Supporting Information Appendix

Figure S1

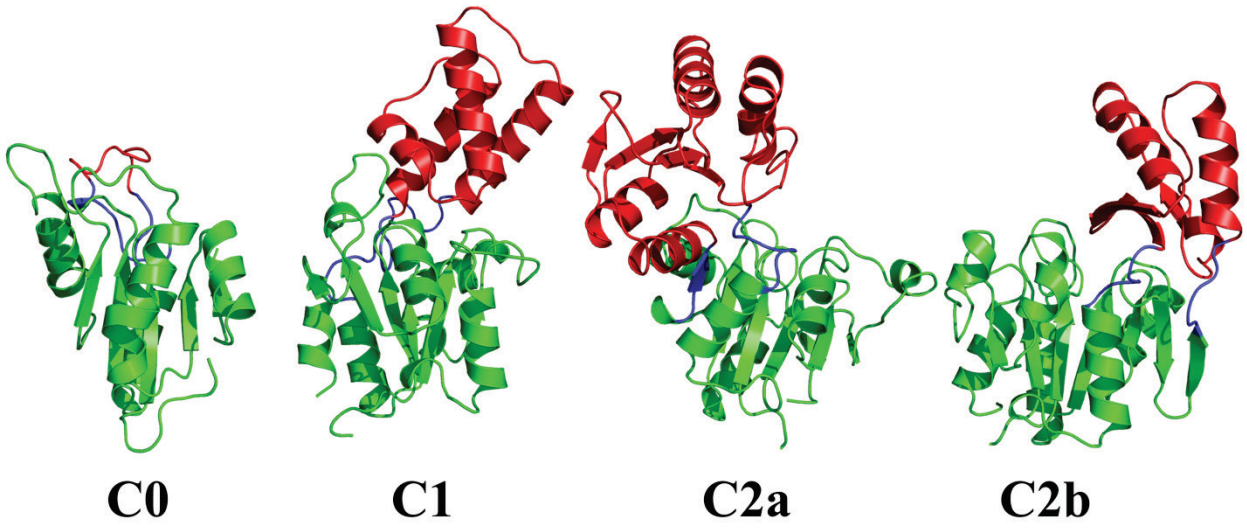

FIGURE S1. Typical HADSF architecture illustrated using representative structures. The different cap domain inserts C0, C1, C2a and C2b are represented by PDB codes 1LTQ, 2HSZ, 2C4N and 1L6R, respectively. Conserved core domain is shown in green, inserted cap domain in red and flexible linker region in blue.

