Amino-acid site variability among natural and designed proteins

Eleisha L. Jackson, Noah Ollikainen, Arthur W. Covert III,
Tanja Kortemme, and Claus O. Wilke

Supporting Figures

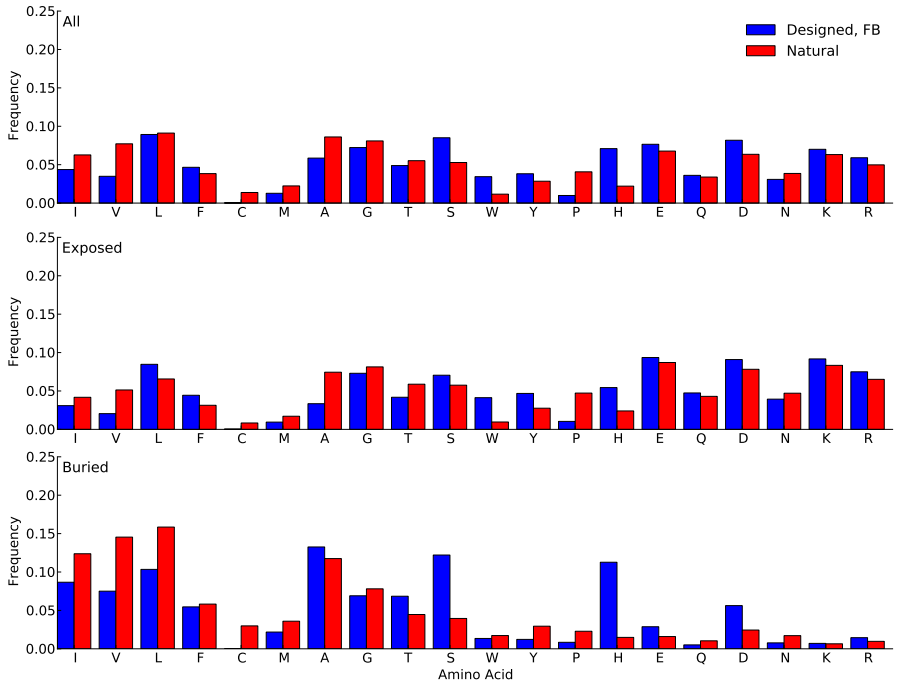

Figure S1. Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only fixed-backbone designs were considered. Top: overall frequencies. Middle: frequencies at exposed sites (defined as sites with RSA > 0.05). Bottom: frequencies at buried sites (defined as sites with RSA ≤ 0.05).

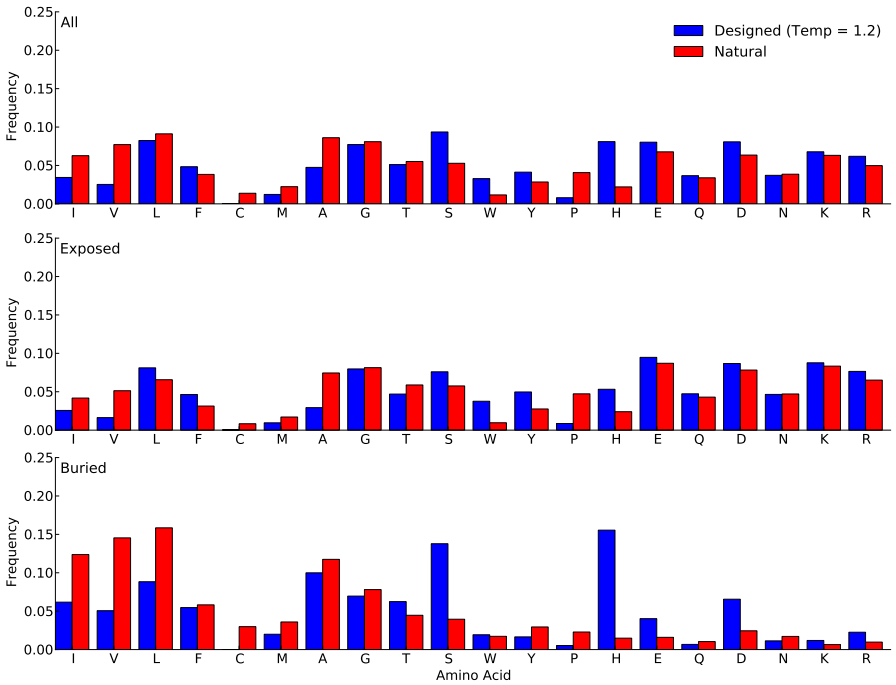

Figure S2. Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the yeast-proteins data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. Top: overall frequencies. Middle: frequencies at exposed sites (defined as sites with RSA > 0.05). Bottom: frequencies at buried sites (defined as sites with RSA ≤ 0.05).

