Supporting Online Material for

The second SH3 domain of ponsin solved from powder diffraction.

Irene Margiolaki[1]*, *Jonathan P. Wright*[1], *Matthias Wilmanns*[2], *Andrew N. Fitch*[1] & *Nikos Pinotsis*[2]*.

1European Synchrotron Radiation Facility, ESRF, BP-220, F-38043, Grenoble, France.

2 European Molecular Biology Laboratory, EMBL-Hamburg c/o DESY, Notkeststrasse 85, D-22603 Hamburg, Germany.

AUTHOR EMAIL ADDRESS (margiolaki@esrf.fr, pinotsis@embl-hamburg.de)

Assessment of Powder Diffraction Data and Rietveld Analysis.

(a) Data evaluation

The least squares matrix from a Pawley refinement gives information about the data quality. Sivia has described how the eigenvalue spectrum of the matrix is related to the effective error bar on the extracted intensities (or linear combinations of intensities when peaks are overlapped)[1]. In order to give an easier comparison with single crystal data processing statistics we have computed the eigenvalues and eigenvectors of a matrix which has rows and columns multiplied by the intensities themselves. In the case of non-overlapping peaks, e.g. single crystal data, this matrix has $(I/\sigma(I))^2$ directly as the diagonal matrix elements with off diagonal elements being zero, and the eigenvalue-eigenvector transformation is not needed. When there are significant peak overlaps we choose linear combinations of intensities which are uncorrelated via the eigenvalue-eigenvector transformation. We propose that the eigenvalues are closely related to $(I/\sigma(I))^2$ for the linear combinations of peaks which have been resolved in the powder experiment. In a conventional eigenvalue-eigenvector decomposition the eigenvectors have unit 2-norm ($\Sigma|vi|^2 = 1$), and we suggest that the eigenvalue represents $(I/\sigma(I))^2$ for the linear combination of peaks corresponding to that eigenvector. In order to generate statistics which are independent of the choice of intensity partitioning we normalize the eigenvectors to have unit 1-norm ($\Sigma|vi| = 1$). The d-spacing for a linear combination of peaks is then computed as the weighted average of the d-spacings of the contributing peaks.

In this case, four data sets of enhanced quality were selected for performing structure refinement. The selected profiles were collected on sample A at 1.252481(32) Å wavelength and contain different levels of radiation damage and therefore marginally different lattice parameters. An effective completeness for single as well as combined data sets is proposed as the fraction of "peaks" having $I/\sigma(I)$ greater than some threshold (3 and 1 were chosen here) and this is tabulated in Table S1 and plotted in Fig. S3. These values indicate that combination of all four data sets results in improved data effective completeness and therefore they were employed in a combined stereochemically restrained Rietveld analysis described in the following section.

(b) Stereochemically restrained Rietveld analysis

Maximum information content of the powder data was achieved by employing four high quality profiles corresponding to slightly different lattice parameters. The selected profiles comprise two data sets collected on sample A, corresponding to 2 and 4 minutes exposure time, with a wavelength of 1.252481(32) Å and two more collected on sample B using wavelengths of 0.8012034(76) and 1.251209(40) Å respectively. The different cell dimensions for the four profiles were taken into account by using a special profile function implemented in GSAS[2]. In this function only one set of lattice parameters is refined and those corresponding to the rest of the profiles are related via a strain ($\Delta d/d$) of the reciprocal metric tensor elements and parameters related to the strain are refined in the least squares procedure. An isotropic temperature factor $U_{iso} = B_{iso}/8\pi^2 = 0.30$ Å^2 was initially used for the description of the thermal motion of all atoms and it was allowed to refine only at the latest stages of the refinement to a value of 0.348(2) Å^2. A band-matrix approximation implemented in GSAS software was employed with a matrix bandwidth of 20 parameters.

In analogy to earlier observations[3], simultaneous refinement of the lowest resolution part (d > 15.3 Å) together with the rest of the pattern usually led to some distortion of the protein structure and a poor fit. Evidently the isotropic solvent distribution via the simple Babinet's principle function, currently implemented in GSAS[2], was insufficient for representing the solvent scattering and a different approach was required. Hence, only the data in the d-spacing range less than 15.3 Å were kept in the refinement procedure for all patterns and two coefficients (As and Bs) were refined to account for the solvent scattering. These coefficients were varied separately for each of the different patterns, and they can therefore account for some of the small differences in peak intensities at low angle between the different profiles. In total, 1980 stereochemical restraints were imposed in order to refine the positions of 544 protein atoms in the asymmetric unit using experimental data between 15.3 and 2.27 Å resolution. The refinement proceeded smoothly for all the 4 profiles leading to good quality of the fit (total agreement factors: Rwp = 3.82%, Rp = 2.86%).

Finally, periodical evaluations of the protein stereochemistry were essential to monitor the progress of the refinement, therefore the validation software PROCHECK[4], WHATCHECK[5] and ERRAT-2[6] were employed. The result from ERRAT-2 is illustrated in Figure S8. In order to improve the model during the course of the refinement we also implemented iterative energy minimization processes using the Swiss-PdbViewer package[7] and WebLab Viewer Pro 3.2 (*Molecular Simulations Inc.*). Details of this refinement are listed in Table 2 and the final fitted powder diffraction profiles are presented in Fig. 1 and Fig. S7.

The use of multiple diffraction patterns having slightly different unit cell parameters in order to increase the data available for structure refinement needs to be justified. Usually a change in unit cell parameters implies a corresponding change in the crystal structure which will be averaged out in this combined fit. Some differences between diffraction patterns have been modeled by allowing the solvent scattering coefficients to differ for each of the four histograms following the observation that the largest observed differences in peak intensities were mainly at low angles. The remaining differences between model and data were then analyzed in terms of the integrated peak intensities computed from the Rietveld refinement. We have computed the correlation coefficient of $\Delta I=(I_{obs}-I_{calc})$ for the four diffraction patterns (Table S2). If the model describes an average of these different datasets then there would be a strong negative correlation of the intensity differences. However, if the differences between model and data are larger than the differences amongst the various datasets then the correlations will be positive. Figure S9 shows plots of these differences for pairs of data 3-4 and 1-4. We observe that the correlation coefficient follows the difference in unit cell parameters between the diffraction patterns, with the smallest value of 0.40207 corresponding to the pair of patterns having the largest difference in lattice parameters associated with two different samples (A and B) and different levels of radiation damage. This positive correlation of 0.40207 in the worst case indicates that the differences between the diffraction patterns are smaller than the residual differences remaining after structure refinement.

