Protein thermal stability probed by statistical potentials

Benjamin Folch\textsuperscript{1}, Marianne Rooman\textsuperscript{1} and Yves Dehouck\textsuperscript{1}

\textsuperscript{1}Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, CP 165/61, Avenue F. Roosevelt 50, 1050 Bruxelles, Belgique. 
\{bfolch, mrooman, ydehouck\}@ulb.ac.be

Abstract: Thermophilic organisms require proteins that maintain their structure and activity in the extreme temperature conditions in which their hosts thrive. In view of evaluating the influence of temperature on the different types of interactions that stabilize protein structures, we developed a new approach based on temperature-dependent mean force potentials. Our results show that the stabilizing weight of hydrophobic interactions remains constant, relatively to the other interactions, as the temperature increases. In contrast, Arg-involving salt bridges were found to be significantly more stabilizing at high temperature. A preference for more compact salt bridge geometries is moreover noticeable in heat-resistant proteins. Since the melting temperatures of proteins ($T_{\text{m}}$) are frequently estimated on the basis of the living temperatures ($T_{\text{env}}$) of their host organisms, we also investigated the relationship between these two quantities, to assess the relevance of such an approximation in the context of protein thermostability analyzes.

Keywords: Living temperature, melting temperature, statistical potentials, salt-bridges, thermal stability.

1 Introduction

During the course of evolution, some microorganisms have had to adapt to the most extreme environments. In particular, (hyper)thermophilic organisms live and grow at unusually high temperatures, sometimes even beyond 100°C. Even though their proteins are strikingly similar to their mesophilic counterparts, both in sequence and in structure, they require specific characteristics to be able to maintain their native conformation and perform their activity under such extreme conditions [1]. The understanding of the mechanisms that allow these proteins to function at high temperatures has crucial industrial applications [2]. Indeed, increasing the thermostability of proteins used as catalytic agents would present several advantages, such as the improvement in productivity that results from the higher speed of chemical reactions, the destruction of some byproducts and the avoidance of microorganism contamination.

The thermostability of a protein is defined by its melting temperature $T_{\text{m}}$, but the number of proteins for which the $T_{\text{m}}$ has been measured is unfortunately still very low. Therefore, in most studies devoted to the identification of the factors that promote thermostability, the thermal resistance of a protein is evaluated from the living temperature of its host organism, $T_{\text{env}}$, instead of $T_{\text{m}}$. This is an approximation, since mesophilic organisms are known to contain proteins that maintain their structure and activity at extremely high temperatures. This approximation is one of the reasons why no general relationship has been found between the thermal resistance of proteins
and specific sequence/structure features such as the percentage of amino acid types and secondary structure, the compactness, the hydrophobicity of the protein surface, or the number of interactions of different types. Besides, it appears more and more likely that thermostability results from a subtle combination of residue arrangements and interactions, rather than from a few decisive characteristics [3], and that there exist different ways of adapting to extreme temperatures [4].

Still, a few factors likely to play a determinant role in protein thermostability have been reported. In particular, salt bridges have been suggested as a means to increase the thermal stability, as the desolvation penalty resulting from the burying of charged groups into the protein core is reduced at high temperatures [5]. Furthermore, some comparisons of proteins from mesophilic and thermophilic proteins indicate a contribution of hydrophobic interactions to protein thermostability [6]. However, since accurate theoretical models dealing with the complexity of protein systems and their temperature dependence are difficult to achieve, and since trends observed in a family of homologous proteins are frequently contradicted in another family, the role of salt bridges and hydrophobic packing in thermostability remains far from settled.

In this study, rather than analyzing a series of known structures and interactions taken out of the protein context, we investigated the temperature dependence of statistical mean force potentials, describing salt bridges and hydrophobic interactions [7]. This approach has the advantage of allowing an objective determination of the temperature dependence of a given interaction in a protein environment. We also analyzed the relationship between \( T_m \) and \( T_{env} \), in view of clarifying the impact of the approximation that consists of evaluating the thermostability of a protein on the basis of the living temperature of its host organism [8].

