Supplementary information 1. Why at least 20 residue type categories are required for robust analysis of evolution on inner tree branches?

This Supplement shows the effects of categories number reduction from 20 (20 categories of residue solvent accessibilities or 20 canonical amino acids) to 8 (secondary structure types). All effects demonstrated in this Supplement can be explained by the reduction of visible mutation numbers.

The comparisons of median lengths (obtained by random delete-half-jackknifing of alignments) of inner branches of all analyzed 512 protein trees obtained based on residue solvent accessibilities (20 residue categories) and secondary structures (8 residue categories) analyses demonstrated in Fig. S1.1 and Fig. S1.2. Both heterotachy and site partitioning models show that increasing in number of branches with near zero branch lengths relates with residue categories number reduction: 2.5 fold increasing of zero-length branches in heterotachy model and 4 fold increasing of zero-length branches in site partitioning model. Additionally, the reduction of residue categories number relates with total inner branch lengths shortening.

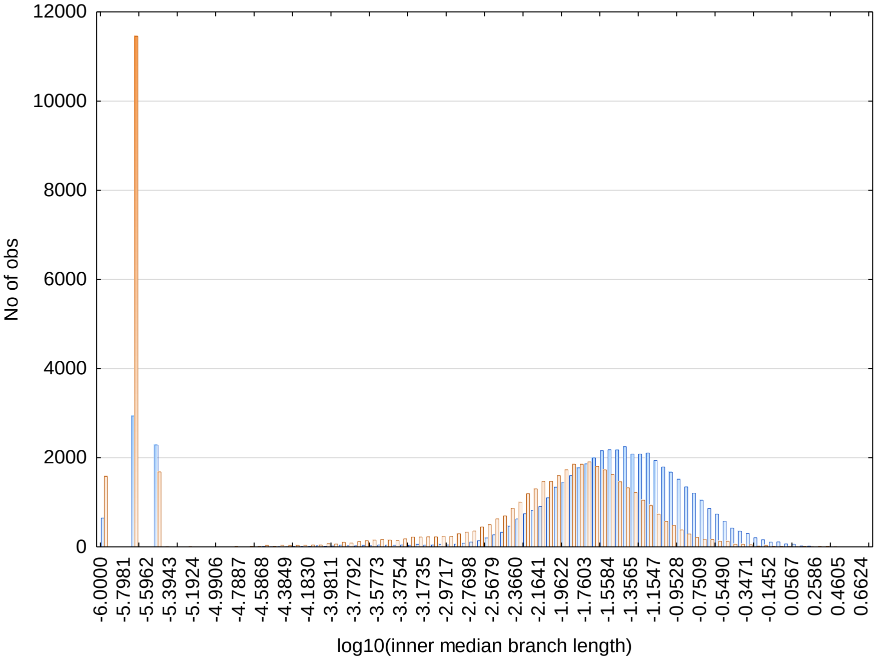

Fig. S1.1 Lengths of inner tree branches in trees based on residue solvent accessibilities (blue) and secondary structures (orange) analyses, heterotachy model.

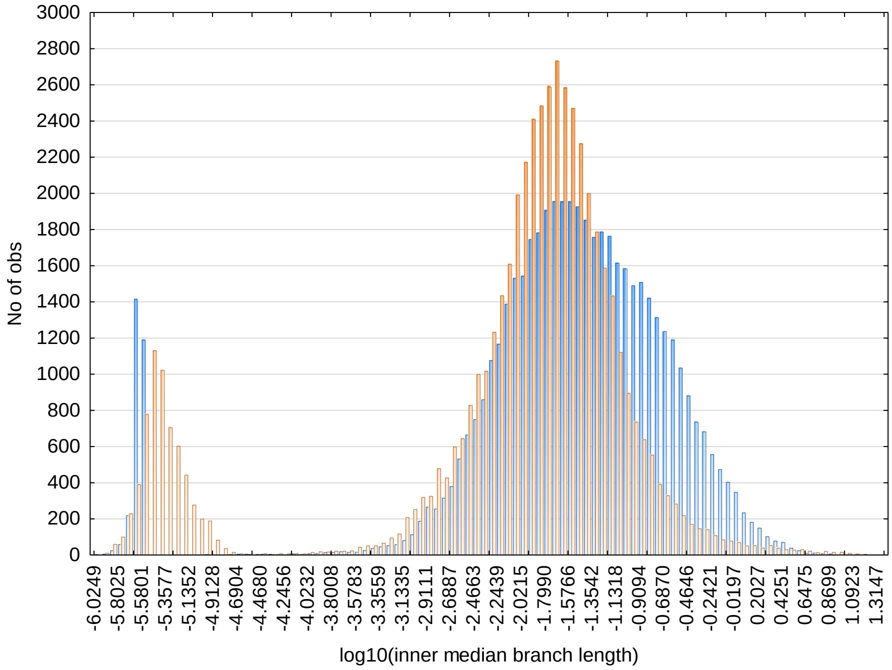

Fig. S1.2 Lengths of inner tree branches in trees based on residue solvent accessibilities (blue) and secondary structures (orange) analyses, site partitioning model.

