PNAS
www.pnas.org

Supplementary Information Appendix for

Consensus sequence design as a general strategy to create hyperstable, biologically active proteins

Matt Sternke, Katherine W. Tripp, & Doug Barrick

Corresponding author: Doug Barrick
Email: barrick@jhu.edu

This PDF file includes:

Supplementary methods
Figs. S1 to S15
Tables S1 to S5
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1816707116

Supplementary methods

Cloning, expression, and purification of consensus proteins

Genes encoding the consensus sequences for each protein family were synthesized by GeneArt (ThermoFischer Scientific) as linear, double-stranded fragments. Gene fragments were designed to include a 5' extension (5'-TAAGAAGGAGATATACATATGGGA -3') for cloning into a modified pET24 vector, the consensus protein sequence open reading frame, a sequence encoding for a C-terminal 6x His-tag, three stop codons, and a 3' extension (extending from the 3' end of the stop codon sequence 5'- GGATCCAGACGTAAGCGCACC -3') for cloning. Gene fragments were cloned into linearized vector between *NdeI* and *BamHI* restriction sites. Amino acid sequences for the resulting consensus protein constructs (including additions for cloning and purification) are shown in Table S2.

Consensus proteins were expressed in *E. coli* BL21(DE3) cells. Cells were grown in Luria broth with 50 μg/mL kanamycin at 37 °C until OD_{600} = 0.6-0.8, and were then induced with 1 mM IPTG. Cells were allowed to grow for a further 4-6 hours at 37 °C, pelleted by centrifugation, and cell pellets were stored at -80 °C.

Consensus NTL9, SH3, HD, SH2, and PGK were purified using a common protocol. Cell pellets were resuspended in either 50 mM $NaPO_4$ buffer (pH 7.0; NTL9, SH3, HD, and SH2) or buffer containing 50 mM Tris-HCl (pH 8.0) and 1 mM TCEP (PGK) with a Pierce EDTA-free protease inhibitor cocktail tablet (Thermo Scientific). Cells were lysed by sonication. Cell lysate was centrifuged to separate soluble and insoluble fractions, from which the supernatant was collected. Proteins were purified using Ni-NTA chromatography followed by ion-exchange chromatography. Purified proteins were dialyzed in either 25 mM $NaPO_4$ (pH 7.0) and 150 mM NaCl (NTL9, SH3, HD, and SH2) or 25 mM Tris-HCl (pH 8.0), 50 mM NaCl, and 0.5 mM TCEP (PGK).

Consensus DHFR and AK were purified denatured in urea, since these proteins prepared under native conditions were found to retain endogenous substrates from the cells. Cell pellets were resuspended in buffer containing either 50 mM $NaPO_4$ (pH 7.0) and 8 M urea (DHFR) or 50 mM Tris-HCl (pH 8.0), 1 mM TCEP, and 8 M urea (AK). Cells were lysed by sonication, centrifuged to separate soluble and insoluble portions. The soluble fractions containing denatured AK and DHFR were loaded onto Ni-NTA columns and were refolded on the column by washing into buffer without urea. Proteins were eluted under native conditions and further purified by ion exchange chromatography. Purified consensus DHFR was dialyzed 25 mM $NaPO_4$ (pH 7.0) and 150 mM NaCl. Purified consensus AK was dialyzed into 25 mM Tris-HCl (pH 8.0), 50 mM NaCl, and 0.5 mM TCEP. All proteins were frozen in liquid nitrogen and stored at -80 °C. Protein concentrations for all experiments were determined by UV-Vis absorbance spectroscopy (1).

NMR spectroscopy

^{15}N- and ^{13}C,^{15}N-isotopically labeled proteins were expressed and purified in *E. coli* BL21(DE3) cells grown in M9 minimal media supplemented with $^{15}NH_4Cl$ and either ^{12}C- or ^{13}C-glucose (Cambridge Isotope Laboratories). Proteins were purified as described above, and were concentrated to 400-800 μM.

^{1}H-^{15}N HSQC spectra were collected for ^{15}N-labeled consensus NTL9, SH3, HD, SH2, DHFR, and AK at 25 °C on Bruker Avance or Avance II 600 MHz spectrometers equipped with cryoprobes. NMR samples for NTL9, SH3, HD, and SH2 contained 150 mM NaCl, 5% D_2O, and 25 mM $NaPO_4$ (pH 7.0). Spectra of consensus DHFR were collected under the same conditions both without and with a 1:1 molar equivalent of methotrexate (Sigma Aldrich). Spectra of consensus AK were collected in 50 mM NaCl, 1 mM TCEP, 5% D_2O, and 25 mM Tris-HCl (pH 7.5) at 25 °C. For ^{15}N-labeled consensus

PGK, ^1H-^{15}N TROSY spectra were collected on a Varian 800 MHz spectrometer at 35 °C. NMR samples of consensus PGK contained 50 mM NaCl, 1 mM TCEP, 5% D_2O, and 25 mM Tris-HCl (pH 7.5).

All data were processed and analyzed using NMRPipe (2) and NMRFAM-SPARKY (3). Backbone assignments for consensus NTL9, SH3, and SH2 were made using standard triple-resonance experiments including HNCACB, HNCO, HN(CA)N on Bruker Avance or Avance II 600 MHz spectrometers. Backbone chemical shift data were used for secondary structure predictions using TALOS-N (4).

Consensus SH3 peptide binding was monitored using ^1H-^{15}N HSQC spectra as described in the main text. At each peptide concentration, chemical shift perturbations for each residue ($\Delta\delta_{NH,i}$) were calculated as a weighted Pythagorean distance:

$$\Delta\delta_{NH,i}=\sqrt{\frac{1}{2}[\Delta\delta^2_{H,i} + 0.14\Delta\delta^2_{N,i}]} \qquad (1)$$

where $\Delta\delta_{X,i}$ is the change in chemical shift of the i^{th} resonance in either ^{15}N or ^1H, relative to the apo-protein chemical shift value (5). The weighting factor of 0.14 accounts for differences in ^1H and ^{15}N chemical shift sensitivities. For the ten peaks showing the greatest changes, chemical shift perturbations at each peptide concentrations were globally fit to the single-site binding equation:

$$\Delta\delta_{NH} = \Delta\delta_{max}\frac{([P]_t + [L]_t + K_d) - \sqrt{([P]_t + [L]_t + K_d)^2 - 4[P]_t[L]_t}}{2[P]_t} \qquad (2)$$

where $\Delta\delta_{max}$ is a local parameter for maximal chemical shift perturbation for each peak, $[P]_t$ is the total cSH3 concentration, $[L]_t$ is the total peptide concentration, and K_d is a global dissociation constant (5). In the fit, $\Delta\delta_{max}$ values were optimized locally (for each of the ten resonances), K_d was optimized globally, and $[P]_t$ was fixed at its known value.

Consensus AK enzyme activity at various temperatures was measured using a direct ^{31}P NMR assay previously used for an AK from *A. aeolicus* (6). Conversion of ADP to ATP and AMP was monitored in real time by ^{31}P NMR using a Bruker Avance III HD 400 MHz spectrometer equipped with a broadband probe. ADP at an initial concentration of 13 mM was rapidly mixed with cAK, and 1D ^{31}P NMR spectra were collected continuously (32 scans per spectrum with an inter-scan delay of 8-10 seconds) until equilibrium was reached. For ATP and ADP, peak areas of each ^{31}P resonance (three for ATP, two for ADP) were globally fitted along with the single AMP peak to obtain forward and reverse rate constants. Forward rate constants were converted to k_{cat} values by dividing by cAK concentration. To maintain the kinetics in a measurable range, we decreased the enzyme concentration as temperature was increased from 1.1 μM at 20 °C to 24 nM at 70 °C.

Sequence analysis

Analysis of curated multiple sequence alignments and consensus sequences was performed using in-house scripts (available upon request). Residues were sorted into groups based on physiochemical properties: charged residues (K, R, D, E), polar uncharged residues (N, C, Q, S, T, H), and nonpolar residues (A, I, L, M, V, F, W, G, P, Y). Sequence net charge was calculated assuming contributions of +1 for all K and R residues, -1 for all E and D residues, and 0 for H residues.

Sequence entropies for positions in the multiple sequence alignment were calculated as described in the main text. In all position-by-position comparisons to naturally-occurring proteins, only positions represented in the consensus sequence were considered.

For a particular sequence feature f, the position of the value of the feature for the consensus sequences (f_c) within distributions for multiple sequence alignments (f_{MSA}) were evaluated using a Z-score,

$$Z = \frac{f_c - < f_{MSA} >}{s_{MSA}} \tag{3}$$

where s_{MSA} is the standard deviation of f_{MSA} values. In this context, the Z-score can be thought of as the number of standard deviations the consensus sequence is from the average from the multiple sequence alignment ($<f_{MSA}>$).

Homology models for all consensus proteins were made using SWISS-MODEL(7), using the structure of the sequence displaying the highest sequence identity to the consensus sequence as a template. Residue-specific solvent accessible surface areas were calculated from these homology models using GETAREA (8). Surface, intermediate, and buried positions were defined as positions at which residues show greater than a 50%, 20%--50%, and less than 20% changes in side-chain solvent accessible surface area in homology models of the consensus proteins relative to an ensemble of Gly-X-Gly tripeptides as calculated by the GETAREA algorithm.

