Thermostability of Salt Bridges versus Hydrophobic Interactions in Proteins Probed by Statistical Potentials

Benjamin Folch,* Marianne Rooman, and Yves Dehouck

Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, Av. F. Roosevelt 50, CP 165/61, 1050 Bruxelles, Belgium

Received July 4, 2007

The temperature dependence of the interactions that stabilize protein structures is a long-standing issue, the elucidation of which would enable the prediction and the rational modification of the thermostability of a target protein. It is tackled here by deriving distance-dependent amino acid pair potentials from four datasets of proteins with increasing melting temperatures (T_m). The temperature dependence of the interactions is determined from the differences in the shape of the potentials derived from the four datasets. Note that, here, we use an unusual dataset definition, which is based on the T_m values, rather than on the living temperature of the host organisms. Our results show that the stabilizing weight of hydrophobic interactions (between Ile, Leu, and Val) remains constant as the temperature increases, compared to the other interactions. In contrast, the two minima of the Arg−Glu and Arg−Asp salt bridge potentials show a significant T_m dependence. These two minima correspond to two geometries: the fork−fork geometry, where the side chains point toward each other, and the fork−stick geometry, which involves the N_ϵ side chain atom of Arg. These two types of salt bridges were determined to be significantly more stabilizing at high temperature. Moreover, a preference for more-compact salt bridges is noticeable in heat-resistant proteins, especially for the fork−fork geometry. The T_m-dependent potentials that have been defined here should be useful for predicting thermal stability changes upon mutation.

INTRODUCTION

Among the huge diversity of microorganisms on Earth, many had to adapt to the most-extreme environments during the course of evolution. In particular, some have developed cold- or heat-resistant proteins that are able to continue performing their biological function at unusual temperatures. A striking feature of these proteins is their resemblance to their mesophilic counterparts, both in sequence and in structure. This makes the identification of the factors promoting thermostability or psychrostability quite complex. Note that the interest in protein thermostability recently has increased considerably, because of the important number of industrial applications that can be developed using proteins of modified thermal stability as catalytic agents.[1] Working at high temperatures, indeed, has several advantages: the improvement in productivity that results from the higher speed of chemical reactions, the destruction of some byproducts, and the avoidance of microorganism contamination.[2,3] On the other hand, the need for heating may vanish with proteins active at lower temperatures, which is both economically and ecologically valuable.[4]

Several authors have analyzed proteins as a function of the living temperature of their host organisms in search for the sequence and structure features that influence their temperature resistance.[5−13] The sought-after characteristics include the percentage of amino acid types and secondary structure, the compactness, the hydrophobicity of the protein surface, the stabilization of α-helix dipoles, and the number of hydrophobic, aromatic, cation-π, and salt bridge interactions.[14] Most results turned out to lack generality. Indeed, the variation of these factors could explain the observed changes in thermostability in some families of homologous proteins; however, no general rule could be derived, which suggests the existence of different ways of adapting to extreme temperatures.

However, a recurrent factor that promotes thermostability seems to be the number of salt bridges. In 1975, salt bridges were suggested as a means to increase the thermal stability, on the basis of sequence comparisons of proteins coming from thermophilic and mesophilic organisms.[15] Further theoretical studies[16] indicated that salt bridges have a destabilizing effect at room temperature, because of the large and unfavorable desolvation penalty incurred for burying charged groups into the core of folded proteins. However, the desolvation penalty has been shown to be reduced at high temperatures.[17]

The role of salt bridges in stability and thermostability continues to be a matter of discussion.[18] For example, energy calculations using an electrostatic continuum model showed that buried salt bridges are more stabilizing than exposed ones; this suggests that the penalty for bringing two charged residues into the core is counterbalanced by the stronger interaction that is due to the avoidance of solvent screening.[19] This result is consistent with the belief that surface-charged residues poorly affect the stability of proteins, because their interactions with the solvent should be similar in the native and unfolded states. However, it has been shown that surface

* Corresponding author phone: 32-2-6503001; Fax: 32-2-6503575; e-mail: bfolch@ulb.ac.be.