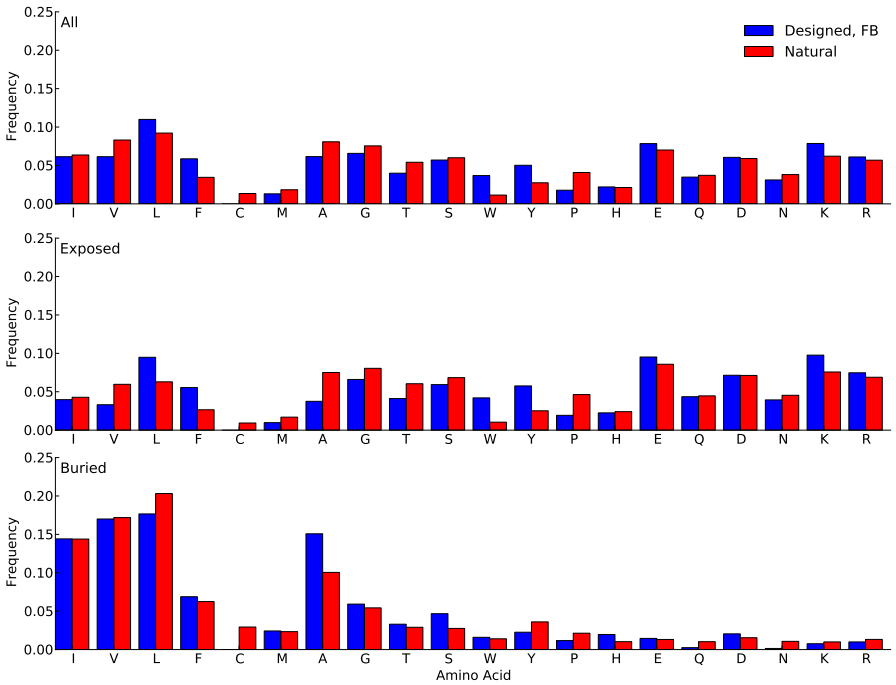

Figure S3. Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only fixed-backbone designs were considered. Top: overall frequencies. Middle: frequencies at exposed sites (defined as sites with RSA > 0.05). Bottom: frequencies at buried sites (defined as sites with RSA ≤ 0.05).

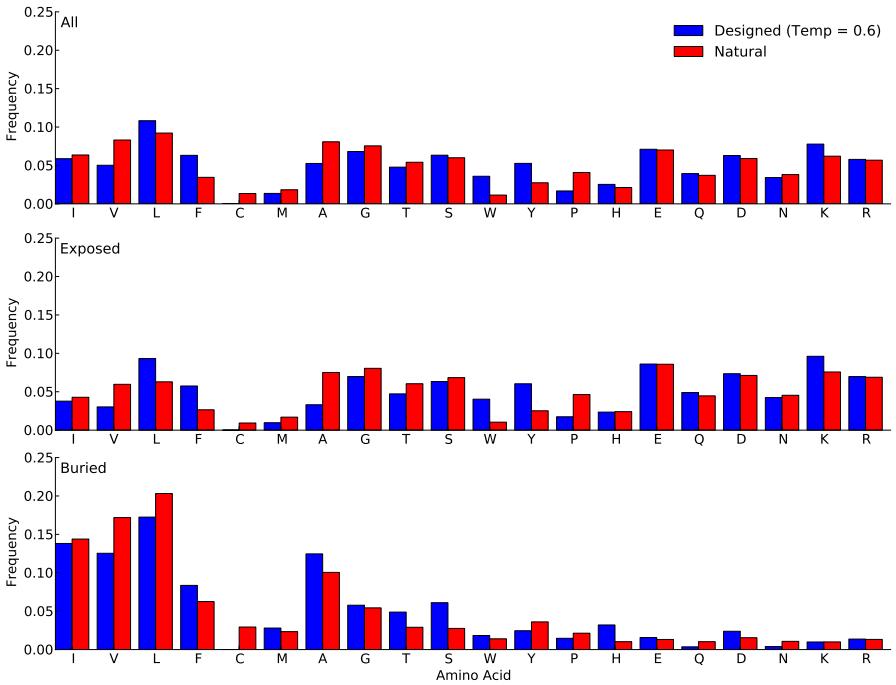

Figure S4. Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 0.6 were considered. Top: overall frequencies. Middle: frequencies at exposed sites (defined as sites with RSA > 0.05). Bottom: frequencies at buried sites (defined as sites with RSA ≤ 0.05).