To determine the number of thermophilic/mesophilic sequences in the MSAs composed of predominantly bacterial sequences (NTL9, DHFR, AK, and PGK), the source organism of each sequence was identified from the database-specific sequence IDs using the UniProt database (9). For the subset of sequences for which a source organism could be identified by UniProt (on average 97% of sequences in the MSAs) and that belonged to the bacteria or archaea domains, we used the BacDive database (accessed on 3/25/19) to determine whether the source bacterial or archael organism was a thermophile or mesophile using the "Temperature range" classification in the

“Culture and growth conditions” data fields under the “Advanced search” (10). Only organisms that contained both genus and species information in BacDive (10381 sequences total) were assigned to the thermophilic/hyperthermophilic or mesophilic classifications to limit ambiguous classifications. Of these sequences, 853 were thermophilic or hyperthermophilic and 9528 were mesophilic organisms were identified by the BacDive database.

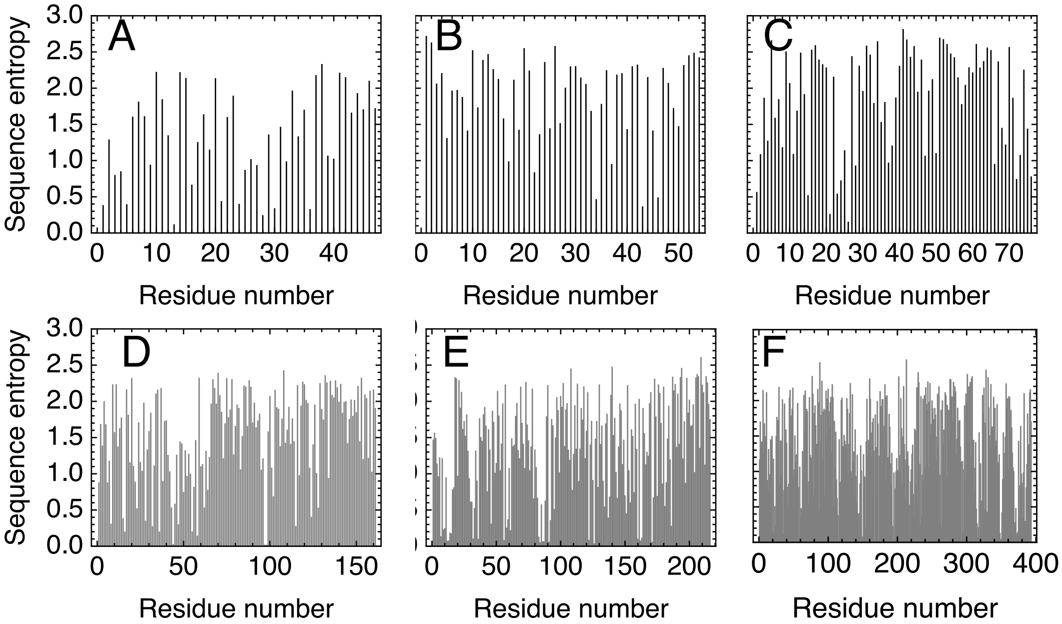

Figure S1. Position-specific conservations of multiple sequence alignments. Sequence entropies shown for all consensus positions in multiple sequence alignments of (A) NTL9, (B) SH3, (C) SH2, (D) DHFR, (E) AK, and (F) PGK.

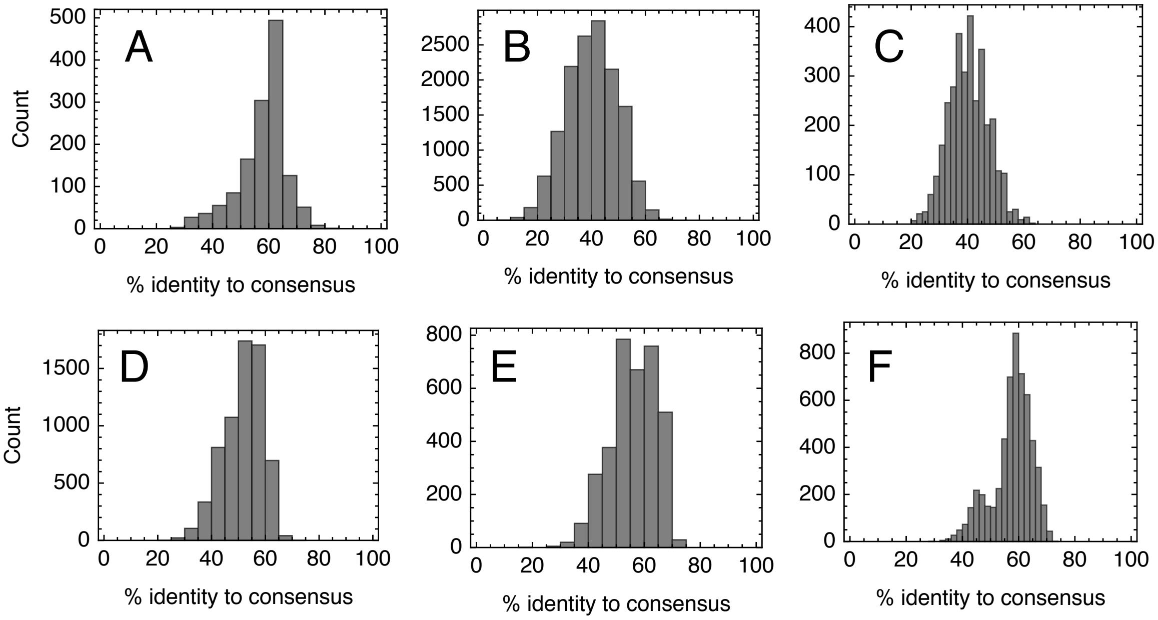

Figure S2. Identities of naturally-occurring sequences to consensus sequence for (A) NTL9, (B) SH3, (C) SH2, (D) DHFR, (E) AK, and (F) PGK.

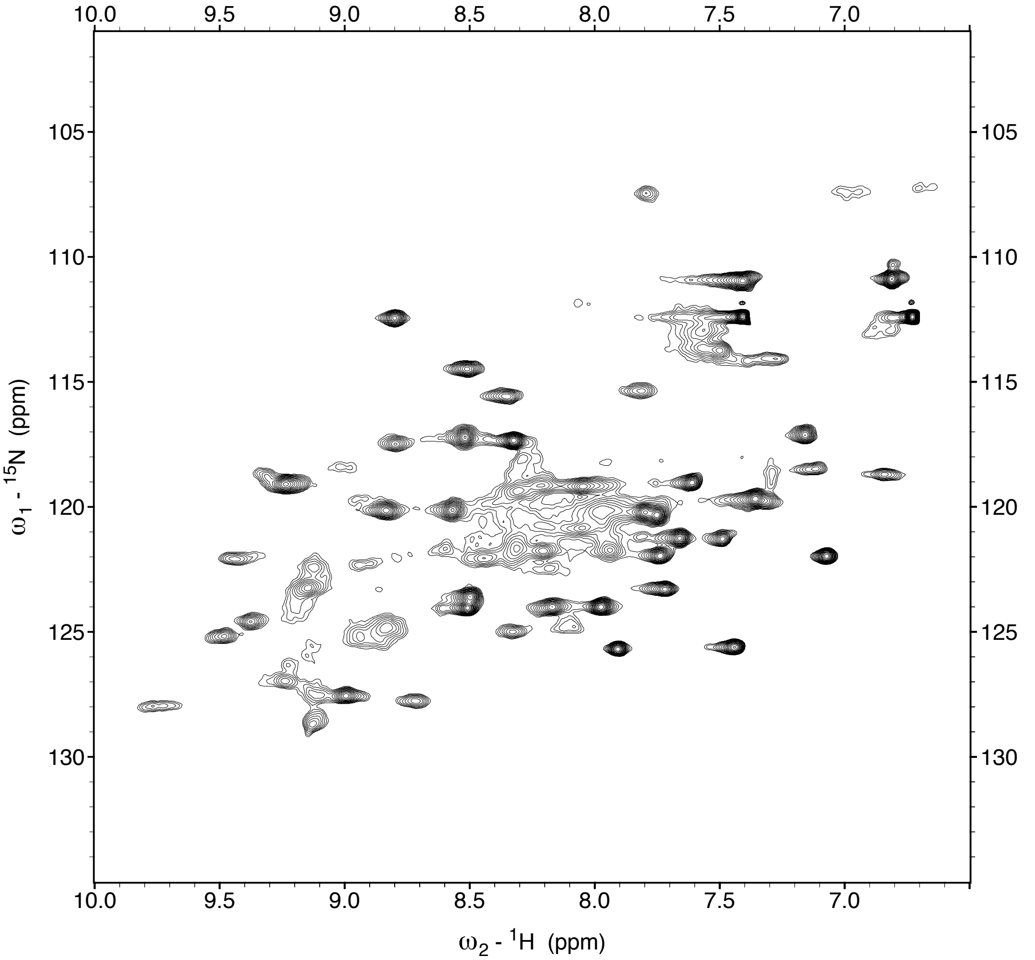

Figure S3. ^{1}H-^{15}N HSQC spectrum of cDHFR in the apo state at 600 MHz. Experimental conditions: 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.0, 25 °C.