Supporting information figure captions

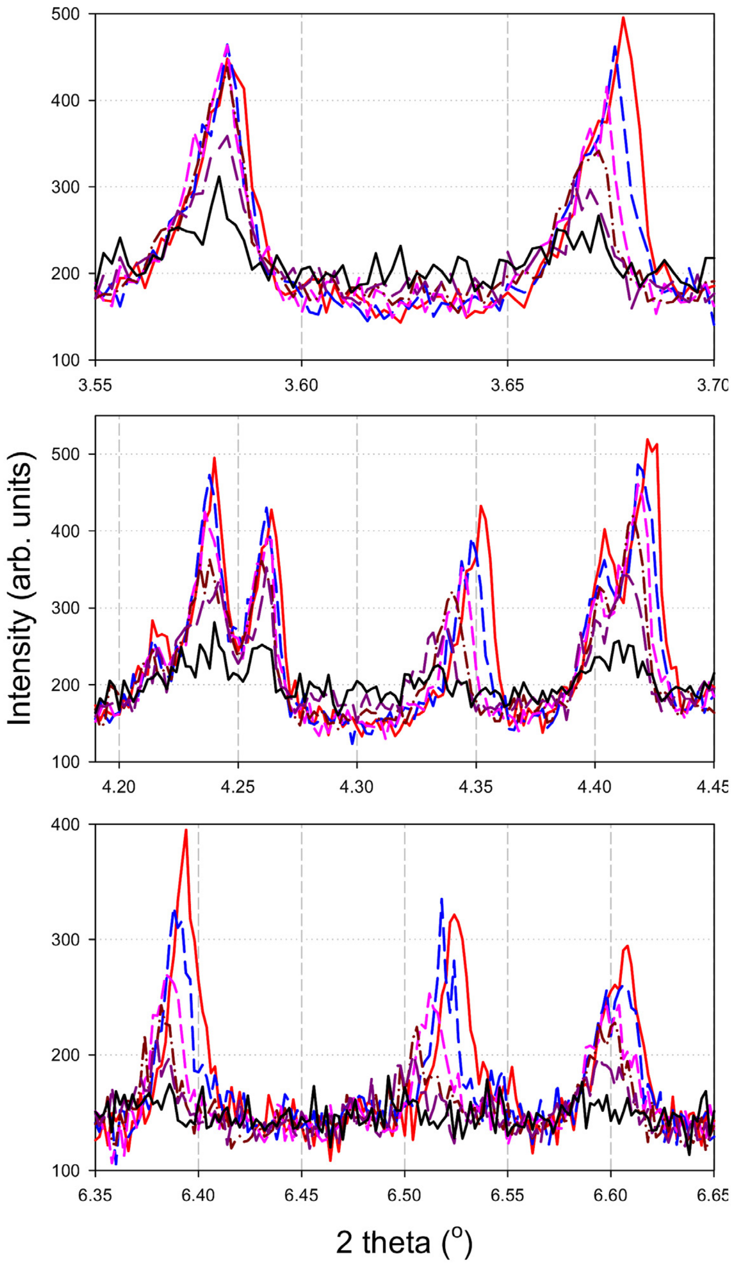


Figure S1. Selected two theta regions of powder diffraction profiles of SH3 domain- sample B (ID31: λ= 0.8012034(76) Å, 295K) showing a gradual evolution of the peak positions and widths with increasing irradiation time. The different colors correspond to different sample irradiation times varying from 2 to 12 minutes.

