OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Stability Curve Prediction of Homologous Proteins Using Temperature-Dependent Statistical Potentials

Fabrizio Pucci*, Marianne Rooman*

Department of BioModeling, BioInformatics & BioProcesses, Université Libre de Bruxelles, Brussels, Belgium

Abstract

The unraveling and control of protein stability at different temperatures is a fundamental problem in biophysics that is substantially far from being quantitatively and accurately solved, as it requires a precise knowledge of the temperature dependence of amino acid interactions. In this paper we attempt to gain insight into the thermal stability of proteins by designing a tool to predict the full stability curve as a function of the temperature for a set of 45 proteins belonging to 11 homologous families, given their sequence and structure, as well as the melting temperature (T_m) and the change in heat capacity (ΔC_P) of proteins belonging to the same family. Stability curves constitute a fundamental instrument to analyze in detail the thermal stability and its relation to the thermodynamic stability, and to estimate the enthalpic and entropic contributions to the folding free energy. In summary, our approach for predicting the protein stability curves relies on temperature-dependent statistical potentials derived from three datasets of protein structures with targeted thermal stability properties. Using these potentials, the folding free energies (ΔG) at three different temperatures were computed for each protein. The Gibbs-Helmholtz equation was then used to predict the protein's stability curve as the curve that best fits these three points. The results are quite encouraging: the standard deviations between the experimental and predicted T_m 's, ΔC_P 's and folding free energies at room temperature (ΔG_{25}) are equal to 13 $^{\circ}C$, 1.3 $kcal/(mol^{\circ}C)$ and 4.1 $kcal/mol$, respectively, in cross-validation. The main sources of error and some further improvements and perspectives are briefly discussed.

Citation: Pucci F, Rooman M (2014) Stability Curve Prediction of Homologous Proteins Using Temperature-Dependent Statistical Potentials. PLoS Comput Biol 10(7): e1003689. doi:10.1371/journal.pcbi.1003689

Editor: Jacquelyn Fetrow, Wake Forest University, United States of America

Received March 6, 2014; **Accepted** May 12, 2014; **Published** July 17, 2014

Funding: Project supported by the Belgian fund for scientific research (FNRS). FP is Postdoctoral Fellow and MR Research Director at FNRS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: fapucci@ulb.ac.be (FP); mrooman@ulb.ac.be (MR)

Introduction

The understanding of the mechanisms used by nature to stabilize proteins against thermal inactivation is still an open issue of primary importance. From a theoretical perspective, such comprehension is fundamental in the study of the adaptive strategies used by the organisms to inhabit extreme environments. Due to evolution, such organisms are not only able to tolerate extreme temperature conditions, that range from less than ten degree Celsius to more than 120 ^{o}C, but require these conditions for their survival. The control of the thermal resistance is also important from an applicative perspective, as it would allow the optimization of a wide series of industrial, bioanalytical and pharmaceutical bioprocesses through the design and manufacture of new and more efficient enzymes [1–3].

In the last decades, different attempts and methods have been developed to obtain proteins of increased thermal stability. Protein engineering methods that include directed evolution methods [4–6] have been quite successful even if their applicability remains limited due to the intensive work required. *In silico* engineering approaches based on sequence conservation or free energy calculation methods have also been developed but with only partial success [7–12].

Recently, we developed a thermal stability prediction tool based on (melting)-temperature dependent statistical potentials that are derived from datasets in which only proteins with given thermostability properties are included [13–15]. The introduction of such potentials in the thermal stability framework is motivated by the fact that the amino acid pair interactions are temperature dependent, which means that some of them are more stabilizing than others in the high temperature regime and less stabilizing at lower temperatures (and *vice versa*) [16–29]. This peculiar approach allowed us to study the thermal properties of proteins without detour through their thermodynamical stability, which is advantageous since it is well known that the two types of stability are poorly correlated.

Proteins use different ways to promote their thermoresistance, which can – in a first approximation – be classified in three main strategies according to the Nojima analysis [30] (for a more recent review see also [31]). Let us start by introducing the stability curve of a protein, which can be described by the Gibbs-Helmholtz equation:

$$\Delta G(T) = \Delta H_R + \Delta C_P(T - T_R) - T\left[\Delta S_R + \Delta C_P \mathrm{Log}\left(\frac{T}{T_R}\right)\right], \quad (1)$$

where $\Delta G(T)$ is the free energy change associated to the folding transition from the unfolded to the native state, ΔH_R and ΔS_R the change in enthalpy and entropy measured at the reference temperature T_R, and ΔC_P the change of the heat capacity across the transition. To obtain this equation, one has to fix the pressure

Author Summary

The prediction of protein stability remains one of the key goals of protein science. Despite the significant efforts of the last decades, faster and more accurate stability predictors on the proteomic-wide scale are currently demanded. The determination and control of protein stability are indeed fundamental steps on the path towards *de novo* design. In this paper we develop a method for predicting the stability curve of proteins. This curve encodes the temperature dependence of the folding free energy (ΔG). Its knowledge is important in the study of protein stability since all the thermodynamic parameters characterizing the folding transition can be extracted from it. Our prediction method is based on temperature-dependent mean force potentials and uses the tertiary structure of the target protein as well as the melting temperature (T_m) and the heat capacity change (ΔC_P) of some other proteins belonging to the same family. From the predicted stability curves, the T_m, the ΔC_P and the ΔG at room temperature can be inferred. The predictions obtained are compared with experimental data and show reasonable performances.

of the system, to consider two-state transitions only, and to take ΔC_P as temperature independent. Usually, the melting temperature T_m, which is the midpoint of the thermal denaturation, is chosen as the reference temperature. Eq.(1) can then be rewritten as:

$$\Delta G(T) = \Delta H_m \left(1 - \frac{T}{T_m}\right) - \Delta C_P \left[(T_m - T) + T \mathrm{Log}\left(\frac{T}{T_m}\right)\right], \quad (2)$$

where ΔH_m is the enthalpy measured at T_m. Sometimes, the reference temperature is taken equal to T_s, the temperature of maximal stability, which yields the equation:

$$\Delta G(T) = \Delta H_s - \Delta C_P \left[(T_s - T) + T \mathrm{Log}\left(\frac{T}{T_s}\right)\right]. \quad (3)$$

The first strategy that a protein can use to increase its thermostability [30] is to make the enthalpy change (ΔH_s) measured at T_s more negative. This yields an overall decrease of ΔG for all temperatures as we can see from Eq.(3) (Figure 1.a). In the second strategy, ΔC_P becomes less negative, which leads to an increase of T_m through a modification of the shape of the curve (see Eq.(2) and Figure 1.b). The last strategy consists in an increase of the maximum stability temperature, T_s, defined at the minimum of the $\Delta G(T)$ curve, where the transition is purely enthalpic. This shifts the curve towards the high temperature region (see Figure 1.c).

It is, in general, not obvious to determine which type of strategy is adopted by a given protein; often several strategies are used in combination [31]. A realistic example of stability curve is depicted in Figure 1.d: the value of the folding free energy $\Delta G(T)$ is plotted both for a thermostable protein, the O^6 -methyl-guanine-DNA methyltransferase from *Thermococcus kodakaraensis* (*Tk*-MGMT) with $T_m = 98.6\ ^oC$, and for its mesostable counterpart, the C-terminal Ada protein from *Escherichia coli* (*Ec*-AdaC) with $T_m = 54.8\ ^oC$, as determined experimentally in [32]. We can clearly see that in this case the three strategies are used simultaneously in the achievement of a higher thermal stability.

The strategies for improving the thermal resistance of a protein sometimes also improve the thermodynamic stability, defined by the folding free energy $\Delta G(T_r)$ at room temperature ($25\ ^oC$), and sometimes not. The first strategy clearly does; for the other two strategies, it depends on the relative values of T_s and T_r (see Figures 1a–c).