Figure S2

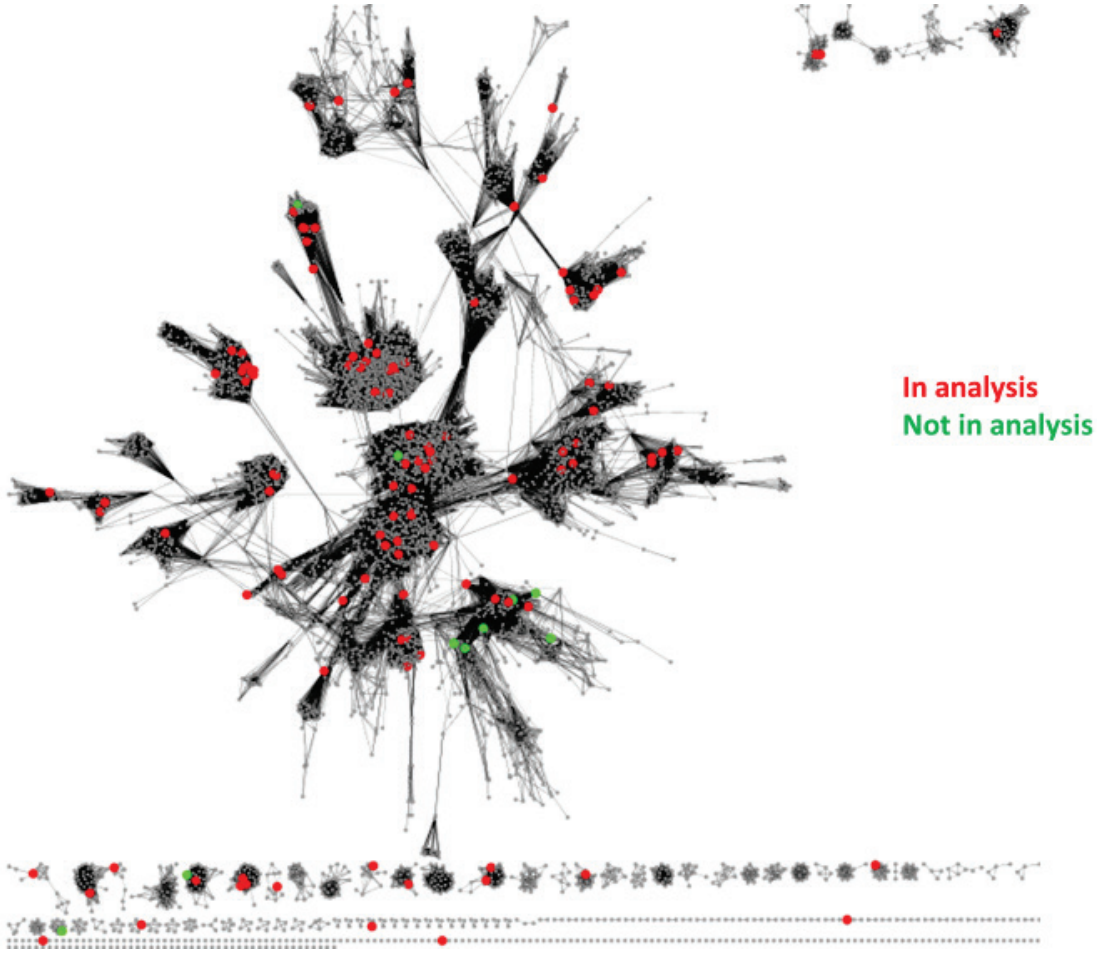

FIGURE S2. Representative sequence similarity network for the HADSF with each node representing all members with sequence identity of ≥ 40% and edges connecting those nodes with BLAST e-value $<10^{-20}$. Each protein structure which fit the selection criteria is highlighted in red and the remaining structures in green. Any bias in this network due to the presence of close homologs can be ruled out as the vast majority of these nodes share < 20% sequence identity to any other node. The network was visualized using Cytoscape version 2.8 with the yFiles organic layout scheme.

Figure S3

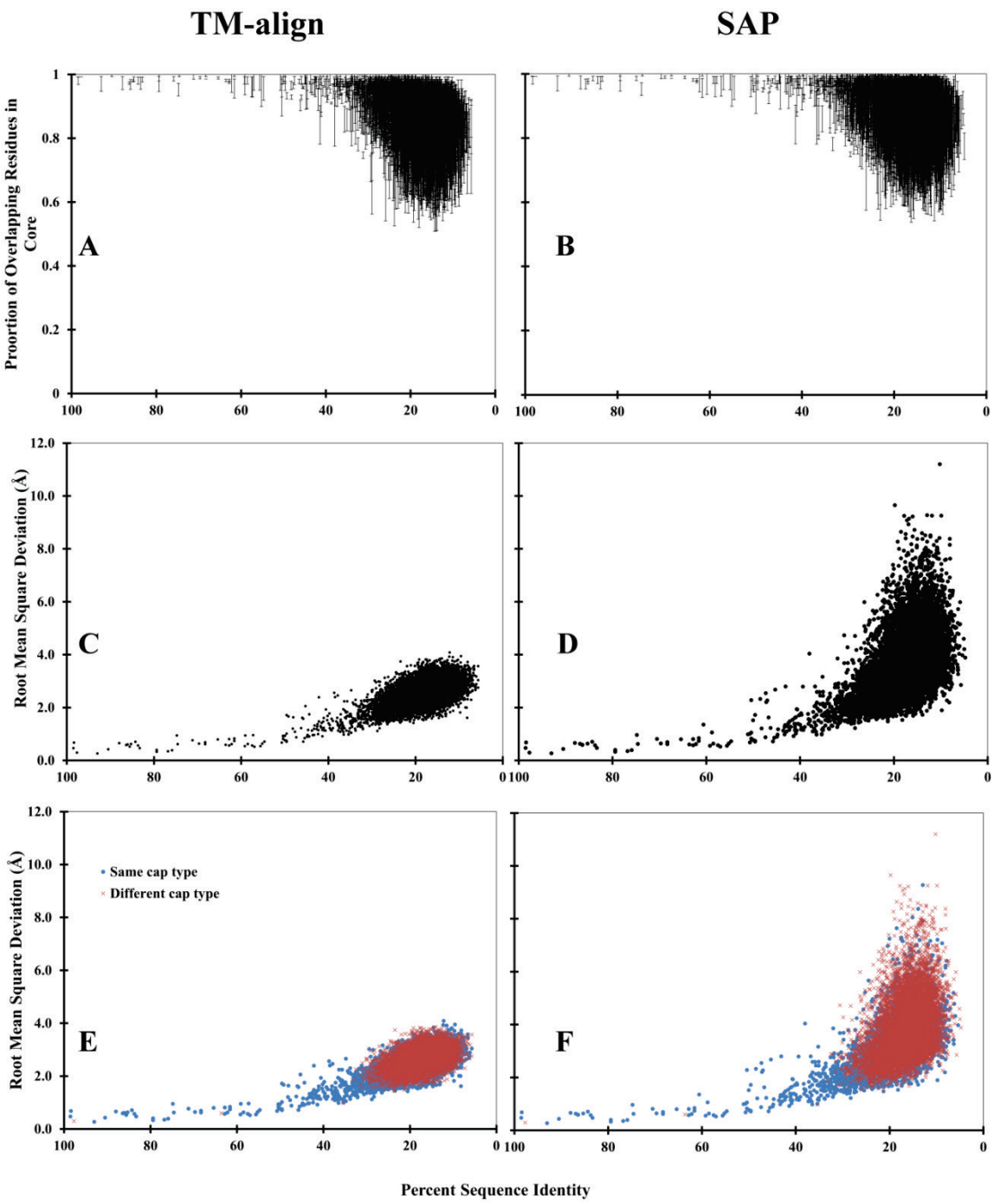

FIGURE S3. Estimating relationship between sequence and structure divergence using two pair-wise structural alignment programs. *A,C and E* represent data for TM-align while *B,D* and *F* represent data generated by SAP.