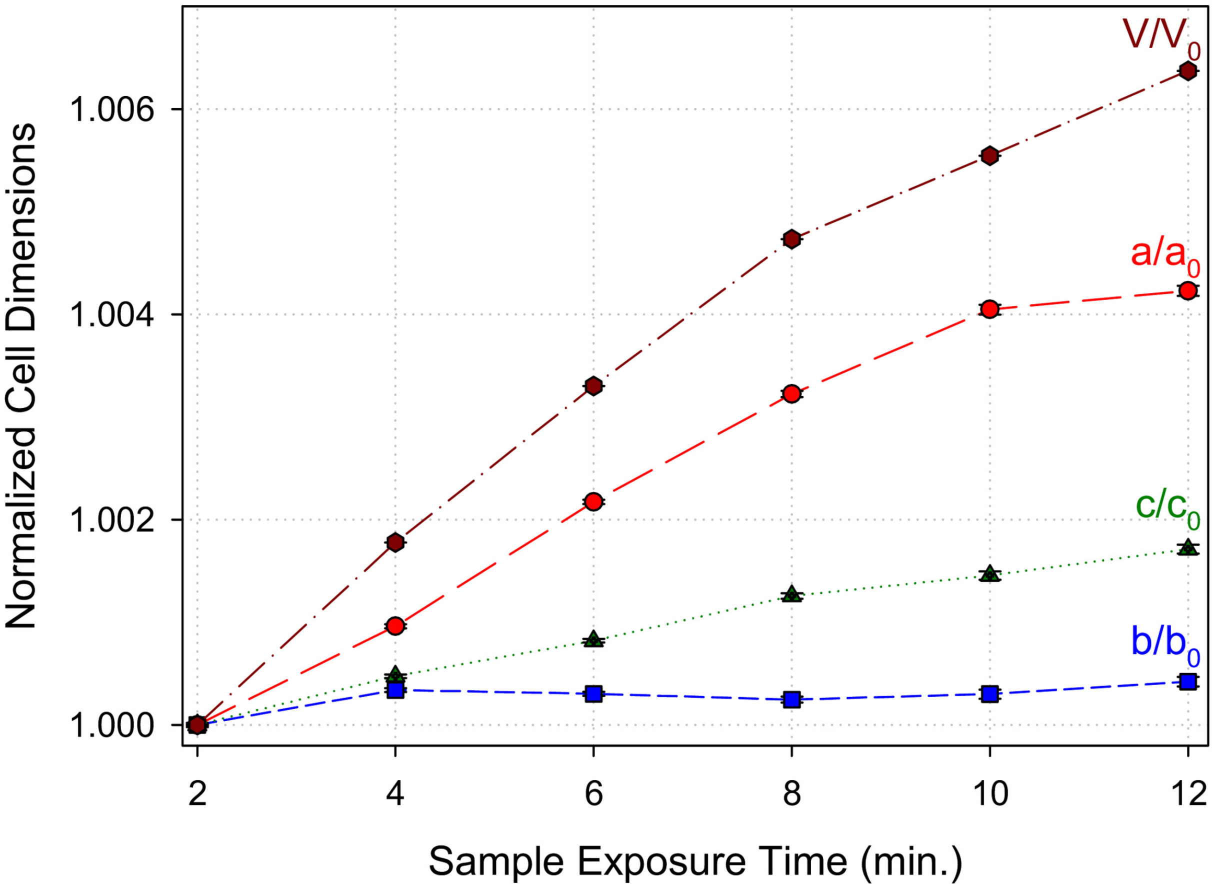


Figure S2. Evolution of the normalized orthorhombic unit cell dimensions and volume of SH3 domain (sample A), extracted from data collected at ID31 (RT, λ= 0.8012034(76) Å), with increasing irradiation time.

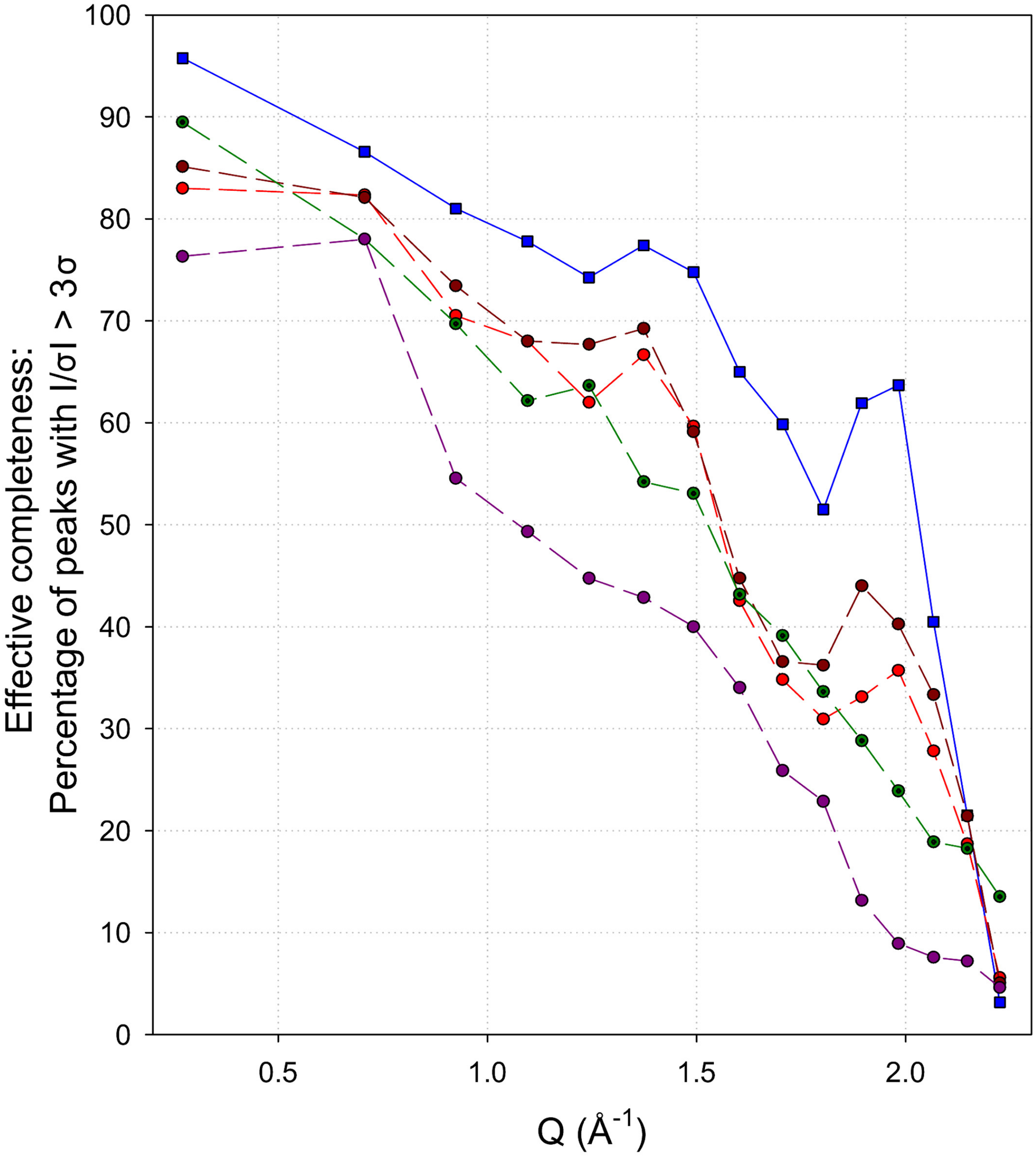


Figure S3. Effective completeness for the powder diffraction data at 3 sigma level (see text) versus Q range (Q= 2p/d). Red, brown, green and magenta colors represent single powder patterns and blue corresponds to a combined fit to all four patterns.