10.1021/ci700237g CCC: $40.75
Published on Web 12/28/2007

Table 1. Characteristics of the Four Overlapping Protein Sets Containing the Most Psychrostable, Mesostable, Thermostable, and Most Hyperthermostable Proteins, as Defined Based on the Melting Temperature (T_m)

			pH		Number of Residues	
group name	number of proteins	$\langle T_m \rangle$ (°C)	$\langle$pH$\rangle$	standard deviation, σ	$\langle$number of residues$\rangle$	standard deviation, σ
$G_{psychro}$	57	54.0	6.8	1.5	259	164
G_{meso}	60	60.8	6.7	1.5	242	141
G_{thermo}	57	66.3	6.6	1.6	253	153
G_{hypth}	55	79.5	6.8	1.1	235	156

charge−charge interactions are also important for protein (thermo)stability.[20] Another important thermostabilizing feature is the formation of salt bridge networks inside proteins, creating an electrostatic cohesion of charged residues.[21−23]

The role of hydrophobic packing in promoting thermostability is not settled either. It is well-known to have a dominant role in protein stability at room temperature; however, its dependence on temperature remains unclear. Comparisons of mesostable and thermostable proteins or proteins from mesophilic or thermophilic organisms indicate either a negligible[24] or a non-negligible[25,26] contribution of hydrophobic interactions to protein thermostability. Furthermore, their thermostabilizing role is suggested to be weak, relative to that of electrostatic interactions, based on energy calculations on a simplified model.[17] Other calculations suggest that hydrophobic interactions reach their maximum strength at ∼75 °C and are entropy-driven at room temperature and enthalpy-driven at high temperatures.[27]

An evolutionary approach to thermostability has recently been proposed,[28] which could explain the sometimes-contradictory trends that have been observed in statistical sequence and structure surveys. Those researchers proposed the existence of two different adaptation strategies to high temperatures, which are used by different organisms. The first is a sequence adaptation undergone by proteins from mesophilic organisms having colonized back a hot environment. The second is a structure adaptation of proteins from *archaea* organisms that appeared in the early history of Earth, when its climate was supposed to be extremely hot.[29] Their structure is supposed to be optimized directly to hot conditions and seems more compact.

In this study, we focus on the temperature dependence of salt bridges and hydrophobic interactions. Rather than analyze a series of known structures and interactions taken out of the protein context, we have chosen quite a different approach, which consists of analyzing the temperature dependence of statistical mean force potentials derived from various protein datasets. This approach has the advantage of allowing an objective determination of whether or not a given interaction in a protein environment is temperature-dependent. The difficulty of this approach lies in the correct definition of the datasets. Noisy or misdefined datasets can indeed hide the sought-after characteristics, especially when these characteristics are quite tiny. Here, we have chosen to define the datasets according to the value of the protein melting temperatures, rather than the living temperatures of the host organisms. This reduces the size of the datasets, because there are many more proteins of known living temperature than of known melting temperature. However, the datasets are much less noisy, because the thermostability of the proteins and the thermophilicity of their host organisms are only imperfectly correlated. Indeed, proteins of mesophilic organisms may be hyperthermostable, although it has been reported that, on average, the melting temperatures (T_m) are ∼24 °C higher than the living temperatures (T_{env}).[30]

METHODS

Protein Dataset. A set of 127 proteins of known X-ray structure and melting temperature T_m was collected from the literature and the ProTherm database.[31] The following criteria were used: (i) T_m was measured in the absence of denaturant, (ii) the X-ray structures present an atomic resolution of 2.5 Å or better, and (iii) the proteins are monomeric. The multimeric proteins, according to the Protein Quaternary Structure database,[32] were excluded, because their denaturation is usually not a one-step mechanism and, therefore, the meaning of the measured T_m values is less precisely defined. When the same T_m measure was performed at different pH values, the measure with the pH closest to 7 was kept.

The 127 proteins so obtained were ranked as a function of increasing T_m value and distributed into four overlapping groups of 70 proteins (group 1, proteins 1−70; group 2, proteins 20−89; group 3, proteins 39−108; and group 4, proteins 58−127). These groups were further refined to avoid redundancies. This was achieved by using the PISCES program[33] to detect, in each group, pairs of proteins with more than 25% sequence identity. These redundant proteins were filtered out in a group-dependent fashion: proteins with higher T_m values were preferentially excluded from the most psychrostable group, whereas they were preferentially retained in the most thermostable group. As a result, the overlap between these two groups is reduced to only 6 proteins. In the two intermediate groups, the selection was made to keep proteins with T_m values close to the group's average, $\langle T_m \rangle$. The final characteristics of the four groups are presented in Table 1, and a list is given as Supporting Information. Note that the average lengths and pH values are very similar in the different groups. The groups are called psychrostable, mesostable, thermostable, and hyperthermostable.