Figure S4

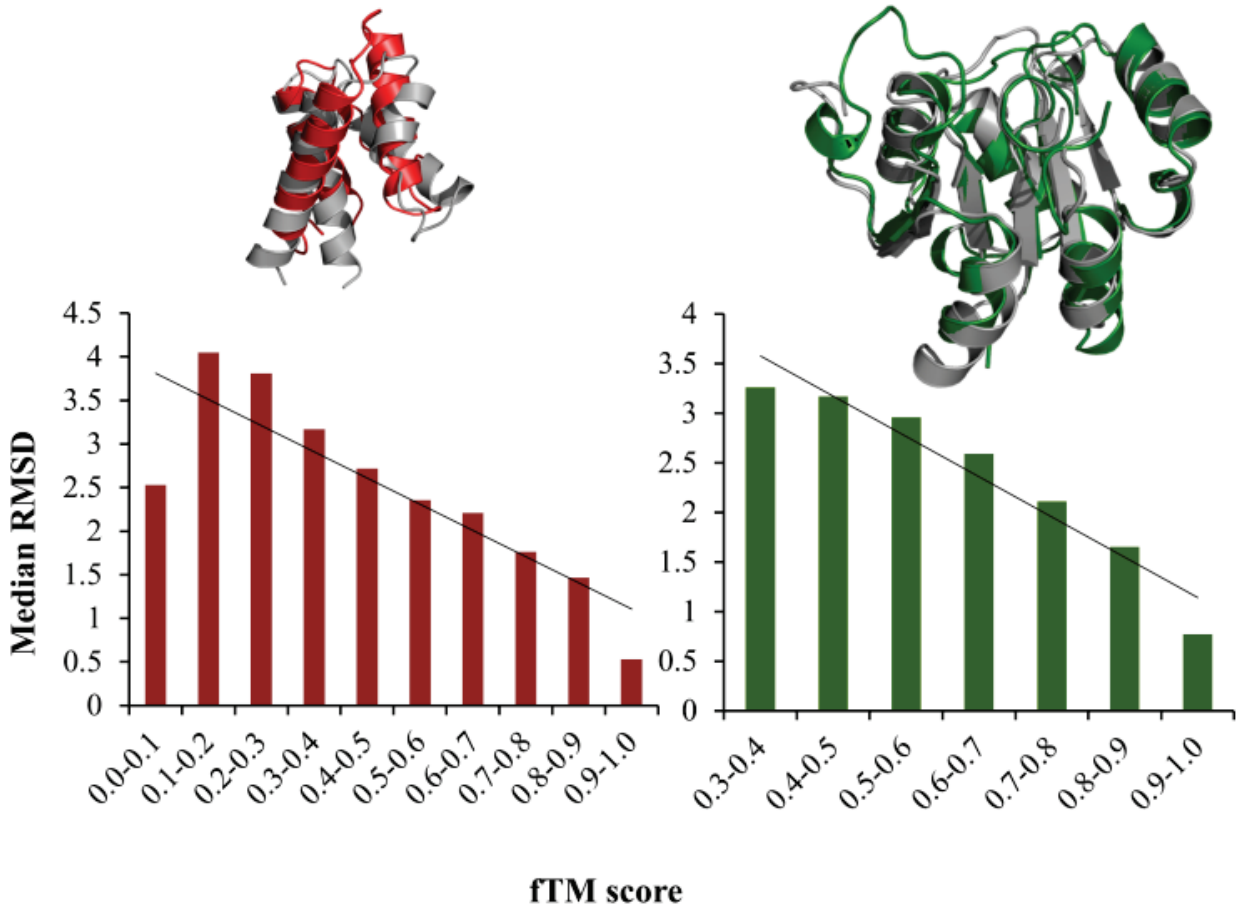

FIGURE S4. Distribution plot of HADSF pairwise cap and core structural comparisons, shown in the left and right panel respectively, illustrating the correlation between RMSD and fTM score (using TM-align).