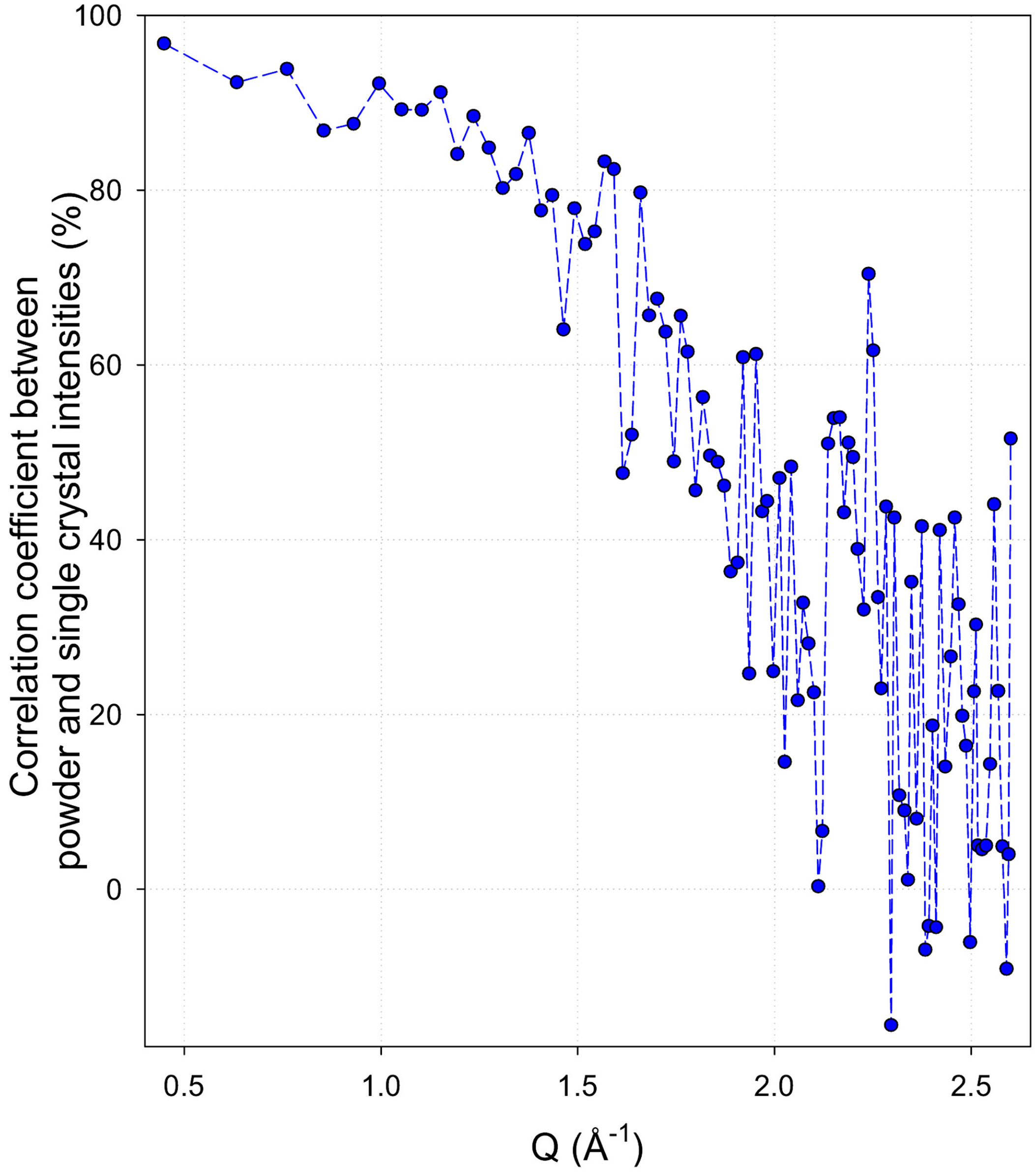


Figure S4. Correlation coefficients between powder and single crystal intensities versus data resolution.

