Improving protein template recognition by using small angle X-ray scattering profiles

Marcelo Augusto dos Reis[1, 2], Ricardo Aparicio[2] and Yang Zhang[1]

[1]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, USA
[2]The Institute of Chemistry, University of Campinas, Campinas, SP, 13083-970, Brazil

SUPPORTING MATERIALS

TABLE S1. Summary of various SAXS profile matching scores in template prioritizations.

Scheme	**Scoring Function: $F_{SAXS}(i,j)$**	**<PCC>[a]**		**<TM-score>**	
		SAXS	**SAXSTER**	**Top1**	**Top5**
I[b]	$\frac{1}{n}\sum_{k=1}^{n}\left\{\frac{I_i(q_k)-I_j(q_k)}{\sigma(q_k)}\right\}^2$	0.34	0.47	0.5414	0.6042
II	$\frac{\sum_{k=1}^{n}\lvert I_i(q_k)-I_j(q_k)\rvert}{\sum_{k=1}^{n}\lvert I_i(q_k)\rvert}$	0.39	0.31	0.5431	0.6063
III	$\frac{\sum_{k=1}^{n}\lvert log[I_i(q_k)]-log[I_j(q_k)]\rvert}{\sum_{k=1}^{n}\lvert log[I_i(q_k)]\rvert}$	0.36	0.31	0.5455	0.6078
IV	$\sum_{k=1}^{n} q_k\{I_i(q_k)-I_j(q_k)\}^2$	0.38	0.36	0.5423	0.6037
V	$\sum_{k=1}^{n} q_k\{log[I_i(q_k)]-log[I_j(q_k)]\}^2$	0.30	0.30	0.5463	0.6085
VI	$\frac{\sum_{k=1}^{n} {q_k}^2\lvert I_i(q_k)-I_j(q_k)\rvert}{\sum_{k=1}^{n} {q_k}^2\lvert I_i(q_k)\rvert}$	0.37	0.30	0.5468	0.6085
VII	$\sum_{k=1}^{n}\{p_i(r)-p_j(r)\}^2$	0.38	0.32	0.5429	0.6060
VIII	$log\left\{1-corr\left(I_i(q),I_j(q)\right)\right\}$	0.36	0.62	0.5430	0.6083
IX[c]	$log\left\{1-corr\left(p_i(r),p_j(r)\right)\right\}$	0.35	0.61	0.5484	0.6095

[a]Average Pearson correlation coefficient between TM-score of the templates and Z-score of the scoring functions based on SAXS score only (SAXS) or SAXS plus threading score (SAXSTER).
[b]σ is the experimental error. In case of theoretical curves $\sigma=I(q)\times(q+0.15)\times0.3$.

[c] The function *corr*(*f*,*g*) is the Pearson Correlation Coefficient between *f* and *g* defined in Eq. 11 in the main text.