To analyze the statistical significance of our results, we also defined 1000 random series of four subsets. To obtain each series, the same 127 proteins were distributed in four overlapping sets, using the same method as described previously, except that the proteins were ranked randomly, rather than according to their T_m value. In addition, a maximal difference of 3 °C was required between the $\langle T_m \rangle$ values of the four overlapping random sets, to avoid creating unwanted T_m dependence.

Distance-Dependent 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 addressing 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 that contribute to the folding free energy of the proteins. More precisely, we derived, from each of the four datasets presented previously (and from each of their 4000 random counterparts), distance-dependent amino acid pair potentials (ΔW) that describe the interaction between two residues in a mean protein environment, as a function of the distance separating them. These potentials are computed from the relative frequencies $F(s_i,s_j,d_{ij})$ of a pair of amino acids of types s_i and s_j at positions i and j along the sequence and separated by a spatial distance d_{ij}, the relative frequencies $F(s_i,s_j)$ of pairs of residues s_i, s_j, independent of their distance, and the relative frequencies $F(d_{ij})$ of distances d_{ij}, independent of the type of residues. These frequencies are assimilated to probabilities P and converted into folding free energies (ΔW_0), using the Boltzmann law:[34−36]

$$\Delta W_0(s_i,s_j,d_{ij}) = -kT \ln\left(\frac{P(s_i,s_j,d_{ij})}{P(s_i,s_j)P(d_{ij})}\right) \cong -kT \ln\left(\frac{F(s_i,s_j,d_{ij})}{F(s_i,s_j)F(d_{ij})}\right) \quad (1)$$

where k is the Boltzmann constant and T is the absolute temperature.

The potentials ΔW_0 are normalized in such a way that the differences in amino acid composition among the subsets do not influence the potentials. The frequencies F appearing in eq 1 are indeed all computed on a given subset, without reference to the others. Thus, even if the abundance of a particular amino acid type in, for example, thermostable proteins, is physically relevant, its effect is overlooked. To take this effect into account, we devised a novel potential, ΔW_{T_m}, in which the melting temperature is considered as a structure descriptor in the same way as the distance d_{ij}:

$$\Delta W_{T_m}(s_{i,sj},d_{ij},T_m) = -kT \ln\left(\frac{P(s_i,s_j,d_{ij},T_m)}{P(s_i,s_j)P(d_{ij})P(T_m)}\right) \quad (2)$$

These potentials are derived from the entire dataset and not from the four subsets separately, but still are dependent on the melting temperature T_m. In practice, the T_m value that is associated with each pair of residues is taken as the average melting temperature $\langle T_m \rangle$ of the subset to which the protein belongs. Note that this approach requires checking that the distribution of spatial distances d_{ij} does not vary too much across the subsets, to avoid mixing the effects of protein size and T_m. We verified that the d_{ij} distributions in the four subsets are similar enough and have a negligible impact on the potentials.

In computing ΔW_0 and ΔW_{T_m}, we focused on nonlocal interactions along the polypeptide chain and dismissed all pairs of residues separated by <8 positions along the sequence. The spatial inter-residue distances d_{ij} were computed between the geometrical centers of the heavy-side-chain atoms of each residue. These side chain centroids are called C_μ.[35] Distances between 3.0 Å and 8.9 Å were grouped in 59 overlapping bins with a width of 0.5 Å, each shifted by 0.1 Å; two additional bins describe distances of <3.0 Å and >8.9 Å. This procedure ensures both a sufficient number of observations in each bin and good resolution of the potentials, because of the bin width of 0.5 Å and their 0.1 Å shift, respectively. Yet, when the number of occurrences of an amino acid pair in a bin was <15, the energy value computed on this bin was considered insignificant and excluded from our analysis. Note that we did not use the common correction for sparse data, where the probabilities $P(s_i,s_j,r_{ij})$ are replaced by linear combinations of $P(s_i,s_j,r_{ij})$ and $P(r_{ij})P(s_i,s_j)$ with coefficients that are dependent on the number of occurrences in the bins,[34] because this procedure has a tendency to reduce the absolute values of the computed energies, making it difficult to interpret differences between energies extracted from several sets of proteins. The potentials obtained were finally smoothed by replacing the energy value corresponding to a distance bin b, $\Delta W(b)$, by a linear combination involving the flanking bins: $[{}^1/_3\Delta W(b-2) + {}^2/_3\Delta W(b-1) + \Delta W(b) + {}^2/_3\Delta W(b+1) + {}^1/_3\Delta W(b+2)]/3$.

