Supporting Information

Skolnick et al. 10.1073/pnas.0907683106

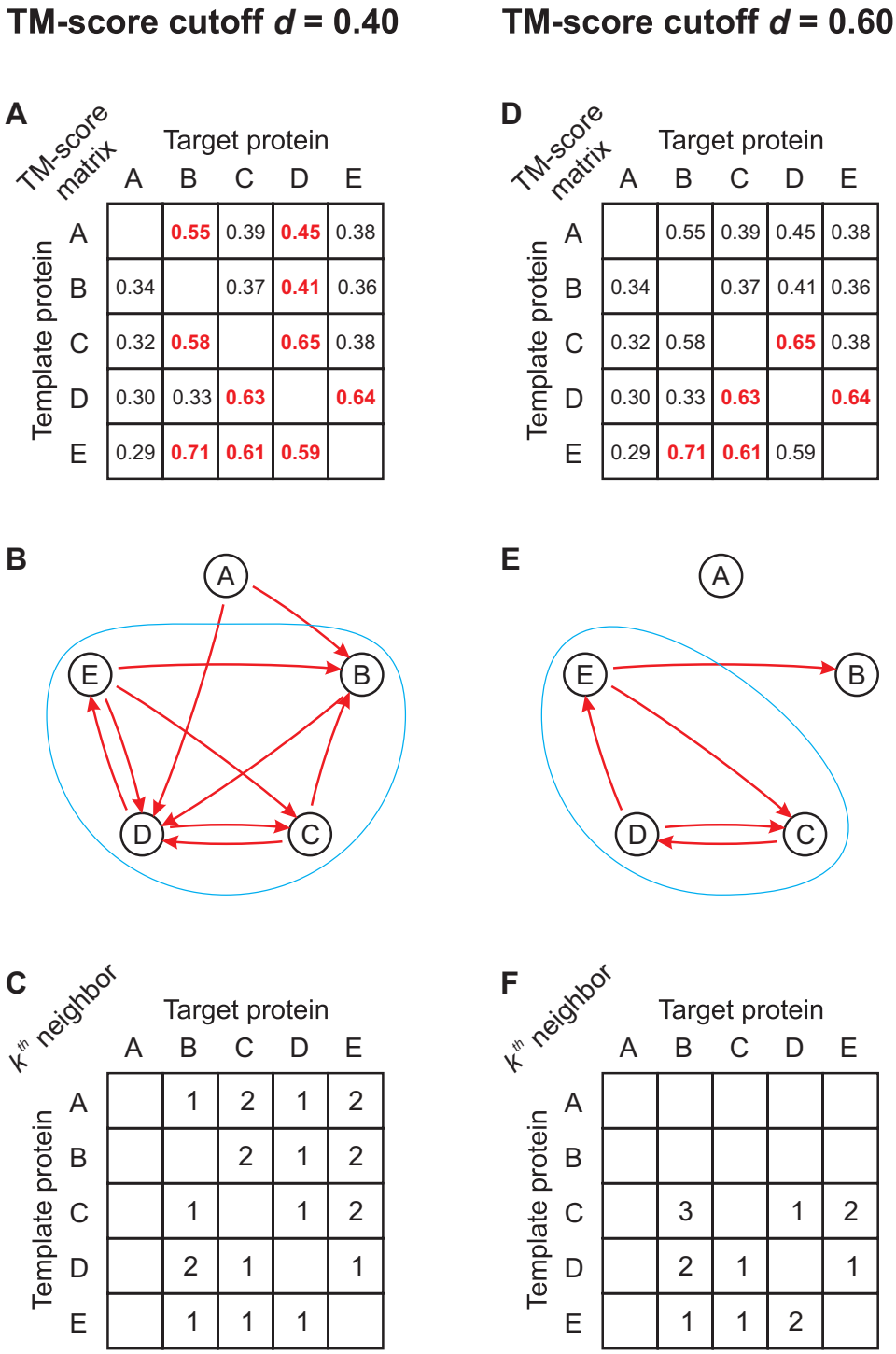

Fig. S1. Graphical representation of structural relationships between proteins. For TM-score cutoffs $d = 0.40$ (*A–C*) and $d = 0.60$ (*D–F*), we show (*i*) the matrix of TM-scores between template and target proteins (*A* and *D*), where the TM-scores greater or equal to the corresponding TM-score cutoff are highlighted in red; (*ii*) the graph derived from the TM-score matrix, representing the structural relationships between template and target proteins (*B* and *E*), with the LSCC of the graph contained inside the blue curve; and (*iii*) the matrix representing the neighboring order k between each template and target protein (*C* and *F*). For the sake of simplicity, this figure does not show the distinction between small proteins (up to 200 residues) or large proteins (200–300 residues, acting as structural bridges), which can be represented as 2 possible colors associated with each vertex in the graph (see ''Random Directed Digraphs'' in *Methods* in the main text).

