Supplementary Information: Thermometer: a webserver to predict protein thermal stability

Mattia Miotto,[1] Alexandros Armaos,[2] Lorenzo Di Rienzo,[1] Giancarlo Ruocco,[1,3] Edoardo Milanetti,[3,1] and Gian Gaetano Tartaglia[2,1]

[1] *Center for Life Nano & Neuroscience, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161, Rome, Italy*
[2] *Department of Biology, Sapienza University, Piazzale Aldo Moro 5, 00185, Rome, Italy*
[3] *Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185, Rome, Italy*

I. DATASET

A. Datasets

Proteins with known melting temperature (T_m) were obtained from the ProTherm database [1]. We selected all wild-type proteins for which the following thermodynamic data and experimental conditions were reported: $T_m \geq 0$ ^oC; $6.5 \leq$ pH ≤ 7.5 and no denaturants. Experimentally determined structures were collected from the PDB (Berman et al. 2) and filtered according to method (x-ray diffraction), resolution (below $3\mathring{A}$) and percentage of missing residues (5% compared to the Uniprot (Pundir et al. 3) sequence). Proteins for which experimentally determined structures were only available in a bound state, i.e. in complex with either a ligand or a ion, were excluded. Proteins were filtered using the CD-HIT software (Huang et al. 4) to remove proteins with chain sequence identity $\geq 40\%$ to each other. The final dataset, hereinafter referred to as the T_m dataset, consisted of 86 proteins. Consistently with previous reported dataset, thermostable proteins ($T_m \geq 70^oC$) represent about a third of the overall dataset (Karshikoff and Ladenstein 5, Parthasarathy and Murthy 6, Kannan and Vishveshwara 7). In order to have a dataset as balanced as possible, we also manually collected a second, independent dataset consisting of proteins from hyperthermophilic organisms with optimal growth at T ≥ 90 ^oC and pH between 6.5 and 7.5. Experimentally determined structures were collected and filtered according to same criteria described above for the T_m dataset, leading to a total of 13 protein structures. This second dataset is referred to as the T_{hyper} dataset. The union of the two dataset, referred as the T_{whole} dataset, accounts of 99 proteins

B. Clustering analysis

We clustered the T_s descriptors using the Euclidean distance and the Ward method as linkage function (Ward 8) via the hclust function of the Stats package of R (Ihaka and Gentleman 9). To better compare the different T_s score between them we normalize the data dividing each T_s score for the maximum of the absolute values. Finally, using the R package "clValid" (Brock et al. 10), we performed an internal validation for the hierarchical cluster considering both the Connectivity, Dunn and Silhouette parameters.

[1] R. Nikam, A. Kulandaisamy, K. Harini, D. Sharma, and M. M. Gromiha, Nucleic Acids Res **49**, D420 (2021).
[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[3] S. Pundir, M. J. Martin, and C. O'Donovan, Methods Mol. Biol. **1558**, 41 (2017).
[4] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, Bioinformatics **26**, 680 (2010).
[5] A. Karshikoff and R. Ladenstein, Protein Eng. **11**, 867 (1998).
[6] S. Parthasarathy and M. R. Murthy, Protein Eng. **13**, 9 (2000).
[7] N. Kannan and S. Vishveshwara, Protein Eng. **13**, 753 (2000).
[8] J. H. Ward, Journal of the American Statistical Association **58**, 236 (1963).
[9] R. Ihaka and R. Gentleman, Journal of Computational and Graphical Statistics **5**, 299 (1996).
[10] G. Brock, V. Pihur, S. Datta, and S. Datta, Journal of Statistical Software **25** (2008).
[11] U. Ermler, M. Merckel, R. Thauer, and S. Shima, Structure **5**, 635 (1997).
[12] H. Tanaka, M. Chinami, T. Mizushima, K. Ogasahara, M. Ota, T. Tsukihara, and K. Yutani, J. Biochem. **130**, 107 (2001).
[13] C. Charron, X. Manival, B. Charpentier, C. Branlant, and A. Aubry, Acta Crystallogr. D Biol. Crystallogr. **60**, 122 (2004).
[14] R. Kawakami, H. Sakuraba, S. Goda, H. Tsuge, and T. Ohshima, Biochim. Biophys. Acta **1794**, 1496 (2009).
[15] S. Yoshinari, T. Shiba, D. K. Inaoka, T. Itoh, G. Kurisu, S. Harada, K. Kita, and Y. Watanabe, Nucleic Acids Res. **37**, 4787 (2009).
[16] M. Blaise, M. Frechin, V. Olieric, C. Charron, C. Sauter, B. Lorber, H. Roy, and D. Kern, J. Mol. Biol. **412**, 437 (2011).
[17] H. M. Park, M. Shin, J. Sun, G. S. Kim, Y. C. Lee, J. H. Park, B. Y. Kim, and J. S. Kim, Proteins **80**, 1895 (2012).
[18] C. Wang, Q. Jia, R. Chen, Y. Wei, J. Li, J. Ma, and W. Xie, Sci Rep p. 33553 (2016).

