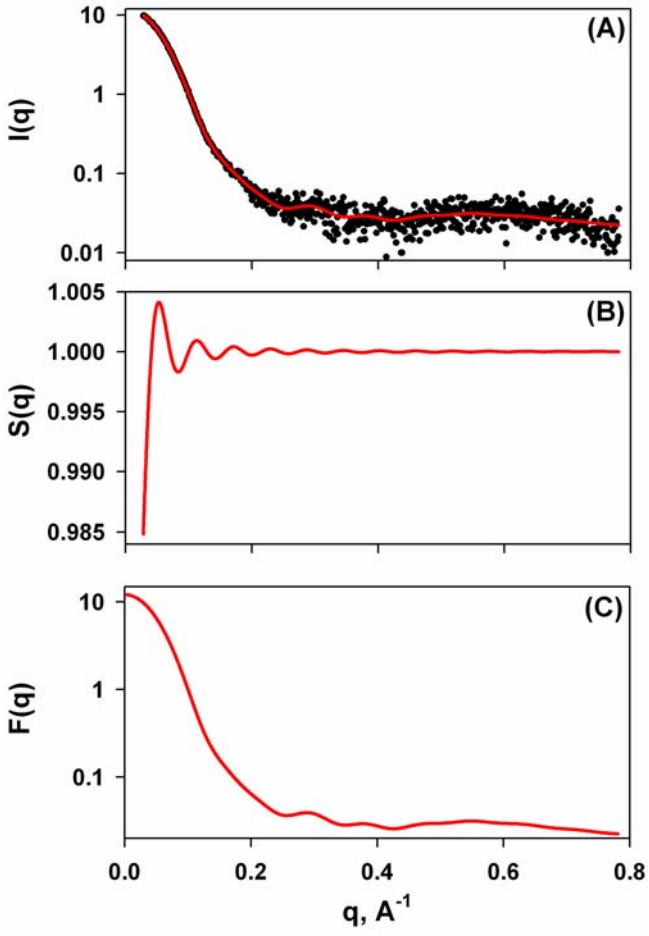
Supplementary Information for the

Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints

Alexander Grishaev[1,*], Vitali Tugarinov[2], Lewis E. Kay[2], Jill Trewhella[3], and Ad Bax[1]

Estimation of the structure factor present in the scattering data of MSG.

In order to quantify the effect of the inter-particle interactions on our experimental data, we have decomposed the data into the form and structure factors with GIFT software (Bergmann et al. 2000), using the Percus-Yevick closure relationship (Percus and Yevick 1958) for the structure factor calculation (Supplementary Figure 1). The obtained structure factor is very close to unity over the entire experimental q-range: its smallest value, at $q_{min} = 0.027$ Å^{-1}, is 0.984 and its highest value, at q=0.052 Å^{-1}, is 1.004. Relative to the ~0.7% experimental noise in our low-q data, the impact of its deviation from unity becomes negligible above 0.035 Å^{-1}



Supplementary Figure 1. Results of the joint extraction of the single-particle form factor F(q) and the inter-particle structure factor S(q) from the MSG scattering data via the GIFT program. The inter-atomic distance distribution was constructed from 20 spline functions with a maximum particle dimension of 80 Å. A Lagrange multiplier of 5.0 was found to provide the optimum balance between the fit quality and solution stability. The Percus-Yevick closure relationship was used for the structure factor calculation. Panel (A) shows the agreement between the experimental data (black dots) and the predicted scattering curve calculated as S(q)*F(q) (red line). Panels (B) and (C) show the extracted structure and form factors, respectively.

