Supporting Information

Quality matters: Extension of clusters of residues with good hydrophobic contacts stabilize (hyper)thermophilic proteins

Prakash Chandra Rathi[1], Hans Wolfgang Höffken[2], Holger Gohlke[1]*

[1]Institute for Pharmaceutical and Medicinal Chemistry, Department of Mathematics and Natural Sciences, Heinrich Heine University, Düsseldorf, Germany

[2]BASF SE, GVC/C - A030, 67056 Ludwigshafen, Germany

*Universitätsstr. 1, 40225 Düsseldorf, Germany, Phone: (+49) 211-81-13662, Fax: (+49) 211-81-13847, Email: gohlke@uni-duesseldorf.de

Supplemental experimental procedures

Preparation of protein structures

All proteins in the dataset were downloaded from the Protein Data Bank (PDB)[1], and if required, the desired chain was extracted from the PDB file. All water molecules, ligands, and ions were removed from the structures. The command line version of the protein preparation wizard[2] of the Schrödinger software (Schrödinger, LLC, New York, NY, 2011) was used to prepare the protomer structures in order to I) add hydrogen atoms, II) add missing side chain atoms, and III) build disulphide bridges. Protein bio-assemblies were downloaded from the PDB and prepared in an identical manner as the protomers.

Calculation of residue-wise energy components

Protomers (and bio-assemblies) were minimized using the Prime module[3,4] version 3.0 of the Schrödinger software (Schrödinger, LLC, New York, NY, 2011) using default settings. Then, residue-wise electrostatic, van der Waals, and hydrophobic interaction energy components were calculated for the minimized structures by Prime. The hydrogen bond (including charge assisted hydrogen bonds) energy was calculated using a geometry-based energy function developed for protein design[5] as implemented in the FIRST software[6]. The energies of all hydrogen bonds for a residue were summed for calculating residue-wise hydrogen bond energies.

Clustering of residues by residue-wise energy components

Residues in a protein structure are clustered together if they are neighbors and if their values of the residue-wise energy components are below a cutoff E_C. Residues are considered neighbors if the distance between the closest pair of atoms is ≤ 4 Å. E_C is increased in a stepwise manner, and the clustering is repeated for each new E_C. As a result, a hierarchical clustering is obtained where clusters become larger as E_C increases. For each E_C value, the fraction of residues that is part of the largest cluster with respect to all protein residues (F_{LC}) is calculated. E_C was increased from an initial value of −30 kcal mol^{-1} to a final value of 0 kcal mol^{-1} with a step size of 0.5 kcal mol^{-1}.

Clustering of hydrophobic residues by inter-residue distances

Residues in a protein structure are clustered together if they belong to the type “hydrophobic” (Ala, Cys, Ile, Leu, Met, Phe, Trp, and Val) and are within a distance cutoff D_C. The distance between the closest atoms of two residues was considered the distance between these residues. D_C is increased in a stepwise-manner, and the clustering is repeated for each new D_C. As a result, a hierarchical clustering is obtained where clusters become larger as D_C increases. For each D_C value, the fraction of residues that is part of the largest cluster with respect to all protein residues (F_{LC}) is calculated. D_C was increased from an initial value of 1 Å to a final value of 5 Å with a step size of 0.05 Å.

Supplemental tables

Table S1. PDB ID and chain identifier of pairs of mesophilic/thermophilic proteins.