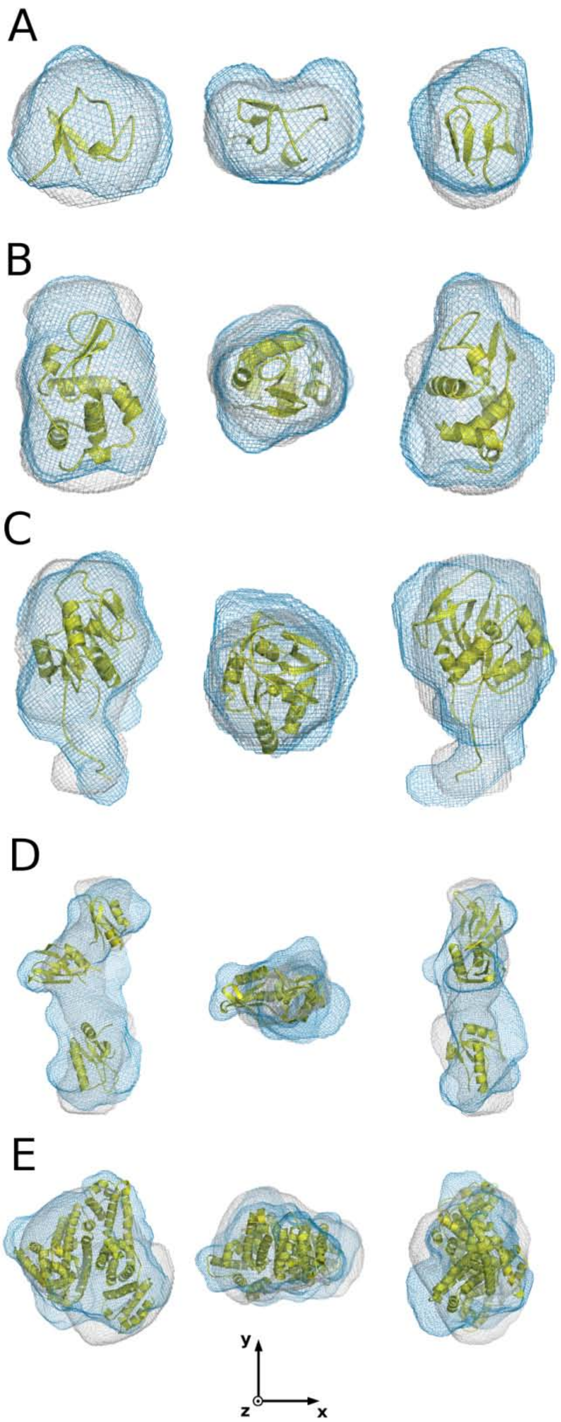

Figure S1. Low-resolution envelopes obtained by DAMMIF and template-based method. DAMMIF and template-based envelopes are shown in gray and blue, respectively, and target structures are in cartoon representation. For each case, the center and right views were rotated counterclockwise by 90^0 around x- and y-axes from the left view, respectively. (A) 1RBDGP; (B) 6LYZ; (C) 1AMIGP; (D) 1U2FKP; (E) HSA.

