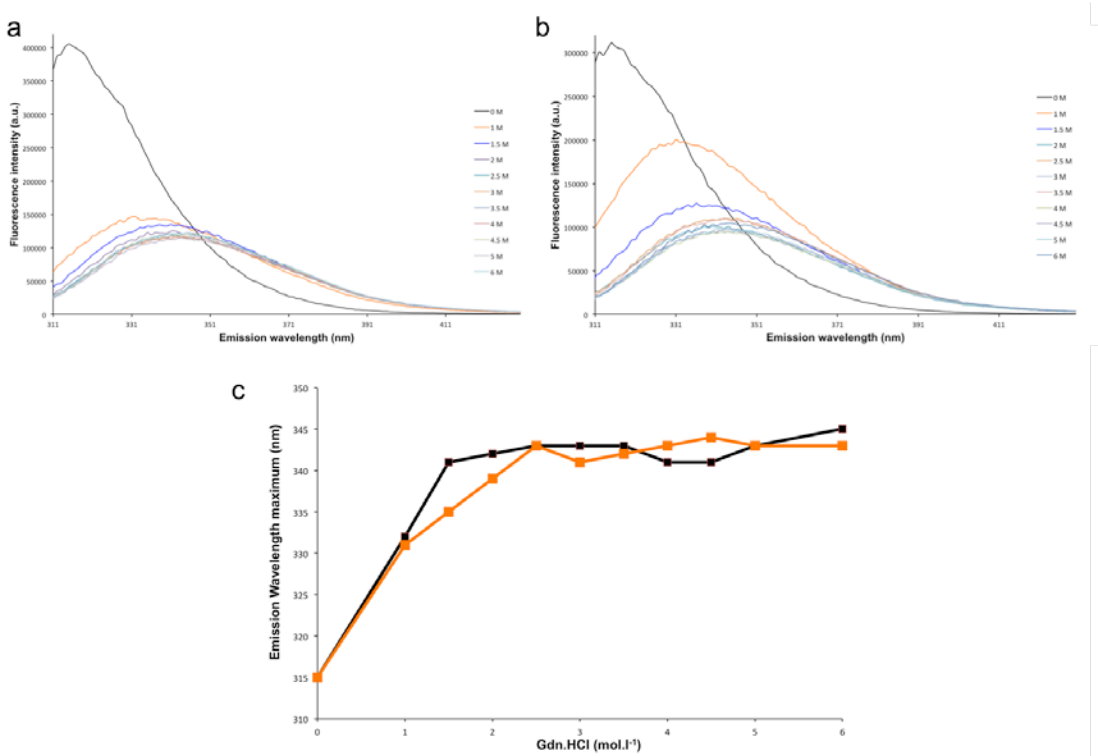
Supplementary Figure 1. CD spectra of native folded RNase (blue), unfolded *cam*-RNase (red) and *cam*-RNase in the presence of 1 M or 2M urea (black and orange, respectively)

RNAse A *CD Samples*	Protein Concentration (mg/ml)	*RNAse A* *% Secondary structure* α-helix	β	turn+other
Native RNAse	0.105	20.1	34.7	45.2
cam-RNAse 10 mM HCl	0.119	0	23.9	76
cam-RNAse 10 mM HCl, 1 M Urea	0.105	-	-	-
cam-RNAse 10 mM HCl, 2 M Urea	0.092	-	-	-
	PDB: 3MZQ	21.0	33.1	46.0
	PCDDBID: CD0000063000	20.9	33.1	45.9

Table 1: Sample protein concentrations and secondary structure analysis derived from CD measurements of native and *cam*-RNAse. Included is a comparison with secondary structure content extracted from the X-ray crystal structure of RNAse (PDB 3MZQ) and from CD spectra deposited in the Circular Dichroism Data Bank (CD0000063000).

Lipase B SAXS Samples		Lipase B fluorescence spectroscopy samples		
Gdn.HCl concentration (M)	Protein Concentration (mg/ml)	Gdn.HCl concentration (M)	Protein Concentration (mg/ml)	Protein Concentration (mg/ml), plus DTT
0	4.65	0	0.48	0.5
1	4.66	1	0.43	0.38
1.5	4.59	1.5	0.37	0.42
2	4.52	2	0.4	0.38
2.5	4.48	2.5	0.37	0.41
3	4.19	3	0.39	0.37
3.5	4.06	3.5	0.39	0.44
4	4.09	4	0.38	0.39
4.5	4.12	4.5	0.41	0.37
5	4.02	5	0.38	0.39
6	4.19	6	0.41	0.38

Table 2: Concentration of Gdn.HCl and Lipase B used for SAXS and fluorescence spectroscopy measurements. Note: the SAXS samples for lipase B under reducing conditions were prepared by adding 1 μl of 1 M DTT to 99 μl of protein. Therefore, within pipetting and spectrophotometric error, it is expected that the reduced lipase B sample concentrations will not differ significantly from the concentrations quoted here for the lipase B SAXS samples in Gdn.HCl.