Thermoph.	Mesoph.	Thermoph.	Mesoph.	Thermoph.	Mesoph.	Thermoph.	Mesoph.
1nw2_A	1fb6_A	1hbn_B	1e6y_B	1odk_B	1vhw_A	2cuy_A	1mla_A
1urd_A	1anf_A	2f2b_A	1z98_A	1uay_A	1e6w_D	2cwd_A	1xww_A
2hm7_A	1lzl_A	2q5b_A	2cj3_A	1ub3_A	1p1x_A	2d1y_C	2zat_A
2sqc_A	1w6j_A	2v08_A	1ls9_A	1ufy_A	1dbf_A	2d29_A	2vig_A
1c9o_A	2es2_A	1ugp_B	2cz1_B	1ug6_A	1e4i_A	2d4e_A	2ve5_D
2b5a_A	1y7y_A	1b06_A	1bsm_A	1ui0_A	2c2q_A	2d4p_A	1s3z_A
1b4b_A	2p5m_A	1mp9_A	1qna_A	1uir_B	2pt9_A	2d5b_A	1pfv_A
1g2w_A	1iye_A	1thm_A	2tec_E	1uj5_A	2f8m_A	2d5c_A	1nyt_A
1gtf_A	1wap_A	3tec_I	1cse_I	1ulr_A	1urr_A	2d5w_A	1zu0_A
1hvx_A	3bh4_A	1h1n_A	1a3h_A	1umd_A	2ozl_A	2d8d_A	1ecm_A
1lqy_A	2okl_A	1i1x_A	1ta3_B	1umd_B	2ozl_B	2dt9_A	2dtj_A
1r2z_A	1xc8_A	1mtp_A	1sek_A	1v37_A	2a6p_A	2e7u_A	2hp1_A
1tqh_A	3dlt_A	2fla_A	1b0y_A	1v6s_A	1hdi_A	2ebj_A	1aug_A
1whi_A	1vqo_K	1bqc_A	1a3h_A	1v8f_A	1n2e_A	2eg4_A	1h4k_X
1y51_A	1ptf_A	1tf4_A	1ga2_B	1v8m_A	3ees_A	2eiy_B	1iye_A
1zdr_A	3dau_A	1tml_A	1dys_A	1v98_A	1fb6_A	2ekp_A	2zat_A
1zin_A	2eu8_A	1tib_A	3tgl_A	1vbi_A	1wtj_A	2fk5_A	1e4c_P
2bkm_A	2qrw_B	1yna_A	2dfb_A	1vc3_B	1uhe_A	2is8_A	1ihc_A
2exi_A	1y7b_A	2dte_A	1uzn_A	1vcd_A	1ktg_A	2j07_A	1owl_A
2tlx_A	1bqb_A	1my6_A	1xre_A	1ve1_A	2pqm_A	2p5y_A	1ek6_A
2bd0_A	1oaa_A	2c41_A	2c2u_A	1vef_A	2oat_A	2prd_A	2bqx_A
1aoh_A	1g1k_A	1esw_A	1x1n_A	1vfj_A	2pii_A	2pwy_A	1i9g_A
1cem_A	1v5d_A	2ng1_A	2qy9_A	1wlu_A	1q4u_A	2qhs_A	1w66_A
1h6y_A	1gny_A	1b5p_A	1asd_A	1wmw_A	2f94_F	2yqu_A	3lad_A
1nbc_A	1g43_A	1bxb_A	1oad_A	1wo8_A	1b93_A	2yvp_A	1g0s_B
1xyz_A	1e0w_A	1gd7_A	2q2i_A	1wur_A	1a8r_A	2z1a_A	1hpu_A
2b59_A	1qzn_A	1iv3_A	1h47_A	1wz8_A	2zqq_A	2z1y_B	1j32_A
2olj_A	1b0u_A	1iz9_A	2hjr_A	1x1o_A	2jbm_C	2zc8_A	1sjd_A
2q8x_A	1uqz_A	1j33_A	1l5o_A	1yya_A	2dp3_A	2zdb_A	3d0s_A
3d60_A	1gyh_A	1j3n_A	1oxh_A	1z54_A	1s5u_A	2zdh_A	1iow_A
1eje_A	3bnk_A	1j3w_A	1vet_B	2b3f_A	1anf_A	3cm0_A	3adk_A
1ep0_A	2ixc_A	1n97_A	1bu7_A	2bhq_A	2o2r_A	3hrx_A	1dci_A
1g5c_A	1ylk_A	1nza_A	2zfh_A	2cuk_A	2gcg_A	3mds_A	1xre_A

Table S2. PDB ID and chain identifier of pairs of mesophilic/hyperthermophilic proteins.