2 Methods

2.1 Protein Datasets

We collected, from the literature and the ProTherm database [9], 127 monomeric proteins of known X-ray structure, with an atomic resolution of 2.5Å at most, and whose \( T_m \) was measured in the absence of denaturant. An average environmental temperature \( T_{env} \), corresponding to the optimal growth or normal living temperature of the species, was also assigned to each of these proteins from the literature and the PGTdb database [10].

We split the 127 proteins into two subsets, one consisting of the 63 proteins of lowest \( T_m \)'s and the other, of the 64 proteins of highest \( T_m \)'s. Both subsets were then refined to remove proteins presenting more than 25% sequence identity with another protein of the same subset. The mean temperature \( \langle T_m \rangle \) in these subsets is equal to 53°C and 81°C. In addition, 1000 random pairs of subsets were generated, with the constraint that their \( \langle T_m \rangle \) differ by no more than 3°C.

2.2 Distance-dependant Pair Potentials

Statistical potentials, derived from datasets of known protein structures, are commonly used in protein structure and stability prediction. They present the advantages of easily dealing with low-resolution protein models and taking the solvent implicitly into account. We used such potentials to evaluate the temperature dependence of different types of interactions contributing to the proteins folding free energy. More precisely, we derived, from the different datasets presented above, distance-dependent amino acid pair potentials (\( \Delta W \)) describing the interaction between two residues in a mean protein environment as a function of the distance separating them [7]:

\[
\Delta W(s_i, s_j, d_{ij}, T_s) = -kT \ln \frac{P(s_i, s_j, d_{ij}, T_s)}{P(s_i, s_j)P(d_{ij})P(T_s)}
\] (1)
where $k$ is the Boltzmann constant, $T$ is set to room temperature, and the different $P$'s are the relative frequencies of observing a pair of residues with certain characteristics in a dataset: $s_i$ and $s_j$ are the nature of these amino acids, $d_{ij}$ is the spatial distance separating the geometrical centers (Cα) of their respective side chains, and $T_S$ is the mean living or melting temperature assigned to the subset to which the protein belongs. We checked that $P(d_{ij})$ does not depend on the subset. However, as it is the case for $P(s_i, s_j)$, using this potential amounts to make the assumption that differences in amino acid composition among the subsets are not random events that should be corrected for, but reflect a sequence adaptation necessary to face the thermal conditions.

3 Protein Melting Temperature versus Organism Living Temperature

As expected, there is a correlation between the $T_m$ of a protein and the $T_{env}$ of its host organism. On the 127 proteins of our database, we found a correlation coefficient of 0.59 (Figure 1). Although significant ($p$-value ~ $10^{-12}$), this correlation is obviously far from perfect [8].

Interestingly, the proteins that have a $T_{env}$ close to 37°C span almost the entire range of $T_m$. In particular, the $T_m$'s of human proteins included in our dataset are comprised between 39.45 and 90°C. The most thermoresistant protein from a mesophilic organism in our set has a $T_m$ beyond 120°C, and comes from the sulfate-reducing bacteria *Desulfovibrio vulgaris*, which can survive in contaminated environments [11]. This example suggests that, besides $T_{env}$, other characteristics of the source organism may have an influence on the $T_m$ of its proteins.

![Figure 1](image_url)

**Figure 1.** Comparison of the melting temperature ($T_m$) of the 127 selected proteins, with the living temperature of their host organism ($T_{env}$). The regression line is also plotted: $T_m = 42.9°C + 0.62T_{env}$. The dotted line corresponds to $T_m = T_{env}$.