Salt Bridge Geometries. Two oppositely charged residues—Asp or Glu and Lys or Arg—can interact and form a salt bridge when they are close enough in space. A maximum cutoff distance of 4 Å is commonly assumed[37] between the charge-carrying atoms, i.e., the oxygen atoms O_{δ_1}, O_{δ_2} for Asp and O_{ϵ_1}, O_{ϵ_2} for Glu, and the nitrogen atoms N_ϵ, N_{η_1}, and N_{η_2} for Arg and N_ζ for Lys. This distance criterion cannot be directly used to interpret the energy profiles, because these are based on a simplified representation where the side chains are represented by the C_μ pseudo-atoms.

To address this issue, using HBPLUS,[38] we identified all salt bridges, respecting the cutoff distance and connecting residues that were separated by at least 8 amino acids along the chain, in a reference dataset that contained 540 monomeric high-resolution (≤2 Å) X-ray structures of protein chains with <20% pairwise sequence identity. The geometrical similarities between different occurrences of a given type of salt bridge (Asp−Lys, Glu−Lys, Asp−Arg, or Glu-Arg) were evaluated through calculation of the root-mean-square (rms) deviation of superimposed atoms, using the U3BEST algorithm.[39] The atoms considered are the charge-carrying atoms, in addition to the carbon atoms C_γ for Asp, C_δ for Glu, and C_ζ for Arg. Based on this similarity measure, all salt bridges of a given type were classified using a hierarchical treelike clustering algorithm. At the bottom of the tree, each salt bridge forms a class on its own. The classes are merged pairwise in such a way that the newly created classes have a minimum average rms deviation between all pairs of members. The clustering stops when the next class to be created presents a jump in the average rms deviation. Here, we chose an rms jump of 0.2 Å as a threshold value. Each of the classes so obtained is represented by the salt bridge that has the lowest average rms deviation, with respect to the other salt bridges in the class.

Finally, we computed the distribution of distances separating the C_μ pseudo-atoms of pairs of residues involved in salt bridges, considering distances between 3.0 Å and 8.9 Å divided into 59 overlapping bins with a width of 0.5 Å and a shift of 0.1 Å, in much the same way as that used to derive the potentials (see previous discussion). The number of salt bridges in each distance bin was normalized by the number of amino acid pairs separated by an inter-C_μ distance in the same range. This distance distribution was smoothed in the same way as the potentials.

Hydrophobic Interactions. Hydrophobic interactions result basically from the tendency of hydrophobic residues to avoid contacts with the polar solvent, which are entropically unfavorable, because of the organization of the solvent molecules around the residue. Therefore, they are indirect, entropy-driven interactions. The hydrophobic residues considered here are the aliphatic residues Ile, Leu, and Val. Aromatic residues were overlooked, because they can make $\pi-\pi$ interactions, in addition to hydrophobic packing, and as we wish to avoid mixing different energy contributions whose temperature dependence may differ. Met was not considered either, because of its rareness and its availability to form hydrogen bonds.

RESULTS

The effects of residue−residue interactions on protein thermostability were investigated with the focus on salt bridges and hydrophobic contacts. The thermostability of a protein is defined by its two denaturation temperatures, between which the native state is thermodynamically stable, with a negative folding free energy ΔG. Cold denaturation was disregarded here; only heat denaturation, at the melting temperature T_m, was considered. There is no simple relationship between $\Delta G(T)$ and T_m, except that ΔG vanishes at $T = T_m$. An empirical correlation between these two quantities has nevertheless been observed; usually, the more thermodynamically stable at a given temperature, the more thermostable a protein.[40] However, this correlation is not sufficiently accurate to achieve reliable thermostability predictions from thermodynamic stability.

Here, we chose a different approach to analyze the weight of the different types of interactions in conferring thermostability, which consists of investigating their influence on the thermodynamic stability in different temperature ranges by means of statistical distance-dependent amino acid pair potentials. Indeed, the statistical folding free energies ΔW may be dependent on specific properties of the dataset from which they are derived. For example, it has been shown that hydrophobic interactions seem to be more favorable in small proteins than in large proteins, because the protein core is, on average, less hydrophobic in large proteins.[41] Indeed, the ratio of hydrophobic to hydrophilic residues does not increase sufficiently in larger proteins to compensate for the larger volume-to-surface ratio. Here, instead of protein size, the melting temperature T_m is used as a dataset property and the thermostabilizing character of the different interactions is determined from the variations in computed folding free energies (ΔW), as a function of the average melting temperature $\langle T_m \rangle$ of the dataset.