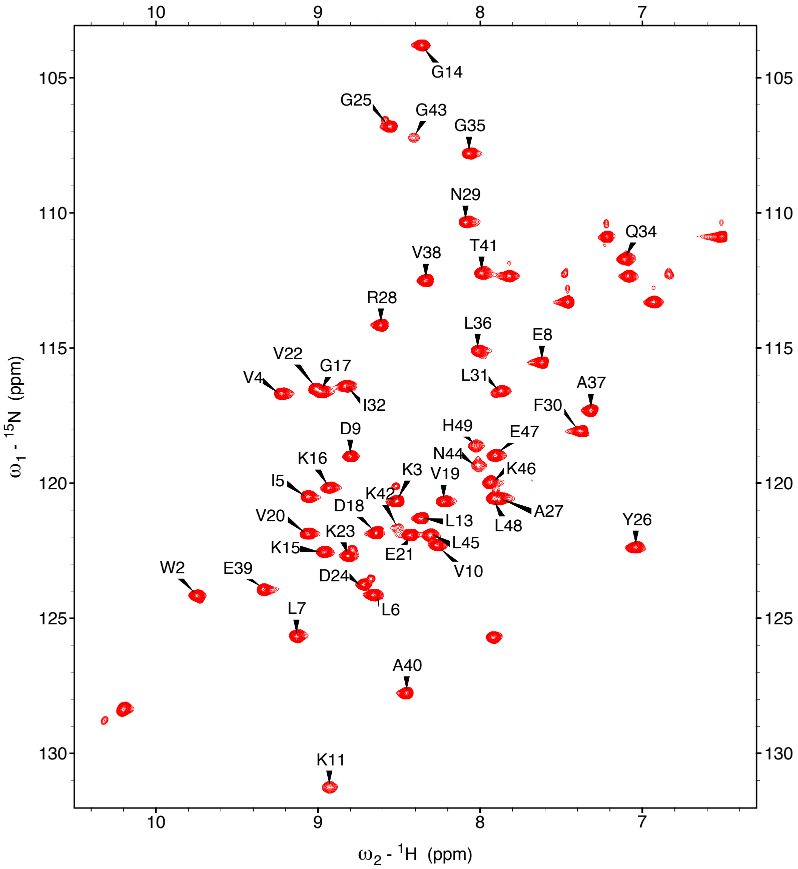

Figure S4. ${}^{1}H$-${}^{15}N$ HSQC spectrum of cNTL9 at 600 MHz. Assigned peaks are label. Experimental conditions: 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.0, 25 °C.

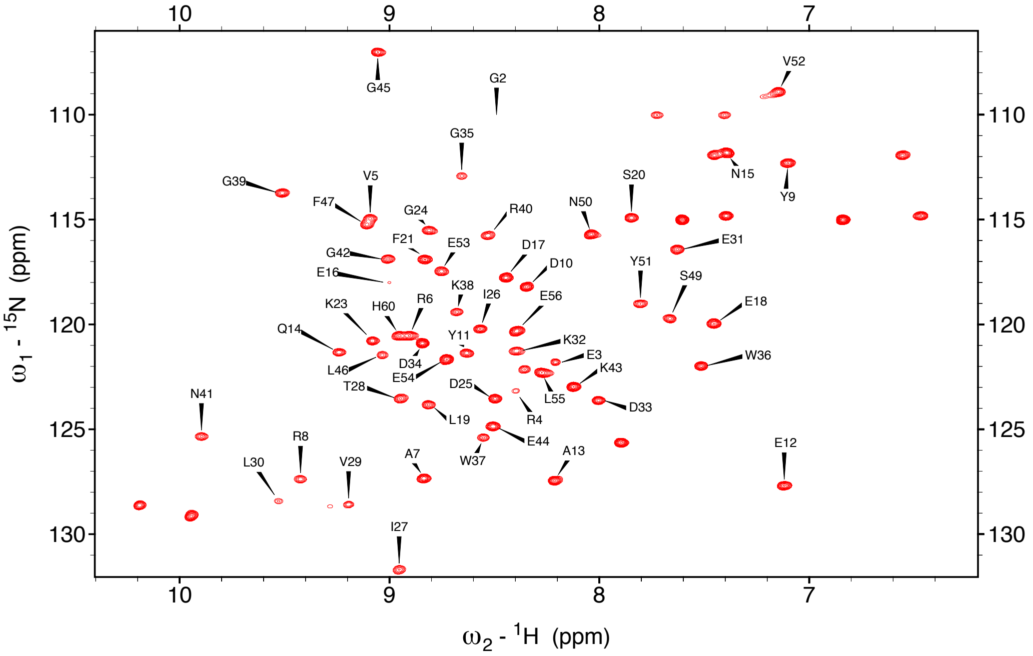

Figure S5. ^{1}H-^{15}N HSQC spectrum of cSH3 at 600 MHz. Assigned peaks are label. Experimental conditions: 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.0, 25 °C.

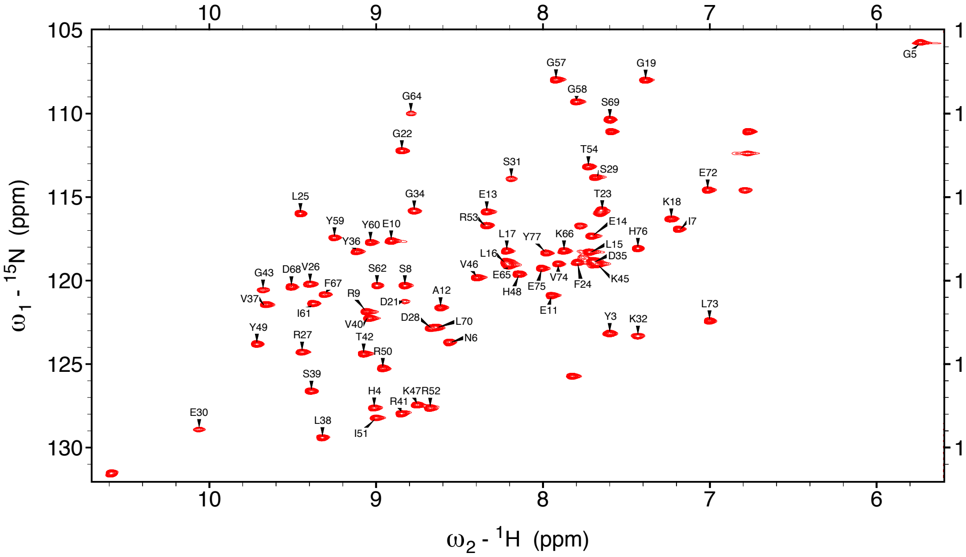

Figure S6. ^{1}H-^{15}N HSQC spectrum of cSH2 at 600 MHz. Assigned peaks are label. Experimental conditions: 25 mM sodium phosphate, 150 mM sodium chloride, pH 7.0, 25 °C.

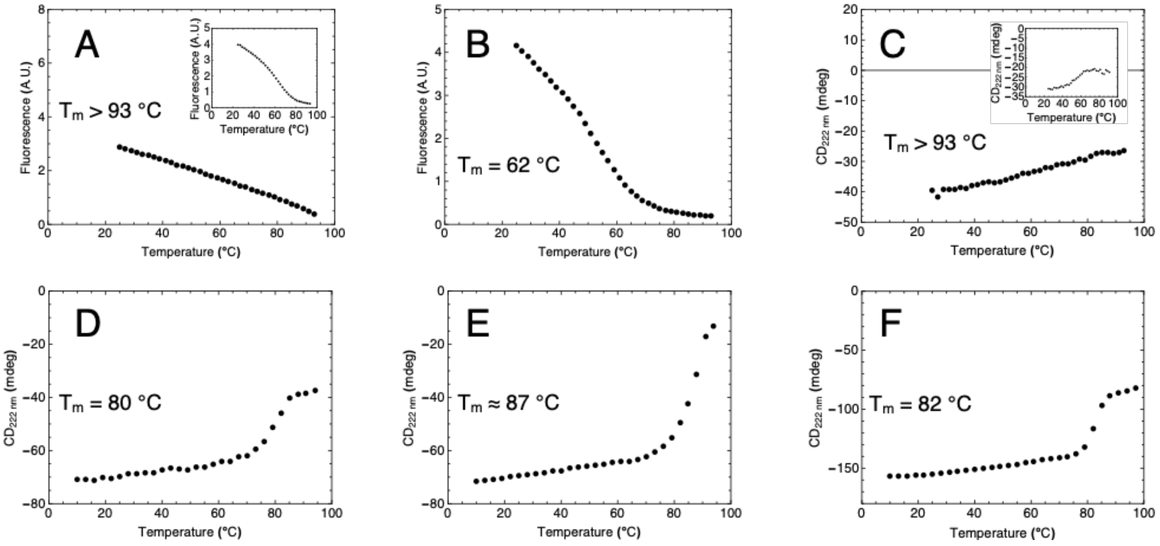

Figure S7. Temperature-induced unfolding curves for (A) cNTL9, (B) cSH3, (C) cSH2, (D) cDHFR, (E) cAK, and (F) cPGK. For cNTL9 and cSH2 (A and C), there are no obvious thermal unfolding transitions up to 93 °C; insets show unfolding transitions in the presence of 4 and 2 M GdnHCl, respectively. Plots show raw values measured using circular dichroism (SH2, DHFR, AK, and PGK) or fluorescence (NTL9, and SH3) spectroscopies.

