SUPPORTING INFORMATION

Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability

Brandon J. Sullivan[1], Tran Nguyen[3], Venuka Durani[2], Deepti Mathur, Samantha Rojas, Miriam Thomas[2], Trixy Syu[3] and Thomas J. Magliery [2,3*]

[1]Ohio State Biochemistry Program, The Ohio State University, Columbus, OH 43210
[2]Department of Chemistry, The Ohio State University, Columbus, OH 43210
[3]Department of Biochemistry, The Ohio State University, Columbus, OH 43210

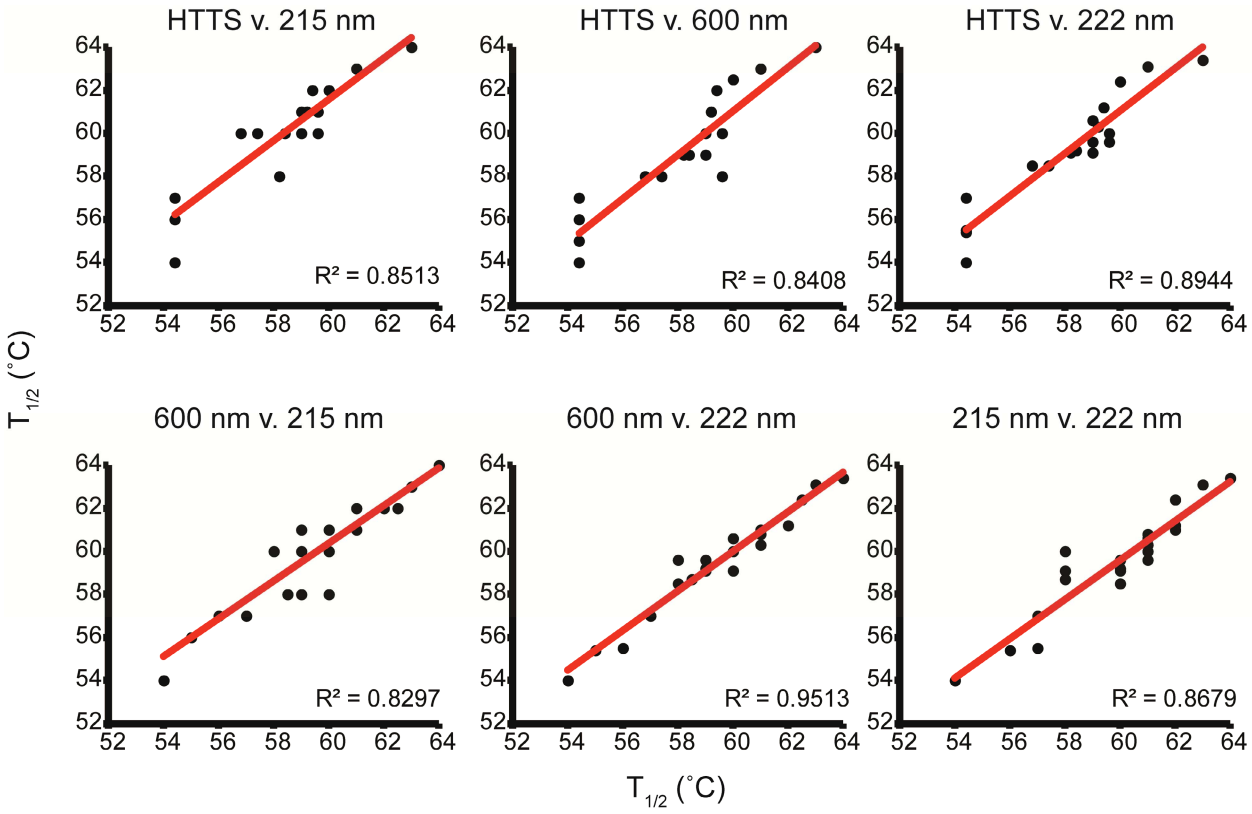
[*]Corresponding author. E-mail address: magliery.1@osu.edu. Departments of Chemistry and Biochemistry, The Ohio State University, 100 W. 18th Ave., Columbus, OH 43210. Phone (614) 247-8425. Fax (614) 292-1685.