Figure S1.3 shows interquartile range (Q3-Q1) of the ln(*L*) measures of inner branches of all analyzed 512 protein trees under heterotachy and site partitioning models based on 8 secondary structure types. The less the residue type categories in analysis the heavier the tails of interquartile range (Q3-Q1) of the ln(*L*) value distributions are (compare with Fig. 3). This fact clearly demonstrates that increasing in branch length variability tightly related with the residue categories number reduction.

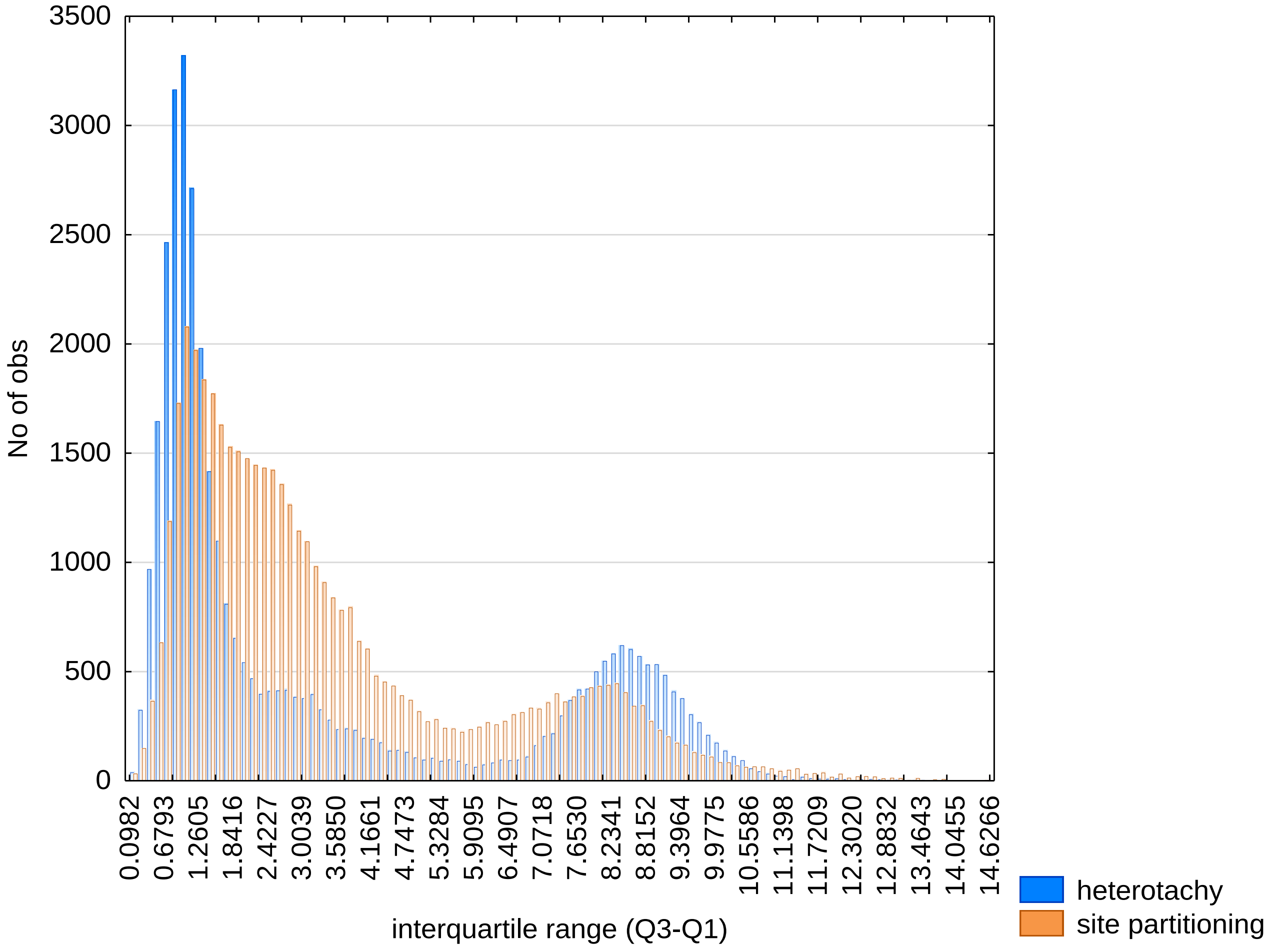

Fig. S1.3 Interquartile range (Q3-Q1) of the ln(*L*) branch measures in heterotachy (A) and site partitioning (B) models.