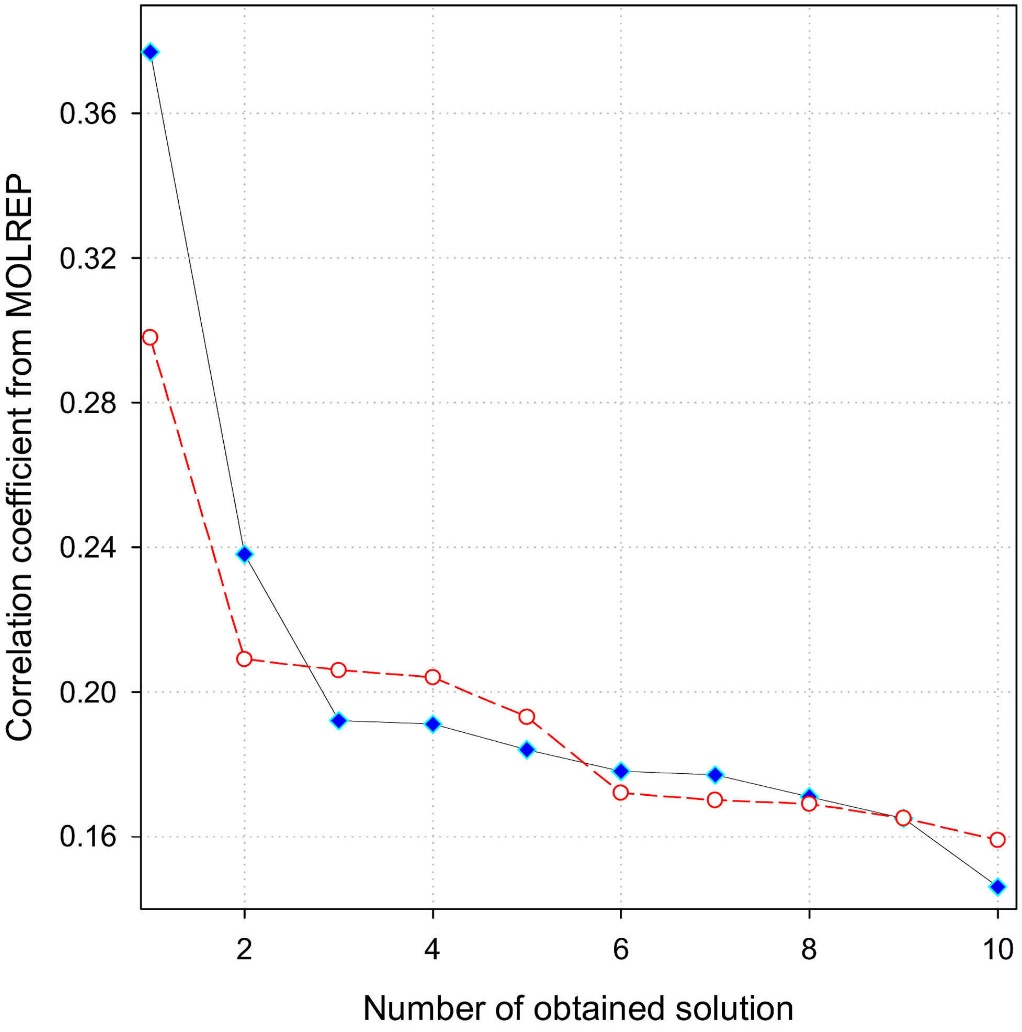


Figure S5. Correlation coefficients computed for molecular replacement using model 1W70 (blue symbols) and 1OOT (red symbols) and intensities extracted from a 4-dataset Pawley