Supplementary Materials

TM-score probabilities for proteins filtered with radius gyration difference.

Figure S1A shows the conditional probabilities of the TM-score for the protein pairs sharing the same fold or different folds from three raw datasets (SCOP, CATH, and the consensus), compared to those from three filtered datasets where protein pairs with a radius of gyration (RG) difference >10% are excluded. The solid ‘squares’, ‘triangles (up)’ and ‘stars’ represent, respectively, the results of the raw datasets from SCOP, CATH and the consensus (SCOP&CATH) with 746,420 2,769,868 and 186,359 same-fold protein pairs. The open ‘squares’, ‘triangles (up)’ and ‘stars’ are those from the filtered datasets with 449,281 (SCOP_RG), 1,360,782 (CATH_RG) and 117,446 (SCOP&CATH_RG) same-fold protein pairs, respectively. The open ‘diamonds’, ‘triangles (down)’ and ‘circles’ correspond to the different fold pairs defined by SCOP, CATH and SCOP&CATH, respectively.

Figure S1B shows the posterior probabilities for proteins having the same fold, given the TM-score of the structure pairs. The solid points represent the results from the raw datasets and the open points are those from the filtered datasets. The ‘squares’, ‘triangles’ and ‘stars’ represent the data with fold defined by SCOP, CATH, and a consensus of SCOP and CATH.

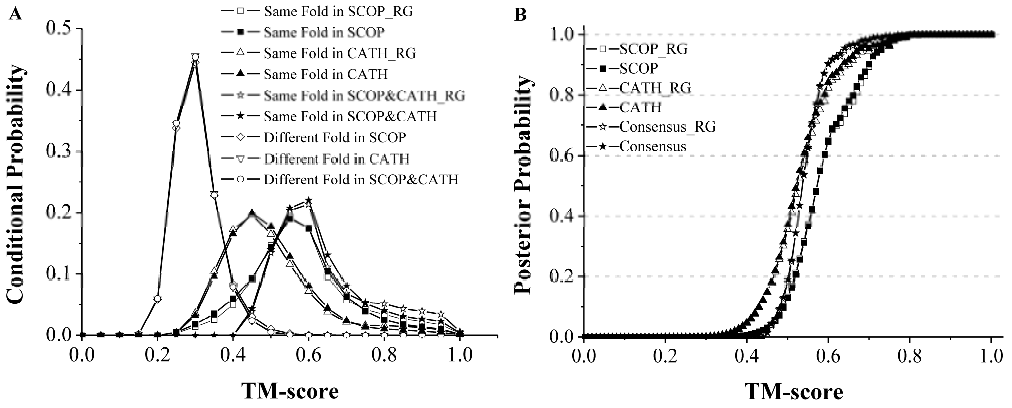

Figure S1. (A) Conditional probabilities of TM-score for protein pairs in the same or different fold families. (B) Posterior probability for protein pairs sharing the same fold given the TM-score in filtered or raw datasets. The structural filter is conducted by the difference in radius of gyration of two proteins in pair.

TM-score probabilities for proteins with long coiled tails truncated.

The data in Figure S2 are similar as those in Figure S1 but here the outlier structures are filtered by truncating the long tails (LT). Thus, the numbers of protein pairs in the raw datasets and the filtered datasets are the same.

Figure S2A shows the conditional probabilities of the TM-score for the same fold and different fold protein pairs in three raw datasets, where the solid ‘squares’, ‘triangles (up)’, and ‘stars’ represent the results from raw datasets and the corresponding open points represent that from the filtered datasets. The open ‘diamonds’, ‘triangles (down)’ and ‘circles’ are the results from the different fold pairs as defined by SCOP, CATH and SCOP&CATH, respectively.

Figure S2B shows the posterior probabilities for proteins in the same fold families. The open points are the results from the filtered datasets, compared to those from the raw datasets (solid points).

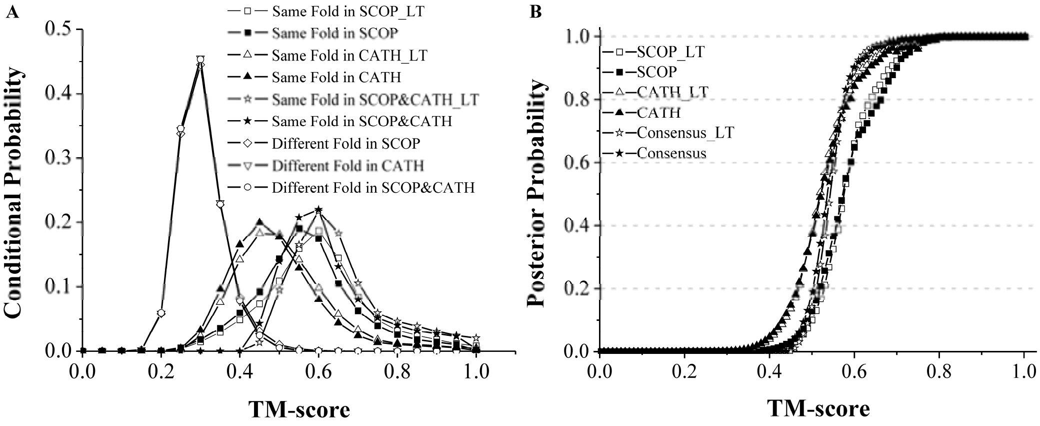

Figure S2. (A) Conditional probabilities of TM-score for protein pairs in the same or different fold families. (B) Posterior probabilities of protein pairs sharing the same fold given the TM-score in filtered or raw datasets. The structural filter is conducted by truncating the outlier long tails.