Hyper-thermoph.	**Mesoph.**	**Hyper-thermoph.**	**Mesoph.**	**Hyper-thermoph.**	**Mesoph.**	**Hyper-thermoph.**	**Mesoph.**
1h2b_A	1n8k_A	3c7b_B	2v4j_B	1zjj_A	2c4n_A	2z30_A	1st3_A
1n7k_A	1p1x_A	3cnu_A	2qzt_B	2cun_A	1hdi_A	1eu8_A	1anf_A
1tyo_A	1pb1_A	3do8_B	1coz_A	2cwp_A	3ers_X	1wst_A	2r2n_A
2fc3_A	1zwz_A	1g6h_A	2ff7_A	2d69_A	1rbl_A	1uxt_A	1euh_A
2yvu_A	2pez_A	1l2t_A	2ff7_A	2dbb_A	2qz8_A	2r91_A	2v8z_A
1c3p_A	1t64_A	1pkh_A	2qxx_A	2dr1_A	1w23_A	1d1g_A	3fq0_A
1hqk_A	2c92_A	1snn_A	1k4i_A	2dxe_A	1npk_A	1inl_C	2o07_B
1mzh_A	1p1x_A	1twi_A	1ko0_A	2e5f_A	1moq_A	1kq3_A	1ta9_B
1tz7_A	1x1n_A	2eb0_B	1k20_A	2e5w_A	2pt9_A	1nf2_A	1rkq_A
1ulz_A	2w70_A	2j9d_C	2pii_A	2ekn_A	2eey_A	1o0x_A	1y1n_A
1wwr_D	2b3j_A	2pa6_A	2akz_A	2hun_B	1r66_A	1o0y_A	1p1x_A
2e55_A	1bd3_A	2yww_A	2fzc_B	3cg3_A	3cfx_A	1o4s_A	1asd_A
2e8e_A	1n2f_A	2z02_A	2gqs_A	1ais_A	1qna_A	1oh4_A	1pmj_X
2ebd_A	1zow_A	2z8u_B	1qna_A	1mxd_A	3bh4_A	1p1m_A	2i9u_A
2egj_A	1s5u_A	1ftr_A	1m5s_A	1b7g_O	1u8f_R	1tmy_A	3chy_A
2ehh_A	3di1_A	1vcv_A	1p1x_A	1io7_A	3bdz_A	1tzx_A	1eyv_A
2ehs_A	1l0i_A	1ml4_A	1ekx_A	1je1_C	1vhw_A	1vbu_A	1ta3_B
2eja_A	3gw0_A	1aj8_A	1csh_A	1nto_A	1n8k_A	1vc1_A	1h4y_A
2hk9_A	1nyt_A	1gtm_A	1bgv_A	1uwr_A	2e3z_A	1vj0_A	1n8k_A
2omd_A	2q5w_E	1jg1_A	1i1n_A	1vph_A	1vmh_A	1vl8_A	1gee_A
2pbq_A	2g4r_A	1nnh_A	12as_A	1xtt_A	1bd3_A	1vlc_A	1cnz_A
2pbr_A	1e9e_A	1pvv_A	1oth_A	2f5g_A	2vjv_A	1vlg_C	1eum_A
2pnf_A	1q7b_A	1vkc_A	2fe7_B	2i6j_A	1fpz_A	1vlh_B	1qjc_A
2r75_1	2vxy_A	1ybz_A	1ecm_A	2var_C	1rkd_A	1vlj_A	1ta9_B
2yvl_A	1i9g_A	2dsk_A	2uy3_A	3f8p_D	1trb_A	1vm7_A	2fv7_A
2yvw_A	1uae_A	1gde_A	2r5e_A	1vgm_A	2h12_B	1vma_A	2qy9_A
2yw2_A	3g8c_A	1iu8_A	1aug_A	1wlt_A	2ixc_A	1vmj_A	1vmh_A
2z1m_A	1rpn_A	1lk5_A	1m0s_A	1wrj_A	1sfe_A	1vp2_A	1ex2_A
1coj_A	1bsm_A	1ub9_A	1r1u_A	1x0u_A	1on3_A	1vq0_A	1vzy_A
1jji_A	1lzl_A	1udd_A	2qcx_B	1x25_A	1qd9_A	1w2t_A	1y4w_A
1lbv_A	2qfl_A	1uku_A	2zfh_A	2e0q_A	1fb6_A	1w3j_A	1e4i_A
1p1l_A	2nuh_A	1v96_B	2h1c_A	2e7x_A	2qz8_A	1wa3_A	1wbh_A
1txg_A	1n1e_A	1w2i_A	1urr_A	2ehg_A	1jl1_A	1wos_A	1wsr_A
1vi6_A	3bch_A	1wqa_A	1k2y_X	2ekl_A	1dxy_A	2e54_A	2oat_A
2a5w_A	2v4j_C	1wr8_A	1s2o_A	2ggs_A	1n2s_A	2fnc_A	1anf_A
2b2h_A	3bhs_A	1wwk_A	1dxy_A	2bo1_A	1vqo_F	2h3h_B	2dri_A
2cyb_A	2yxn_A	1wy1_A	2zhz_A	1mgt_A	1sfe_A	2p3n_A	2qfl_A
3c7b_A	2v4j_A						