Evaluation of the uncertainty in the relative orientations of the MSG domains. Relative orientations of the four domains that constitute MSG were determined from the orientations of the molecular alignment tensors for each of the domains, obtained by SVD fits to the combined N-H^N RDC and ^{13}C' RCSA data. The relative weights of RDCs and RCSAs reflected the respective rms values in the SVD fits to the 1D8C structure. The uncertainties of the absolute orientations of each domain were determined as the rmsds between the orientations of the alignment tensors fitted to the model (error-free) data and the ones fitted to the data corrupted by additional noise and model imperfections, as described below. All SVD fits involving RCSA data were performed assuming uniform ^{13}C' CSA tensors (Cornilescu and Bax 2000). Since the site-specific values of the ^{13}C CSA tensors are expected to vary slightly (Wei et al. 2001,Markwick and Sattler 2004,Loth et al. 2005), the model ^{13}C' RCSAs were adjusted for CSA tensor variation by including a random 5 ppm rms variation of the individual in-peptide plane tensor component magnitudes and a random 3° rms variation of the orientation of these components within the peptide plane with respect to the average ^{13}C' tensor. In addition, 10,000 random noise additions were applied to the model N-H^N RDCs and ^{13}C' RCSA data. The rotation matrices (**R**) between the "noise-free" and "noise–added" fitted frames of the alignment tensors for each domain were converted into the corresponding angles as $\cos\theta=(\mathrm{Tr}(\mathbf{R})-1)/2$. The latter were transformed into the variances of the inter-domain orientations as sums of the variance for the domain in question and for the core domain.

The simulated site-specific variation of the ^{13}C' CSA results in ~7 ppb rms variation in the predicted ^{13}C' RCSAs, in addition to the ~20 ppb intrinsic data uncertainty for a combined rms error of ~21 ppb, close to the SVD-fitted rms of ~22 ppb.

Thus, the residual rms in the fit of the RCSAs is dominated by the measurement uncertainty and not by the structural noise in the X-ray model. On the other hand, the residual rms of the fit of N-H^N RDCs (~5.8 Hz compared to a measurement error of only 2 Hz) is dominated by structural noise in the X-ray model. After correction for the measurement error, this residual rms corresponds to errors of ~8° for the N-H^N vector orientations (Zweckstetter and Bax 2002), a reasonable value given the 2.0 Å resolution of the 1D8C X-ray structure. The Monte Carlo analysis yielded the following lower bounds on the rms uncertainties of the peripheral domain orientations with respect to the TIM barrel core: N-terminal domain – 1.8°; α/β domain – 1.4°; and C-terminal domain – 1.7°.

Protein fold recognition from the scattering data and a structural database.

Scattering data–driven fold recognition from a structure database was performed using the DARA server (Sokolova et al. 2003). The submitted scattering intensity data were analyzed by the server in two modes, shape analysis from lower-q data with q_{max} of 0.15 Å^{-1}, and a combination of shape and domain architecture analysis from the entire scattering curve up to a q_{max} of 0.781 Å^{-1}. In both cases, the submitted scattering data were compared against the curves predicted from the structures of 507 proteins with molecular masses between 76.3 and 87.7 kDa, which include the structure of MSG (1N8I model). In both modes the server returns the 10 best matches, ordered by decreasing similarity between the submitted and predicted scattering patterns.

