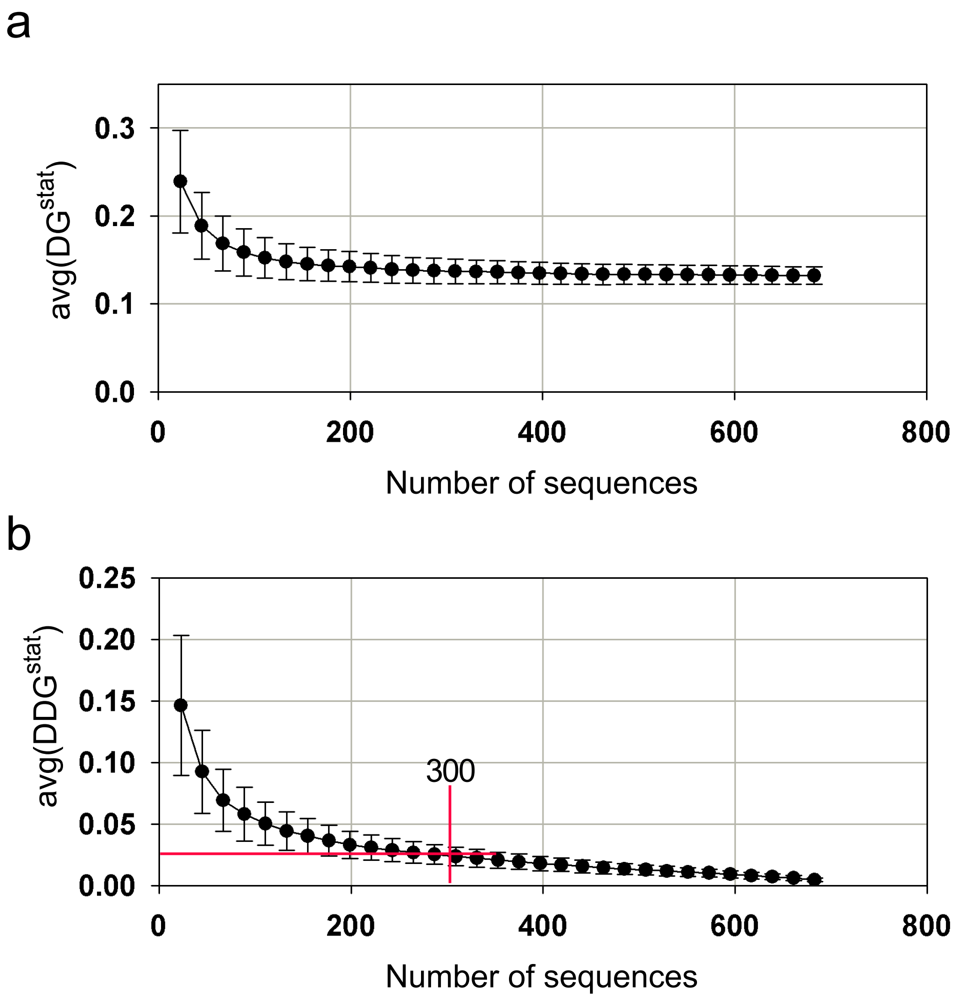
a
0.3
0.2
0.1
0.0
avg(DG^{stat})
0
200
400
600
800
Number of sequences
b
0.25
0.20
0.15
0.10
0.05
0.00
avg(DDG^{stat})
300
0
200
400
600
800
Number of sequences

Fig. 9. Statistical evaluation of the G protein MSA and criteria for selecting perturbations for the SCA. (*a*) The average conservation score [$avg(\Delta G^{stat})$] of the five least conserved sites in the G protein MSA as a function of random elimination of sequences from the parent alignment. A condition of SCA is that the alignment be large and diverse enough to accurately represent the evolutionary constraint in the amino acid distributions at sites rather than representing just the history of the evolutionary process or inadequate sampling of G protein family members. Practically, we interpret this to mean that (*i*) many sites display amino acid distributions that have relaxed close to their mean frequencies in nature, and (*ii*) random elimination of sequences from the alignment shows no significant change to the amino acid distributions at even unconserved sites. The graph shows that the average conservation score at the least conserved sites in the G protein alignment is invariant to random elimination of up to 80% of sequences. (*b*) The average statistical coupling [$avg(\Delta\Delta G^{stat})$] for the five least unconserved sites on random elimination of sequences from the parent alignment. The goal of SCA is to expose functional rather than historical relationships between positions in the MSA. To eliminate the latter, we require that perturbations at sites produce subalignments that are also large and diverse such that they remain a representative ensemble of the parent alignment. If so, the perturbation should not globally alter conservation at sites in the alignment, and unconserved sites (which by definition are not evolutionarily constrained) should remain unconserved and display coupling energies close to zero. The graph shows an empirical approach to determine a minimum subalignment size by plotting the impact of random elimination of sequences from the MSA on a set of unconserved sites. In this work, we conservatively chose a minimum subalignment cutoff of 300 sequences (44% of the parent MSA), corresponding to a total of 33 perturbations that are used to build the statistical coupling matrix (Fig. 1*A*).