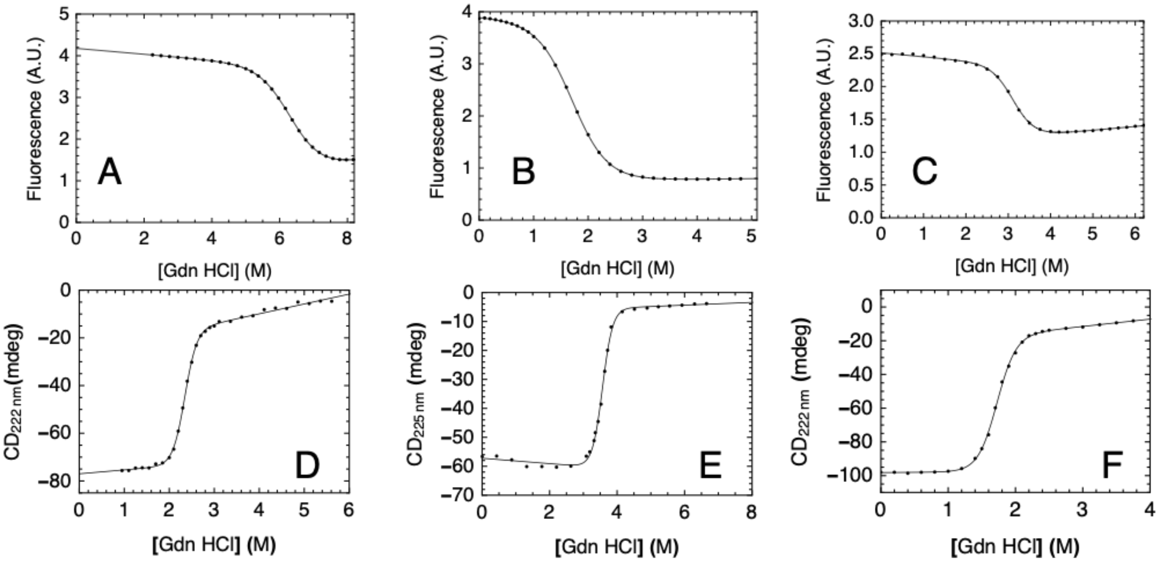

Figure S8. Guanidine hydrochloride-induced unfolding curves for (A) cNTL9, (B) cSH3, (C) cSH2, (D) cDHFR, (E) cAK, and (F) cPGK. Plots show raw values measured using circular dichroism (cDHFR, cAK, and cPGK) or fluorescence (cNTL9, cSH3, cSH2) spectroscopies. Solid lines are obtained from fitting a two-state model to the data. Parameters obtained from two-state fits to raw data were used to convert to fraction folded. Experimental conditions are as noted in main text.

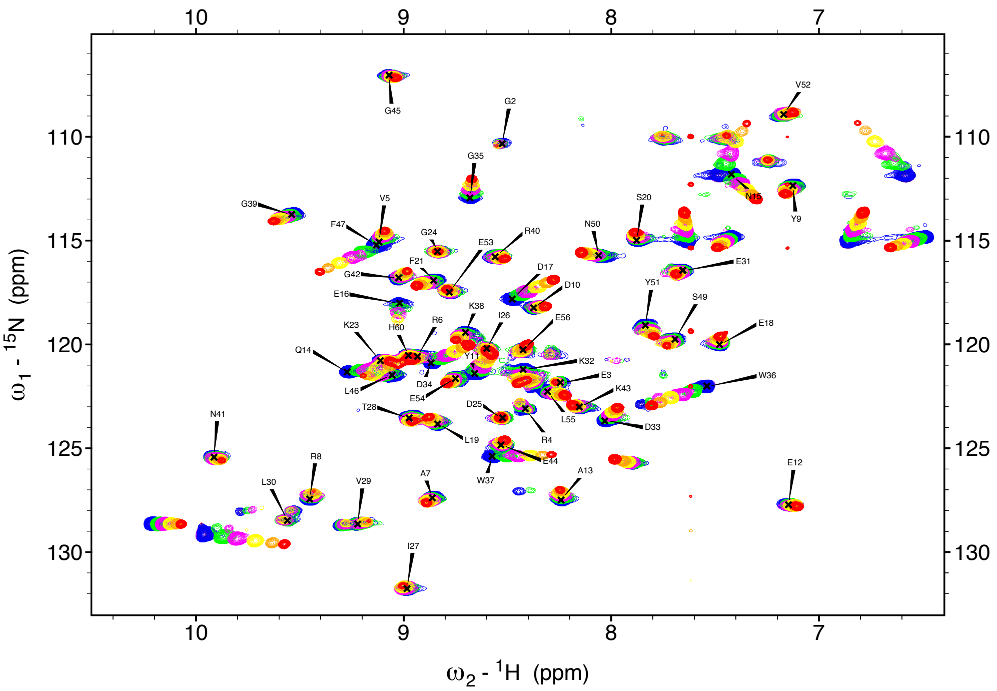

Figure S9. ^{1}H-^{15}N HSQC peptide titration of cSH3. Experiments were collected at 0- (purple), 0.05- (dark blue), 0.125- (blue), 0.5- (dark green), 1.25- (green), 2.5- (magenta), 5- (yellow), 10- (orange), and 20-fold (red) saturation of nonisotopically labeled peptide concentrations. 200 μM consensus SH3 was used for all experiments. Residues assigned in apo-spectrum are labeled.

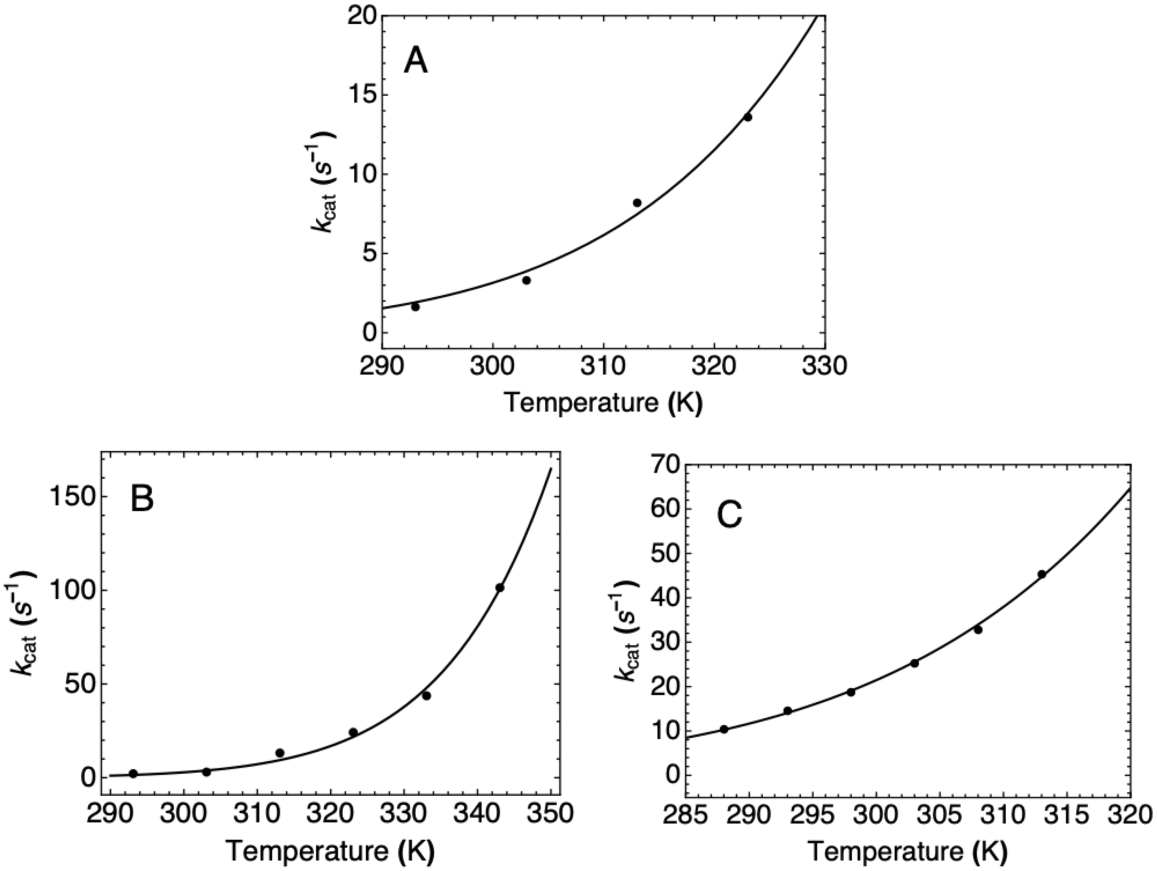

Figure S10. Temperature dependence of (A) cDHFR (in the DHF + NADPH to THF + $NADP^+$ direction), (B) cAK (in the 2ADP to ATP + AMP direction), and (C) cPGK (in the 3-PG + ATP to 1,3-BPG + ADP direction). Activities for cDHFR andc PGK were measured using absorbance spectroscopic assays; activities for cAK were measured using ^{31}P NMR assay (see Supplementary methods, main text). Solid lines are obtained by fitting an Arrhenius model to the data. Experimental conditions are given in the main text.