Among the other reasons that could explain the discrepancies between $T_m$ and $T_{env}$, we may consider the different functions and locations of proteins within a given organism, and the fact that some proteins may be stabilized *in vivo* by the formation of intermolecular complexes. A stronger relationship between $T_m$ and $T_{env}$ may thus be expected when focusing on a single family of homologous proteins. This is indeed the case for the few protein families included in our dataset, even though the small number of proteins within each family often precludes the demonstration of a statistical significance. For example, five adenylate kinases from different organisms, with $T_{env}$ ranging from 15 to 60°C, are included in our dataset. The correlation coefficient between $T_m$ and $T_{env}$ is equal to 0.81 within this family, but with a relatively large $p$-value (0.096).
To analyze the weight of the different types of interactions in conferring thermostability, we investigated their influence on the thermodynamic stability in different temperature ranges by means of statistical distance-dependent amino acid pair potentials [7]. We focused on hydrophobic interactions and salt bridges, and derived these potentials from two subsets of our dataset: the first containing proteins with low $T_m$ and the second proteins with high $T_m$. To assess the significance of our results, the same potentials were also derived from 1000 random series of two subsets, as described in Methods. Our approach gives the opportunity to evaluate quantitatively the effective impact of temperature on the different types of interactions. Although the limited size of our dataset is currently still an issue, such quantitative assessments will be necessary to any predictive application, as for example the prediction of residue substitutions likely to modify the thermostability of a protein.

4.1 Hydrophobic Interactions

Hydrophobic residues tend to avoid contact with the polar solvent and to regroup in the protein core. The interactions between these residues are thus mostly indirect, entropy-driven, interactions. We consider here the aliphatic residues Ile, Leu and Val. Aromatic residues were overlooked as they can also make $\pi-\pi$ interactions and we wish to avoid mixing different energy contributions whose temperature dependence may differ.

The potentials $W$ describing hydrophobic interactions present a broad minimum, situated at an inter-C$_\alpha$ distance of about 5-6 Å, depending on the length of the side-chains. They are depicted in Figure 2 for the Leu-Leu and Ile-Ile pairs. Most of these potentials appear to be mainly independent of $T_m$. Indeed, the differences between the potentials extracted from proteins of high or low $T_m$ are very limited and larger differences are observed in most random series (see Table 1). A manifest exception is the Ile-Ile pair, for which an energy difference of 0.26 kcal/mol between the minima of the potentials derived from the two datasets is observed. Such a signal is only observed in about 4% of the random series and, therefore, we cannot rule out its significance.

Since statistical mean force potentials only give information about the relative tendencies of the different interactions, we may conclude that hydrophobic interactions keep basically the same weight relative to the other amino acid pair interactions at all temperatures considered, except perhaps the Ile-Ile pair.
Table 1. $\Delta W_1$ and $\Delta W_2$ correspond to the energy difference between potentials derived from proteins with high and low $T_m$ or $T_{env}$, at the first and second minimum, respectively. $\text{a} P_1$ and $P_2$ are the probabilities of observing such a difference in a random series. $\text{b}$ The probabilities $P_1$ and $P_2$ are computed assuming that Asp-Arg and Glu-Arg pairs should have the same behaviour.

