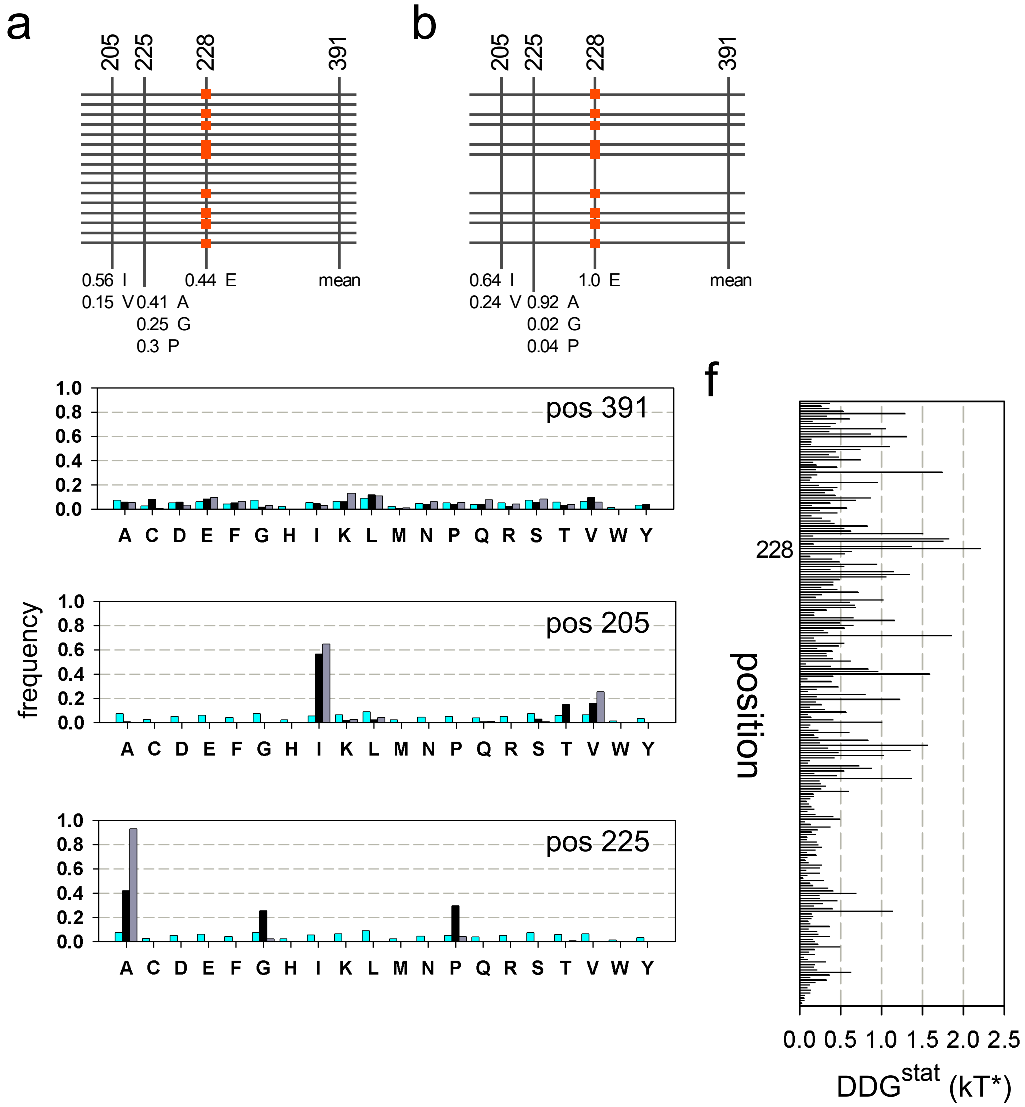
a
205
225
228
391
0.56 I
0.15 V
0.41 A
0.25 G
0.3 P
0.44 E
mean
b
205
225
228
391
0.64 I
0.24 V
0.92 A
0.02 G
0.04 P
1.0 E
mean
c
pos 391
d
pos 205
frequency
e
pos 225
1.0
0.8
0.6
0.4
0.2
0.0
A C D E F G H I K L M N P Q R S T V W Y
f
228
position
0.0 0.5 1.0 1.5 2.0 2.5
DDG^{stat} (kT*)

Fig. 8. The statistical coupling analysis (SCA) in a multiple sequence alignment of 717 members of the G protein family (http://hhmi.swmed.edu/Labs/rr). Briefly, each position j in the multiple sequence alignment (MSA) is described as a 20-element vector of individual amino acid frequencies (e.g., $\overrightarrow{f_j}$ =[$f_j^{ala}, f_j^{cys}, f_j^{asp}, \ldots, f_j^{tyr}$]). The frequency vector is then converted to a vector of statistical energies ($\overrightarrow{\Delta G_j^{stat}}$ =[$\Delta G_j^{ala}, \Delta G_j^{cys}, \Delta G_j^{asp}, \ldots, \Delta G_j^{tyr}$]), where each term is the value for amino acid x at site j and is given by $\Delta G_j^x = kT^* \ln P_j^x$. P_j^x is the binomial probability of observing amino acid x at site j given its mean frequency in all natural proteins. To measure coupling between a perturbation at i and any site j, we calculate the difference energy vector, $\overrightarrow{\Delta\Delta G_{j,i}^{stat}} = \overrightarrow{\Delta G_j^{stat}} - \overrightarrow{\Delta G_{j|i}^{stat}}$, where $\overrightarrow{\Delta G_j^{stat}}$ is the statistical energy vector of site j in the parent alignment, and $\overrightarrow{\Delta G_{j|i}^{stat}}$ is that of site j in the subalignment derived from the perturbation at i. The scalar coupling energy ($\Delta\Delta G_{j,i}^{stat}$) is the magnitude of this difference vector and reports the combined effect of perturbation at i on all amino acids at position j. If sites i and j are evolutionarily independent, the coupling energy is zero and is consistent with lack of interaction, but if the coupling energy is nonzero, the two sites interact to the extent measured by $\Delta\Delta G_{j,i}^{stat}$. Calculation of $\Delta\Delta G_{j,i}^{stat}$ for all sites j given a perturbation at i is a mapping of how all sites in the protein feel the effect of perturbing i. (*a*) A schematic representation of the alignment, showing the frequencies of the dominant amino acids