refinement. The top 10 rotation-function peaks were each used to generate 10 translation- function peaks with *MOLREP* program.

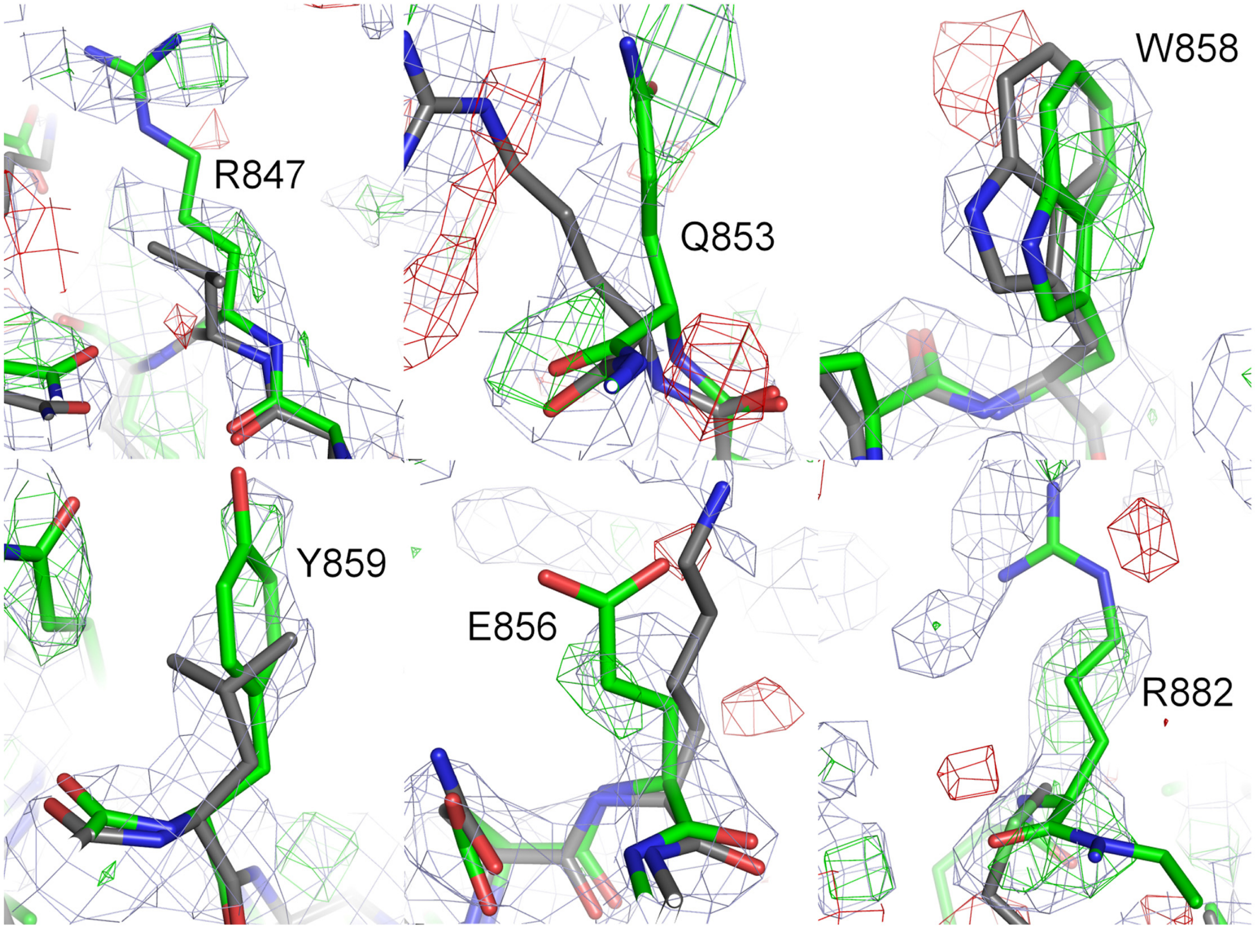
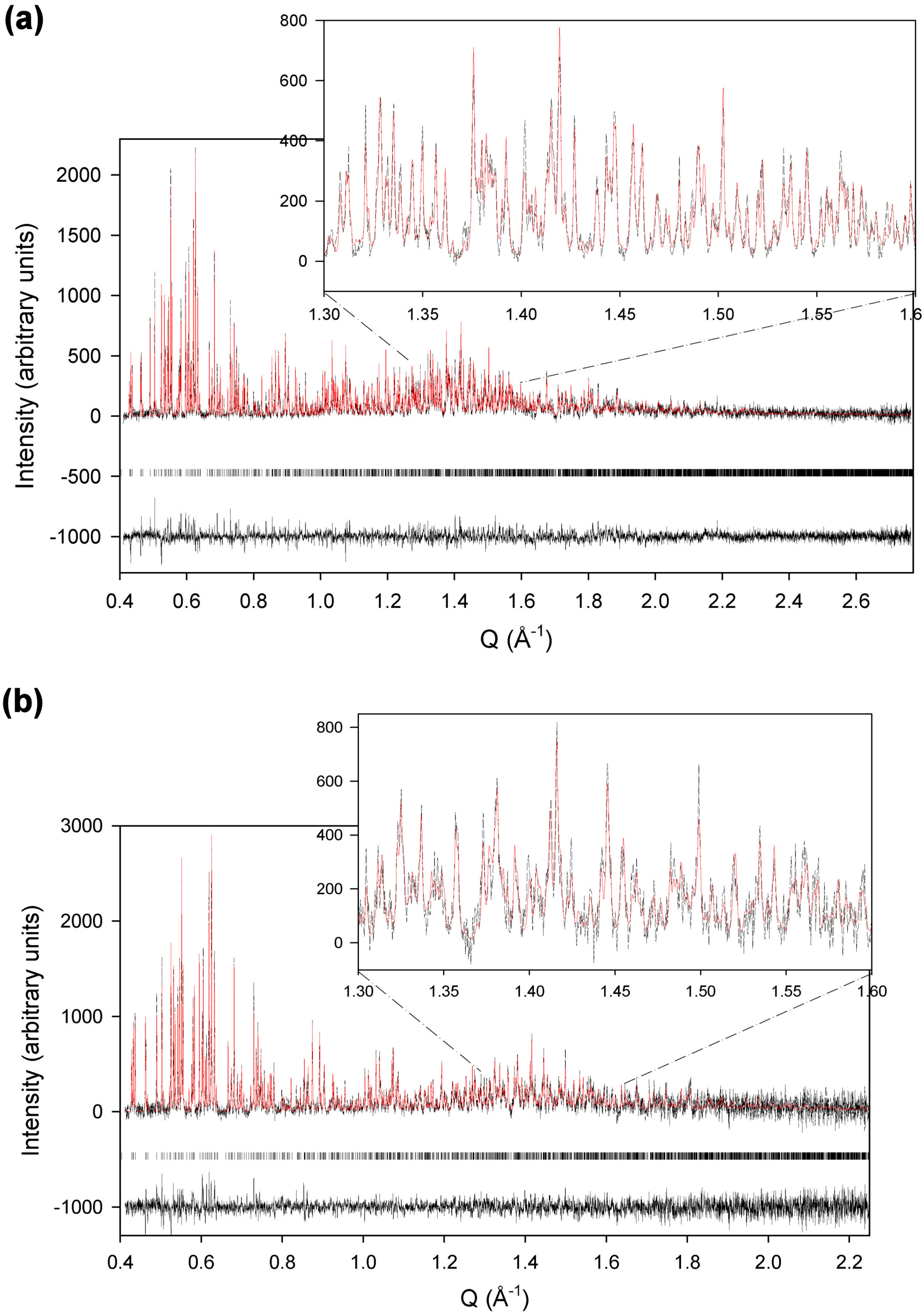