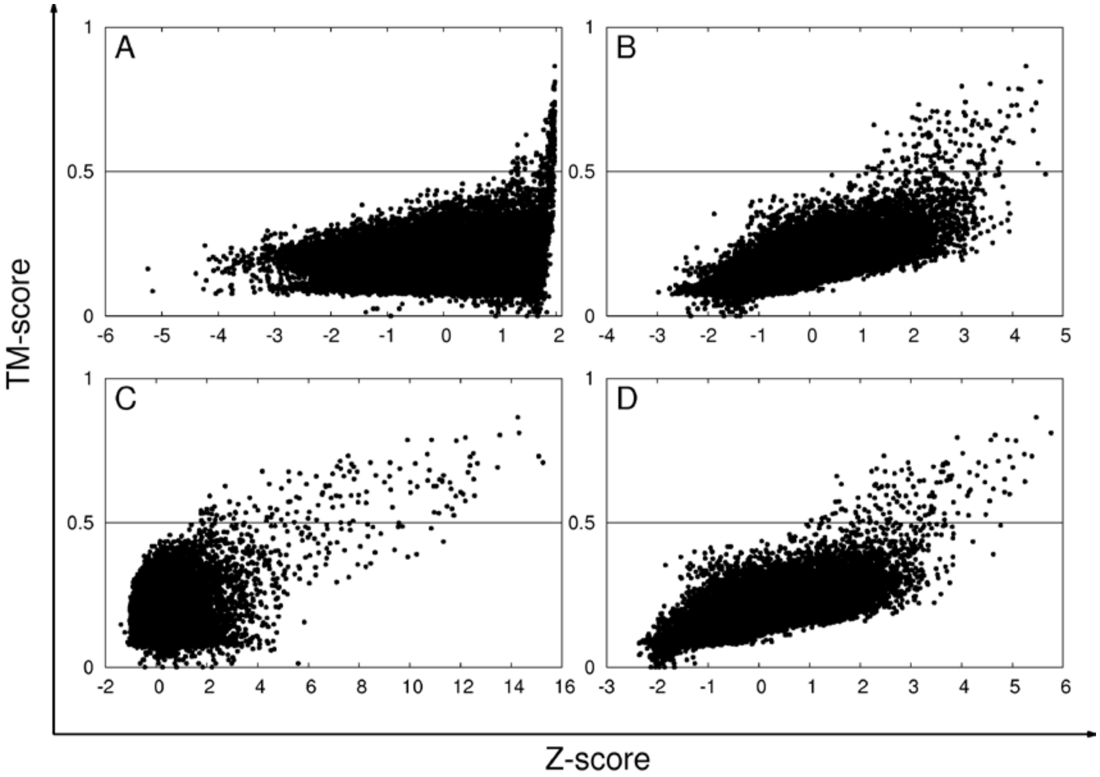

Figure S2. Illustration of TM-score versus Z-score for the target protein 1IC2A. Left column shows data using SAXS term only and Right column using SAXS+MUSTER terms. (A) and (B) use the SAXS function Scheme I whereas (C) and (D) use Scheme IX in Table S1.

Gap size and SAXS sensitivity

When using random walks (RW) to model unaligned regions from threading alignments, we considered up to 1,000 walks per step until all missed residues are filled in the template structure. This number of walks appeared sufficient to complete the structure construction for all gaps encountered in our benchmark test.

An import issue to be explored is the sensitivity of the SAXS score to the gap size with structures built by RW. To illustrate this issue, we consider the *Lysozyme* protein (PDB ID: 6LYZ) as a model. First, from the crystal structure we removed the residues of the N-terminal one by one and then reconstructed the structure by RW in order to mimic threading gaps of several sizes. For each model structure, we calculated its respective SAXS profile and calculated χ against the experimental curve. Similarly, the approach was performed in the gaps opened at the C-terminal. To test the gap effect at the core region of the protein, we took the closest amino acid to the center of the mass (Leu56) as the starting residue, and increased the gap in both sides towards to the N- and C-terminals. Such gap modeling procedure was repeated by a number of times with the average χ values versus the gap size shown in Fig. S3. As expected, the models by RW become less reliable as the unaligned gaps increase. A similar tendency of χ dependence on gap sizes was observed in both core and tail regions. Overall, our results suggest that models with gap filled by <10 continuous residues are indistinguishable from our coarse-grained simulation. But when the gap size is larger than 10 residues, the RW modeling becomes less reliable than the coarse-grained simulations as demonstrated by the χ values.

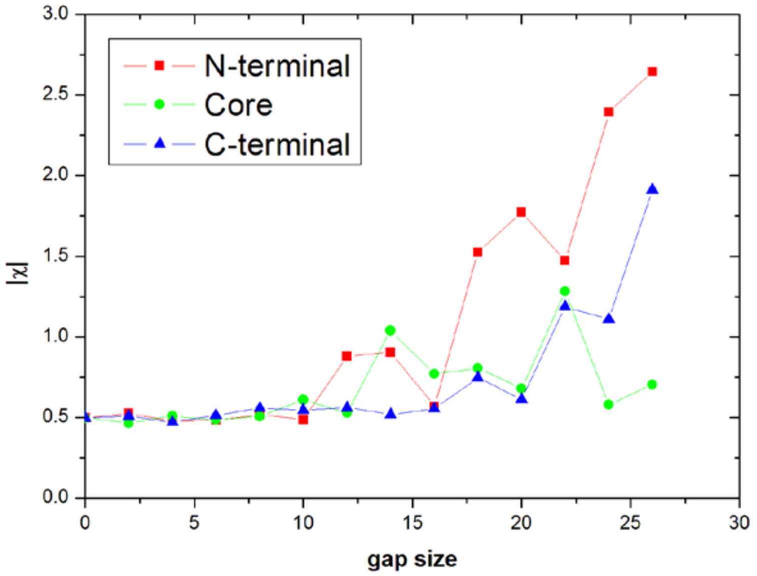

Figure S3. Distribution of χ between calculated and experimental SAXS profiles for *Lyzozyme* protein. The calculated SAXS profiles were obtained from structures with different gap sizes filled by Random Walks.