Table S3. *p*-values regarding equality in discrimination accuracies between mesophilic and (hyper)thermophilic protomers for clustering based on different residue-wise energy components *versus* a random discrimination.[a]

	Hydrogen bond	**van der Waals**	**Hydrophobic interaction**	**Random**
Hydrogen bond	- [b]	0.3245	<0.0001	0.2196
van der Waals	0.0005	- [b]	<0.0001	0.8056
Hydrophobic interaction	0.0042	0.5829	- [b]	<0.0001
Random	0.0610	<0.0001	<0.0001	- [b]

[a] The discrimination analysis is based on clustering by residue-wise energy components. The *p*-values were computed by a bootstrap hypothesis of equality generating 10000 bootstrap samples. Values in shaded shells correspond to mesophilic/hyperthermophilic protomers, other values correspond to mesophilic/thermophilic protomers.

[b] Not determined.

Table S4. *p*-values regarding equality in discrimination accuracies between mesophilic and (hyper)thermophilic protein bio-assemblies for clustering based on different residue-wise energy components *versus* a random discrimination. [a]

	Hydrogen bond	**van der Waals**	**Hydrophobic interaction**	**Random**
Hydrogen bond	- [b]	1.0000	<0.0001	0.6020
van der Waals	0.0572	- [b]	<0.0001	0.6068
Hydrophobic interaction	0.0534	1.0000	- [b]	<0.0001
Random	0.1141	0.0004	0.0007	- [b]

[a] The discrimination analysis is based on clustering by residue-wise energy components. The *p*-values were computed by a bootstrap hypothesis of equality generating 10000 bootstrap samples. Values in shaded shells correspond to mesophilic/hyperthermophilic protein bio-assemblies, other values correspond to mesophilic/thermophilic protein bio-assemblies.

[b] Not determined.

Table S5. Discrimination between mesophilic and (hyper)thermophilic protomers when clustering residues of type “hydrophobic” by inter-residue spatial distance.

Pairs	**Discrimination accuracy[b]**	***p*-value[a]**
Mesophilic/thermophilic	53.03	0.6175
Mesophilic/hyperthermophilic	61.74	0.0369

[a] The *p*-values were computed by a bootstrap hypothesis of equality between the given discrimination accuracy and a random discrimination (50% correct discrimination) generating 10000 bootstrap samples.

[b] In %.

Table S6. Thermostability of *E. coli* DHFR mutants.