Four overlapping groups of protein X-ray structures of increasing $\langle T_m \rangle$, representing more psychrostable, mesostable, thermostable, and more hyperthermostable proteins (see Table 1), as well as 1000 random series of four overlapping sets, were defined as described in the Methods section. From each of these datasets, two types of distance-dependent amino acid pair potentials, noted as ΔW_0 and ΔW_{T_m}, were derived using eqs 1 and 2. The term ΔW_{T_m} takes into account the variations in amino acid composition across the datasets, which reflect the adaptation of the sequences to the temperature, whereas ΔW_0 is normalized in such a way that it is not dependent on it. The variations of these potentials were analyzed for the amino acid pairs that are able to form salt bridges or hydrophobic interactions.

Hydrophobic Interactions. We focused on the hydrophobic interactions between the aliphatic residues Ile, Leu, and Val (see the Methods section). The potentials ΔW_{T_m} were computed from the four overlapping datasets, for all pairs involving these residues. They are depicted in Figures 1a and b for the Ile−Leu and Ile−Ile pairs, respectively. These potentials present a broad minimum, situated at an inter-C_μ distance of ∼5.75 Å for the longest side chains (Ile and Leu) and ∼5 Å for the shortest one (Val). The value of ΔW_{T_m} at the minimum is most negative for the longest and, thus, most hydrophobic side chains (approximately −0.9 kcal/mol), and slightly higher for the interactions that involve Val (approximately −0.8 kcal/mol).

The potentials ΔW_{T_m} for aliphatic residue pairs are mainly independent of $\langle T_m \rangle$. Indeed, the differences between the ΔW_{T_m} curves of $G_{psychro}$ and G_{hypth} are very limited, and larger differences are observed in most random series (see Table 2). Moreover, when some differences are observed, there is no progression from $G_{psychro}$ to G_{meso} to G_{thermo} to G_{hypth}. There is only one exception—the Ile−Ile interaction (Figure 1b)—which shows an energy gap of 0.17 kcal/mol between the minima of the potentials derived from the two extreme datasets ($G_{psychro}$ and G_{hypth}) and a monotonous progression of the energies from the most psychrostable group to the most thermostable one. Such a signal is only observed in ∼7% of the random series; therefore, we cannot rule out its significance.

However, there are two arguments against it. First, the variation of Ile−Ile energies with $\langle T_m \rangle$ results only from an increase of the composition in Ile, from 4.9% in $G_{psychro}$ to 5.3% in G_{hypth}. Indeed, it is not observed in the ΔW_0 potentials in which changes in amino acid composition are normalized (result not shown). Moreover, the increase in Ile observed here contradicts previous studies that addressed the temperature adaptation of sequence composition[7,10,12] and may thus be suspected to be insignificant. The second argument is the absence of physical explanation for the difference in behavior of Ile−Ile, compared to the other aliphatic pairs. If we take for granted that there is no basic difference between them, we may evaluate the probability of finding in the random subsets, for at least one of the six aliphatic pairs, an energy gap at least equal to that observed for Ile−Ile, as well as a progression among the curves. As indicated in Table 2, this probability is as high as 28%. This result casts doubt on the significance of the Ile−Ile T_m dependence and suggests that it may be a mere statistical fluctuation. This issue will have to be settled when larger datasets become available.

Thus, we may conclude that the strength of hydrophobic interactions remains constant when the temperature increases, except perhaps for the Ile−Ile pair. It must be stressed that this conclusion is somewhat hasty, because statistical mean force potentials only give information about the relative tendencies of the different interactions. More precisely, our results show that hydrophobic interactions keep basically the same weight, relative to the other amino acid pair interactions, at all temperatures considered.

Salt Bridges. Salt bridges are electrostatic interactions between two amino acids of opposite charge (i.e., between Asp or Glu and Lys or Arg). Histidines were not considered,