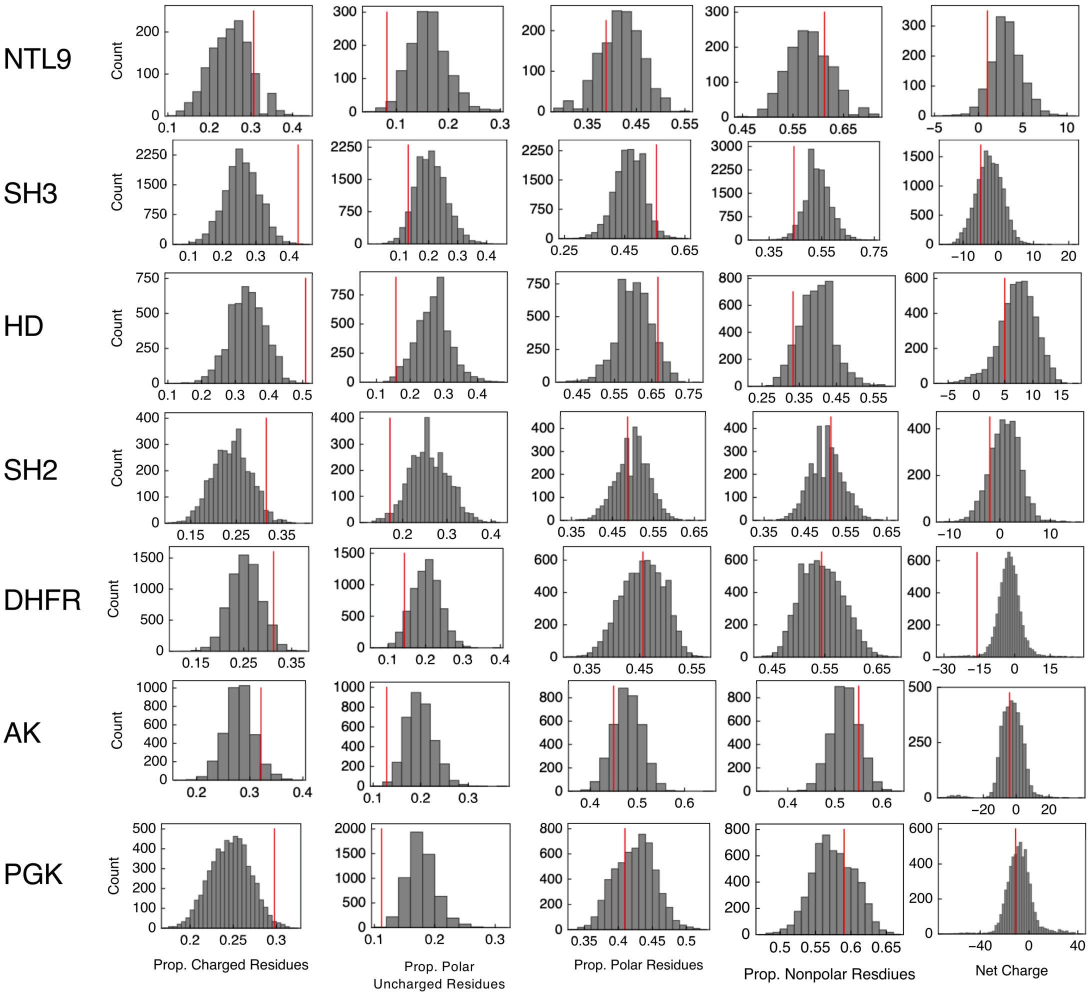

Figure S11. Sequence biases of consensus sequences. Distributions of proportions of sequence made up of charged, polar uncharged, total polar, and nonpolar residues, and net charge of extant sequences in final multiple sequence alignments. Red lines indicate where the consensus sequence lies for each parameter.

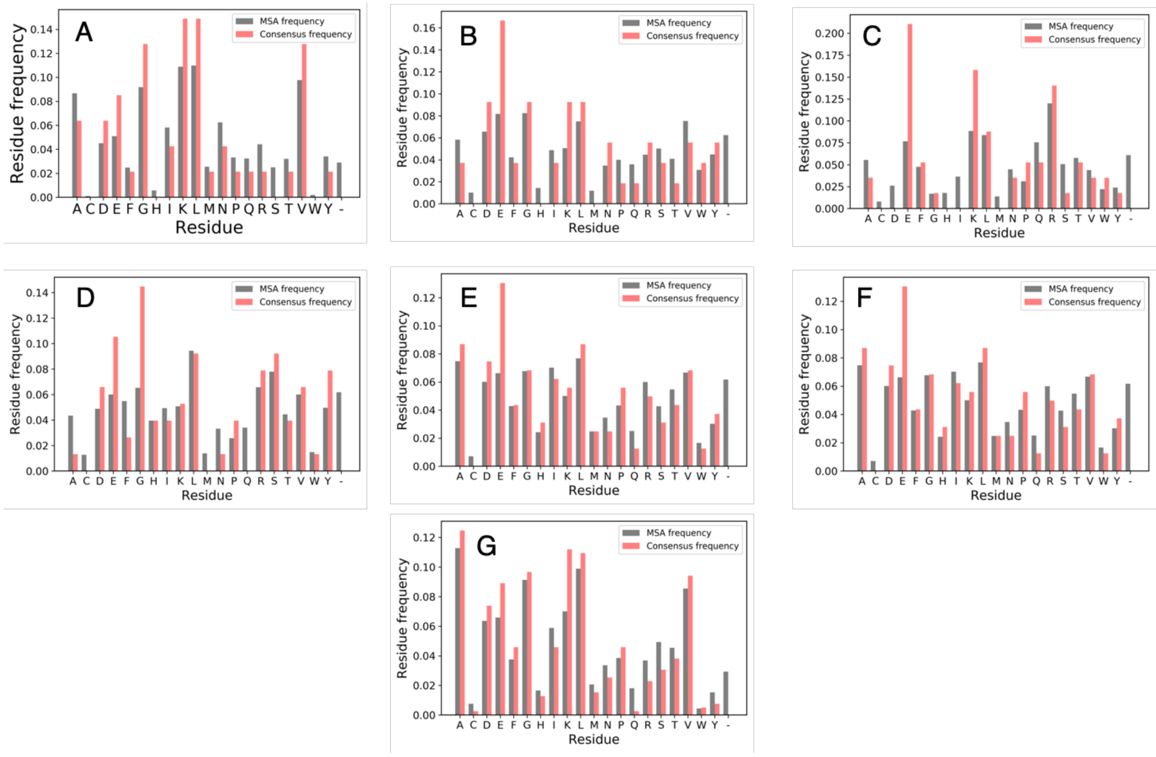

Figure S12. Residue frequencies for extant sequences in MSA (black) and consensus sequence (red) for (A) NTL9, (B) SH3, (C) HD, (D) SH2, (E) DHFR, (F) AK, and (G) PGK.

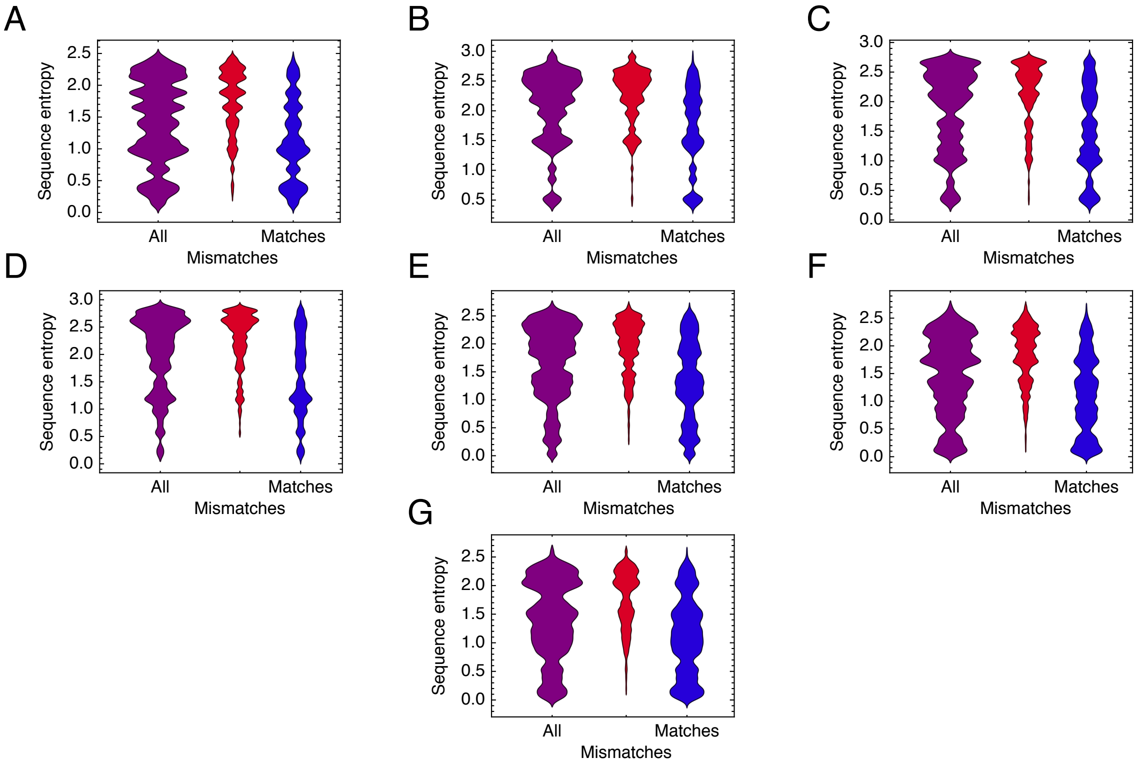

Figure S13. Positional conservation bias of consensus mismatches. Distributions of sequence entropy values for all positions in MSA (“all”; purple), positions at which extant sequences differ from the consensus sequence (“mismatches”; red), and positions at which extant sequences match the consensus sequence (“matches”; blue) for consensus (A) NTL9, (B) SH3, (C) HD, (D) SH2, (E) DHFR, (F) AK, and (G) PGK.