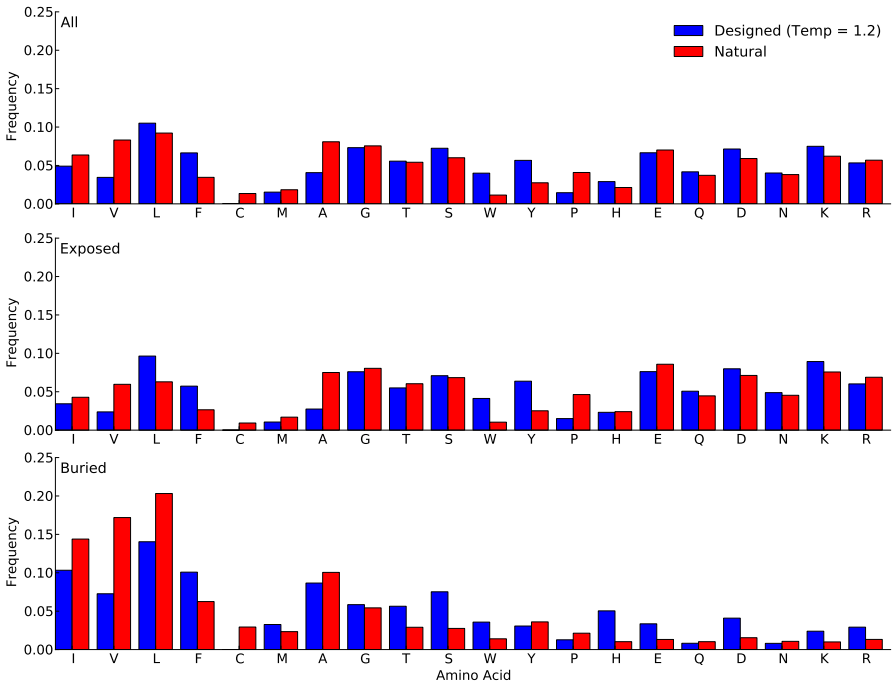

Figure S5. Amino-acid frequencies in designed and natural proteins. Frequencies were calculated over all sites in all proteins belonging to the protein-domains data set. For designed proteins, only flexible-backbone designs with design temperature 1.2 were considered. Top: overall frequencies. Middle: frequencies at exposed sites (defined as sites with RSA > 0.05). Bottom: frequencies at buried sites (defined as sites with RSA ≤ 0.05).