Figure S5

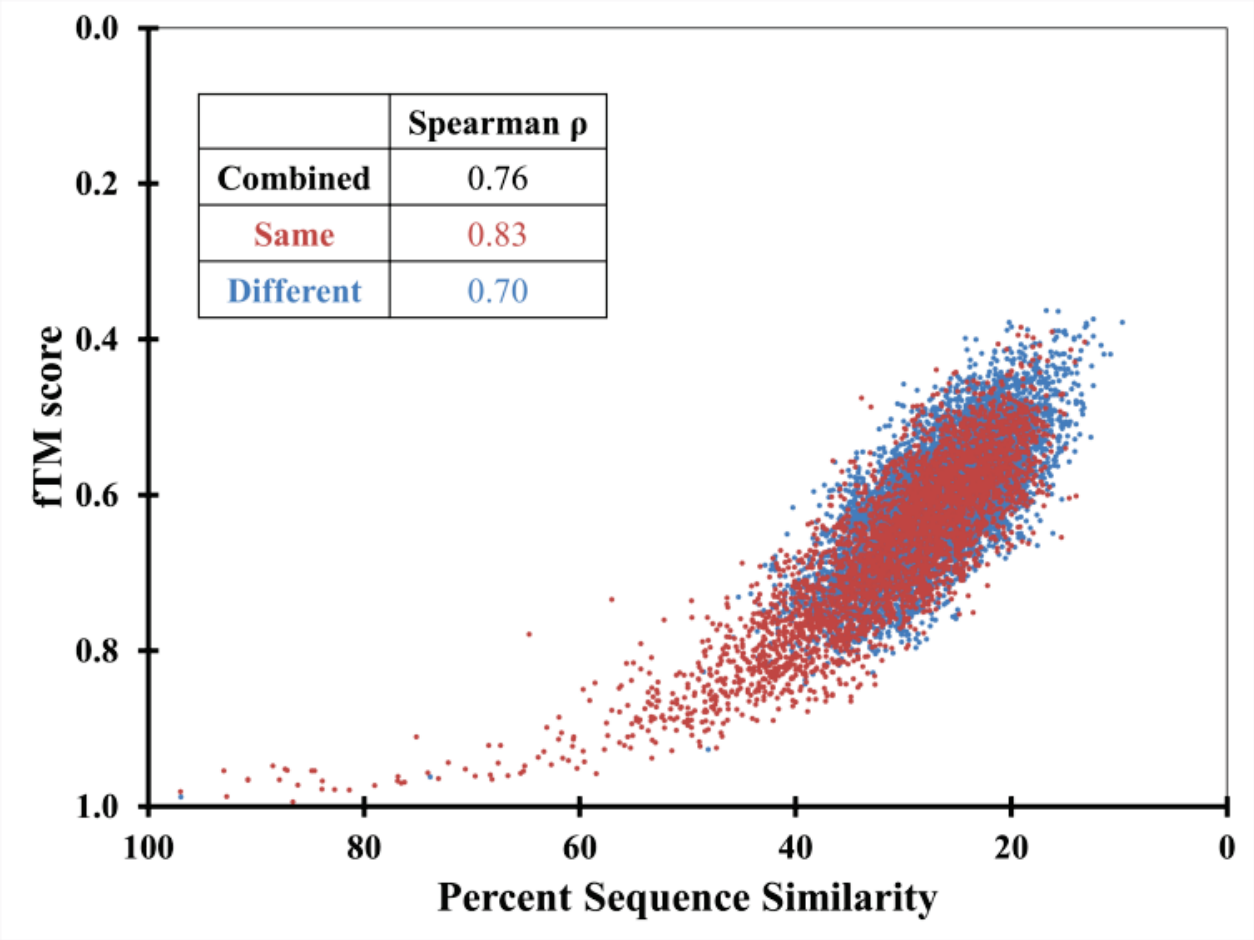

FIGURE S5. Plot of sequence divergence (percent sequence similarity) versus structure divergence (fTM score using TM-align) for HADSF core domains in the same and different cap types (inset: Spearman's rank correlation coefficient for the two sets). Similar residues are classified as aromatic (F,Y,W), aliphatic (A,V,I,L), positive (R,K,H), negative (D,E) and, polar (N,Q,C,M,S,T).

