Supplementary note for "Rajapaksa *et al.*'s *Lossless compression and mapping local sequence to structure in proteins*"

Supplementary Figure S1:

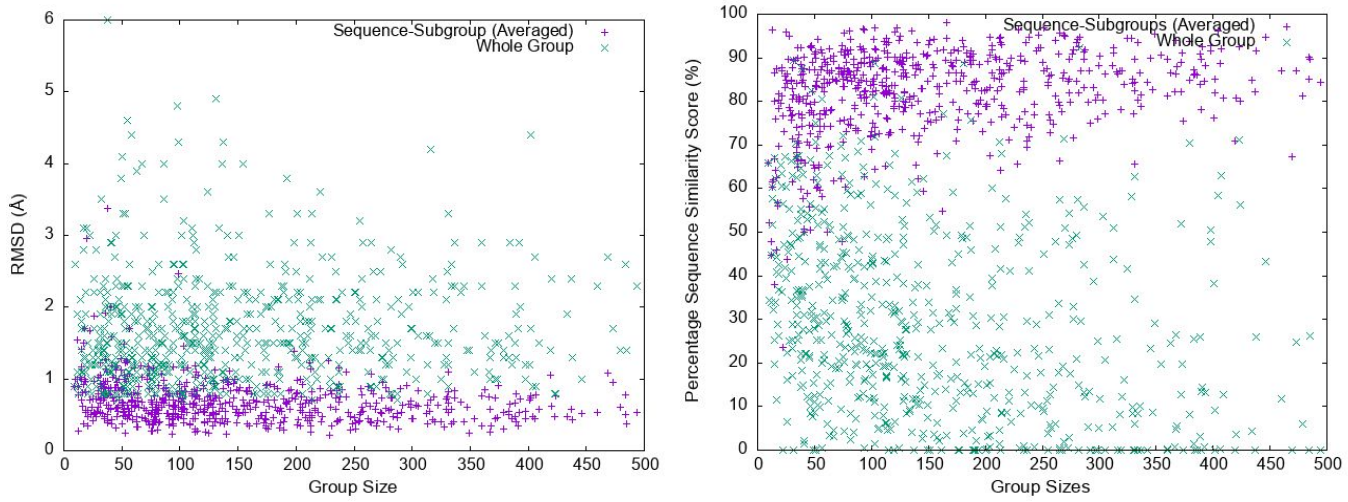

Fig. S1 (a) Comparison of the RMSDs of superpositions of conserved cores in 638 of 1493 Proçodic concept-groups. Green marks in the plot show RMSDs when the members in the respective concept-groups are structurally aligned as a whole. Magenta marks show the weighted-average RMSDs when subsets of members are structurally aligned based on their inferred subgroups informed by variant A method. The improved RMSDs of the latter confirm that the local structures within the inferred subgroups are better in fit than the whole. Note: X-axis shows the increasing order of the concept-group sizes for all groups with less than 500 members. For groups with >=500 members, the multiple structural alignment program we used (MUSTANG) does not scale to align those large groups in practical time. (b) Comparison of the percentage sequence similarity for the same 638 Proçodic concept-groups. Green marks in the plot show the sequence similarity when the corresponding amino acid sequences are aligned as a whole. Magenta marks show the weighted-average sequence similarity when the sequences are aligned based on their inferred representative subgroups informed by variant A approach. The drastic improvement in the similarity score is an evidence of a strong sequence signal in each subgroup.

Figure S1 shows that the inferred subgroups that result from our variant A method, consistently contain a strong local sequence-structure signal, as demonstrated by their high sequence-similarity scores (magenta marks in Fig. S1(b)) and low RMSD values (magenta marks in Fig. S1(a)) within members of the subgroup.