at four sites, 205, 225, 228, and 391. (*b*) The alignment after making a perturbation at position 228 (by restricting it to Glu). Because the parent alignment showed 44% Glu at 307, the subalignment now contains $0.44 * 717 = 314$ sequences. (*c*) The frequency distribution of amino acids at position 391 in the parent (black bars) and 228E subalignment (grey bars). For comparison, the mean frequencies of amino acids in 36,498 proteins in the nonredundant database are shown in blue in *c--e*. A postulate of SCA is that the lack of evolutionary constraint at any position should cause the observed amino acid frequencies at that site to approach these mean frequencies; as a corollary, the evolutionary constraint is the degree to which the observed distribution deviates from these mean frequencies. Position 391 is not

conserved (the distribution in the parent alignment is close to that in all proteins in nature) and remains unconserved on making the perturbation at position 228. Thus it shows a low conservation score ($\Delta G_{391}^{stat} = 0.11kT^*$) and a low coupling score to the 228E perturbation ($\Delta\Delta G_{391,228E}^{stat} = 0.06kT^*$). (*d*) Position 205 is moderately conserved, because it shows a strong bias for Ile and Val above their mean frequencies (black bars, $\Delta G_{205}^{stat} = 1.05kT^*$) but is nevertheless uncoupled to the 228E perturbation ($\Delta\Delta G_{205,228E}^{stat} = 0.3kT^*$),because the distribution is nearly unchanged in the subalignment (compare black and grey bars). (*e*) Position 225 shows significant conservation ($\Delta G_{225}^{stat} = 0.62kT^*$) but also a significant redistribution of its amino acid frequencies on the 228E perturbation. Thus, this position is evolutionarily coupled to 228 ($\Delta\Delta G_{225,228E}^{stat} = 1.75kT^*$). (*f*) The complete set of statistical coupling values for the 228E perturbation for all sites i ($\Delta\Delta G_{i,228E}^{stat}$) gives a global map of how this perturbation impacts evolution at other sites and represents an evolution-based prediction of the thermodynamic interactions between 228 and these other sites. The statistical coupling matrix (Fig. 1*A*) is a concatenation of many such perturbation experiments; thus the data in *f* represent one column in the matrix.