Figure S6

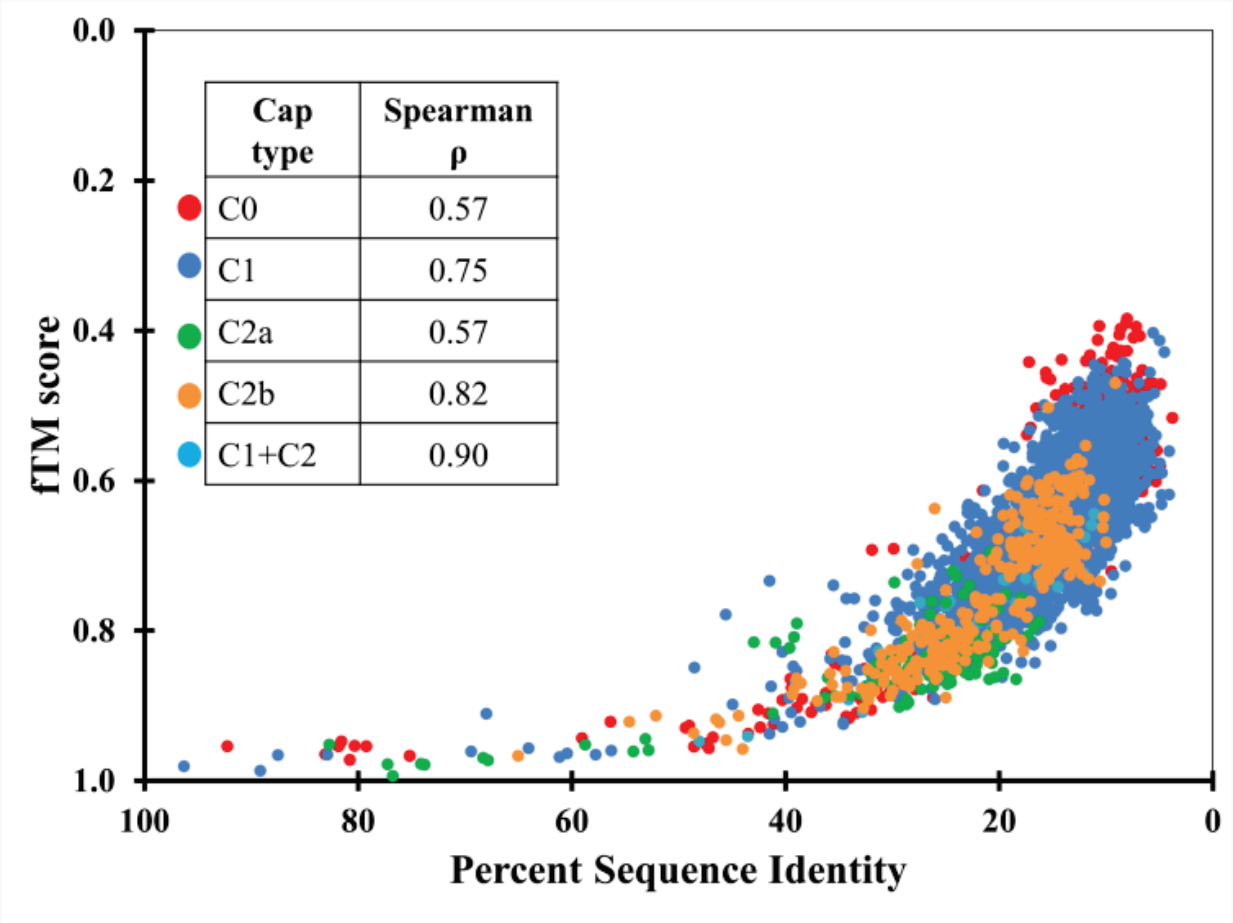

FIGURE S6. Plot of sequence divergence (percent sequence identity) versus structure divergence (fTM score using TM-align) for HADSF core domains with the same cap type. Data points are colored according to the cap type. Inset shows corresponding Spearman's rank correlation coefficients (all p-values < 10^{-10}).

Figure S7

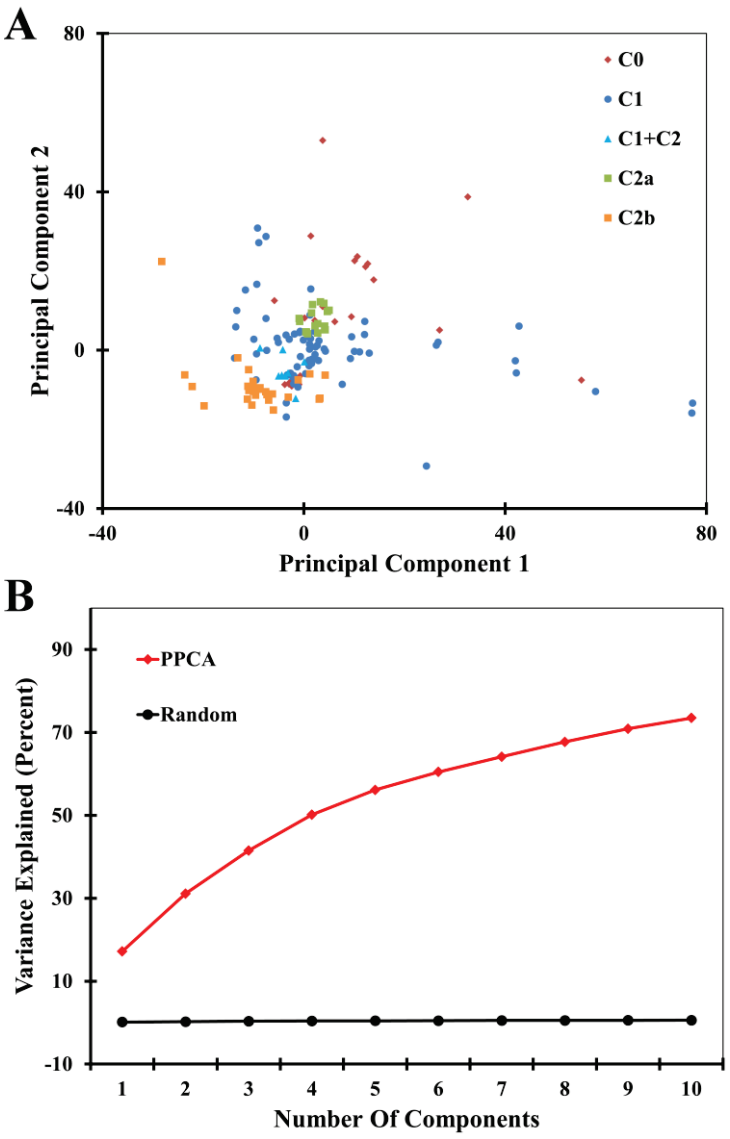

FIGURE S7 Primary Components from Probabilistic Principal Component Analysis using Staccato. *(A)* shows core domain structural data projected onto Principal Component 1 (PC1) plotted against data projected onto Principal Component 2 (PC2). Core domains are colored according to corresponding cap type. *(B)* shows the plot of cumulative variance described by the principal components (red) and random (black).