Figure S6. Selected regions of the 2Fo-1Fc (blue at 1σ) and 1Fo-Fc (red at -2.5σ and green at 2.5σ) electron density maps, as determined directly after the molecular replacement. The residues represented in grey stick carbon atoms correspond to the molecular replacement model used for the calculation of the maps, while the residues in green color carbon atom sticks represent the final refined model.


(a)
Intensity (arbitrary units)
2000
1500
1000
500
0
-500
-1000
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Q (Å^{-1})
800
600
400
200
0
1.30
1.35
1.40
1.45
1.50
1.55
1.60
(b)
Intensity (arbitrary units)
3000
2000
1000
0
-1000
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
Q (Å^{-1})
800
600
400
200
0
1.30
1.35
1.40
1.45
1.50
1.55
1.60

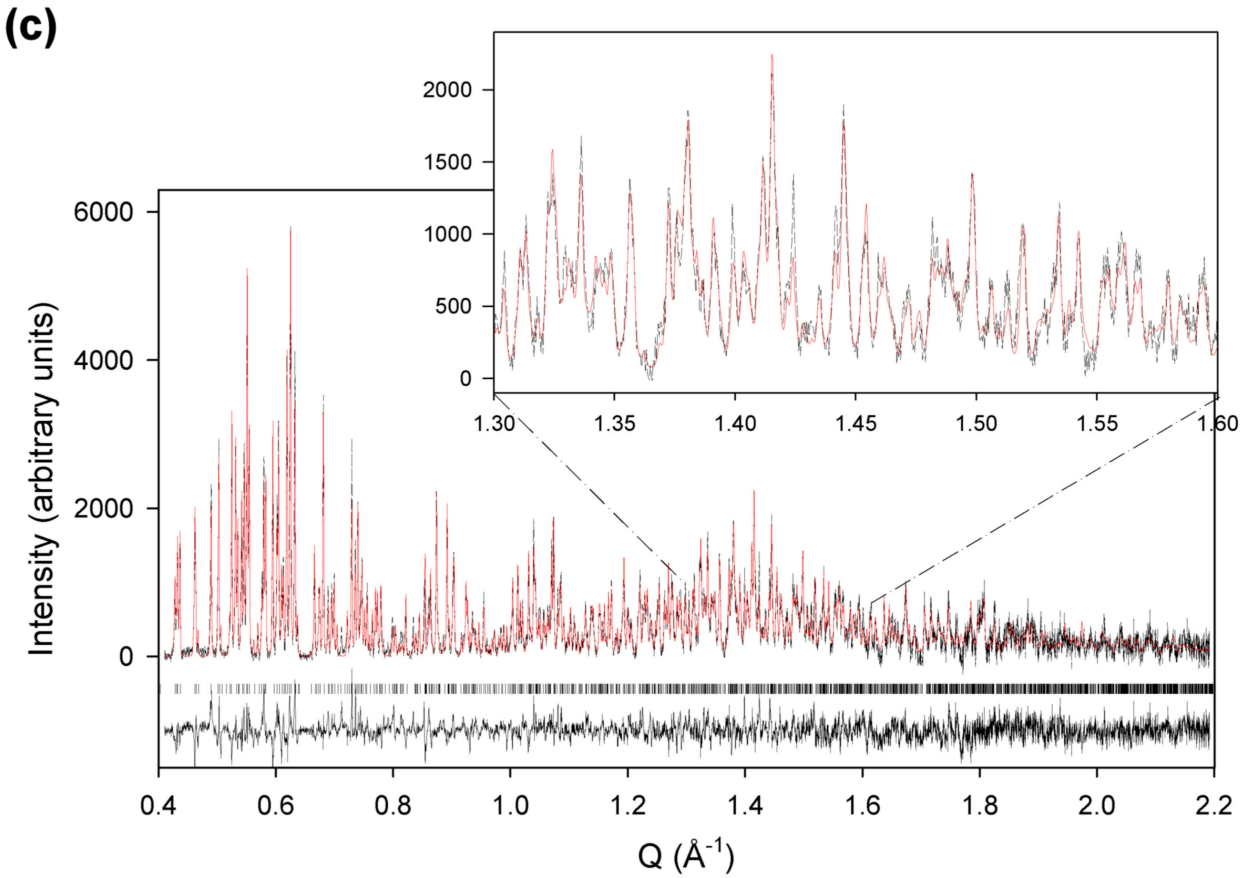


Figure S7. Final fits of the rest of the four data sets employed for stereochemically restrained Rietveld analysis. The data were collected on two samples (A and B) at 295K (ID31, (a) sample A, λ = 1.252481(32) Å, (b) sample B, λ = 1.251209(40) Å and (c) sample B, λ = 0.8012034(76) Å). The dashed black, red and lower black lines represent the experimental data, calculated pattern and the difference between experimental and calculated profiles respectively. The vertical bars correspond to Bragg reflections compatible with the refined orthorhombic structural model. The insets correspond to magnifications of the observed and calculated profiles in the Q region between 1.3 and 1.6 Å^{-1}. The background intensity has been subtracted for clarity.

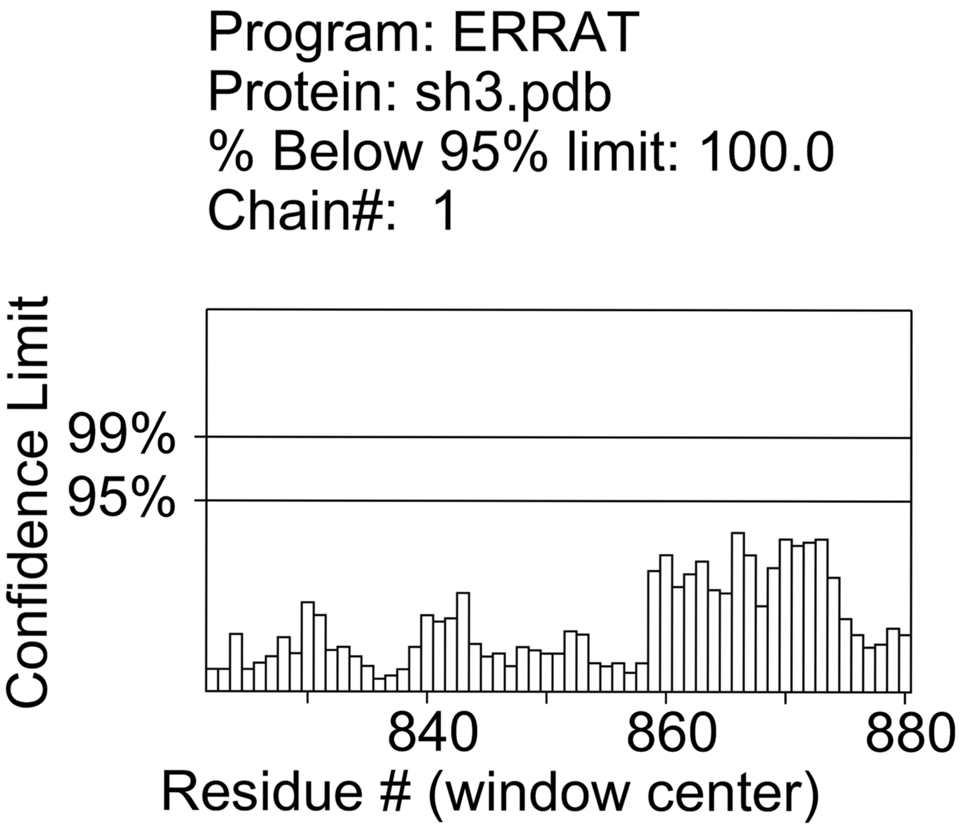


Figure S8. ERRAT-2 (*16*) results of the refined conformation of the SH3.2 domain as it was derived from the four data set restrained Rietveld refinement (below 95%).