It is unfortunately quite difficult to get accurate predictions of thermal stability. The results described in the literature are in general family-dependent and sometimes even contradictory [16–29]. Indeed, the temperature-dependent nature of the amino acid interactions makes the thermal stability analyses quite intricate and the mechanism behind it difficult to unravel. Predicting the thermodynamic stability is not easy either. There are no methods for predicting the thermodynamic stability of a given protein, with the notable exception of molecular dynamic simulations, which are however very time-consuming and not applicable on a large or medium scale. Only methods for predicting thermodynamic stability changes upon point mutations ($\Delta\Delta G(T_r)$) have been developed and reach good scores [33–43]. No predictions of the enthalpy ΔH or entropy ΔS do exist either. In contrast, the prediction of ΔC_P is relatively easy since it is strongly correlated to the change of accessible surface area upon unfolding [44–46].

In this paper we go a step further than previous analyses aiming at evaluating either T_m, $\Delta G(T_r)$ or ΔC_P. We indeed present a method for predicting the whole stability curve $\Delta G(T)$ of a protein from its sequence and structure, in the temperature range that is relevant for such systems ($\approx 0 - 150^oC$), using as main tool the temperature-dependent statistical potentials developed and tested in [13]. We would like to emphasize that this is, to our knowledge, the first prediction method that outputs the complete stability curve. To get a satisfactory performance, we used in the predictions some information about proteins belonging to the same homologous family, and more precisely their T_m and ΔC_P. The predicted stability curve yields an estimation of the melting temperature T_m, the thermodynamic stability $\Delta G(T_r)$, the temperature of optimal stability T_s, the ΔC_P, as well as the enthalpy ΔH and the entropy ΔS at certain temperatures. We present our results in cross validation for a set of 45 proteins belonging to eleven homologous families (for the list of their PDB codes [47] and their characteristics, see Table S1 of Supporting Material). The predicted values are compared with the experimentally determined values when available, and the different strategies used by the proteins for thermal stabilization are investigated and discussed.

Methods

T-dependent statistical potentials

In this section we describe the main tools used in this analysis, namely the statistical potentials, and how they have been optimized for the current investigation. The main steps of our approach are schematically illustrated in Figure 2.

The statistical potentials are well known since some seminal papers [48–50]. They are derived from the frequency of associations between certain sequence and structure elements in a dataset of experimentally determined native protein structures. Even though such potentials have been extensively and successfully used in the analysis of the thermodynamic stability of proteins, they have only recently been applied in the thermal stability context, where the temperature dependence of the amino acid interactions must be taken into account [13–15]. To deal with this, potentials that depend on the melting temperature were

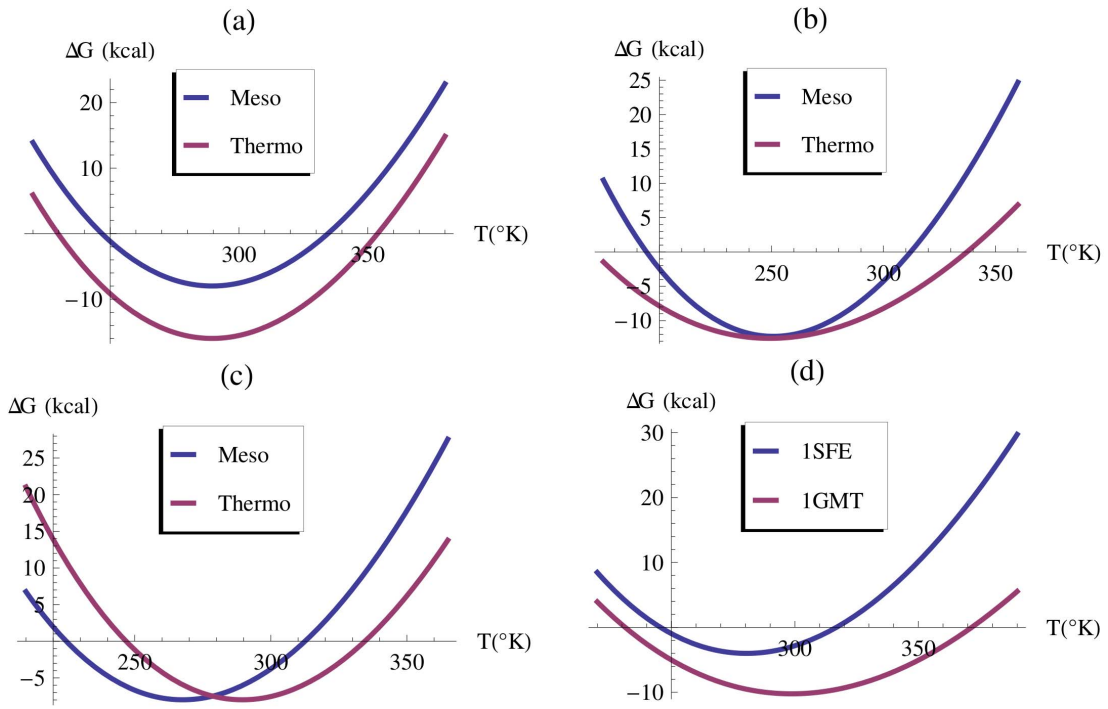

Figure 1. Stability curves of thermostable and mesostable proteins. (a,b,c) Different strategies of thermal adaptation of hypothetic proteins. (d) Comparison between the stability curve of *Tk*-MGMT (PDB [47] code 1GMT) and its mesophilic counterpart *Ec*-AdaC (PDB code 1SFE) [32].)
doi:10.1371/journal.pcbi.1003689.g001

derived from different datasets in which only proteins with given thermal properties were included. Three such datasets were considered [15]: a set containing only mesostable proteins, denoted S^{∇} and characterized by a mean value of the melting temperature of its entries ($\bar{T}_m^{\nabla}$) of about 50^oC, a thermostable ensemble, denoted S^{Δ}, with $\bar{T}_m^{\Delta}\tilde{8}0^oC$, and a reference set containing both mesostable and thermostable proteins, denoted $S^{\diamondsuit}$, with $\bar{T}_m^{\diamondsuit}\tilde{6}5^oC$. The list of proteins belonging to these datasets are given in Table S0–S11 and Table S13 of the Supplementary Material of [15].

From these different datasets, statistical potentials were derived using the standard formalism of the inverse Boltzmann law [13,14]:

$$\Delta W(s,c,\bar{T}_m) \cong -kT \ln \frac{F(s,c,\bar{T}_m)}{F(s,\bar{T}_m)F(c,\bar{T}_m)}, \qquad (4)$$

where $F(s,c,\bar{T}_m)$ is the relative frequency of observation of the sequence element s associated to the structure element c, and $F(s,\bar{T}_m)$ and $F(c,\bar{T}_m)$ are the frequencies of observation of the sequence element s and of the structure element c, respectively. In this computation, s corresponds either to the amino acid type a_i of residue i along the polypeptide chain, or to the amino acid types (a_i,a_j) of residues i and j, while c is either the backbone torsion angle domain t_k of residue k, as defined in [51], or the spatial distance d_{ij} between the residues i and j. The former are called torsion potentials and the latter distance potentials.

While the torsion potentials describe local interactions along the chain and are a measure of the propensity of a given amino acid to adopt certain backbone torsion angles, the distance potentials describe the tertiary interactions and measure the propensity of amino acids to be separated by a certain spatial distance d. The values of the distance between two residues, defined as the distance between the geometrical centers of the heavy side chain atoms, range between 3.0 and 8.0 A and were grouped into 25 bins of 0.2 A width, with two additional bins that contain distances larger than 8.0 A and smaller than 3.0 A, respectively.

Note that we have made the $\bar{T}_m$ -dependence of the frequencies explicit to stress that these are computed from a dataset associated with specific thermal properties, characterized by $\bar{T}_m$. As a consequence, the potentials are $\bar{T}_m$ -dependent and reflect the thermal characteristics of the dataset from which they are derived.