<table>
<thead>
<tr>
<th>Subset definition</th>
<th>$T_m$</th>
<th>$\Delta W_1$ $^a$ (kcal/mol)</th>
<th>$P_1$ $^b$ (%)</th>
<th>$\Delta W_2$ $^a$ (kcal/mol)</th>
<th>$P_2$ $^b$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ile-Ile</td>
<td>$T_m$</td>
<td>0.26</td>
<td>3.9</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ile-Leu</td>
<td>$T_m$</td>
<td>0.11</td>
<td>23.0</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Ile-Val</td>
<td>$T_m$</td>
<td>0.18</td>
<td>10.8</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Leu-Leu</td>
<td>$T_m$</td>
<td>0.06</td>
<td>56.4</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Leu-Val</td>
<td>$T_m$</td>
<td>0.03</td>
<td>79.2</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Val-Val</td>
<td>$T_m$</td>
<td>0.11</td>
<td>35.4</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Asp-Lys</td>
<td>$T_m$</td>
<td>-0.01</td>
<td>96.0</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Glu-Lys</td>
<td>$T_m$</td>
<td>0.2</td>
<td>18.8</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Asp-Arg</td>
<td>$T_m$</td>
<td>0.37</td>
<td>1.5</td>
<td>0.20</td>
<td>10.8</td>
</tr>
<tr>
<td>Glu-Arg</td>
<td>$T_m$</td>
<td>0.26</td>
<td>5.8</td>
<td>0.38</td>
<td>1.7</td>
</tr>
<tr>
<td>Asp/Glu-Arg$^c$</td>
<td>$T_m$</td>
<td>/</td>
<td>0.3</td>
<td>/</td>
<td>0.2</td>
</tr>
<tr>
<td>Asp-Arg$^d$</td>
<td>$T_{env}$</td>
<td>0.30</td>
<td>5.6</td>
<td>0.13</td>
<td>31.2</td>
</tr>
<tr>
<td>Glu-Arg$^d$</td>
<td>$T_{env}$</td>
<td>0.22</td>
<td>11.8</td>
<td>0.26</td>
<td>10.7</td>
</tr>
<tr>
<td>Asp/Glu-Arg$^d$</td>
<td>$T_{env}$</td>
<td>/</td>
<td>1.5</td>
<td>/</td>
<td>3.8</td>
</tr>
</tbody>
</table>

Salt Bridges

Salt bridges are electrostatic interactions between two amino acids of opposite charge, i.e. between Asp or Glu and Lys or Arg. For the Lys-involving salt bridges, the potentials show a broad minimum around 4.5Å, indicating the favorable nature of these interactions. The influence of $T_m$ on the depth of this minimum is relatively small and, moreover, such energy differences often occur in the random series, especially for Asp-Lys (Table 1). It should however be noted that the percentage of lysine residues is smaller in the group of proteins with high $T_m$ than in the group with low $T_m$. This is in contradiction with previous studies, which report instead a significant increase in Lys content in thermostable proteins [12], and may be due to the limited size of our dataset.

In the case of salt bridges involving arginine, the potential presents two minima (Figure 3), which correspond to different salt bridge geometries [7]. These minima are shifted towards smaller inter-residue distances for proteins with a higher temperature resistance, suggesting the preference for more compact geometries, in agreement with previous studies [13]. In addition, both minima are deeper for the potentials derived from proteins with high $T_{env}$, which indicates that these salt bridges
Protein thermal stability probed by statistical potentials

are more favorable at higher temperatures, relative to other interactions. The relevance of this result is supported by the low probabilities of observing such differences in random series (Table 1), and by the fact that the behavior is similar for Glu and Asp, as expected. We may hence estimate the probability of having a random set with such an energy difference in both Asp-Arg and Glu-Arg curves. This probability is as small as 0.3% for the first minimum, and 0.2% for the second.

To further investigate the relationship between $T_m$ and $T_{env}$, we also derived the potentials corresponding to Glu-Arg and Asp-Arg salt bridges from subsets containing proteins with high or low $T_{env}$ (Figure 3). Although the same general trends are observed as for the $T_m$-based sets, the results are less clear-cut: the differences in energy at the minima are smaller, and the probabilities to observe equivalent or larger differences in random series are two to nine times larger (Table 1). This example nicely illustrates that when proteins are divided according to their $T_{env}$, the presence of proteins with very different $T_m$ is likely to provoke a decrease of the signal’s strength, which may be even more problematic for less pronounced sequence/structure determinants of thermostability. However, even if the correlation between $T_m$ and $T_{env}$ is rather poor, $T_{env}$ remains informative for sufficiently strong tendencies.

Acknowledgements

We acknowledge support from the Belgian State Science Policy Office through an Interuniversity Attraction Poles Programme (DYSCO), from the Belgian Fund for Scientific Research (F.R.S.) through an FRFC project, and from the BioXpr bioinformatics company. BF benefits from a FRIA grant of the F.R.S., and YD from a First-Postdoc grant of the Walloon region (PROMeTHe project). MR is Research Director at the F.R.S.

References