Figure S8

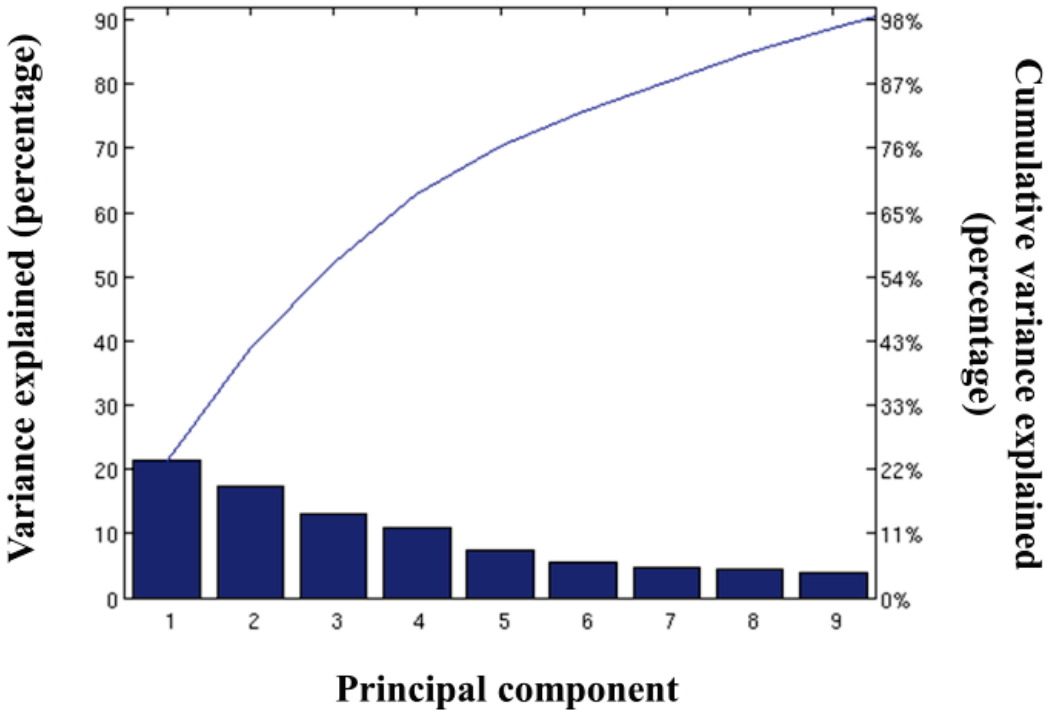

FIGURE S8. Pareto plot illustrating the cumulative variance explained by inclusion of the principal components.

Figure S9

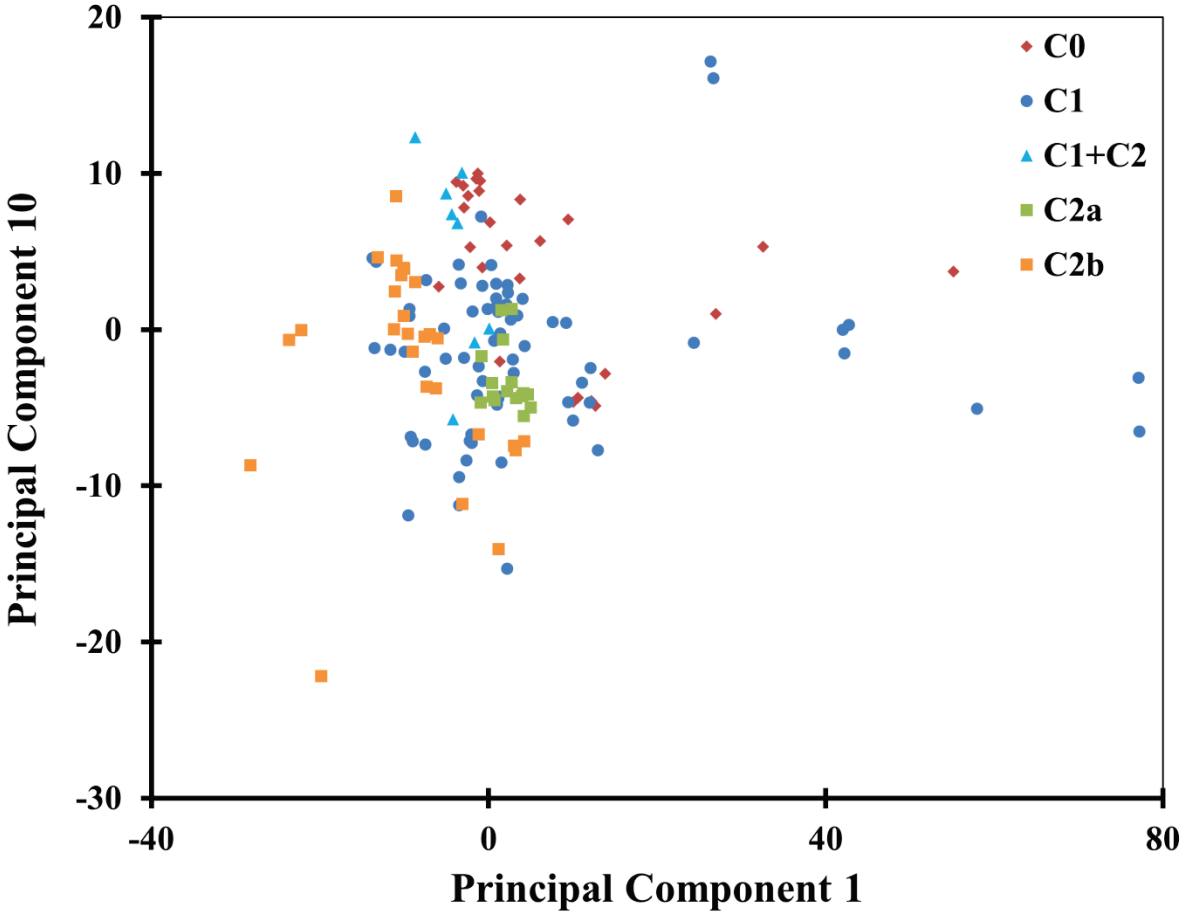

FIGURE S9. Core domain structural data projected onto Principal Component 1 (PC1) plotted against data projected onto Principal Component 10 (PC10). Core domains are colored according to corresponding cap type.