Due to the smallness of the dataset, some techniques are required to smooth the potentials and improve their performances. A first modification that has been performed is a correction for sparse data consisting in rewriting the frequencies as [52]:

$$F(s,c,\bar{T}_m) \rightarrow \frac{\sigma F(s)F(c,\bar{T}_m) + gF(s,c,\bar{T}_m)}{\sigma + g}, \qquad (5)$$

where σ is an adjustable parameter chosen to be equal to 10 for the distance potentials and to 20 for the torsion potentials (based on preliminary tests), and where g is equal to $n(s,\bar{T}_m) \times n(c,\bar{T}_m)/n(\bar{T}_m)$. This correction ensures that the potentials tend to zero when the number of observations in the data set is too small. A second trick that has been used consists, for a given bin i, in summing the number of occurrences of the neighboring bins giving them a decreasing weight:

$$n^i = \left[\frac{n^{i-2}}{3} + \frac{n^{i-1}}{2} + \ldots n^i + \frac{n^{i-1}}{2} + \frac{n^{i-2}}{3}\right] \qquad (6)$$

where n^i is the number of occurrences in bin i.

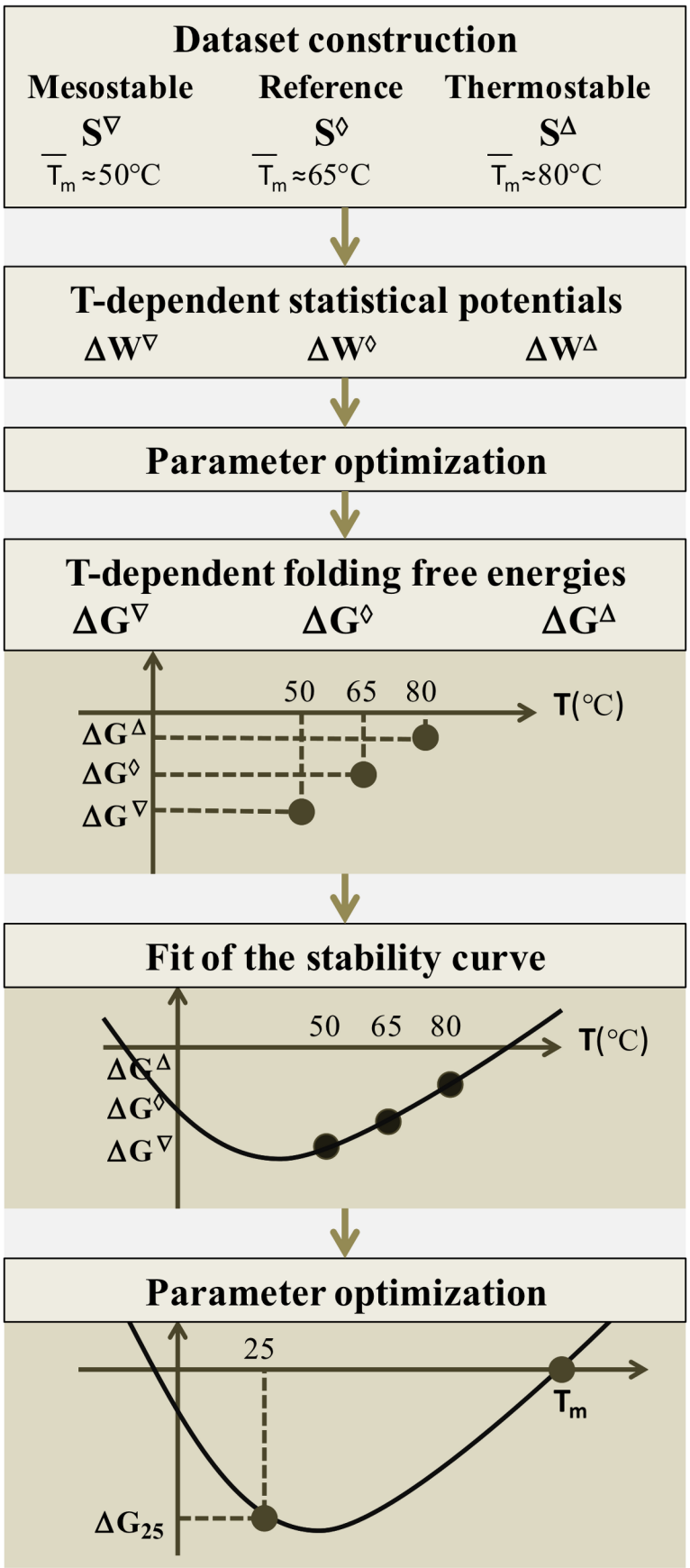

Figure 2. Flowchart of the protein stability curve prediction method.
doi:10.1371/journal.pcbi.1003689.g002

Families of homologous proteins

Predicting the stability curve of proteins from their sequence and structure alone is quite a difficult task. To slightly simplify the problem, we focused on families of homologous proteins, and make predictions that take into account some informations from the other family members. We therefore searched the full protein set S for families of homologous proteins with at least three members of known T_m. We found 11 such families containing both mesostable and thermostable proteins. They are: α -amylase, acylphosphatase, lysozyme, myoglobin, β -lactamase, α -lactalbumin, adenylate kinase, cell 12A endoglucanase, cold shock protein, cytochrome P450 and ribonuclease. The complete list of the 45 proteins belonging to these families is given in Table S1 of Supporting Material.

Some quantities (such as the number of residues, ΔC_P, etc.) remain approximately constant inside a given family. This obviously makes the prediction method simpler to build. Such family-dependent analysis remains nevertheless quite intricate, since the thermostability properties of the proteins of a given family are sometimes very different.

In order to improve the performance of our method, the datasets S^∇, S^Δ and $S^\diamondsuit$ have been further enlarged by adding proteins that belong to the protein family considered but whose T_m was estimated from their environmental temperature instead of being experimentally determined; note that the pairwise sequence identity within each set was kept below 25% to avoid biasing the potentials (see [15] for details about the dataset construction procedure). Strictly speaking, this modification makes the datasets and the corresponding potentials family dependent.

Computation of the folding free energy at different temperatures

The folding free energy ΔG of a given protein is computed at the temperatures $\bar{T}_m^\nabla$, $\bar{T}_m^\diamondsuit$ and $\bar{T}_m^\Delta$ from the (melting-)temperature dependent potentials defined in the previous subsections. More precisely, we have:

$$\Delta G(\bar{T}_m^\nabla) = \frac{1}{\mathcal{N}}\left[b_0 \sum_{i,j=1}^{N} \Delta W(a_i,a_j,d_{ij},\bar{T}_m^\nabla) + b_1 \sum_{i,j=1}^{N} \Delta W(a_i,d_{ij},\bar{T}_m^\nabla) + b_2 \sum_{i,j,k=1}^{N} \Delta W(a_i,a_j,t_k,\bar{T}_m^\nabla) + b_3 \sum_{i,k=1}^{N} \Delta W(a_i,t_k,\bar{T}_m^\nabla) \right], \tag{7}$$

$$\Delta G(\bar{T}_m^\diamondsuit) = \frac{1}{2\mathcal{N}}\left[(b_0+c_0) \sum_{i,j=1}^{N} \Delta W(a_i,a_j,d_{ij},\bar{T}_m^\diamondsuit) + (b_1+c_1) \sum_{i,j=1}^{N} \Delta W_f(a_i,d_{ij},\bar{T}_m^\diamondsuit) + (b_2+c_2) \sum_{i,j,k=1}^{N} \Delta W(a_i,a_j,t_k,\bar{T}_m^\diamondsuit) + (b_3+c_3) \sum_{i,k=1}^{N} \Delta W(a_i,t_k,\bar{T}_m^\diamondsuit) \right], \tag{8}$$