Mutant	**ΔΔ*G* (H_2O) [kcal mol^{-1}]**[a]	**Mutant**	**ΔΔ*G* (H_2O) [kcal mol^{-1}]**[a]
G15A	**0.70**	W74F	-1.20
P21L	-0.10	V88I	**0.75**
W22L	**0.10**	V88A	**0.39**
L24V	-1.90	G95A	**1.30**
D27N	**1.40**	T113V	-1.20
L28R	**1.72**	D122A	-1.60
W30M	-2.03	E139Q	-0.42
W30Y	-2.16	E139K	-1.00
W30A	-2.33	S148K	-0.26
W30R	-2.49	S148P	-0.26
W30N	-2.52	S148V	-0.33
W30S	-2.74	S148A	-0.47
W30H	-2.78	S148T	-0.51
W30E	-2.89	S148E	-0.52
F31V	-1.50	S148R	-0.75
F31A	-1.90	S148N	-0.89
T35A	-1.10	I155V	-0.58
P39C	-3.00[b]	I155L	-2.27
V40I	-0.85	I155L	-2.80
V40L	-1.35	I155E	-3.26
V40A	-1.55	I155R	-3.28
V40R	-1.72	I155T	-3.30
V40M	-2.00	I155K	-3.35
V40F	-2.15	I155Y	-3.64
V40N	-2.17	I155A	-3.82
V40S	-2.52	I155Q	-3.86
V40H	-3.27	I155S	-3.93
G43A	-0.40	I155A	-4.00
L54V	**0.40**	I155D	-4.10
P66A	**1.30**	I155W	-4.31

[a] ΔG (mutant) ─ ΔG (wildt-ype) where ΔG is the free energy of unfolding in water, determined by denaturant (urea; guanidine hydrochloride; glutathione disulfide/glutathione; guanidinium thiocyanate) denaturation of proteins and extrapolation of the data to zero concentration of the denaturant. A positive value (marked in bold) indicates that the mutant is more thermostable than the wild-type.

[b] T_m (mutant) ─ T_m (wild-type) in K where T_m is the melting temperature identified as a mid-point temperature at which half of the protein is unfolded in a thermal unfolding method.

Table S7. Additional validation of weak spot prediction.

Protein	Mutations[a]	% correct prediction	Comment	Ref.
B. subtilis adenylate kinase	**Stabilizing mutations** *L3I*, G17A, D23K, **K69R**, **G73S**, **D75S**, I99S, *Y103M*, K105R, E114Q, **D118E**, V119E, *M121I*, E122A, S169T, Q180A, **D184A**, S187D, **E188S**, **G190E**, *Y191V*, *A193V*, *Y205F*, **D210V**, *L211I*, **K217Q**	34.61% (47.36% excluding mutations involving the exchange of one hydrophobic residue with another)	Mutations in two multiple-mutants that led to an increase of 11.6ºC and 12.5ºC in the melting temperature T_m compared to the wild type	[7]
E. coli maltose binding protein	**Destabilizing mutations** **V8G**, **W10A**, G19C, **I59A**, **I108A**, **L115A**, **L147A**, **P159A**, **I161A**, **L192A**, **L195A**, **I226A**, **A276G**, **Y283D**, **V347A**, **L361A**	93.75%	Single point mutations that led to a decrease in the T_m in a range of 0.1 to 7.5ºC or in the free energy of unfolding ΔG in a range of 0.3 to 5.5 kcal mol^{-1} compared to the wild type	[8-11]

[a] A correctly predicted mutation site is marked in bold. A mutation in italic involves the exchange of one hydrophobic residue with another.