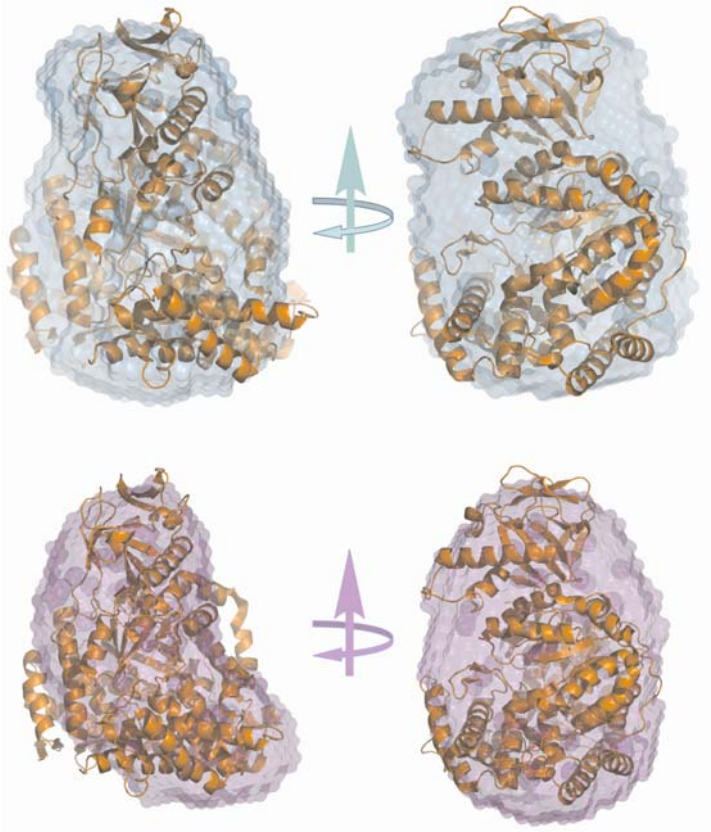
When run in the shape analysis mode, DARA misidentified our data and ranked the correct match to the MSG structure as the second-best, after a very different Yrbl phosphatase (PDB entry 1K1E). The misidentification is likely due to program's use of a fixed q_{max} = 0.15 Å^{-1}, since the scattering curve predicted from 1K1E shows significant differences with respect to the experimental MSG curve above 0.15 Å^{-1}. On the other hand, when using the entire experimental q-range in the shape/domain architecture analysis mode, DARA correctly identified MSG as the best match (χ=1.151 for the 1N8I entry), followed by an aminotransferase (χ=1.509 for the 1M32 entry). The latter is an example of a protein structure that agrees with our scattering data fairly well and has an overall shape that is very similar to the MSG, but whose domain architecture is distinctly different.

Low-resolution shape reconstructions from the scattering data.

A total of 100 independent shape reconstructions were performed with DAMMIN software (Svergun 1999), fitting the scattering data between 0.027 $Å^{-1}$ and 0.220 $Å^{-1}$, and with both real-space and inverse-space versions of GASBOR (Petoukhov and Svergun 2003) that use the entire angular range up to 0.781 $Å^{-1}$. Both methods represent the macromolecule as a collection of uniformly and densely packed spheres. The $P(r)$ functions used to obtain the target $I(q)$ for the DAMMIN and the inverse-space GASBOR, or as a target for the real-space GASBOR reconstructions were calculated from the experimental data using GNOM (Svergun 1992) with d_{max} set to 80 Å. The individual reconstructions were averaged using DAMAVER (Volkov and Svergun 2003) and superimposed on the X-ray model 1D8C with SUPCOMB (Kozin and Svergun 2001). Normalized spatial discrepancy, or NSD (Kozin and Svergun 2001), was used to quantify the agreement between the obtained individual reconstructions and the Cα coordinates of the 1D8C X-ray model. This measure was developed for comparing models of different intrinsic resolutions where metrics such as rmsd are inapplicable, with values smaller than 1.0 indicating close structural proximity.

The P(r) distribution from the data with higher q_{max}, while being quite similar to the one generated from the medium-q_{max} data, exhibits additional fine detail, presumably reflecting the expanded fitted angular range (Figure 1B, main text). However, it proved difficult to detect visually the benefits of fitting the expanded q-range when comparing the two reconstructions with the high-resolution X-ray structure of MSG (Supplementary Figure 2). In fact, the higher-q_{max} models from GASBOR gave somewhat worse

agreement (normalized spatial discrepancy (Kozin and Svergun 2001), or NSD = 1.176±0.014 for the inverse-space mode and NSD = 1.146±0.030 for the real-space mode) with the C^{α}-only geometry of 1D8C than the medium-q_{max} models from DAMMIN (NSD = 0.958±0.030).