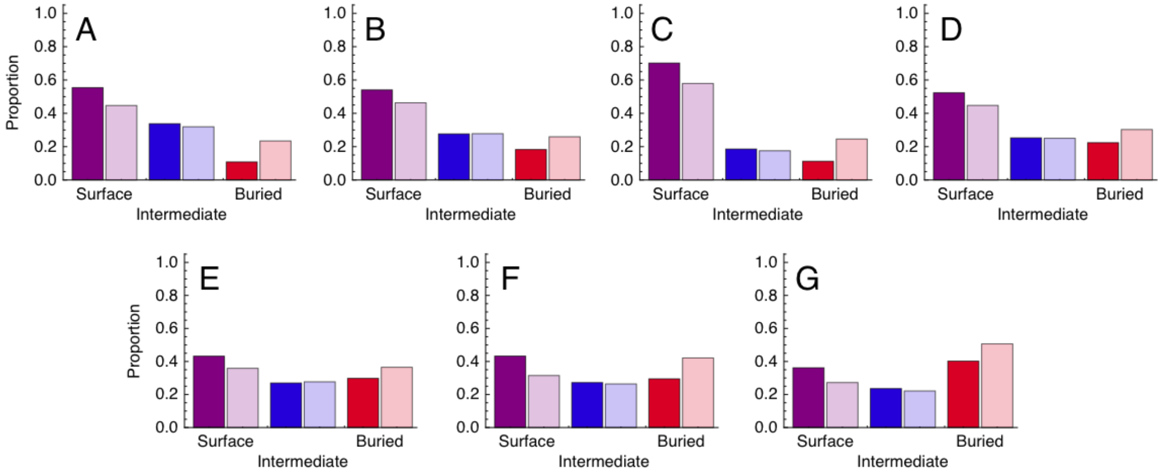

Figure S14. Surface exposure of consensus mismatches. Dark bars show the proportion of mismatch residues that are at surface (purple), intermediate (blue), and buried positions (red) for different consensus protein targets. Light bars show the proportion of all residues at surface, intermediate, and buried positions. (A) NTL9, (B) SH3, (C) HD, (D) SH2, (E) DHFR, (F) AK, and (G) PGK. The degree of burial is determined as described in the main text.

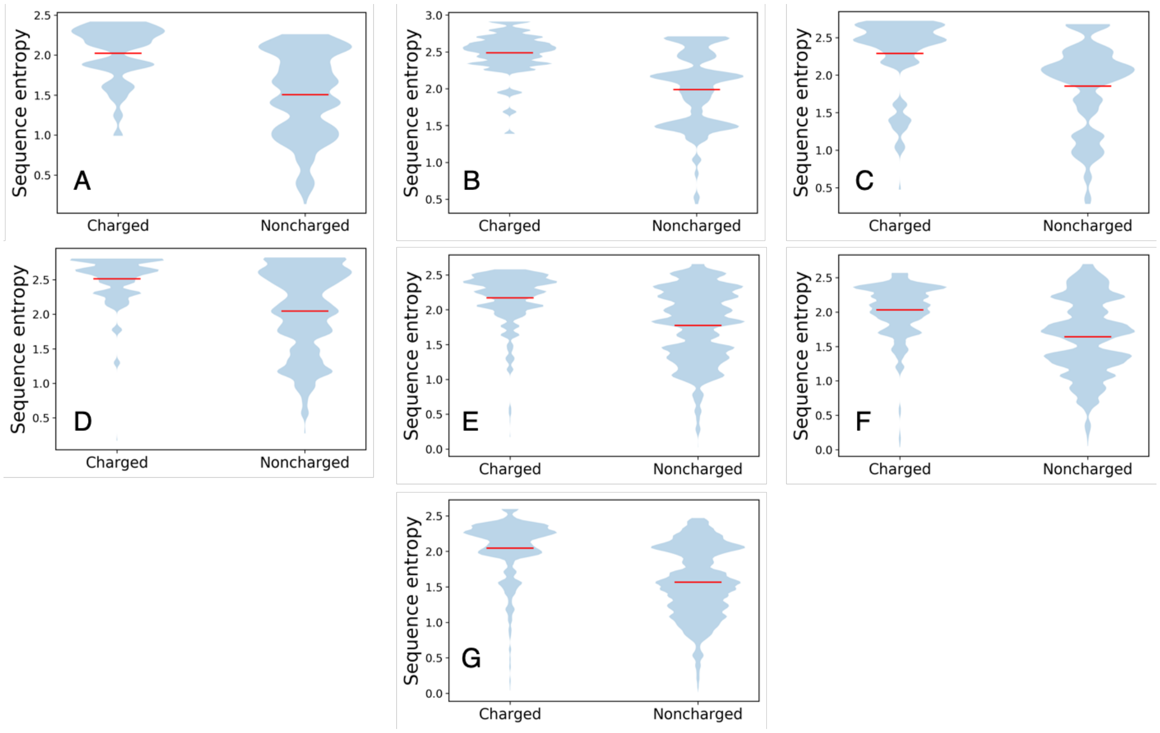

Figure S15. Sequence entropy distributions of consensus substitutions to from uncharged to charged residues (left) and substitutions among uncharged residues (right) for (A) NTL9, (B) SH3, (C) HD, (D) SH2, (E) DHFR, (F) AK, and (G) PGK. The red line signifies the mean of each distribution.

Table S1. Data for sequence sets used for consensus sequence generation.

Protein family	Database Accession identity Date accessed	Number of sequence in initial set	Number of sequences after curation	Average pairwise sequence identity	Phylogenetic distribution of sequences
NTL9	Pfam PF01281 5/20/16	1911	1355	42%	Ar: 0% Bac: 83.9% Eu: 16.1%
SH3	SMART SM00326 4/4/17	54382	14474	26%	Ar: 0.0% Bac: 4.3% Eu: 95.7%
SH2	Pfam PF00017 5/20/16	10051	3326	28%	Ar: 0.0% Bac: 0.4% Eu: 99.6%
DHFR	InterPro IPR001796 5/2/17	16038	6542	37%	Ar: 2.3% Bac: 92.4% Eu: 5.3%
AK	InterPro IPR007862 2/9/17	13905	3534	43%	Ar: 4.1 % Bac: 74.5% Eu: 21.5%
PGK	InterPro IPR001576 4/5/17	17724	5581	45%	Ar: 6.2% Bac: 82.7% Eu: 11.1%

Databases used to obtain sequence sets for each protein family are noted along with the database-specific accession identity of each protein family and the date sequence sets were obtained. The number of sequences in the initial set gathered directly from the database is reported along with the number of sequences in the final set used for consensus sequence generation after sequences were removed based on sequence length and sequence identity (see main text). Reported average pairwise identity is the average of the identities of all pairwise comparisons between sequences in the final sequence set used. Phylogenetic distributions represent the percentage of sequences in the final sequence set classified as archeal (Ar), bacterial (Bac), and eukaryotic (Eu) by the database used to obtain the sequence set.

Table S2. Consensus sequence constructs generated for protein families.

Protein family	Consensus sequence
NTL9	MGWKVILLEDVKGLGKKGDVVEVKDGYARNFLIPQGLAVEATKGNLKELHHHHHH
SH3	MGERVRARYDYEAQNEDELSFKKGDIITVLEKDDGWWKGRNGKEGLFPSNYVEELEHHHHHH
SH2	MGWYHGNISREEAEELLLKGPDGTFLVRDSESKPGDYVLSVRTGGKVKHYRIRRTDGGGYYISGGEKFDSLPELVEHYHHHHHH
DHFR	MGISLIVAVAENGVIGKDNDLPWHLPEDLKHFKELTMGHPVIMGRKTFESIGRPLPGRRNIVLTRDPDYQAEGAEVVHSLEEALALAKEAEEVFVIGGAEIYAQALPLADRLYLTEIDADFEGDTFFPEIDSEWKEVSREEHPADEKNGYDYTFVTYERKKHHHHHH
AK	MGWRIILLGPPGAGKGTQAKRIVEKYGIPHISTGDMLRAAIKAGTELGKKAKSYMDAGELVPDEIVIGLVKERLAQPDCNGFLLDGFPRTIPQAEALDELLKELGVKLDAVIELDVPDEELVERLSGRRVCPAKCGRTYHVKFNPPKVEGVCDVCGEELIQRDDDKEETVRKRLEVYHEQTAPLIDYYKKKGLLVTVDGTGSIDEVFADILAALGKKKHHHHHH
PGK	MGNKTIDDLDLKGKRVLVRVDFNVPLKDGKITDDTRIRAALPTIKYLLEKGAKVILMSHLGRPKGEVDPKFSLAPVAKRLSELLGKPVKFADDCVGEEAEAAVAALKPGEVLLLENLRFHKGEEKNDPEFAKKLASLGDVYVNDAFGTAHRAHASTVGVAKFLPAAAGFLMEKELEALGKALENPERPFVAILGGAKVSDKIGVIENLLDKVDKLIIGGGMANTFLKAQGYEVGKSLVEEDKLDTAKELLEKAKEKGVKIVLPVDVVVADEFSADAETKVVPVDEIPDDWMGLDIGPKTVELFAEAIKDAKTIVWNGPMGVFEFEPFAKGTKAVAKAIAEATGAFSIVGGGDTAAAVNKLGLADKFSHISTGGGASLEFLEGKELPGVAALEDKHHHHHH

Consensus sequences generated for each protein family are shown. Consensus sequence derived from multiple sequence alignments are shown in black. Residues added to consensus sequences for cloning (N-terminal MG), quantification (N-terminal W, added if consensus sequence did not contain a tryptophan), or purification (C-terminal His-tag) purposes are shown in red.

Table S3. Stabilities of naturally-occurring proteins reported in literature.