Figure S10

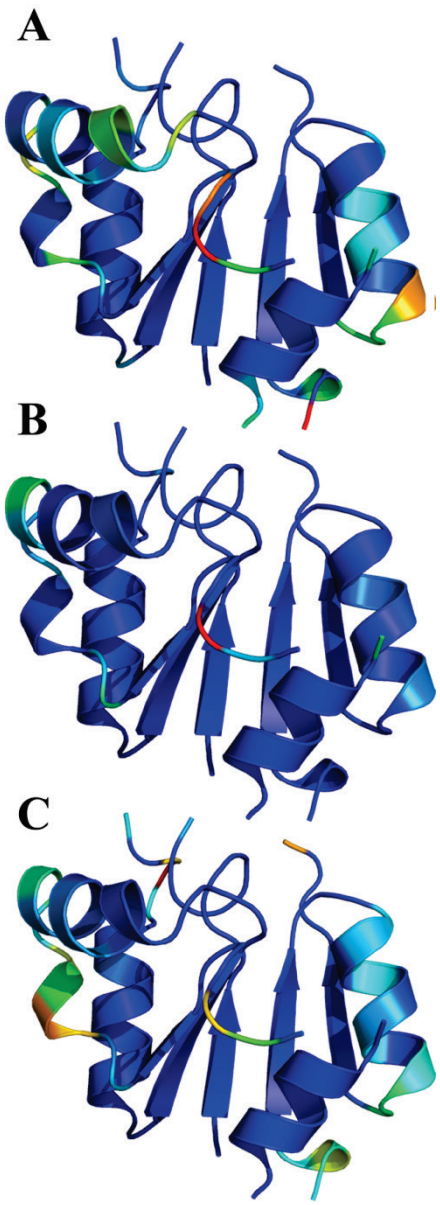

FIGURE S10. Eigenvectors corresponding to principal components from Probabilistic Principal Component Analysis mapped onto representative core domain 2HSZ, chain A. (*A)*, *(B)* and, *(C)* show maps for principal component 1, 2 and, 3 respectively. Structures are colored as a color ramp according to corresponding eigenvector values with blue denoting the lowest value and red the highest.