Figure S1.4 shows fraction of inner branches in all analyzed 512 protein trees with absolute median ln(*L*) measure > 6 under heterotachy model based on 20 RSA and 8 secondary structure types. In order to fix incongruence between branch lengths based on analysis of amino acids and secondary structures, we used threshold of abs(ln(L))>10. We select this threshold because the meaningful minimum of branch length is 5E-5 (see Construction and content), therefore the case when abs(ln(L)) = 10 reflects the comparison of minimum branch length with branch length closer to maximum. Figure S1.4 clearly demonstrates that reducing number of residue type categories associated with increasing number of branches having incongruence between length obtained based on amino acid alignments analysis, and length obtained based on secondary structure alignments analysis.

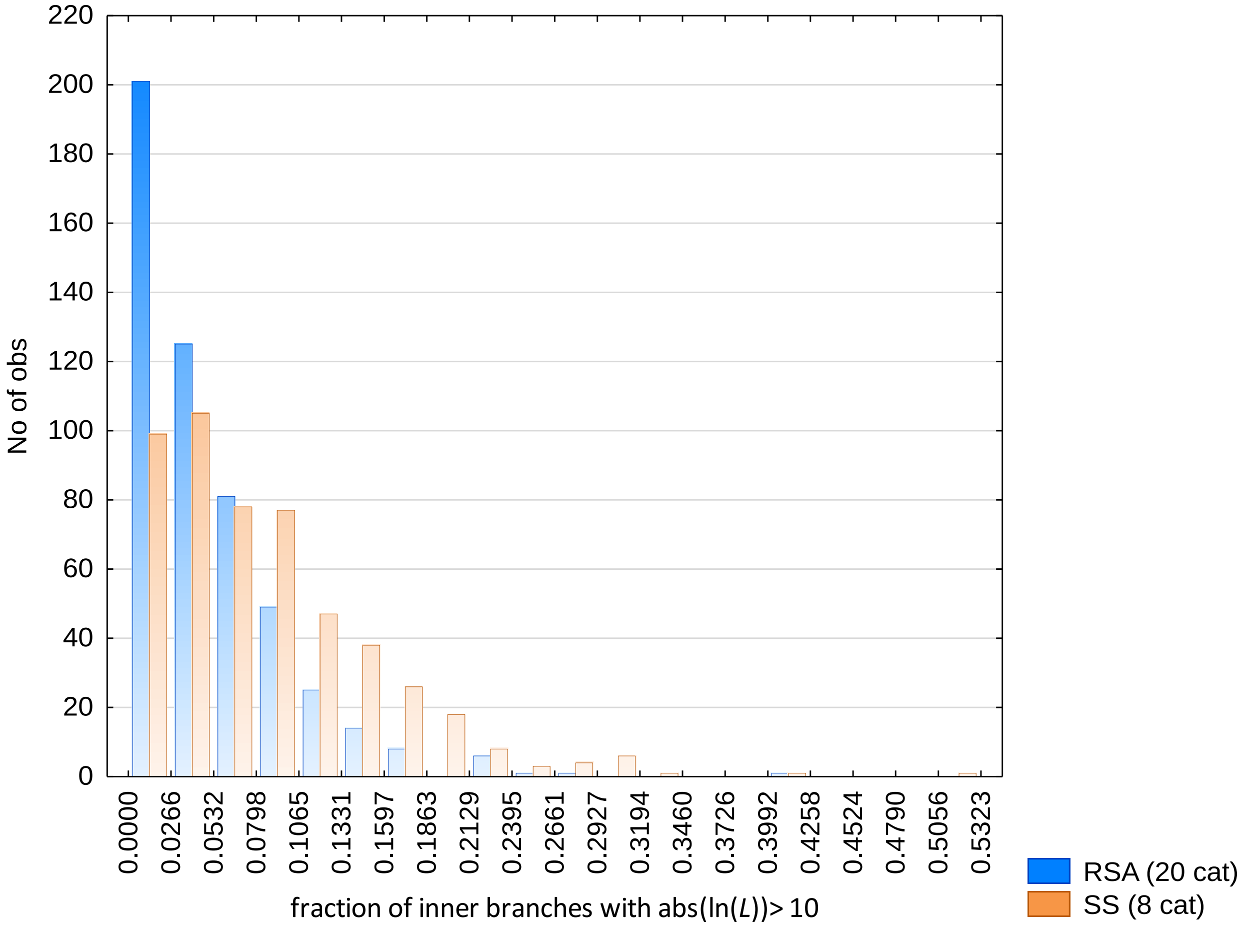

Fig. S1.4 The fraction of inner branches with absolute median ln(*L*) measure > 10, heterotachy model.