Protein	Organism	Denaturant used	$\Delta G°_{H2O}$ (kcal/mol)	Mean $\Delta G°_{H2O}$ (kcal/mol)	*m*-value (kcal/mol/M)	Reference
NTL9	Consensus	Gdn HCl	-7.9	-3.2	1.23	This study
	E. coli	Gdn HCl	-1.98		1.21	(11)
	B. stearothermophilus	Urea	-4.5		NR	(12)
SH3	Consensus	Gdn HCl	-3.2	-3.5	1.88	This study
	Human Fyn	Gdn HCl	-5.0		1.5	(13)
	Human BTK	Gdn HCl	-2.6		1.18	(14)
	C. elegans Sem-5	Gdn HCl	-4.1		1.7	(15)
	Human Src	Gdn HCl	-4.1		1.6	(16)
	D. melanogaster drk	Gdn HCl	-2.18		1.39	(17)
	Human PI3K	Gdn HCl	-3.23		2.33	(18)
	Chicken α-spectrin	Urea	-3.6		0.74	(19)
	Yeast Abp1p	Gdn HCl	-3.1		1.64	(20)
	Mouse Crk-II	Urea	-3.33		0.66	(21)
HD	Consensus	Gdn HCl	-8.1	-2.8	1.6	(22)
	D. melanogaster Engrailed	Gdn HCl	-3.62		1.39	(22)
	Rat TTF-1	Urea	-1.62		NR	(23)
	Human Hesk-1	Urea	-4.43		0.9	(24)
	Rat ISL-1	Urea	-1.67		1.17	(25)
	D. melanogaster Antp	Gdn HCl	-2.85		0.61	(26)
	D. melanogaster TTF-1	Gdn HCl	-2.59		0.65	(26)
SH2	Consensus	Gdn HCl	-7.2	-3.6	2.28	This study
	Human CSK	Gdn HCl	-6.57		2.19	(27)
	Human BTK	Gdn HCl	-2.95		1.34	(28)
	Human Stat5b	Gdn HCl	-1.2		NR	(29)
DHFR	Consensus	Gdn HCl	-10.1	-3.4	4.32	This study
	E. coli	Gdn HCl	-6.97		4.1	(30)
	Human	Gdn HCl	-4.2		5.87	(31)
	M. profunda	Urea	-1.89		1.03	(32)
	S. benthica DB21MT-2	Urea	-2.08		1.1	(33)
	S. benthica DB6705	Urea	-1.89		1.12	(33)
	S. frigidmarina	Urea	-1.98		0.96	(33)
	S. oneidensis	Urea	-1.6		1.4	(33)
	S. putrefaciens	Urea	-1.98		1.65	(33)
	S. violacea	Urea	-1.91		0.86	(33)

	B. stearothermophilus	Gdn HCl	-7.4		NR	(34)
	L. casei	Urea	-4.6		2.4	(35)
	Mouse	Urea	-4.4		NR	(36)
AK	Consensus	Gdn HCl	-13.6	-5.3	3.83	This study
	Yeast	Gdn HCl	-5.47		7.29	(37)
	B. subtilis	Gdn HCl	-3.4		2.9	(38)
	Chicken	Gdn HCl	-3.8		4.7	(39)
	Pig	Gdn HCl	-3.9		4.8	(39)
	E. coli	Gdn HCl	-9.8		2.9	(40)
PGK	Consensus	Gdn HCl	-7.5	-6.1	4.34	This study
	Yeast	Gdn HCl	-3.63		5.86	(41)
	T. thermophiles	Gdn HCl	-6.32		2.7	(41)
	E. coli	Urea	-9.3		NR	(42)
	Human PGK1	Urea	-8.3		3.4	(43)
	Horse	Gdn HCl	-2.87		NR	(44)

Data were gathered by searching the PubMed database for free energies of folding for naturally-occurring sequences of respective protein families. The search was limited to free energy values determined by chemical denaturation experiments and to proteins that appeared monomeric. “Mean “ folding free energy denotes average of values for extant homologues. NR denotes values that were not reported in the studies.

Table S4. Temperature dependence of consensus enzyme activities.

	Temperature	k_{cat}	E_{act}	A_0	Reference
	(°C)	(s^{-1})	(kcal mol^{-1})	(s^{-1})	
DHFR					
Consensus	---	---	12.4	3.3×10^9	This study
	20	1.7			
	30	3.3			
	40	8.2			
	50	13.6			
	[a]*60*	[a]*24.0*			
B. stearothermophilus	60	4.8	NR	NR	(45)
T. maritima	60	0.76	[b]11.1	NR	(46)
AK					
Consensus	---	---	16.9	5.8×10^{12}	This study
	20	1.7			
	30	3.3			
	40	12.9			
	50	23.8			
	60	43.6			
	70	101.6			
A. aeolicus	50	[c]500	NR	NR	(6)
PGK					
Consensus	---	---	10.5	1.0×10^9	This study
	15	10.6			
	20	14.7			
	25	18.9			
	30	25.3			
	35	32.9			
	40	45.4			
	[a]*60*	[a]*128*			
	[a]*70*	[a]*203*			
	[a]*75*	[a]*254*			
T. maritima	40	231	9.1	NR	(47)
T. brockii	40	[c]600	NR	NR	(48)
M. fervidus	60	150	9.80	NR	(49)
P. woesei	70	113	16.7	NR	(49)
T. thermophilus	75	1015	NR	NR	(50)

Temperature dependence of consensus enzyme activities. DHFR (in direction of tetrahydrofolate formation) and PGK (in the direction of 1,3-bisphosphoglycerate formation) turnover numbers (k_{cat}) were measured using absorbance spectroscopy. AK (in the direction of ATP and AMP formation) turnover numbers were measured using a real-time ^{31}P NMR . Activation energies (E_{act}) and pre-exponential factors (A_0) were determined by fitting an Arrhenius model to the data (Figure S10). NR denotes that values were not reported in the studies.

[a]denotes k_{cat} values that were extrapolated using fitted parameters in an Arrhenius model

[b]Arrhenius plot shows two linear regimes with different slopes with a break point around 25 °C . The E_{act} value reported is from the linear regime above 25 °C.

[c]denotes values that were estimated from graphs in references

Table S5. Thermophilic/mesophilic sequence composition for predominantly bacterial protein families.

Protein family	Bacterial + archeal sequences in set	Sequence source organisms identified by BacDive	Thermophilic (T), mesophilic (M), and unclassified (U) sequences	Average identity to consensus for thermophilic (T) and mesophilic (M) sequences
NTL9	1136	761	T: 66 (4.8%) M: 510 (37.6%) U: 185	T: 54% M: 55%
	MKVILLEDVKGLGKKGDVVEVKDGYARNFLIPQGLAVEATKGNLKEL			
DHFR	6195	3059	T: 63 (1.0%) M: 1884 (28.8%) U: 1112	T: 47% M: 48%
	MISLIVAVAENGVIGKDNDLPWHLPEDLKHFKALTMGHPVIMGRKTFESIGRPLPGRRNIV LTRDPDYQAEGAEVVHSLEEALALAKEAEEVFVIGGAEIYAQALPLADRLYLTEIDADFEG DTFFPEIDSEWKEVSREEHPADEKNGYDYTFVTYERKK			
AK	2768	1445	T: 159 (4.4%) M: 835 (23.6%) U: 451	T: 56% M: 57%
	MRIILLGPPGAGKGTQAKRIVEKYGIPHISTGDMLRAAIKAGTELGKKAKSYMDAGELVPD EIVIGLVKERLAQPDCNGFLLDGFPRTIPQAEALDELLKELGVKLDAVIELDVPDEELVERI SGRRVCPASCGRTYHVKFNPPKVEGVCDVCGEELIQRDDDNEETVRKRLEVYHEQTAP LIDYYKKKGLLVTVDGTGSIDEVFADILAALGKKK			
PGK	4961	2361	T: 234 (4.2%) M: 1374 (24.6%) U: 753	T: 56% M: 57%
	MNKTIDDLDLKGKRVLVRVDFNVPLKDGKITDDTRIRAALPTIKYLLEKGAKVILMSHLGRP KGEVDPKFSLAPVAKRLSELLGKPVKFADDCVGEEAEAAVAALKPGEVLLLENLRFHKGE EKNDPEFAKKLASLGDVYVNDAFGTAHRAHASTVGVAKFLPAAAGFLMEKELEALGKAL ENPERPFVAILGGAKVSDKIGVIENLLDKVDKLIIGGGMANTFLKAQGYEVGKSLVEEDKL DTAKELLEKAKEKGVKIVLPVDVVVADEFSADAETKVVPVDEIPDDWMGLDIGPKTVELFA EAIKDAKTIVWNGPMGVFEFEPFAKGTKAVAKAIAEATGAFSIVGGGDTAAAVNKLGLAD KFSHISTGGGASLEFLEGKELPGVAALEDK			

The numbers of bacterial and archeal sequences are as reported in Table S1. Sequence source organisms identified by BacDive indicates the number of sequences in each MSA whose source organisms could be in the BacDive database (10). T and M indicate the number of sequence parent organisms that were classified as thermophilic and mesophilic by the BacDive database. U indicates the number of sequence parent organisms that had no growth temperature classification or had ambiguous classifications. For each family, the sequence is the consensus obtained if thermophilic sequences are removed from the MSA. Residues that differ from the consensus sequence of the full MSA are highlighted in red.

References

1. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T (1995) How to measure and predict the molar absorption coefficient of a protein. *Protein Sci Publ Protein Soc* 4(11):2411–2423.