$$\Delta G(\bar{T}_m^\Delta) = \frac{1}{\mathcal{N}}\left[c_0 \sum_{i,j=1}^{N} \Delta W(a_i,a_j,d_{ij},\bar{T}_m^\Delta) + c_1 \sum_{i,j=1}^{N} \Delta W(a_i,d_{ij},\bar{T}_m^\Delta) + c_2 \sum_{i,j,k=1}^{N} \Delta W(a_i,a_j,t_k,\bar{T}_m^\Delta) + c_3 \sum_{i,k=1}^{N} \Delta W(a_i,t_k,\bar{T}_m^\Delta) \right], \tag{9}$$

where $i \neq j, j \pm 1$ for the distance potentials, $k-8 \leq i < j \leq k+8$ for the torsion potentials, and the parameters $\mathbf{P} = (b_0,b_1,b_2,b_3,c_0,c_1,c_2,c_3)$ are positive real numbers. The normalization coefficient $\mathcal{N}$ is defined as:

$$\mathcal{N} = \frac{1}{4}\sqrt{(b_0+c_0)^2+(b_1+c_1)^2+(b_2+c_2)^2+(b_3+c_3)^2} \quad . \tag{10}$$

The temperatures ($\bar{T}_m^\nabla$, $\bar{T}_m^\Delta$, $\bar{T}_m^\diamondsuit$) correspond to the average melting temperatures of the mesostable, thermostable and average datasets. The real T -dependence of the folding free energies is obviously related to these melting temperatures. However, it would be a very strong (and obviously wrong) assumption to suppose that the average melting temperatures and the real temperatures are equal. Rather, as will be seen in the next subsection, a scale parameter must be introduced to relate the $\bar{T}_m$'s to the real T.

The strategy for identifying the parameter values $\mathbf{P}$ consists in maximizing the anticorrelation between the melting temperature and the difference in free energies $\Delta\Upsilon = \Delta G(\bar{T}_m^\Delta) - \Delta G(\bar{T}_m^\nabla)$. Indeed, $\Delta\Upsilon$ has been shown to be much more correlated to the melting temperature than the folding free energy $\Delta G^\diamondsuit$ [15]. The optimization is performed on all proteins with known T_m (listed in Table S1), excluding those of the protein family f that we want to predict:

$$\mathbf{P}_f = \arg\max_{\mathbf{P}} [\text{Correlation}(\Delta\Upsilon, T_m)] \quad . \tag{11}$$

The subscript f indicates the family-dependent nature of the coefficients since their optimization is performed without the proteins of f. This avoids the overestimation of the performance, and amounts to cross validation. All the optimizations described in this paper are performed using the ordinary least square regression method implemented in *Mathematica* 7.0.

Extrapolation of the full stability curve

In the next steps of the computation, we estimate the full stability curve given by Eq.(2) from the three values of the folding free energies given by Eqs(7–9), for the set of 45 proteins from the 11 protein families. Let us assume for the moment that the $\bar{T}_m$ -dependence is the true T -dependence of the potentials. Under this assumption, the stability curve can easily be obtained: it is has the form (2) and depends on the thermodynamic quantities (ΔH_m, T_m and ΔC_P), viewed as parameters, which are identified to best fit the three data points:

$$\{\bar{T}^\nabla, \Delta G(\bar{T}^\nabla)\}, \{\bar{T}^\diamondsuit, \Delta G(\bar{T}^\diamondsuit)\}, \{\bar{T}^\Delta, \Delta G(\bar{T}^\Delta)\}. \tag{12}$$

However, this simple approach does not give accurate predictions, both because the T_m - and T -dependences differ and because the error on these three points, which are moreover quite close along the T -axis, leads to large errors on the whole curve. Three different issues must be solved to get reasonable stability curves.

The first issue concerns the sign of the second derivative of the curve. In a few cases (less than 10%), this sign is wrong, which implies that the curve is upside-down and the protein seems unfolded in the physiological temperature range. This error is related to the fact that the three points given in Eq.(12) are too close along the T axis; this is due to the limited number of known proteins with very low or very high T_m. The shape of the curve

depends thus strongly on the relative position of the average point $\{\bar{T}^{\diamond},\Delta G(\bar{T}^{\diamond})\}$ relative to the mesostable and thermostable points $\{\bar{T}^{\nabla},\Delta G(\bar{T}^{\nabla})\}$ and $\{\bar{T}^{\Delta},\Delta G(\bar{T}^{\Delta})\}$. Sometimes even a small variation of these values can lead to the inversion of the shape of the curve.

To overcome this problem, we imposed a fourth point in the fitting procedure, in addition to those given in Eq.(12). This point is taken at a temperature of 0°K, where we impose $\Delta G(0)$ to be equal to the average of the $\Delta G(0)$'s of the other proteins that belong to the same family. This quantity has no physical interpretation, as the inverse bell shape of the stability curve may not be extrapolated to zero temperature; indeed, we have in reality $\Delta G(0)=0$. This trick is however quite useful to impose the correct sign of the second derivative of the curve in the physiological temperature range.

This procedure has been applied when the predicted curve is upside-down, but also when the value of $\Delta G(0)$ deviates by more than one standard deviation from the mean $\Delta\bar{G}(0)$ computed inside the family f. This leads to an overall improvement of the results since it smooths out possible errors on the average point $\{\bar{T}^{\diamond},\Delta G(\bar{T}^{\diamond})\}$, which is amplified in the curve derivation procedure.

The second issue is the determination of the overall scaling factor $\mathcal{M}$ of the curve. When more than one value of ΔC_P was experimentally determined within the considered family f, we fix $\mathcal{M}$ for the protein p in the family f as the ratio:

$$\mathcal{M}_p=\sum_{p'\neq p}\frac{\Delta C_P^{\exp}(p')}{\Delta C_P^{\mathrm{curve}}(p')} \quad , \tag{13}$$

where $\Delta C_P^{\mathrm{curve}}$ is extracted from the predicted curves as the coefficient of the $T\ln T$ term, $\Delta C_P^{\exp}$ is the experimental value and the sum is over the proteins belonging to f excluding p; this again amounts to obtain predictions in cross validation. If only one or no ΔC_P values were available for the family, we took as normalization factor the mean of the $\mathcal{M}$ values found for the other families, excluding the largest and smallest values. This is a rough approximation since this quantity is expected to be strongly family dependent. However, despite the crude approximations made, the final result shows a fair performance that will certainly improve when more data or an independent ΔC_P determination will be available.

The last issue concerns the real temperature dependence of the potentials. Strictly speaking, the $\bar{T}_m$ -dependence of the potentials is different from the real T -dependence, even though they are obviously related. Indeed, the temperature resistant interactions can be expected to play a fundamental role in the stabilization in the high temperature regime and *vice versa* in the low temperature region (see [16–20] for the temperature dependence of the amino acid interactions). The assumption that we made is that the real T value at which the potentials are calculated is related to the value of $\bar{T}_m$ by a multiplicative factor that we call $\bar{g}$, which is assumed to be different for each protein. The strategy for fixing it is the following: once the function $\Delta G(T)$ has been estimated for all the proteins p of a given family f, we determined the temperature $\hat{T}_{m,p}$ at which it is zero. We identified $\bar{g}_p$ for a protein $p\in f$ so as to minimize the cost function:

$$\sum_{p'\neq p}(\bar{g}_p\hat{T}_{m,p'}-T_{m,p'}^{\exp})^2. \tag{14}$$

Since we are working in cross validation, the sum is over the proteins $p'\neq p$ that belong to family f. For a given protein p, the folding free energy is thus given as $\Delta G(g_pT)$.