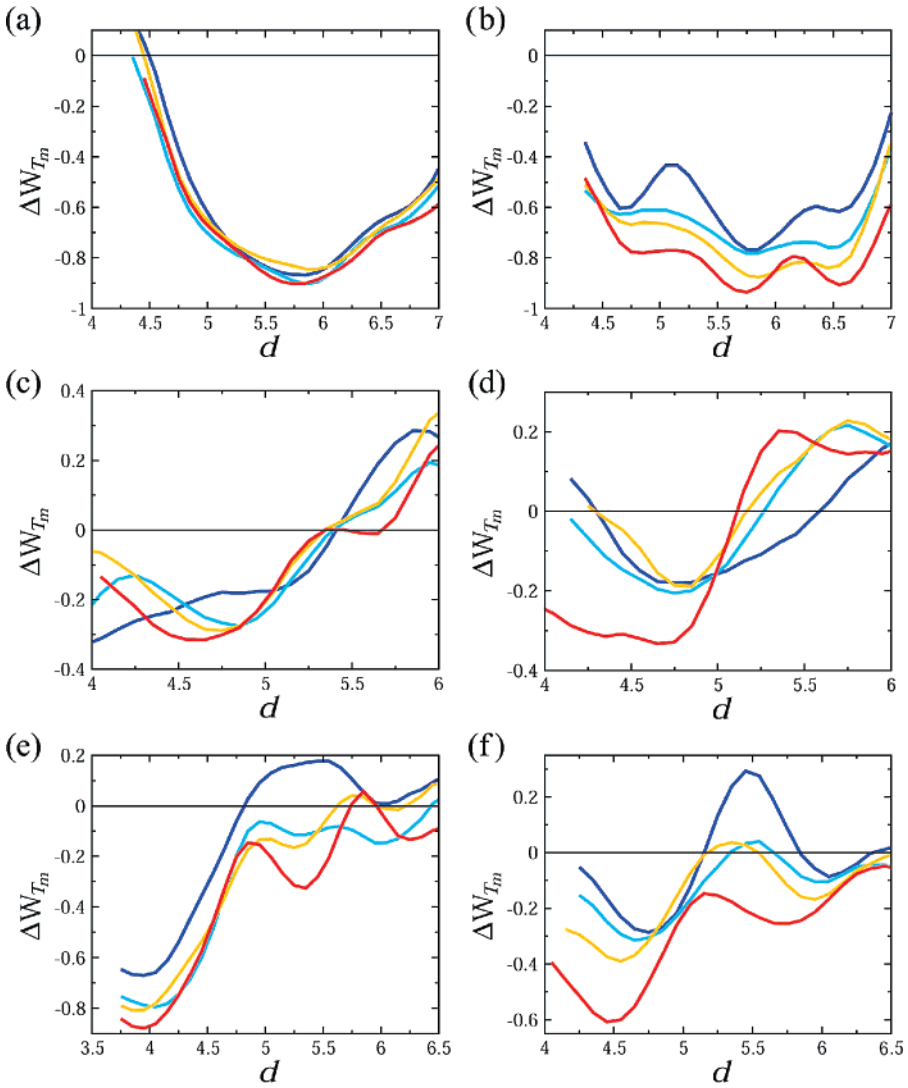

Figure 1. Folding free-energy contributions ΔW_{T_m} (given in units of kcal/mol), as a function of the distance d (in Å) between the side chain centroids C_μ, for the four protein groups $G_{psychro}$ (dark blue), G_{meso} (light blue), G_{thermo} (orange), and G_{hypth} (red): (a) Ile−Leu interaction, (b) Ile−Ile interaction, (c) Asp−Lys interaction, (d) Glu−Lys interaction, (e) Asp−Arg interaction, and (f) Glu−Arg interaction.

because they are not always charged under physiological conditions and carry an aromatic moiety that enables them to form $\pi-\pi$ interactions. As described in the Methods section, all occurrences of salt bridges in our dataset were classified according to their interaction geometry, using a treelike clustering algorithm and the rms deviation of superimposed atoms as a similarity measure. The results of this clustering, as well as the average and standard deviation of the distribution of inter-C_μ distances corresponding to each class, are given in Table 3.

Consider first the salt bridges involving lysine. The clustering algorithm groups them all into a single class, based on the rms deviation. Their inter-C_μ distance distribution shows a large peak at ∼4.5 Å, as depicted in Figure 2a for the Glu−Lys salt bridges. Similarly, the amino acid pair potentials ΔW_{T_m} of Glu−Lys and Asp−Lys show a broad minimum at ∼4.5 Å, indicating the favorable nature of these interactions (see Figures 1c and 1d).