2. Delaglio F, et al. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6(3):277–293.

3. Lee W, Tonelli M, Markley JL (2015) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. *Bioinforma Oxf Engl* 31(8):1325–1327.

4. Shen Y, Bax A (2015) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol Clifton NJ* 1260:17–32.

5. Williamson MP (2013) Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc* 73:1–16.

6. Rogne P, Sparrman T, Anugwom I, Mikkola J-P, Wolf-Watz M (2015) Realtime (31)P NMR Investigation on the Catalytic Behavior of the Enzyme Adenylate kinase in the Matrix of a Switchable Ionic Liquid. *ChemSusChem* 8(22):3764–3768.

7. Biasini M, et al. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res* 42(Web Server issue):W252-258.

8. Fraczkiewicz Robert, Braun Werner (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem* 19(3):319–333.

9. UniProt: a worldwide hub of protein knowledge. (2019) *Nucleic Acids Res* 47(D1):D506–D515.

10. Reimer LC, et al. (2019) BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res* 47(D1):D631–D636.

11. Sato S, Xiang S, Raleigh DP (2001) On the relationship between protein stability and folding kinetics: a comparative study of the N-terminal domains of RNase HI, E. coli and Bacillus stearothermophilus L9. *J Mol Biol* 312(3):569–577.

12. Kuhlman B, Luisi DL, Young P, Raleigh DP (1999) pKa values and the pH dependent stability of the N-terminal domain of L9 as probes of electrostatic interactions in the denatured state. Differentiation between local and nonlocal interactions. *Biochemistry* 38(15):4896–4903.

13. Maxwell KL, Davidson AR (1998) Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. *Biochemistry* 37(46):16172–16182.

14. Chen YJ, et al. (1996) Stability and folding of the SH3 domain of Bruton's tyrosine kinase. *Proteins* 26(4):465–471.

15. Lim WA, Fox RO, Richards FM (1994) Stability and peptide binding affinity of an SH3 domain from the Caenorhabditis elegans signaling protein Sem-5. *Protein Sci Publ Protein Soc* 3(8):1261–1266.

16. Grantcharova VP, Baker D (1997) Folding dynamics of the src SH3 domain. *Biochemistry* 36(50):15685–15692.

17. Crowhurst KA, Tollinger M, Forman-Kay JD (2002) Cooperative interactions and a non-native buried Trp in the unfolded state of an SH3 domain. *J Mol Biol* 322(1):163–178.

18. Guijarro JI, Morton CJ, Plaxco KW, Campbell ID, Dobson CM (1998) Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J Mol Biol* 276(3):657–667.

19. Ventura S, et al. (2002) Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat Struct Biol* 9(6):485–493.

20. Rath A, Davidson AR (2000) The design of a hyperstable mutant of the Abp1p SH3 domain by sequence alignment analysis. *Protein Sci Publ Protein Soc* 9(12):2457–2469.

21. Muralidharan V, et al. (2006) Solution structure and folding characteristics of the C-terminal SH3 domain of c-Crk-II. *Biochemistry* 45(29):8874–8884.

22. Tripp KW, Sternke M, Majumdar A, Barrick D (2017) Creating a Homeodomain with High Stability and DNA Binding Affinity by Sequence Averaging. *J Am Chem Soc*. doi:10.1021/jacs.6b11323.

23. Damante G, et al. (1994) Analysis of the conformation and stability of rat TTF-1 homeodomain by circular dichroism. *FEBS Lett* 354(3):293–296.

24. Torrado M, et al. (2009) Role of conserved salt bridges in homeodomain stability and DNA binding. *J Biol Chem* 284(35):23765–23779.

25. Behravan G, Lycksell PO, Larsson G (1997) Expression, purification and characterization of the homeodomain of rat ISL-1 protein. *Protein Eng* 10(11):1327–1331.

26. Tell G, et al. (1999) Comparative stability analysis of the thyroid transcription factor 1 and Antennapedia homeodomains: evidence for residue 54 in controlling the structural stability of the recognition helix. *Int J Biochem Cell Biol* 31(11):1339–1353.

27. Liu D, Cowburn D (2016) Combining biophysical methods to analyze the disulfide bond in SH2 domain of. *Biophys Rep* 2(1):33–43.

28. Tzeng Shiou-Ru, et al. (2008) Stability and peptide binding specificity of Btk SH2 domain: Molecular basis for X-linked agammaglobulinemia. *Protein Sci* 9(12):2377–2385.

29. Chia DJ, et al. (2006) Aberrant folding of a mutant Stat5b causes growth hormone insensitivity and proteasomal dysfunction. *J Biol Chem* 281(10):6552–6558.

30. Villafranca JE, Howell EE, Oatley SJ, Xuong NH, Kraut J (1987) An engineered disulfide bond in dihydrofolate reductase. *Biochemistry* 26(8):2182–2189.

31. Chunduru SK, et al. (1994) Methotrexate-resistant variants of human dihydrofolate reductase. Effects of Phe31 substitutions. *J Biol Chem* 269(13):9547–9555.

32. Ohmae E, et al. (2012) Pressure dependence of activity and stability of dihydrofolate reductases of the deep-sea bacterium Moritella profunda and Escherichia coli. *Biochim Biophys Acta* 1824(3):511–519.

33. Murakami C, et al. (2011) Comparative study on dihydrofolate reductases from Shewanella species living in deep-sea and ambient atmospheric-pressure environments. *Extrem Life Extreme Cond* 15(2):165–175.

34. Schulenburg C, Stark Y, Kunzle M, Hilvert D (2015) Comparative laboratory evolution of ordered and disordered enzymes. *J Biol Chem* 290(15):9310–9320.

35. Wallace LA, Robert Matthews C (2002) Highly divergent dihydrofolate reductases conserve complex folding mechanisms. *J Mol Biol* 315(2):193–211.

36. Endo T, Schatz G (1988) Latent membrane perturbation activity of a mitochondrial precursor protein is exposed by unfolding. *EMBO J* 7(4):1153–1158.

37. Spuergin P, Abele U, Schulz GE (1995) Stability, activity and structure of adenylate kinase mutants. *Eur J Biochem* 231(2):405–413.

38. Counago R, Wilson CJ, Pena MI, Wittung-Stafshede P, Shamoo Y (2008) An adaptive mutation in adenylate kinase that increases organismal fitness is linked to stability-activity trade-offs. *Protein Eng Des Sel PEDS* 21(1):19–27.

39. Tian GC, Sanders CR 2nd, Kishi F, Nakazawa A, Tsai MD (1988) Mechanism of adenylate kinase. Histidine-36 is not directly involved in catalysis, but protects cysteine-25 and stabilizes the tertiary structure. *Biochemistry* 27(15):5544–5552.

40. Burlacu-Miron S, Perrier V, Gilles AM, Pistotnik E, Craescu CT (1998) Structural and energetic factors of the increased thermal stability in a genetically engineered Escherichia coli adenylate kinase. *J Biol Chem* 273(30):19102–19107.

41. Nojima H, Ikai A, Oshima T, Noda H (1977) Reversible thermal unfolding of thermostable phosphoglycerate kinase. Thermostability associated with mean zero enthalpy change. *J Mol Biol* 116(3):429–442.

42. Young TA, Skordalakes E, Marqusee S (2007) Comparison of proteolytic susceptibility in phosphoglycerate kinases from yeast and E. coli: modulation of conformational ensembles without altering structure or stability. *J Mol Biol* 368(5):1438–1447.

43. Pey AL (2013) The interplay between protein stability and dynamics in conformational diseases: the case of hPGK1 deficiency. *Biochim Biophys Acta* 1834(12):2502–2511.

44. Betton JM, Desmadril M, Mitraki A, Yon JM (1984) Unfolding-refolding transition of a hinge bending enzyme: horse muscle phosphoglycerate kinase induced by guanidine hydrochloride. *Biochemistry* 23(26):6654–6661.

45. Kim HS, Damo SM, Lee S-Y, Wemmer D, Klinman JP (2005) Structure and hydride transfer mechanism of a moderate thermophilic dihydrofolate reductase from Bacillus stearothermophilus and comparison to its mesophilic and hyperthermophilic homologues. *Biochemistry* 44(34):11428–11439.

46. Loveridge EJ, Rodriguez RJ, Swanwick RS, Allemann RK (2009) Effect of dimerization on the stability and catalytic activity of dihydrofolate reductase from the hyperthermophile Thermotoga maritima. *Biochemistry* 48(25):5922–5933.

47. Schurig H, et al. (1995) Phosphoglycerate kinase and triosephosphate isomerase from the hyperthermophilic bacterium Thermotoga maritima form a covalent bifunctional enzyme complex. *EMBO J* 14(3):442–451.

48. Thomas TM, Scopes RK (1998) The effects of temperature on the kinetics and stability of mesophilic and thermophilic 3-phosphoglycerate kinases. *Biochem J* 330 (Pt 3):1087–1095.

49. Hess D, Kruger K, Knappik A, Palm P, Hensel R (1995) Dimeric 3-phosphoglycerate kinases from hyperthermophilic Archaea. Cloning, sequencing and expression of the 3-phosphoglycerate kinase gene of Pyrococcus woesei in Escherichia coli and characterization of the protein. Structural and functional comparison with the 3-phosphoglycerate kinase of Methanothermus fervidus. *Eur J Biochem* 233(1):227–237.

50. Varley PG, Pain RH (1991) Relation between stability, dynamics and enzyme activity in 3-phosphoglycerate kinases from yeast and Thermus thermophilus. *J Mol Biol* 220(2):531–538.