Results

The prediction of the mechanisms used by proteins to enhance their thermoresistance is a highly non-trivial issue. The principal mechanisms of this stabilization can be schematically described in terms of three strategies (see Figures 1a–c). The first consists in a global decrease of the folding free energy $\Delta G(T)$ at all temperatures, which automatically implies an increase of the melting temperature. The second strategy consists of less negative values of ΔC_P, which broadens the stability curve. In the third strategy the temperature of maximal stability T_s undergoes a shift towards the high temperature region. It is not simple to understand which mechanism is used by each protein and if it is used alone or in combination [31]. Moreover, different proteins of the same family can reach higher thermostability through completely different mechanisms.

In order to gain understanding into the thermal stability enhancement strategies and to obtain some quantitative predictions, we designed a method to predict the full stability curve of 45 proteins that belong to 11 homologous families (see Methods section). The results are the 45 stability curves given explicitly in Table S3 and plotted in Figure 3.

To make the analysis quantitative, we extracted from these predicted stability curves three independent thermodynamic parameters that define the transition, namely T_m, ΔC_P and ΔG at 25 $^{\circ}C$, and compared them with the experimental values. For the melting temperature, the experimental values are known for all 45 entries while for the other two quantities, they are known for 17 and 16 proteins, respectively (see Table S2). We report in Table 1 the standard deviation between the computed and the experimental values, as well as the correlation coefficient between the two quantities with the corresponding P-values.

Let us start with the analysis of the melting temperature whose values are simply extracted from the protein stability curves $\Delta G(T)$ by looking for the zero of Eq. (2), since by definition:

$$\Delta G(T_m)=0. \tag{15}$$

The value of the standard deviation between the experimental and the so computed T_m's is, in cross validation, equal to about 13 $^{\circ}C$ and reduces to 10 $^{\circ}C$ when the 10% worst predicted entries are excluded (Table 1). This value is comparable with the one found previously with a different method [15], with the notable difference that we predict here simultaneously the whole stability curve. In Figure 4.a, the predicted versus the experimental T_m's are plotted; the corresponding correlation coefficient r_{T_m} is found to be equal to 0.69 (P-value 10^{-7}), and to increase to 0.76 upon exclusion of the 10% worst predicted proteins.

We also computed the ΔC_P for all the proteins belonging to the eleven homologous families. In this prediction, the identification of the normalization factor $\mathcal{M}_p$ defined in Eq. (13) is fundamental. Unfortunately, we do not have enough input data, *i.e.* experimental ΔC_P's, to identify this parameter inside each family: only for 17 entries is the ΔC_P known, with moreover often quite large experimental errors (of the order of 10–20%). When performing predictions in cross-validation, we have thus to fix $\mathcal{M}_p$ independently of the other proteins of the family (using the procedure explained in Methods) for more than half of the entries, which inevitably gives rise the errors.

The standard deviation between the experimental and the predicted values of ΔC_P is reported in Table 1. It is equal to 1.3 $kcal/(mol^{\cdot}C)$ and reduces to 0.8 $kcal/(mol^{\cdot}C)$ when the two worst predicted proteins are excluded. The experimental and