The depth of these minima is dependent only slightly on the value of $\langle T_m \rangle$. The difference in depth between $G_{psychro}$ and G_{hypth} amounts to −0.03 kcal/mol for Asp−Lys and 0.15 kcal/mol for Glu−Lys, which does not seem to be relevant, because, in the random series, the mean energy gap almost vanishes, with a standard deviation higher than 0.1 kcal/mol (see Table 2). However, note that the percentage of lysine residues decreases from 6.5% in $G_{psychro}$ to 5.9% in G_{hypth}. This contradicts the previous studies, which, instead, have reported a significant increase in Lys content[7,10,12] and may be due to the limited size of our dataset. As a consequence, the gap observed between the ΔW_0 curves is larger than that observed for the ΔW_{T_m} potential, for both Asp−Lys and Glu−Lys (results not shown). At this stage, we thus conclude that the ΔW_{T_m} potentials do not show any significant T_m dependence for salt bridges that involve Lys; however, because of the unusual variation in Lys composition in our datasets, these results should be re-examined when larger protein ensembles will be available.

Consider now the salt bridges that involve an arginine. In this case, the clustering reveals two well-separated classes of salt bridges, for both Asp−Arg and Glu−Arg, which correspond to distinct values of the inter-C_μ distances (see Table 3 and Figure 2b). The representative member of each of the two classes of Glu−Arg salt bridge geometries is depicted in Figure 3. In the first class, the C_μ atoms are quite close to each other and the two oxygen atoms of Asp/Glu interact with the N_ϵ and N_{η_1} or N_{η_2} atoms of Arg. This

Table 2. Comparison of the Potentials (ΔW_{T_m}) Derived from the Two Extreme Subsets $G_{psychro}$ and G_{hypth} for Hydrophobic and Salt Bridge Residue Pairs

	Difference in ΔW_{T_m}[b] (kcal/mol)			Probability (%)		
residue pair[a]	$\Delta\Delta W_{T_m}$	$\langle\Delta\Delta W_{T_m}\rangle_R$	standard deviation, σ	P_G[c]	P_{GP}[d]	P'_{GP}[d]
Ile−Ile	0.17	0.00	(0.10)	8.9	7.1	28[e]
Ile−Leu	0.03	0.00	(0.08)	67.7	−	28[e]
Ile−Val	0.08	0.01	(0.09)	38.8	−	28[e]
Leu−Leu	0.03	0.00	(0.10)	78.8	−	28[e]
Leu−Val	0.04	0.00	(0.10)	72.3	−	28[e]
Val−Val	0.04	0.02	(0.10)	69.6	−	28[e]
Asp−Lys	−0.03	0.00	(0.11)	75.3	−	28[e]
Glu−Lys	0.15	0.00	(0.14)	27.7	−	28[e]
Asp−Arg 1	0.24	0.00	(0.14)	9.0	6.8	0.1[f]
Glu−Arg 1	0.32	0.01	(0.13)	1.2	1.1	0.1[f]
Asp−Arg 2	0.32	0.00	(0.14)	1.3	0.9	0.1[f]
Glu−Arg 2	0.17	0.01	(0.09)	8.4	6.3	0.1[f]

[a] Designations 1 and 2 refer to the fork−stick and fork−fork minima, respectively. [b] Difference between the minimal ΔW_{T_m} energies of $G_{psychro}$ and G_{hypth} for the T_m-dependent datasets (column 2) and the random datasets (column 3); a positive value indicates that the minimum is deeper in G_{hypth}. [c] Probability of observing, in a random series, a gap larger or equal (in absolute value) than that between $G_{psychro}$ and G_{hypth}. [d] Probability of observing, in a random series, a gap larger or equal (in absolute value) than that between $G_{psychro}$ and G_{hypth}, as well as a monotonous progression of the energies corresponding to the four subsets. P_{GP} is only computed when the progression is observed in the T_m-dependent groups. In the last column, several pairs of residues are considered simultaneously. [e] For aliphatic pairs, P'_{GP} is the proportion of random sets where at least one of the six pairs displays behavior similar to that of Ile−Ile in the nonrandom groups. [f] For salt bridges that involve Arg, P'_{GP} is the proportion of random sets where both pairs display a behavior similar to that in the nonrandom groups.

Table 3. Salt Bridge Classes Obtained by a Treelike Clustering Algorithm on a Reference Dataset

salt bridge pair	number of occurrences	average root-mean-square deviation within the class, $\langle$ rms$\rangle$ (Å)	average inter-C_μ distance, $\langle d\rangle$ (Å)	standard deviation, σ (Å)
Asp−Lys	730	0.38	4.52	0.78
Glu−Lys	758	0.40	4.78	0.87
Asp−Arg				
fork−stick	780	0.67	4.31	0.66
fork−fork	508	0.64	5.47	0.73
Glu−Arg				
fork−stick	728	0.66	4.61	0.64
fork−fork	516	0.64	5.67	0.82

geometry will be called fork−stick, because the "fork" of Asp/Glu interacts with the "stick" of the "fork" of Arg. The second class corresponds to a so-called fork−fork salt bridge, with more distant C_μ atoms, collinear side chains, and N_{η_1} and N_{η_2} interacting atoms.