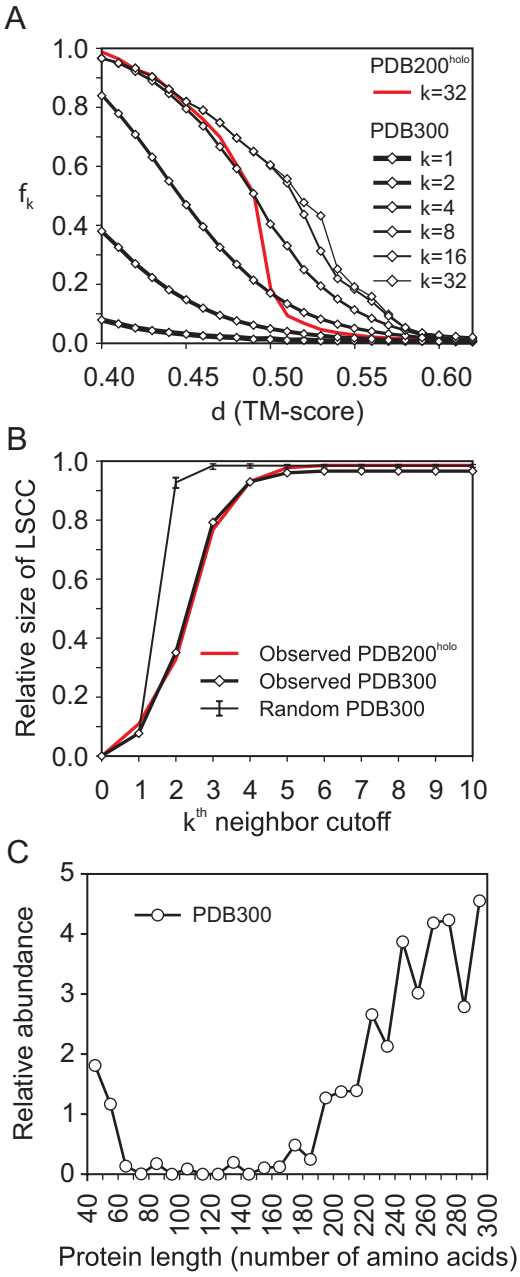

Fig. S2. (*A*) Mean fraction of proteins in the PDB300 set (black lines with diamonds) or in the PDB200^{holo} set (red line) that are no more than k^{th} neighbors (f_k) and whose first neighbors have a TM-score $\geq d$. (*B*) Relative size of the LSCC in the PDB300 set (black lines) or in the PDB200^{holo} set (red line) as a function of the k^{th} neighbor cutoff, at $d = 0.40$. The thick lines corresponds to the values observed in the original PDB300 set (black line with diamonds) or PDB200^{holo} set (red line), whereas the thin line indicates the median values obtained from 2,000 randomly generated digraphs with the same number of nodes, edges, and first-order connectivity per node as in the original PDB300 set (error bars indicate minimum and maximum values from the 2,000 random graphs). (*C*) Length distribution of proteins not belonging to the LSCC, at $d = 0.40$, relative to all proteins in the PDB300 set. The relative abundance is calculated as the fraction of the total number of proteins excluded from the LSCC that fall in a given interval of protein length divided by the fraction of the total number of proteins members of the set falling in that same interval of protein length.

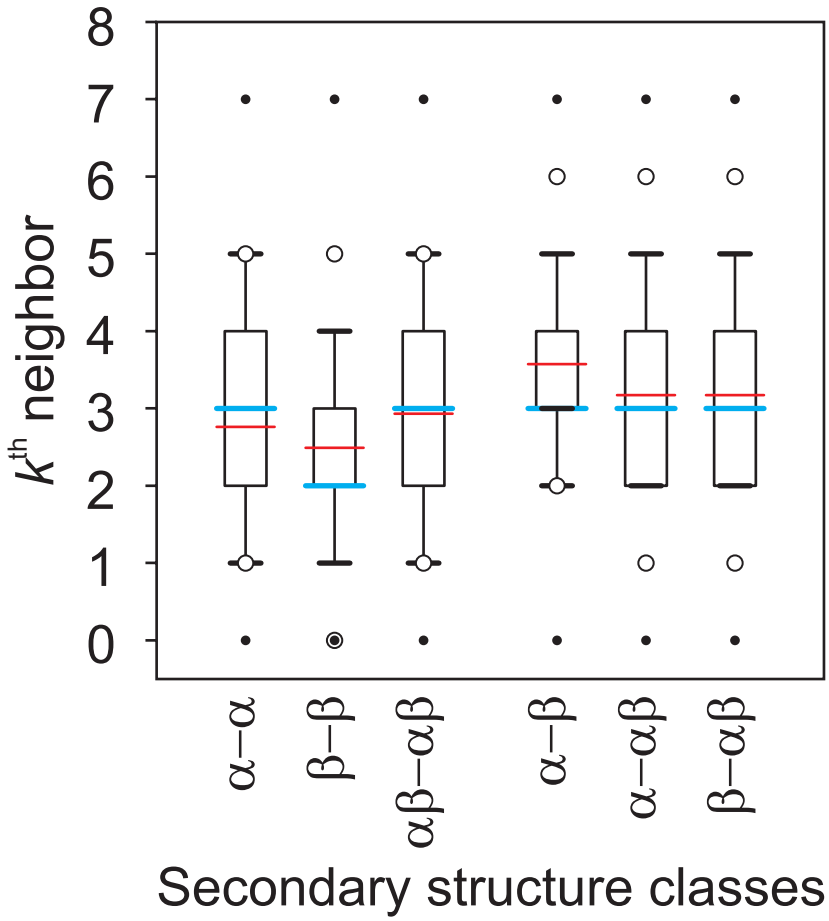

Fig. S3. Distribution of the length of the shortest path *k*, linking all proteins pairs in the PDB200x set, discriminated by secondary structure class. The statistics represented in each box-and-whisker plot are the first percentile (filled circle, *Bottom*), fifth percentile (○, *Bottom*), 10th percentile (whisker, *Bottom*), 25th percentile (box, *Bottom*), median (thick blue line), average (thin red line), 75th percentile (box, *Top*), 90th percentile (whisker, *Top*), 95th percentile (○, *Top*), and 99th percentile (filled circle, *Top*).

Other Supporting Information Files

SI Appendix