Host protein sequence

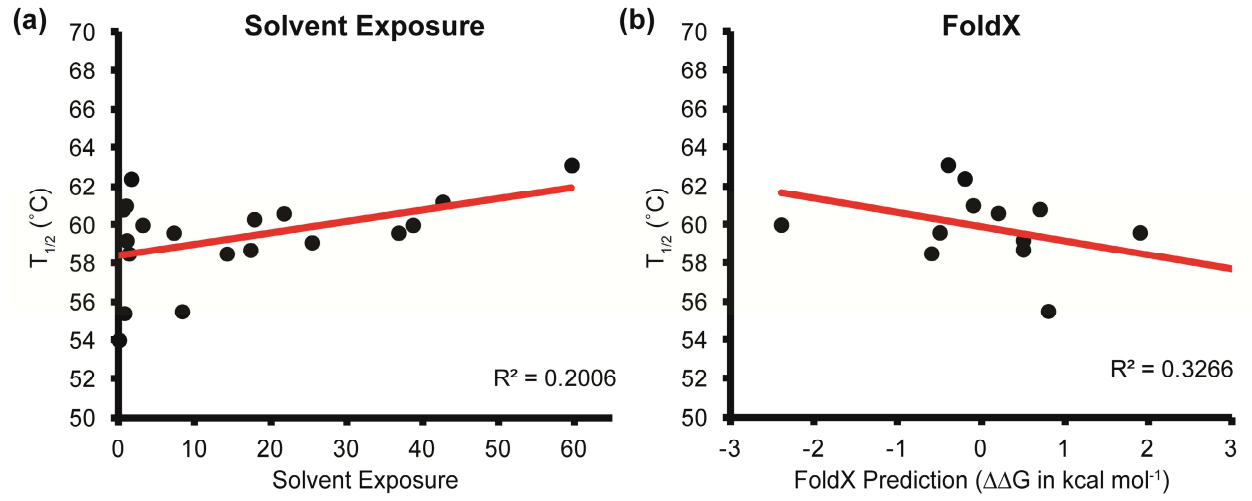
S.c. TIM sequence

```
                    2
MAHHHHHHGGENLYFQGSSGARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQ
VTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVGVILCIGETLEEKK
AGKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSA
NGSNAVTFKDKADVDGFLVGGASLKPEFVDIINSRN*
```

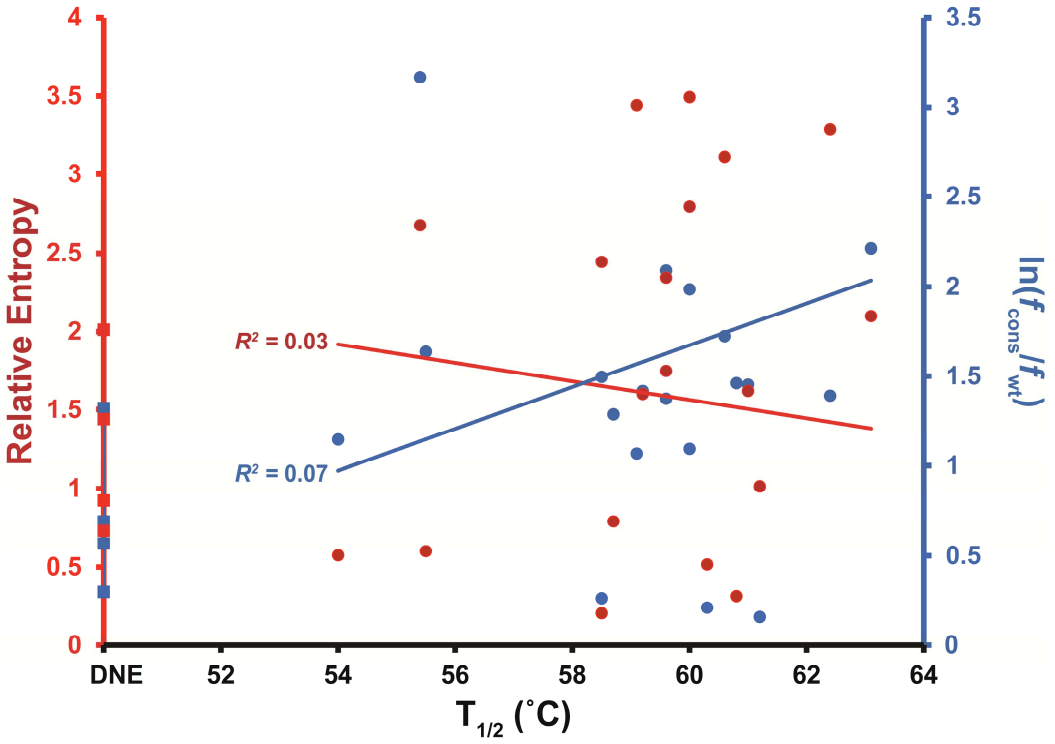
The TEV cleavable sequence is shown underlined. A GSSG linker was cloned downstream of the TEV recognition sequence to improve cleavage efficiency. The native *S.c.* TIM sequence begins MARTFFV..., where the initiating Met is labeled as residue 1. Our characterized variants begin with GSSGARTFFV..., where the second Gly is labeled as residue 1. Both sequences are numbered such that Ala at position 2 is consistent.