	PDB	Tm [C^o]	Type
1	1a3y	69.1	M
2	1ako	42.6	M
3	1b8e	78.0	T
4	1bd8	51.9	M
5	1bk7	64.4	M
6	1bni	54.0	M
7	1bpi	104.0	T
8	1btl	56.4	M
9	1c5g	67.5	M
10	1c9o	76.9	T
11	1cec	70.4	T
12	1chk	52.2	M
13	1cm2	63.4	M
14	1cm7	34.0	M
15	1cmb	54.0	M
16	1csp	52.8	M
17	1czd	52.3	M
18	1div	77.6	T
19	1ekg	58.2	M
20	1ew4	53.8	M
21	1fsf	66.9	M
22	1fvk	76.8	T
23	1gtm	114.4	T
24	1gwy	67.0	M
25	1h7m	93.5	T
26	1h09	51.4	M
27	1hix	66.7	M
28	1hk0	80.0	T
29	1i4n	90.3	T
30	1ino	58.0	M
31	1j2v	148.5	T
32	1j4s	66.0	M
33	1jji	99.0	T
34	1jyd	68.9	M
35	1ke4	54.6	M
36	1lnw	55.6	M
37	1mjc	59.0	M
38	1msi	46.6	M
39	1npk	62.0	M
40	1onc	88.5	T
41	1orc	57.0	M
42	1pii	51.0	M
43	1poh	63.4	M
44	1qhe	56.0	M
45	1r56	68.9	M
46	1rg8	39.8	M
47	1rgg	49.0	M
48	1rhg	60.9	M
49	1rn1	48.9	M
50	1rop	68.7	M
51	1rtb	61.3	M
52	1sfp	78.6	T
53	1spp	60.1	M
54	1stn	52.7	M
55	1tca	57.7	M
56	1tpe	51.4	M
57	1udv	157.5	T
58	1v6h	112.7	T
59	1y4y	88.3	T
60	1ypr	67.9	M
61	1zdr	66.2	M
62	2cro	56.0	M
63	2dri	57.5	M
64	2gd1	78.5	T
65	2izp	83.7	T
66	2lzm	62.2	M
67	2prd	86.0	T
68	2sil	57.0	M
69	2y3z	87.0	T
70	2zta	77.9	T
71	3chy	57.8	M
72	3d2a	63.4	M
73	3dfq	61.2	M
74	3enj	47.8	M
75	3ssi	82.2	T
76	4ake	51.8	M
77	4blm	67.0	M
78	4g03	66.4	M
79	4gcr	70.4	T
80	4lyz	80.0	T
81	4n9h	66.4	M
82	5fb6	68.7	M
83	5pep	52.0	M
84	2x9b	67.2	M
85	3kvd	84.1	T
86	3n4y	100.2	T

TABLE I: Table of the 86 proteins of the T_m dataset, collected from ProTherm database.

[19] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, Nucleic Acids Res. **43**, D364 (2015).

Name	Organism	PDB	Ref.
Formylmethanofuran	Methanopyrus kandleri	1ftr	[11]
pyrrolidone carboxyl peptidase	Pyrococcus furiosus	1iof	[12]
L7Ae sRNP core protein	Pyrococcus abyssi	1pxw	[13]
malate dehydrogenase	Aeropyrum pernix	2d4a	[14]
D-Tyr-tRNA(Tyr) deacylase	Aquifex aeolicus	2dbo	To Be Published
hypothetical protein (Aq-1549)	Aquifex aeolicus	2e8f	To Be Published
3-oxoacyl-[acyl-carrier-protein] synthase III	Aquifex aeolicus	2ebd	To Be Published
aq-1716	Aquifex aeolicus	2p68	To Be Published
3-dehydroquinate dehydratase	Aquifex aeolicus	2ysw	To Be Published
splicing endonuclease	Pyrobaculum aerophilum	2zyz	[15]
archaeal asparagine synthetase A	Pyrococcus abyssi	3p8t	[16]
Cas6	Pyrococcus furiosus	3ufc	[17]
tRNA methyltransferase Trm5a	Pyrococcus abyssi	5hjj	[18]

TABLE II: Table of Hyperthermophiles proteins manually collected on the PDB bank [19].