Figure S11

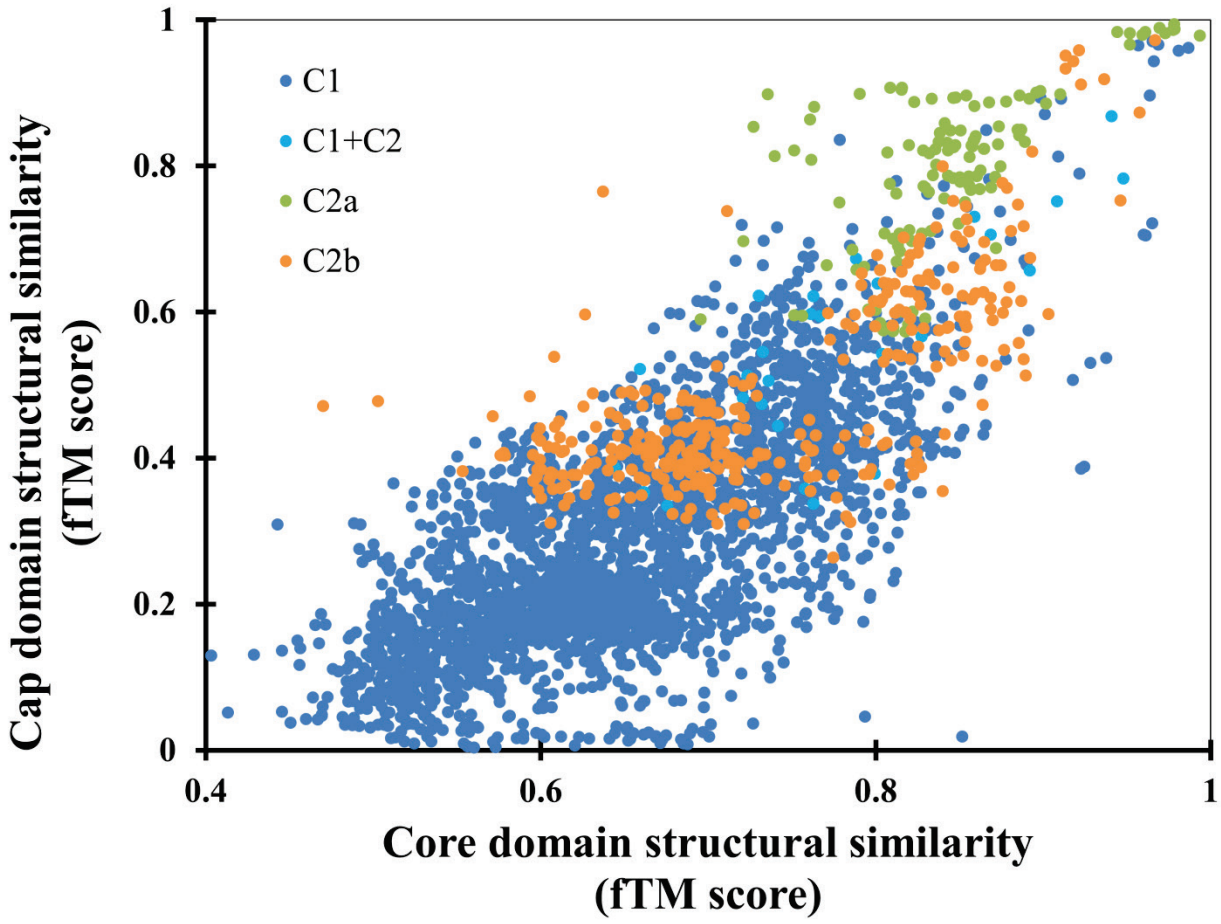

FIGURE S11. Correlation between cap domain and core domain structural similarity split by corresponding cap type. Each point represents a pair of proteins with the core domain fTM score along the x-axis and cap domain fTM score along the y-axis calculated using TM-align.

Figure S12

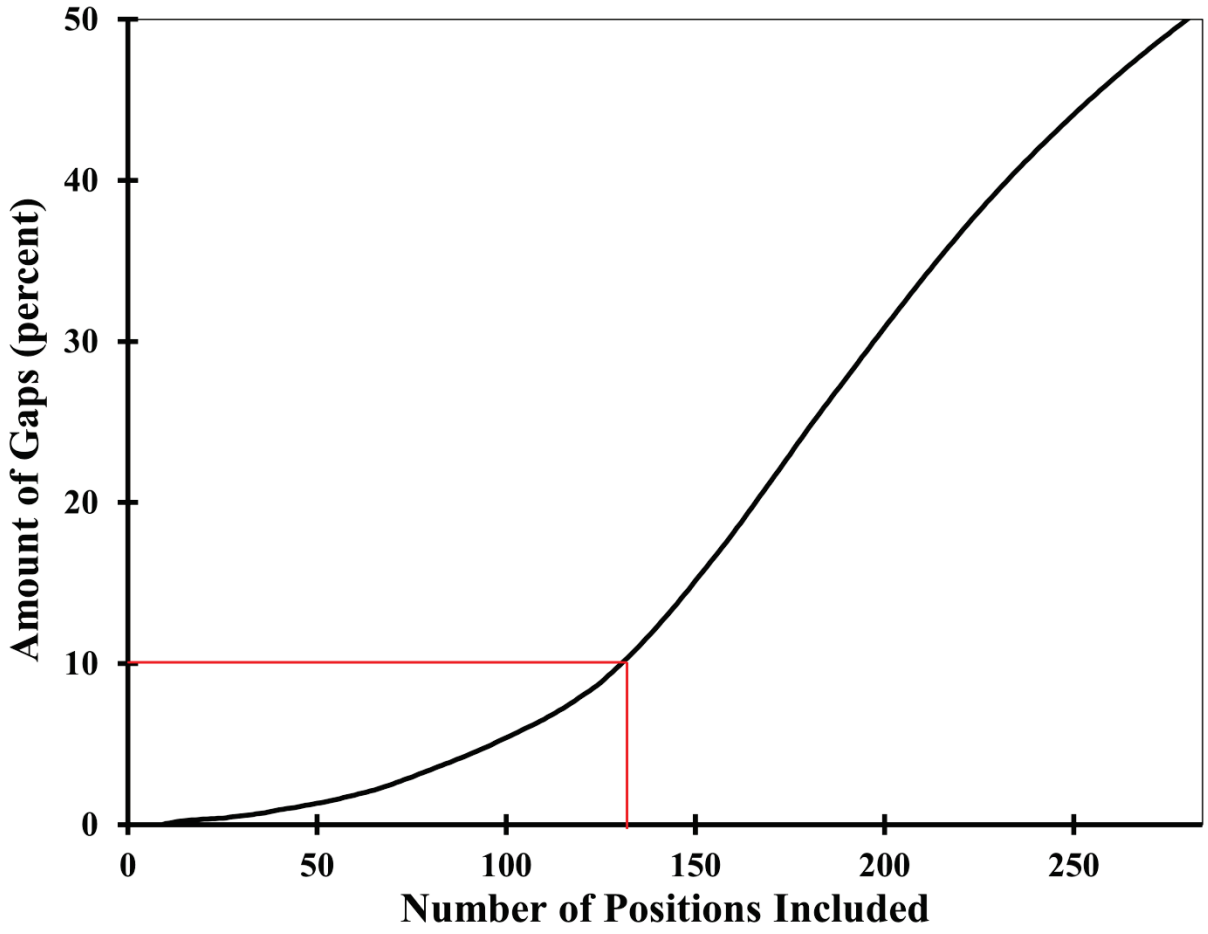

FIGURE S12. Plot displaying the relative number of gaps in the multiple sequence alignment as a function of number of columns in the alignment.