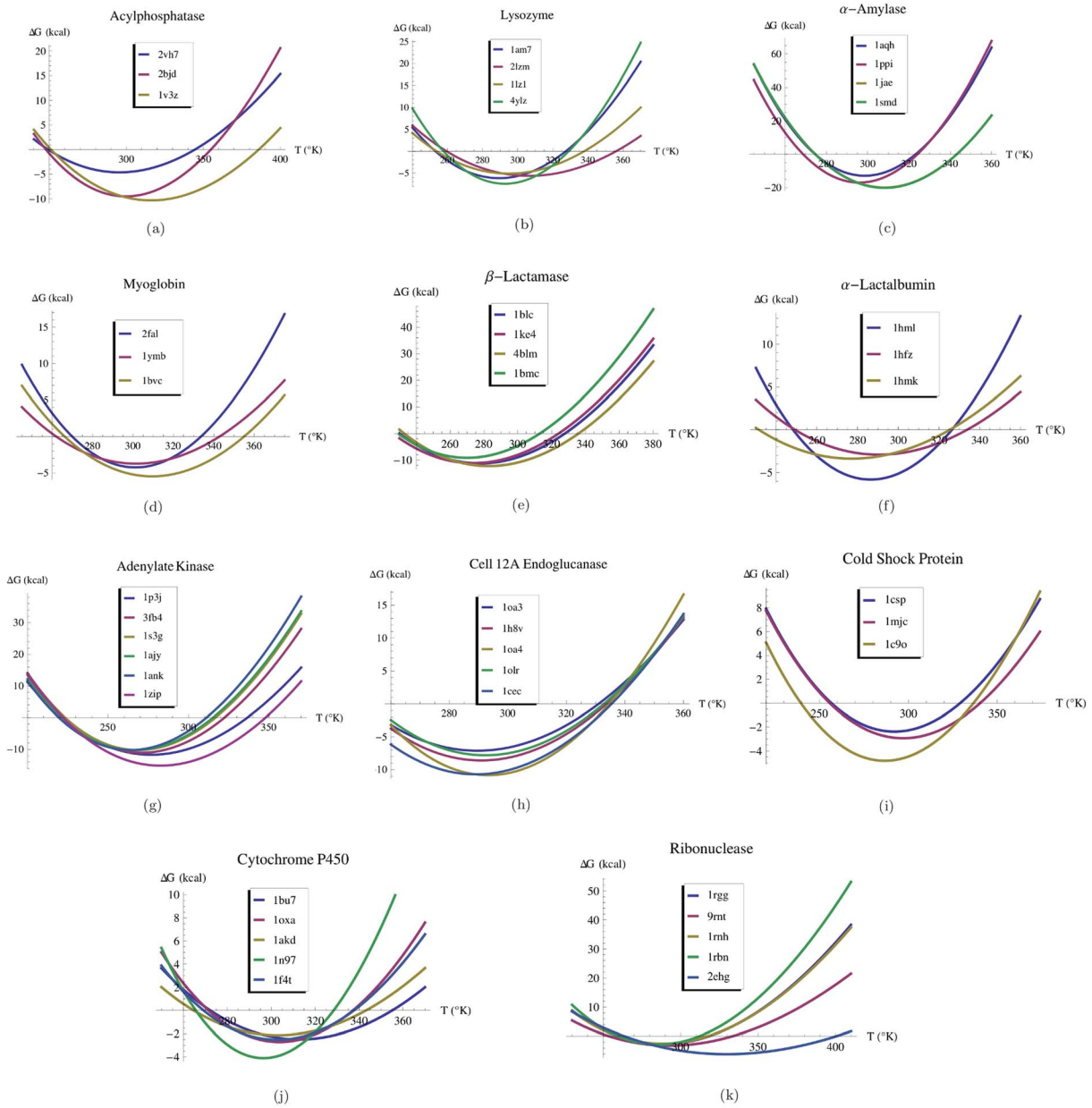

Figure 3. Predicted stability curves of the 45 proteins considered, which belong to 11 homologous families. The PDB codes, the host organisms and their environmental temperatures of all the proteins are given in the following list: (**a**) 2vh7 (*Homo sapiens*, 37 $^{\circ}C$), 2bjd (*Sulfolobus solfataricus*, 80 $^{\circ}C$), 1v3z (*Pyrococcus horikoshii*, 98 $^{\circ}C$). (**b**) 1am7 (*Bacteriophage lambda*, 37 $^{\circ}C$), 2lzm (*Escherichia coli*, 37 $^{\circ}C$), 1lz1 (*Homo sapiens*, 37 $^{\circ}C$), 1am7 (*Gallus gallus*, 41 $^{\circ}C$). (**c**) 1aqh (*Alteromonas haloplanktis*, 26 $^{\circ}C$), 1ppi (*Sus scrofa*, 39 $^{\circ}C$), 1jae (*Tenebrio molitor*, 28 $^{\circ}C$), 1smd (*Homo sapiens*, 37 $^{\circ}C$). (**d**) 2fal (*Aplysia limacina*, 17 $^{\circ}C$), 1ymb (*Equus caballus*, 38 $^{\circ}C$), 1bvc (*Physeter catodon*, 35 $^{\circ}C$). (**e**) 1blc (*Staphylococcus aureus*, 34 $^{\circ}C$), 1ke4 (*Escherichia coli*, 37 $^{\circ}C$), 4blm (*Bacillus licheniformis*, 43 $^{\circ}C$), 1bmc (*Bacillus cereus*, 30 $^{\circ}C$). (**f**) 1hml (*Homo sapiens*, 37 $^{\circ}C$), 1hfz (*Bos taurus*, 38 $^{\circ}C$), 1hmk(*Capra hircus*, 39 $^{\circ}C$). (**g**) 1p3j (*Bacillus subtilis*, 37 $^{\circ}C$), 3fb4 (*Jeotgalibacillus marinus*, 18 $^{\circ}C$), 1s3g (*Bacillus globisporus*, 15 $^{\circ}C$), 1aky (*Saccharomyces cerevisiae*, 28 $^{\circ}C$), 1ank (*Escherichia coli*, 37 $^{\circ}C$), 1zip (*Bacillus stearothermophilus*, 51 $^{\circ}C$). (**h**) 1oa3 (*Hypocrea schweinitzii*, 40 $^{\circ}C$), 1h8v (*Thrichoderma reesei*, 35 $^{\circ}C$), 1oa4 (*Streptomyces sp. 11ag8*, 30 $^{\circ}C$), 1olr (*Humicola grisea*, 50 $^{\circ}C$), 1cec (*Clostridium thermocellum*, 60 $^{\circ}C$). (**i**) 1csp (*Bacillus subtilis*, $^{\circ}C$), 1mjc (*Escherichia coli*, 37 $^{\circ}C$), 1c9o (*Bacillus caldolyticus*, 70 $^{\circ}C$). (**j**) 1bu7 (*Bacillus megaterium*, 30 $^{\circ}C$), 1oxa (*Saccharopolyspora erythraea*, 31 $^{\circ}C$), 1akd (*Pseudomonas putida*, 30 $^{\circ}C$), 1n97 (*Thermus thermophilus*, 68 $^{\circ}C$), 1f4t (*Sulfolobus solfataricus*, 78 $^{\circ}C$). (**k**) 1rgg (*Streptomyces aureofaciens*, 28 $^{\circ}C$), 9rnt (*Aspergillus Oryzae*, 49 $^{\circ}C$), 1rnh (*Escherichia coli*, 37 $^{\circ}C$), 1rbn (*Bos taurus*, 38 $^{\circ}C$), 2ehg (*Sulfolobus tokodaii*, 80 $^{\circ}C$).
doi:10.1371/journal.pcbi.1003689.g003

predicted values are plotted in Figure 4.b; the correlation $r_{\Delta C_P}$ between the two quantities is equal to 0.92 (P-value 10^{-7}), but falls down to 0.41 upon exclusion of the two worst predictions.

We chose as last independent quantity that can be extracted from the predicted curves the folding free energy at 25 $^{\circ}C$ (ΔG_{25}). The considerations made in the previous paragraph about the normalization factor $\mathcal{M}_p$ are valid for this quantity too and thus we cannot expect a perfect correlation between the predicted and experimental values due to the lack of data. We found indeed a standard deviation of 4.1 $kcal/mol$ between predicted and

Table 1. Standard deviation (σ) and linear correlation coefficient (r) between the experimental and predicted thermal and thermodynamic parameters.

Parameter	σ	$\sigma^\star$	r	r *	N (N *)	P-value
T_m	13.4 $^\circ$C	10.2 $^\circ$C	0.69	0.76	45 (40)	10^{-7}
ΔC_P	1.3 kcal/(mol $^\circ$C)	0.7 kcal/(mol $^\circ$C)	0.92	0.41	17 (15)	10^{-7}
ΔG_{25}	4.1 kcal/(mol)	2.6 kcal/(mol)	0.42	0.69	16 (14)	0.05

In the computation of σ^* and r^*, the 10% worst predicted proteins are excluded. N is the number of proteins for which experimental data are available and the results are computed.

doi:10.1371/journal.pcbi.1003689.t001

measured ΔG_{25} 's, which reduces to 2.6 $kcal/mol$ when the two worst predicted proteins are excluded. The correlation coefficient $r_{\Delta G_{25}}$ between the experimental and the predicted values is 0.4 (P-value 0.05) and 0.7 upon exclusion of the two worst predictions. These results are shown in Table 1 and plotted in Figure 4.c. A list of values of T_m, ΔC_P, and ΔG_{25} predicted from the 45 stability curves, as well as the corresponding experimental values where available, are reported in Table S2 of Supporting Material.

A further outcome that can be derived from the predicted stability curves is a better understanding of the strategies used within each protein family to reach a higher thermal stability. In particular, we can evaluate quantitatively the correlation between the thermodynamic and thermal stabilities: the linear anticorrelation between T_m and ΔG_{25} (usually taken as the descriptor of the thermodynamic stability) is relatively high and is of the order of 0.7 when the two worst predicted families are excluded. The increase of the thermodynamic stability thus remains the principal mechanism for the thermal stability enhancement. The reason for this is that single amino acid substitutions can cause much easier an increase of the number of thermodynamically stabilizing interactions, such as hydrogen bonds and hydrophobic interactions, than for example a shift of the optimal stability temperature T_s towards higher T, for which more complex amino acid substitutions are in general necessary. This result, which has already been obtained on the basis of experimental data [31,53], is here derived purely on the basis of our predictions.

The other two mechanisms for enhancing the thermostability, discussed in the previous sections, turn out to be important too even though they show a lower correlation with the melting temperature. In particular, the shift of the maximum stability temperature T_s has a linear correlation coefficient of about 0.5 with T_m and the change in heat capacity ΔC_P an anticorrelation coefficient of about 0.3, when excluding the two worst predicted families.

These predicted values can be compared with experimental data for the few proteins for which the full stability curve has been determined and thus similar correlation coefficients between T_m and T_s, and between T_m and ΔC_P can be computed (see for example [53]). Notably, the experimental correlation coefficients $r_{T_m,\Delta G}$ and $r_{T_m,\Delta C_P}$ are equal to 0.6 and 0.2, respectively, and are thus quite close to the correlation coefficient predicted by our method. The shift of T_s towards higher T appears thus to be a preferred method for enhancing the thermostability compared to the change in ΔC_P. In other words, the reduction of the conformational entropy in the denaturated state or its increase in the native state seems easier to achieve compared to a change of ΔC_P.

Discussion

The full understanding of protein thermal stability remains a challenge in protein science despite the large amount of research on this topic the last decades. As a matter of fact, it is globally more intricate to understand than the thermodynamic stability. Indeed, besides the problem due to the marginal stabilization achieved by a delicate balance of opposite forces, it poses the additional – and not the least – issue of the temperature dependence of the amino acid interactions, which is barely known.

We have designed a method based on (melting)temperature-dependent statistical potentials to deepen the thermal stability investigation. The basic idea behind this approach is simple and consists in constructing different datasets in which only proteins with given thermal properties were considered. Mean force potentials were extracted from sequence-structure frequencies computed from these datasets, following the standard statistical potential formalism, and hence reflect their thermal characteristics. They actually represent the amino acid interactions at some temperature that is related to the average T_m of the proteins in the dataset. The folding free energy of a given protein at a given temperature was estimated on the basis of these T_m -dependent potentials. More precisely, three different datasets with different average $\bar{T}_m$'s were constructed, from which three folding free energies at these $\bar{T}_m$'s were computed for each protein. The identification of the protein's full stability curve was accomplished by the identification of the modified Gibbs-Helmholtz equation (2) that best fits these three points.