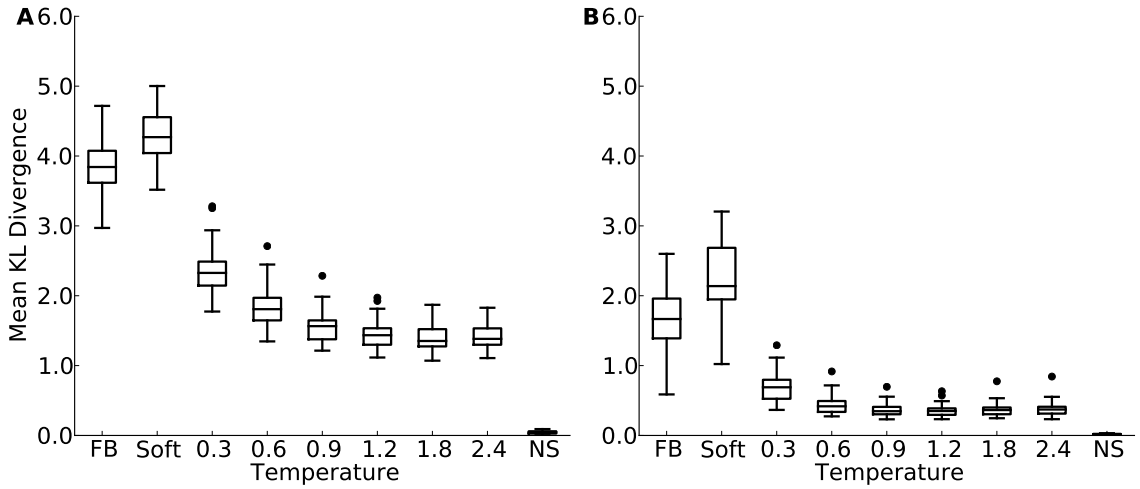

Figure S6. Mean Kullback-Leibler (KL) divergence for designed and natural proteins, shown for the yeast-proteins data set. A higher KL divergence indicates that the amino-acid distributions at sites in designed proteins are less similar to the corresponding distributions in the natural proteins. "FB" refers to fixed backbone design, and "NS" refers to the control case where natural sequences are compared to themselves. (A) KL divergence calculated from the relative frequencies of the 20 amino acids. (B) KL divergence calculated from rank-ordered frequency distributions. The most common amino acid in the reference distribution is compared to the most common amino acid in the focal distribution, the same is done for the second-most common amino acid, and so on, irrespective of the type of amino acids.

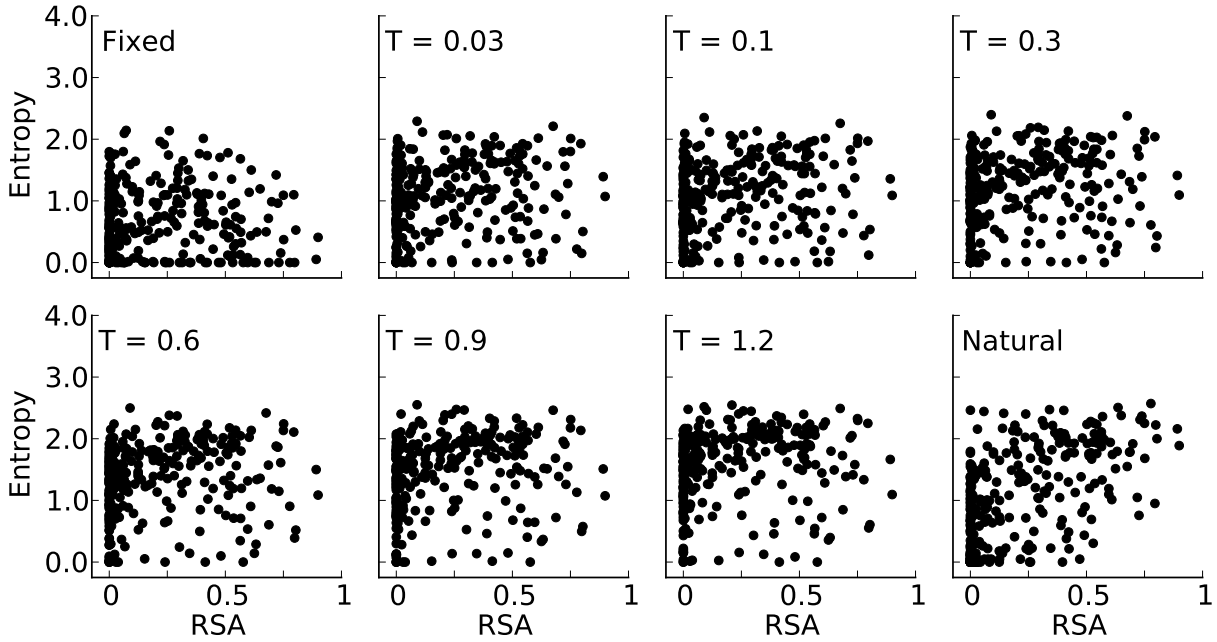

Figure S7. Site entropy versus Relative Solvent Accessibility (RSA) for designed and natural sequence alignments of the protein S-formylglutathione hydrolase (PDB: 1PV1, chain A). Natural sequences exhibit a clear trend of higher site variability at higher RSA values. The flexible backbone designs exhibit a similar trend but the fixed backbone designs do not.