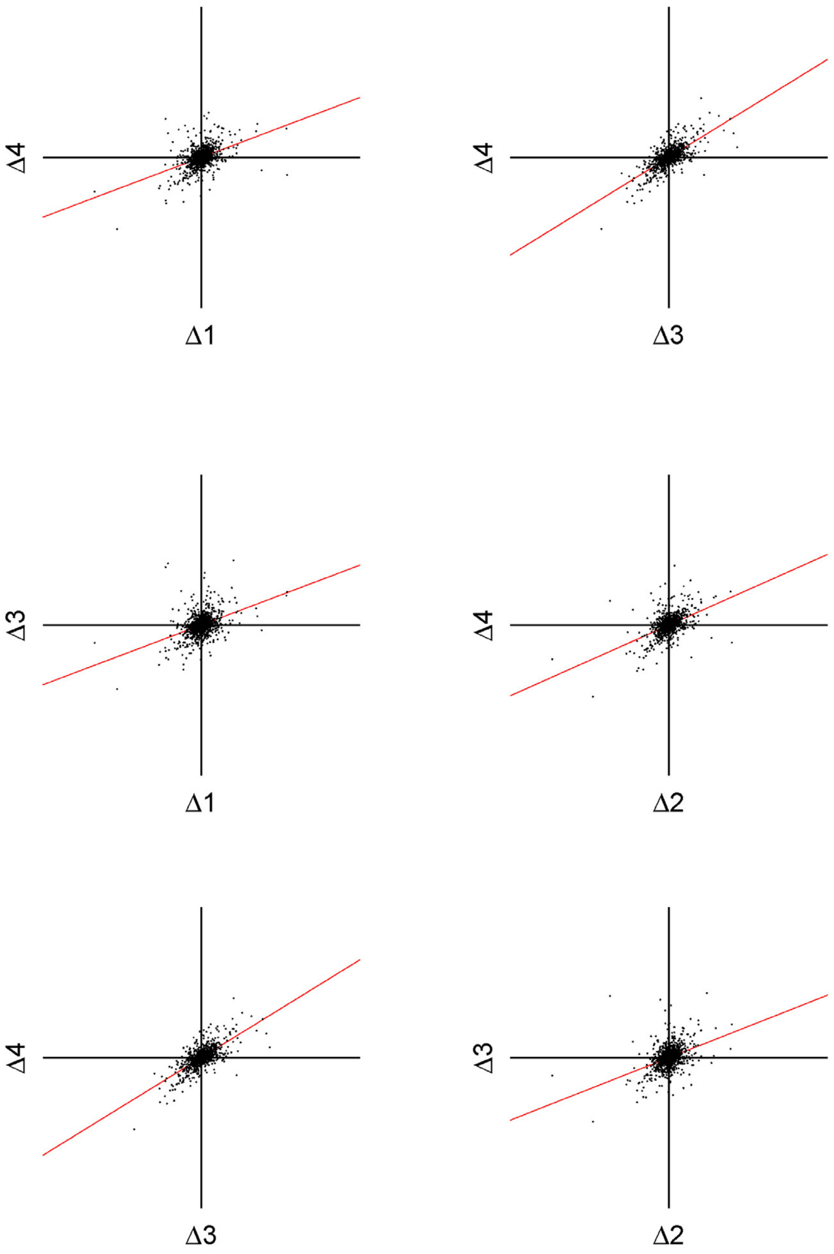


Figure S9. Scatter plots of the regression of the difference between the observed and calculated intensities for the four profiles involved in the multi-dataset Rietveld refinement (see text and table S2). The red lines indicate linear fits.

Tables

Table S1. Effective completeness for single (upper) and 4 combined data sets (lower) as the fraction of I/σ(I) "peaks". The individual columns from left to right correspond to minimum and maximum resolution (d_max and d_min respectively), total number of peaks in d-spacing range (Total), number of peaks with I/σ(I) at specific regions of sigma levels (I/σ(I) < 1σ , 1σ < I/σ(I) < 3σ, I/σ(I) > 3σ), number of unique d-spacings in the corresponding range (Poss) and percentages representing the completeness for I/σ(I) > 1σ and I/σ(I) > 3σ (1σ% and 3σ% respectively).

Single Pattern								
d_max	d_min	Total	I/σ(I)			Poss	I/σ(I)	
			<1 σ	1-3σ	>3σ		1σ%	3σ%
35.995	10.323	47	3	5	39	46	93.62	82.98
10.323	7.454	68	4	8	56	67	94.12	82.35
7.454	6.13	78	16	7	55	79	79.49	70.51
6.13	5.328	100	23	9	68	100	77	68
5.328	4.776	100	28	10	62	101	72	62
4.776	4.367	117	26	13	78	116	77.78	66.67
4.367	4.047	114	30	16	68	116	73.68	59.65
4.047	3.789	141	66	15	60	139	53.19	42.55
3.789	3.574	135	68	20	47	135	49.63	34.81
3.574	3.392	139	77	19	43	140	44.6	30.94
3.392	3.236	154	80	23	51	153	48.05	33.12
3.236	3.099	154	85	14	55	153	44.81	35.71
3.099	2.978	169	102	20	47	170	39.64	27.81
2.978	2.871	171	118	21	32	170	30.99	18.71
2.871	2.774	161	145	7	9	162	9.94	5.59

Combined fit to four patterns								
d_max	d_min	Total	I/σ(I)			Poss	I/σ(I)	
			<1 σ	1-3σ	>3σ		1σ%	3σ%
35.995	10.323	47	1	1	45	46	97.87	95.74
10.323	7.454	67	6	3	58	67	91.04	86.57
7.454	6.13	79	7	8	64	79	91.14	81.01
6.13	5.328	99	12	10	77	100	87.88	77.78
5.328	4.776	101	17	9	75	100	83.17	74.26
4.776	4.367	115	18	8	89	114	84.35	77.39
4.367	4.047	115	20	9	86	117	82.61	74.78
4.047	3.789	140	36	13	91	141	74.29	65
3.789	3.574	137	45	10	82	135	67.15	59.85
3.574	3.392	134	43	22	69	138	67.91	51.49
3.392	3.236	155	35	24	96	151	77.42	61.94
3.236	3.099	157	30	27	100	155	80.89	63.69
3.099	2.978	168	65	35	68	171	61.31	40.48
2.978	2.871	172	94	41	37	170	45.35	21.51
2.871	2.774	157	146	6	5	158	7.01	3.18

Table S2. Correlation coefficients of intensity differences at the end of the four pattern Rietveld refinement.

	ΔI_1	ΔI_2	ΔI_3	ΔI_4
ΔI_1	1			
ΔI_2	0.87721	1		
ΔI_3	0.402283	0.40207	1	
ΔI_4	0.441404	0.494194	0.707728	1

References

[1] Sivia, D. S. J. Appl. Crystalogr. **2000**, 33, 1295-1301.

[2] Larson, A. C.; Von Dreele, R. B. “General Structure Analysis System (GSAS)” (Los Alamos National Laboratory Report LAUR, **2004**).

[3] Von Dreele, R. B. J. Appl. Crystallogr. **2007**, 40, 133-143.

[4] Laskowski, R. A.; Macarthur, M. W.; Moss, D. S.; Thornton, J. M. J. Appl. Crystallogr. **1993**, 26, 283-291.

[5] Hooft, R. W. W.; Vriend, G.; Sander, C.; Abola, E. E. Nature 1996, 381, 272-272.

[6] Colovos, C.; Yeates, T. O. Protein Sci. **1993**, 2, 1511-1519.

[7] Guex, N.; Peitsch, M. C. Electrophoresis **1997**, 18, 2714-2723.