Before concluding with future perspectives, let us summarize briefly the performance of the method and the main errors that affect it. The standard deviations between the experimental and computed quantities are equal, in cross-validation, to 13 $^\circ C$, 1.3 $kcal/(mol^\circ C)$ and 4.0 $kcal/mol$ for the melting temperature, the ΔC_P and the folding free energy at 25 $^\circ C$, respectively. These results can be considered as rather good especially if one considers the three main sources of error that we have encountered. The first source is certainly the lack of data. As already stressed in the main text and in [15], we do not have enough experimentally resolved proteins with known T_m to build larger datasets and thus more accurate potentials, even though we introduced some tricks to partly overcome this problem. This issue will certainly be improved when more experimental data will be available. The second source of error is related to the presence of ligands in some of the analyzed families, which contribute strongly to the protein stabilization but which we unfortunately cannot take into account with our statistical potentials. Finally, the measurement errors are sometimes quite significant, especially due to the fact that the experiments are not performed exactly in the same environmental conditions. These different issues taken together significantly increase the error on the predictions.

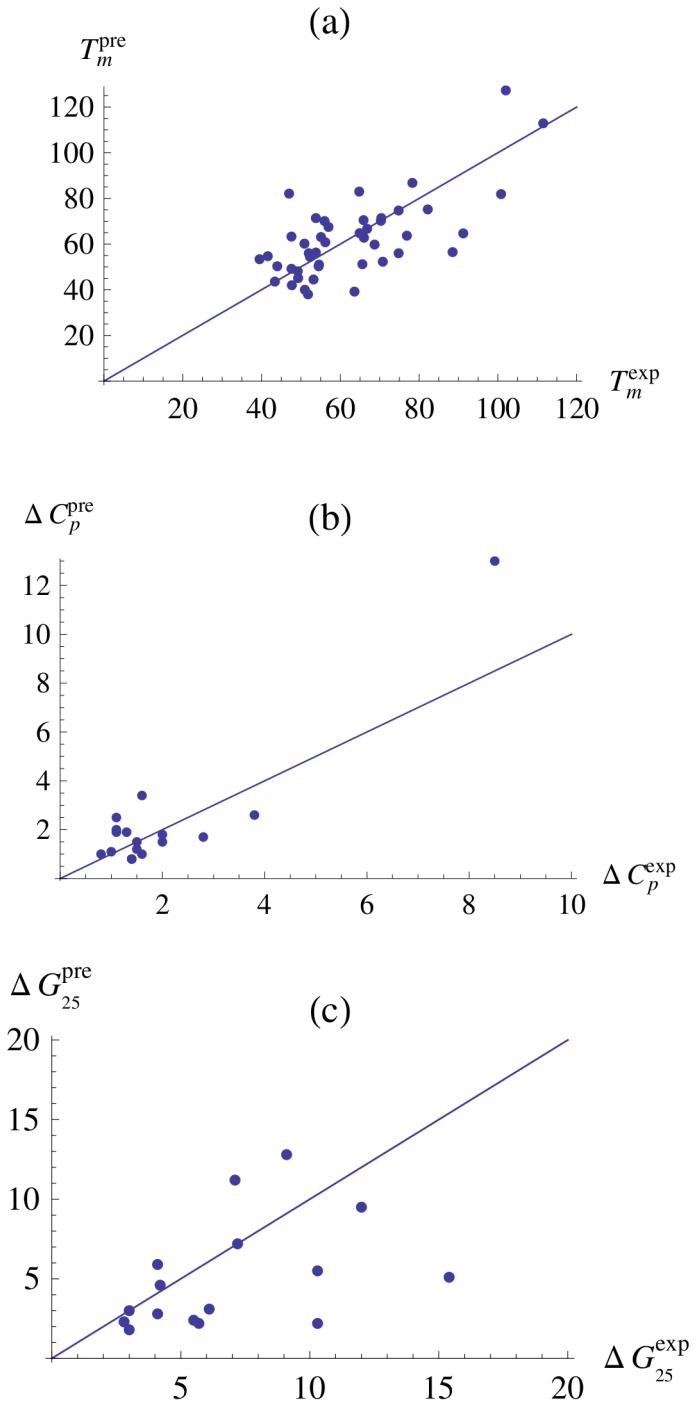

Figure 4. Comparison between: (a) the experimental and predicted melting temperatures (in °C), (b) the experimental and computed ΔC_P (in kcal/(mol °C)) and (c) the experimental and the predicted ΔG_{25} (in kcal/mol), for the set of 45 proteins belonging to the 11 homologous families. The straight lines correspond to the bisector of the first quadrant (y = x).

A noteworthy result that can be deduced from our predictive approach is that the preferred mechanism for enhancing the thermostability is an increase of the thermodynamic stability, in agreement with previous results based on experimental data [31]. Unfortunately, this does not allow us to construct an accurate predictor for the melting temperature on the basis of the thermodynamic stability only [15], since the other thermostabilizing mechanisms turned out to be important too – although to a lesser extent. Taking these other mechanisms into consideration as we did in this paper led us to a prediction method with much better performances, which we moreover hope to further improve in the near future. Furthermore, the analysis of the thermal stability optimization strategies has also shown that it is not possible to determine a unique molecular cause or a thermodynamic effect that explains the complexity of the thermal resistance

modulation for the different families, since different strategies are used in combination.

We would like to underline the main strength of our approach that is the possibility to predict at once all the thermodynamic parameters that characterize the protein folding transition. We can indeed predict with our method both the thermodynamic and thermal stabilities in a large temperature range. As far as we know this is the only method that is able to do that, and moreover it does so in a fast and relatively accurate way. Neither the standard statistical potential formalism nor the molecular dynamics simulations or the coarse-grained computational approaches to protein folding are able to consider explicitly the temperature dependence of the amino acid interactions and give predictions for both kinds of stabilities.

However, some points of the present analysis can still be improved, and we plan to do so in a future investigation. In particular, we will try to supply to the lack of data by enlarging the dataset of proteins whose thermal properties have been measured experimentally and subdivide it in more than three subsets so as to be able to get more reliable fits of the stability curves.

Two different ways can be explored to enlarge the datasets. The first consists in adding proteins with known structure but unknown melting temperature. To decide to which of the thermal ensembles these additional proteins belong, one could estimate their T_m from the method presented in this paper or from the environmental temperature of their host organism. The other strategy consists in the use of proteins with known melting temperature, whose structures are unknown but could be obtained by comparative modeling techniques. This approach is motivated by earlier analyses that tested modeled structures for the prediction of thermodynamic stability changes upon point mutations on the basis of standard statistical potentials [54]. Indeed, predictions applied on modeled structures have been shown to undergo a surprisingly small accuracy loss compared to experimental structures owing to the coarse-grained structural representation on which the potentials are based. This finding lets foresee an increase of the overall accuracy of our T_m prediction method due to the enrichment of the datasets with modeled structures. But it also foreshadows the applicability of the resulting prediction method to low-resolution or modeled structures, with good performances. This undoubtedly increases the potentialities and interest of our approach.

We expect the enlargement of the datasets to play an important role in the reduction of the prediction errors, since it will allow us to define more than three datasets and thus to compute the folding free energies of a target protein at more than three different temperatures. This should definitely reduce the consequence of the errors on the predicted points in the $(\Delta G,T)$ -plane when fitting the stability curve through those points. Moreover, larger datasets will allow us to consider more types of statistical potentials (for example potentials that depend simultaneously on amino acid types, interresidue distances and backbone torsion angle domains [52]), which are now forbidden for statistical significance reasons.

