



SUPPLEMENTARY FIG. S3. Euclidean distance between “grand” GTR rate matrix parameters (i.e., the rate matrix parameters estimated using the GTR model with the complete FRG dataset) and GTR rate matrix parameters estimated using subsets of the data. Rate matrix parameters optimized on sites defined using protein structure are presented as blue diamonds. The other points represent distances between the grand GTR model and rate matrix parameter estimates optimized using random samples (ranging in size from 500-55,000 aligned amino acid sites) that were drawn from the concatenated FRG dataset. For each data subset size, we generated 10 random samples; the distance between the rate matrix for structurally defined sites and the grand GTR model always exceeds the distance for random samples of sites of comparable size.