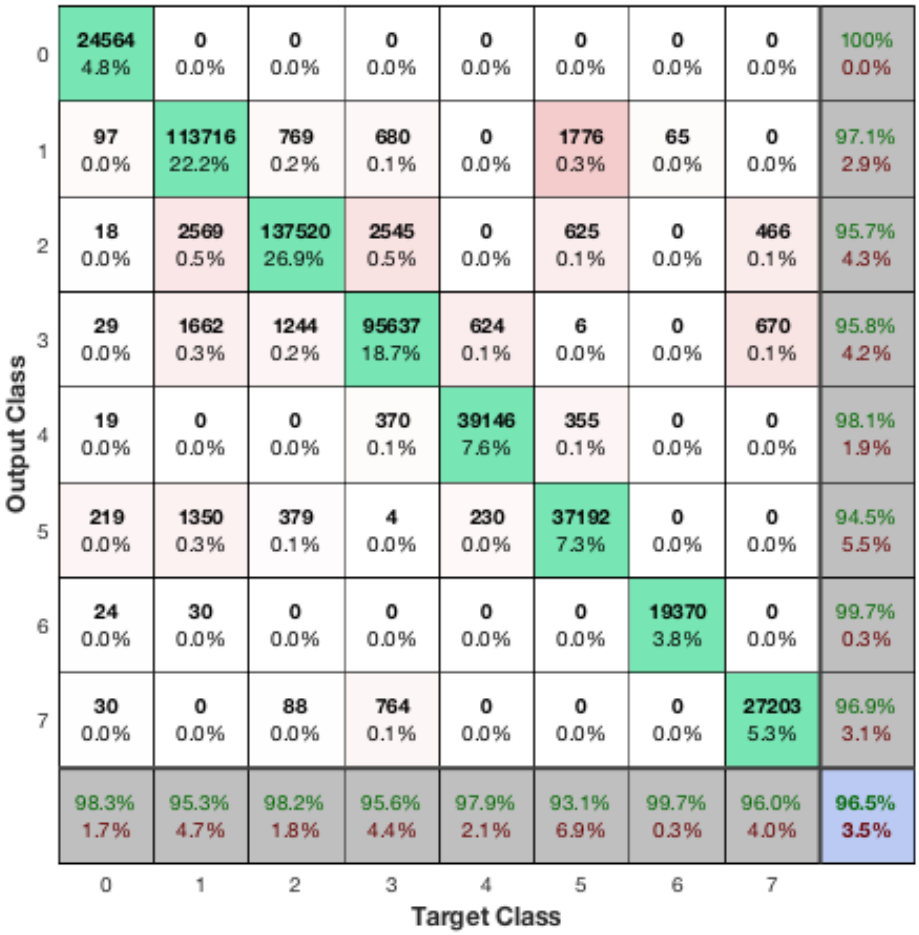
Supplementary Figure 2. a. Tryptophan fluorescence intensities vs emission wavelengths through a Gdn.HCl concentration gradient (0-6 M) for Lipase B with no DTT present in solution. b. With additional 10 mM DTT added to solution. c. The shift in emission wavelength maximum of Lipase B as a function of Gdn.HCl concentration.

Preparation and SAXS data of bovine serum albumin.

Lyophilised bovine serum albumin (Sigma: # 05470) was dissolved in 50 mM HEPES, pH 7.5, and 0.22 micron pore spin-filtered. The final sample concentration was 2.25 mg/ml evaluated at $Abs_{280\ nm}$ using an $E_{0.1\%}$ = 0.646 ml/mg (Gasteiger, et al., 2005). An aliquot of 0.22 micron filtered HEPES buffer was used as the solvent blank for the SAXS measurements. The subsequent SAXS data and collection parameters of both un-subtracted and subtracted SAXS data frames can be found in SASBDB entry SASDBK3.

Class Label	Class Weight
Unknown	1
Compact	2
Extended	2
Flat	2
Ring	2
Compact-hollow	4
Hollow-sphere	2
Random-chain	2

Table 3: Empirical class weights for k-nearest-neighbour shape classification.



Supplementary Figure 3: Leave-One-Out cross validation results for shape classification with recall and precision percentages in the margins. Class labels are (0) unknown, (1) compact, (2) extended, (3) flat, (4) ring, (5) compact-hollow, (6) hollow-sphere, (7) random-chain.

Class Label	PDB		SASBDB	
Unknown	25	0. 02 %	2	0.05 %
Compact	122.913	74.05 %	149	37.16 %
Extended	5.382	3.24 %	36	8.98 %
Flat	9.734	5.86 %	119	29.68 %
Ring	154	0.09 %	3	0.08 %
Compact hollow	26.909	16.21 %	25	6.23 %
Hollow sphere	125	0.08 %	0	0.00 %
Random Chain	740	0.45 %	67	16.71 %
Total	165.982	100.00 %	401	100.00 %

Table 4: Absolute and relative shape counts as depicted in main Figure 2(a) and 2(b).

References

1. Gasteiger, E., C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel and A. Bairoch. 2005. The Proteomics Protocols Handbook Humana Press.
2. Micsonai, A., F. Wien, L. Kernya, Y. H. Lee, Y. Goto, M. Réfrégiers and J. Kardos. 2015. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc. Natl. Acad. Sci. U.S.A.
3. Wang, Y., J. Trewhella and D. P. Goldenberg. 2008. Small-Angle X-ray Scattering of Reduced Ribonuclease A: Effects of Solution Conditions and Comparisons with a Computational Model of Unfolded Proteins. J Mol Biol. 377:1576-1592.