Supplemental figures

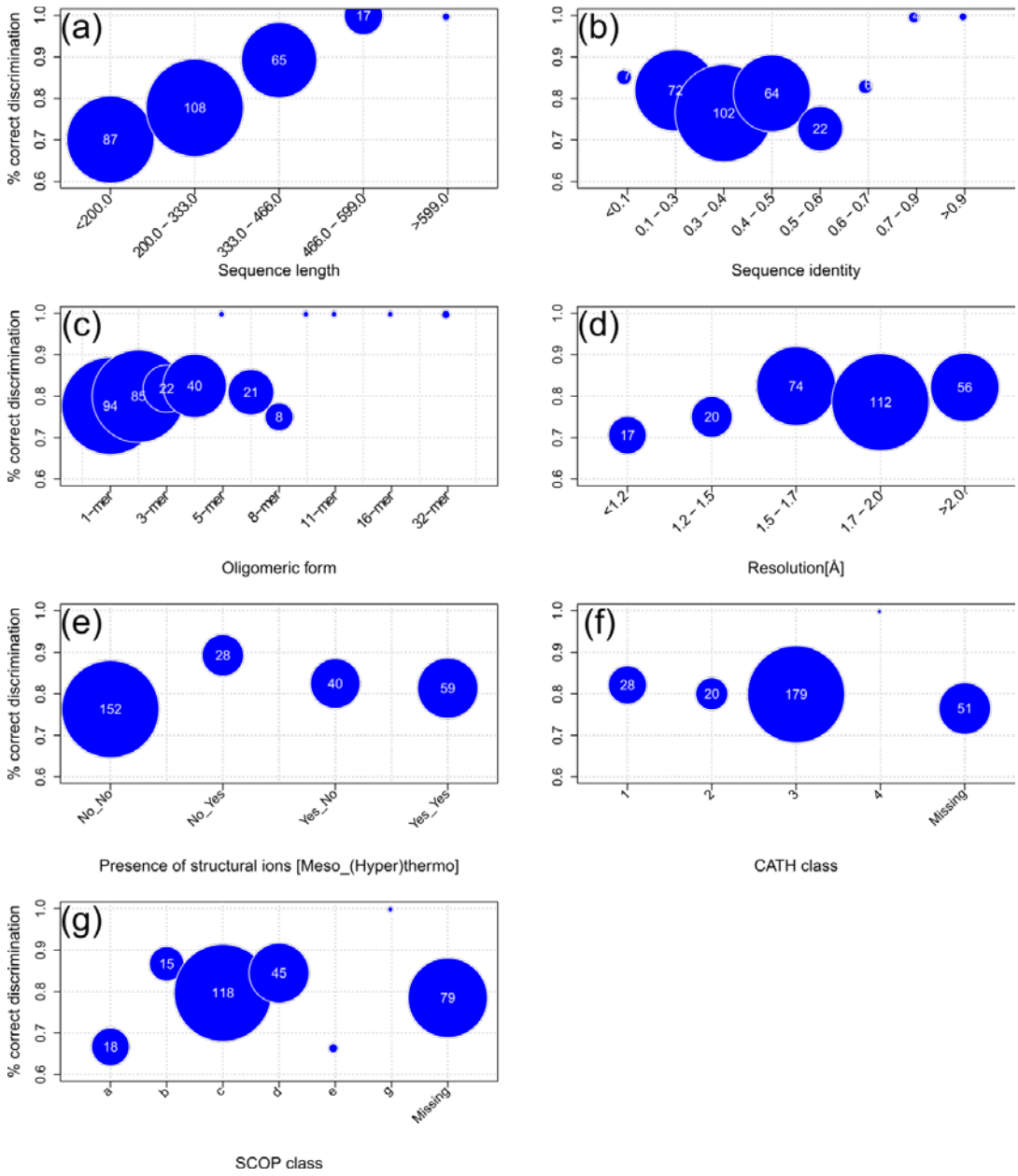

Figure S1. Discrimination accuracy between mesophilic and (hyper)thermophilic protomers, based on clusters of residues with good hydrophobic interaction energies, grouped according to the sequence length (a), sequence identity (b), oligomeric form (c), resolution of the X-ray structure (d), presence of structural ions in the structure (e), CATH class (f), and SCOP class (g). Mesophilic/(hyper)thermophilic pairs were grouped according to the property of the mesophilic protein chain in the structure unless indicated otherwise in the abscissa label of the plot. The size of a circle represents the number of pairs in each group (also indicated by the number in the circle), and the circle's position on the ordinate indicates the percentage of correct discrimination for these pairs.