Note finally that the current version of our prediction method is family-dependent, as the datasets vary slightly from one family to another and the optimization of some parameters is performed inside the families (see Methods section). We would like to stress that this procedure does in no way bias the predictions. All our tests are indeed performed in pure cross validation. Rather, this procedure improves the predictions by exploiting relevant information that characterizes the homologous families. Another promising improvement of our prediction method, which would make it applicable to any target protein of known structure, consists in extending the current version without too much accuracy loss to the more general case that ignores any reference to homologous proteins.

In conclusion, although there is still room for improvements and generalizations, our approach has opened a novel and original way for designing fast and accurate predictors of protein stability at different temperatures.

Supporting Information

Table S1 List of 45 proteins with known melting temperature analyzed in this study.
(PDF)

Table S2 Predicted and experimental values of the thermodynamic and thermal parameters for the set of 45 proteins.
(PDF)

Table S3 Analytic expression of the predicted stability curves (in kcal/mol) for the set of 45 proteins.
(PDF)

Author Contributions

Conceived and designed the experiments: FP MR. Performed the experiments: FP. Analyzed the data: FP MR. Contributed reagents/materials/analysis tools: FP MR. Wrote the paper: FP MR.

References

1. Haki GD, Rakshit SK (2003) Developments in industrially important thermostable enzymes: a review. Bioresour Technol 89: 17–34.
2. Cruins ME, Janssen AE, Boom RM (2001) Thermozymes and their applications. Appl Biochem Biotechnol 90: 155–186.
3. Frokjaer S, Otzen DE (2005) Protein drug stability: a formulation challenge. Nat Rev Drug Discov 4: 298–306.
4. Eijsink VG, Gaseidnes S, Borchert TV, van den Burg B (2005) Directed evolution of enzyme stability Biomol Eng 22: 21–30.
5. Counago R, Chen S, Shamoo Y (2006) In vivo molecular evolution reveals biophysical origins of organismal fitness. Mol Cel 22: 441–449.
6. Wijma HJ, Floor RJ, Janssen DB (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. Curr Opin Struct Biol 23: 17.
7. Korkegian A, Black ME, Baker D, Stoddard BL (2004) Computational Thermostabilization of an Enzyme, Science 308: 857–860.
8. Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL (2007) Full-sequence computational design and solution structure of a thermostable protein variant. J Mol Biol 372: 1–6.
9. Seeliger D, de Groot BL (2010) Protein thermostability calculations using alchemical free energy simulations. Biophys J 98: 2309–16.
10. Bae E, Bannen RM, Phillips GN Jr. (2008) Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proc Natl Acad Sci U S A 105: 9594–7.
11. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, et al. (2004) Relationship between local structural entropy and protein thermostabilty. Proteins 57: 684691.
12. Ku T, Lu P, Chan C, Wang T, Lai S, et al. (2009) Predicting melting temperature directly from protein sequences. Comput Biol Chem 33: 445–450.
13. Folch B, Dehouck Y, Rooman M (2010) Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. Biophys J 98: 667–77.
14. Folch B, Rooman M, Dehouck Y (2008) Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. J Chem Inf Model 48: 119–127.
15. Pucci F, Dhanani M, Dehouck Y, Rooman M (2014) Protein thermostability prediction within homologous families by temperaturedependent statistical potentials PLoS One 9: e91659.
16. Vogt G, Woell S, Argos P (1997) Protein thermal stability, hydrogen bonds, and ion pairs. J Mol Biol 269: 631–43.
17. Kumar S, Tsai CJ, Nussinov R (2001) Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 40: 14152–65.
18. Kumar S, Tsai CJ, Nussinov R (2000) Factors enhancing protein thermostability. Protein Eng 13: 179–91.
19. Kumar S, Nussinov R (1999) Salt bridge stability in monomeric proteins. J Mol Biol 293: 1241–55.

20. Kumar S, Nussinov R (2002) Close-range electrostatic interactions in proteins. Chembiochem 3: 604–17.
21. Haney PJ, Stees M, Konisky J (1999) Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the genus Methanococcus. J Mol Biol 274: 28543–28458.
22. Cambillau C, Claverie JM (2000) Structural and genomic correlates of hyperthermostability. J Biol Chem 275: 32383–32386.
23. Melchionna S, Sinibaldi R, Briganti G (2006) Explanation of the stability of thermophilic proteins based on unique micromorphology. Biophys J 90: 4204–4212.
24. Chakravarty S, Varadarajan R (2002) Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 41: 8152–61.
25. Berezovsky IN (2001) The diversity of physical forces and mechanisms in intermolecular interactions. Phys Biol 8: 035002.
26. Ma BG, Goncearenco A, Berezovsky IN (2010) Thermophilic Adaptation of Protein Complexes Inferred from Proteomic Homology Modeling. Structure 18: 819828.
27. Elcock AH (1998) The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. J Mol Biol 284: 489–502.
28. Berezovsky IN, Zeldovich KB, Shakhnovich EI (2007) Positive and Negative Design in Stability and Thermal Adaptation of Natural Proteins. PLoS Comput Bio 3: e52.
29. Thompson MJ, Eisenberg D (1999) Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J Mol Biol 290: 595–604.
30. Nojima H, Hon-Nami K, Oshima T, Noda H (1978) Reversible thermal unfolding of thermostable cytochrome c-552. J Mol Biol 122: 3342.
31. Razvi A, Scholtz JM (2006) Lessons in stability from thermophilic proteins. Protein Sci 15: 15691578.
32. Shiraki K, Nishikori S, Fujiwara S, Hashimoto H, Kai Y, et al. (2001) Comparative analyses of the conformational stability of a hyperthermophilic protein and its mesophilic counterpart. Eur J Biochem 268: 41444150.
33. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes : a study of more than 1000 mutations. J Mol Biol 320: 369–387.
34. Parthiban V, Gromiha MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34: W239–W242.
35. Seeliger D, De Groot DL (2010) Protein thermostability calculations using alchemical free energy simulations. Biophys J 89: 2309–16.
36. Masso M, Vaisman I (2008) Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics 24: 2002–2009.
37. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0 : predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33: W306–W310.
38. Huang LT, Gromiha MM, Ho SY (2007) Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. J Mol Model 13: 879–890.
39. Cheng J, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62: 1125–1132.
40. Potapov V, Cohen M, Schreiber G (2007) Assessing computational methods for predicting protein stability change upon mutation using tree model. J Mol Model 13: 879–890.
41. Ozen A, Gonen M, Alpaydan E, Haliloglu T (2009) Machine learning integration for predicting the effect of single amino acid substitutions on protein stability. BMC Struct Biol 9: 66.
42. Dehouck Y, Grosfils A, Folch B, Filis D, Bogaerts P, et al. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks : PoPMuSiC-2.0. Bioinformatics 25: 2537–43.
43. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M (2011) PoPMuSiC 2.1 : a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 12: 151.
44. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Protein Science 4: 2138–2148.
45. Livingstone JR, Spolar RS, Record MT (1991) Contribution to the thermodynamics of protein folding from the reduction in water-accessible surface area. Biochemistry 30: 4237–4244.
46. Spolar RS, Livingstone JR, Record MT (1992) Use of liquid hydrocarbon and amide transfer data to estimate contributions to thermodynamic functions of protein folding from the removal of nonpolar and polar surface from water. Biochemistry 3: 3947–3955.
47. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
48. Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9: 945950.
49. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18: 534552.
50. Sippl MJ (19990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213: 859883.
51. Kocher JP, Rooman M, Wodak S (1994) Factors influencing the ability of knowledge based potentials to identify native sequence-structure matches. J Mol Biol 235: 15981613.
52. Dehouck Y, Gilis D, Rooman M (2006) A new generation of statistical potentials for proteins. Biophys J 90: 40104017.
53. Robertson AD, Murphy KP (1997) Protein Structure and the Energetics of Protein Stability. Chem Rev 97: 12511268.
54. Gonnelli G, Rooman M, Dehouck Y (2012) Structure-based mutant stability prediction on protein of unknwon structure. J Biotechnol 161: 287293.