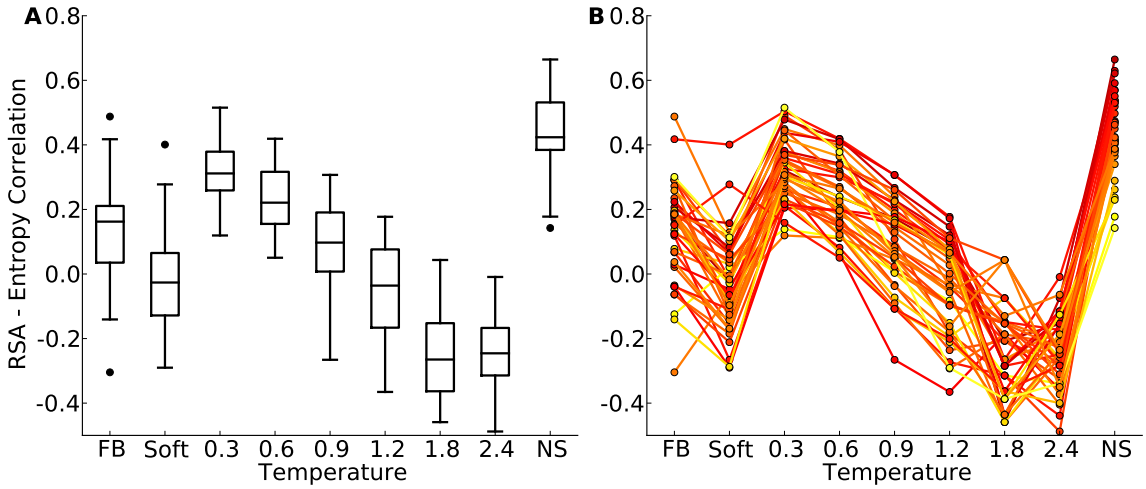

Figure S8. Distributions of correlation coefficients between site entropy and RSA, for the yeast-proteins data set. "FB" indicates fixed-backbone design, "Soft" indicates soft backbone design, and "NS" indicates natural sequences. (A) Distributions represented as boxplots. (B) Correlation coefficients for individual proteins. Lines connect identical structures in the different design conditions. The color shading represents the strength of the correlation for the natural sequence alignment. In general, natural proteins display a stronger correlation between site entropy and RSA than designed proteins.

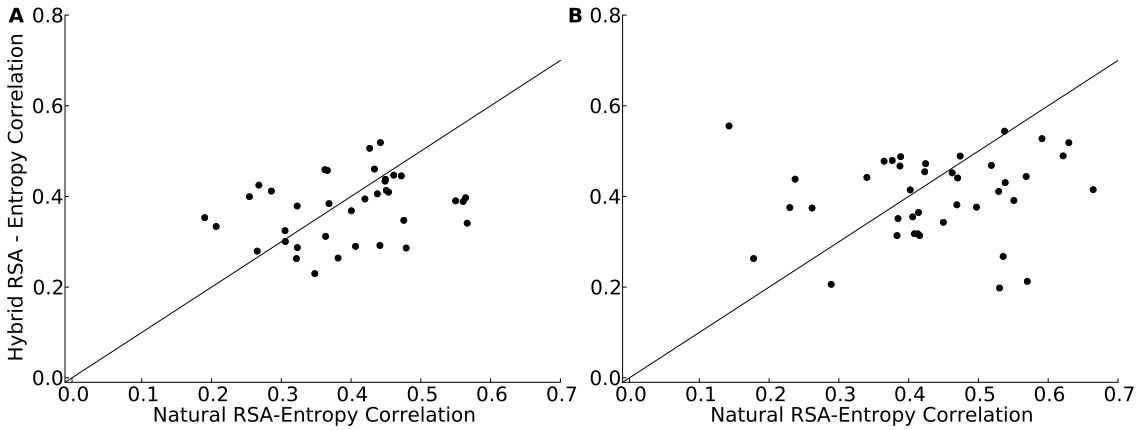

Figure S9. Correlation coefficients between RSA and site entropy for hybrid designs and natural proteins. For the hybrid designs, buried and partially buried sites were taken from proteins designed with a fixed backbone (yeast proteins) or a temperature of $T = 0.6$ (protein domains). Exposed residues were taken from proteins designed with a temperature of $T = 0.1$ (yeast proteins) or $T = 1.8$ (protein domains). The solid line indicates $y = x$. Note that while the range of correlation values in hybrid designs generally matches the range of values in natural alignments, predictions for specific proteins are not that accurate.