Supplementary Figure 2. Superimposition of the low-resolution shapes generated by fitting the data in the q-interval between 0.027 $Å^{-1}$ and 0.220 $Å^{-1}$ (top panels) and in the q-interval between 0.027 $Å^{-1}$ and 0.781 $Å^{-1}$ (bottom panels) with the1D8C X-ray structure of MSG. The two views, generated with PyMOL (DeLano 2002), are related by ~140° rotations around the axis shown.

Supplementary Table 1. Structural statistics for the alternative order of the first two stages of the structure refinement (H-bonding PMF followed by intermediate angle scattering data).

Refinement Stage	rmsd to 1D8C [a]				
	all	core	N-term	α/β	C-term
Model 0 [b]	4.50±0.57	3.71±0.30	1.63±0.15	2.31±0.28	3.49±0.43
Model I' [b]	4.57±0.53	3.60±0.27	1.34±0.13	2.27±0.25	3.27±0.36
Model II' [b]	3.58±0.21	3.10±0.19	1.47±0.13	2.31±0.24	3.00±0.27

[a] Rmsd values in Å are calculated over the backbone C/N/C^{α} atoms of residues 3-722 (all residues); 116-132,266-295,334-550 (core domain); 3-88 (N-terminal domain); 135-262,296-333 (α/β domain); 589-722 (C-terminal domain); calculated with MolMol v 2.1K (Koradi et al. 1996).

[b] Model 0 is obtained by fitting NMR data and the light-scattering-derived R_G; Model I' is obtained by adding to the data for Model 0 the H-bonding pseudo-potential; Model II' is obtained by adding to the data for Model I' the scattering data within the *q*-range of 0.027-0.220 $Å^{-1}$

References

Bergmann A, Fritz G, Glatter O (2000) Solving the generalized indirect fourier transformation (GIFT) by Boltzmann simplex simulated annealing (BSSA). J Appl Cryst 33: 1212-1216

Cornilescu G, Bax A (2000) Measurement of proton, nitrogen, and carbonyl chemical shielding anisotropies in a protein dissolved in a dilute liquid crystalline phase. J Am Chem Soc 122: 10143-10154

DeLano WL (2002) The PyMOL molecular graphics system.

Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: A program for display and analysis of macromolecular structures. J Mol Graph 14: 51-55

Kozin MB, Svergun DI (2001) Automated matching of high- and low-resolution structural models. J Appl Cryst 34: 33-41

Loth K, Pelupessy P, Bodenhausen G (2005) Chemical shift anisotropy tensors of carbonyl, nitrogen, and amide proton nuclei in proteins through cross-correlated relaxation in NMR spectroscopy. J Am Chem Soc 127: 6062-6068

Markwick PRL, Sattler M (2004) Site-specific variations of carbonyl chemical shift anisotropies in proteins. J Am Chem Soc 126: 11424-11425

Percus JK, Yevick GJ (1958) Analysis of classical statistical mechanics by means of collective coordinates. Phys Rev 110: 1-13

Petoukhov MV, Svergun DI (2003) New methods for domain structure determination of proteins from solution scattering data. J Appl Cryst 36: 540-544

Sokolova AV, Volkov VV, Svergun DI (2003) Prototype of a database for rapid protein classification based on solution scattering data. J Appl Cryst 36: 865-868

Svergun DI (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J Appl Cryst 25: 495-503

Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys J 76: 2879-2886

Volkov VV, Svergun DI (2003) Uniqueness of ab initio shape determination in small-angle scattering. J Appl Cryst 36: 860-864

Wei YF, Lee DK, Ramamoorthy A (2001) Solid-state C-13 NMR chemical shift anisotropy tensors of polypeptides. J Am Chem Soc 123: 6118-6126

Zweckstetter M, Bax A (2002) Evaluation of uncertainty in alignment tensors obtained from dipolar couplings. J Biomol NMR 23: 127-137