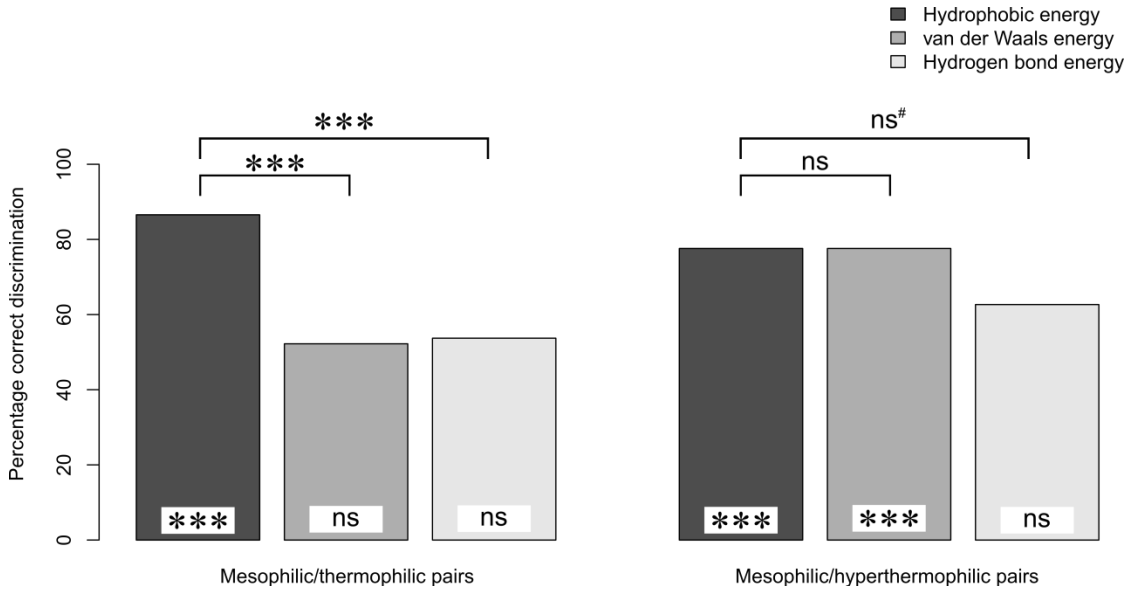

Figure S2. Discrimination accuracy between mesophilic and (hyper)thermophilic protein bio-assemblies based on clusters of residues with good residue-wise energy components. Lines connecting two bars indicate if the difference in discrimination accuracies for the two respective energy components is statistically significant. Marks at the bottom of a column indicate if the discrimination accuracy is significantly different from a random discrimination (50%). The statistical significance of the differences in discrimination accuracies is computed by a bootstrap hypothesis test of equality generating 10000 bootstrap samples; the significance levels are marked by ***: $p < 0.001$; **: $p < 0.01$; ns: $p > 0.05$. The p-value between hydrophobic and hydrogen bond energies in the case of mesophilic/hyperthermophilic pairs is 0.0534 (see ns$^{\#}$ in the figure).

Supplemental references

1. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.
2. Sastry, G. M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013,** *27*, 221-234.
3. Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **2002,** *320*, 597-608.
4. Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J.; Honig, B.; Shaw, D. E.; Friesner, R. A. A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf.* **2004,** *55*, 351-367.
5. Dahiyat, B. I.; Gordon, D. B.; Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci.* **1997,** *6*, 1333-1337.
6. *FIRST, a program for analysing flexibility of networks*, http://flexweb.asu.edu/ (accessed January 17, 2014).
7. Bae, E.; Bannen, R. M.; Phillips, G. N., Jr. Bioinformatic method for protein thermal stabilization by structural entropy optimization. *Proc. Natl. Acad. Sci. USA* **2008,** *105*, 9594-9597.
8. Chun, S. Y.; Strobel, S.; Bassford, P., Jr.; Randall, L. L. Folding of maltose-binding protein. Evidence for the identity of the rate-determining step in vivo and in vitro. *J. Biol. Chem.* **1993,** *268*, 20855-20862.
9. Diamond, D. L.; Strobel, S.; Chun, S. Y.; Randall, L. L. Interaction of SecB with intermediates along the folding pathway of maltose-binding protein. *Protein Sci.* **1995,** *4*, 1118-1123.
10. Prajapati, R. S.; Lingaraju, G. M.; Bacchawat, K.; Surolia, A.; Varadarajan, R. Thermodynamic effects of replacements of Pro residues in helix interiors of maltose-binding protein. *Proteins* **2003,** *53*, 863-871.
11. Chang, Y.; Park, C. Mapping transient partial unfolding by protein engineering and native-state proteolysis. *J. Mol. Biol.* **2009,** *393*, 543-556.