Figure S1.5 shows the same as for Figure S1.4 but under site partitioning model. This figure do not show association between reducing number of residue type categories and increasing number of branches having incongruence between lengths obtained based on amino acids and secondary structures. This is due to significantly higher variability of ln(*L*) measure in site partitioning model comparing to heterotachy model (see Figures 3 and S1.3). In the case of ln(*L*) high variability the median of ln(*L*) for the particular branch tends to median value for general set of ln(*L*) values for all branches.

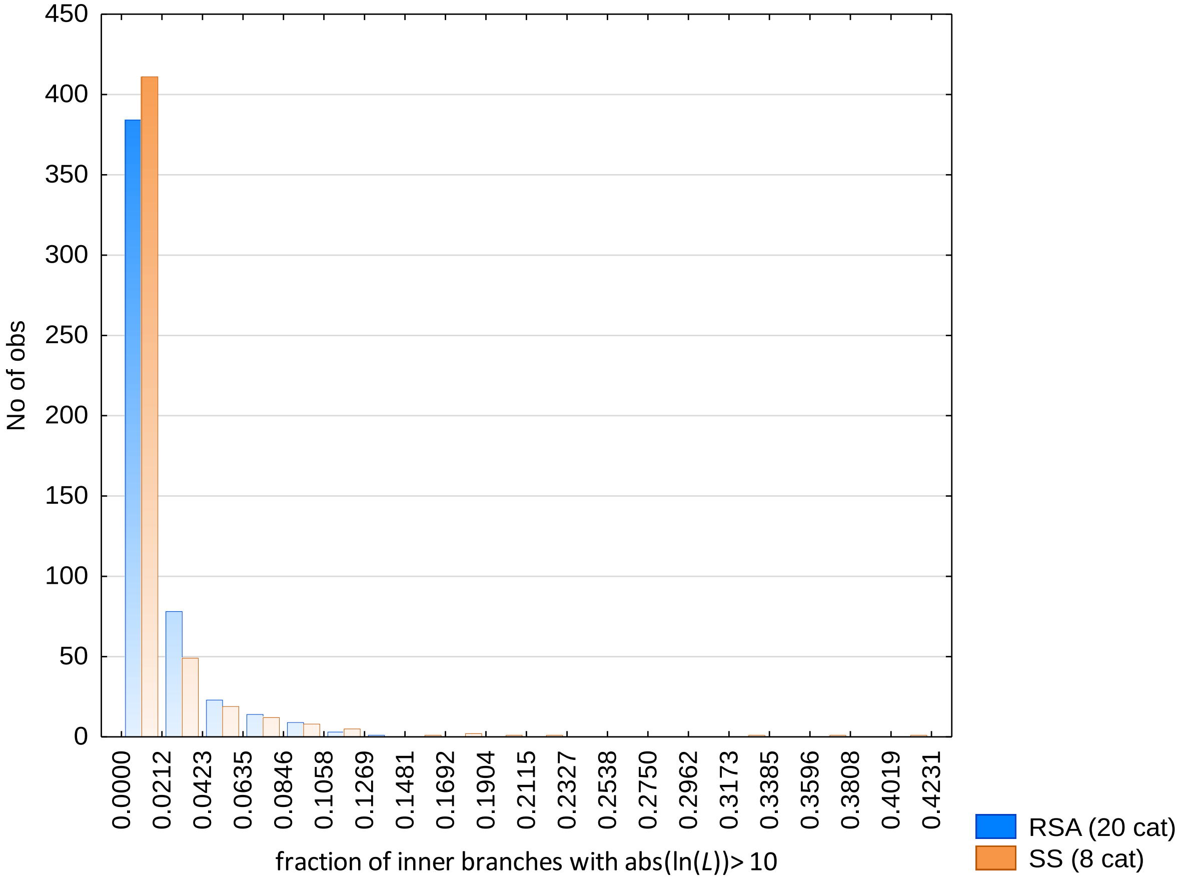

Fig. S1.5 The fraction of inner branches with absolute median ln(*L*) measure > 10, site partitioning model.