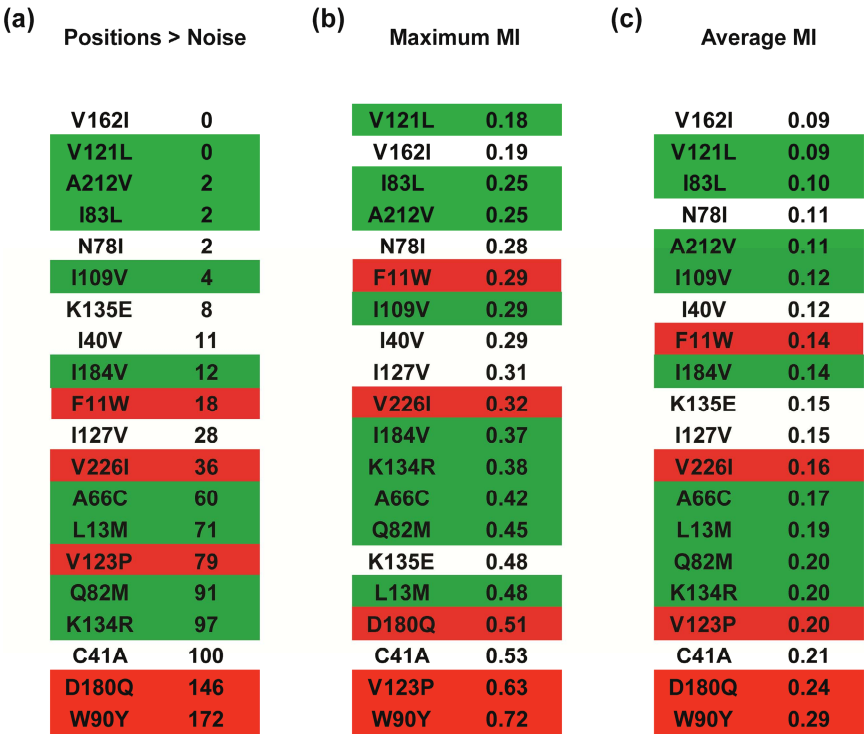
Supplemental Figure 1. The $T_{1/2}$ values are compared between four complementary methods: (1) Loss of CD ellipticity at 222 nm for α-helices, (2) Loss of CD ellipticity at 215 nm for β-strands, (3) Diffraction of light at 600 nm for detection of precipitation products and (4) High-Throughput Thermal Scanning. Here, the $T_{1/2}$ values are plotted for comparison.



Supplemental Figure 2. (a) The percent solvent exposure for each mutation is plotted against the $T_{1/2}$. (b) The computationally predicted ΔΔG from FoldX is plotted against the $T_{1/2}$.



Supplemental Figure 3. The relative entropy and ln (f_{cons}/f_{wt}) are plotted as a function of the $T_{½}$. Neither metric is independently effective at deciphering stabilizing versus destabilizing mutations. The data points shown as squares did not express (DNE). These values were not considered when calculating the R^2 values.



Supplemental Figure 4. Sequence correlation details. Of the 103 possible consensus mutations to *S.c.* TIM only 20 have relative entropies greater than 1.42. (F11W is included here even though its relative entropy is greater than 3.) These 20 mutations are shown here arranged by several measures of correlation. We have individually characterized the effects of stability for 14 of these mutations and that data is shown as red and green for less and more stable, respectively. Mutual information scores for the TIM database range from 0 to 0.83, but MI values less than 0.23 are essentially zero (i.e., noise) based on the scrambled MSA control. (a) Many positions are correlated to each position of mutation above the noise threshold. The strongest site-to-site correlation (b) and the average correlation (c) value for all 240 positions resulting in similar rank-order lists. Note that W90Y, V123P, D18Q and C41A are at or near the bottom of all three lists (i.e., are most correlated).