Unlike Asp−Lys and Glu−Lys potentials, the amino acid pair potential ΔW_{T_m} values for Asp−Arg and Glu−Arg salt bridges show two minima instead of just one (see Figures 1e and 1f), which is consistent with the results of the clustering: each minimum corresponds to a separate class of salt bridge geometries. The first minimum, which is, by far, the deepest, occurs at inter-C_μ distances of ∼4.0 Å for Asp−Arg and ∼4.5−4.8 Å for Glu−Arg and corresponds to fork−stick salt bridge geometries. This minimum is deeper for the most thermostable proteins, reaching −0.60 kcal/mol for Glu−Arg (versus −0.28 kcal/mol in $G_{psychro}$), and −0.94 kcal/mol for Asp−Arg (versus −0.70 kcal/mol in $G_{psychro}$). In addition, a monotonous progression of the energies derived from the four sets of proteins is observed. The probability of observing such a gap and progression in a random set is low: ∼1% for Glu−Arg and ∼7% for Asp−Arg (see Table 2). The relevance of this result is further supported by the fact that the behavior is similar for Glu and Asp, as expected. Hence, we may estimate the probability of having a random set with such an energy gap and progression in both Asp−Arg and Glu−Arg curves. This probability is as small as 0.1%.

Finally, note that the observed influence of T_m on the Asp−Arg and Glu−Arg free energies results, in part, from a higher percentage of Asp, Glu, and Arg residues in thermostable proteins, in agreement with previous studies.[7,10,12] However, this is not the only factor: a similar, albeit smaller, trend is also observed with the ΔW_0 potentials.

The second minimum, corresponding to the fork−fork salt bridge geometry, is not as deep and is situated at inter-C_μ distances of ∼5.4−6.0 Å for Asp−Arg and ∼5.7−6.0 Å for Glu−Arg. The dependence of its depth on $\langle T_m\rangle$ is also very clear, as visible in Figures 1e and 1f and in Table 2: Asp−Arg and Glu−Arg fork−fork salt bridges are more favorable in thermostable proteins, with a nice energy progression from the psychrostable to the hyperthermostable proteins. The probability to observe such a signal at the second minimum in a random set is equal to 0.9% and 6.3% for Asp−Arg and Glu−Arg, respectively. If we assume that Asp and Glu have similar behaviors and consider Asp−Arg and Glu−Arg together, this probability decreases to 0.1%, which is similar to that observed for the first minimum (see Table 2).

Another remarkable trend is the shift of the curves representing the Glu/Asp−Arg potential toward smaller inter-C_μ distances when $\langle T_m\rangle$ increases, especially for the fork−fork minimum, where the increase is on the order of 0.5 Å (see Figures 1e and 1f). This observation is in agreement with previous reports of an increased compactness of

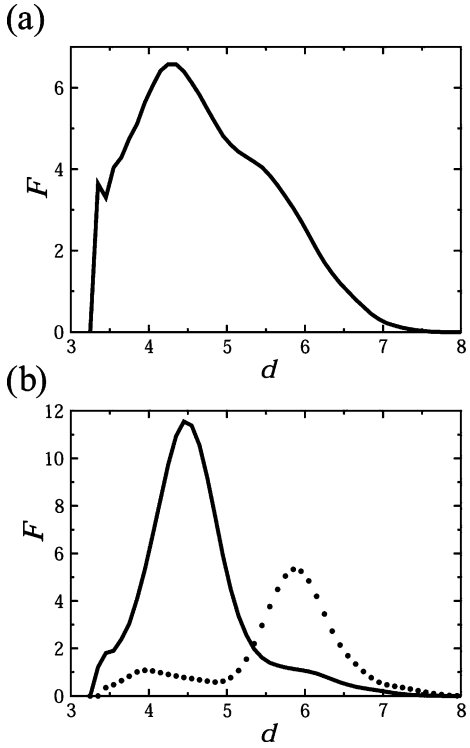

Figure 2. Fraction F (expressed in per thousand) of salt bridges in each inter-C_μ distance bin d (in Å), normalized by the total number of amino acid pairs in the same distance range, smoothed as defined in the Methods section: (a) Glu−Lys salt bridges and (b) the fork−stick (—) and fork−fork (· · ·) classes of Glu−Arg salt bridges.

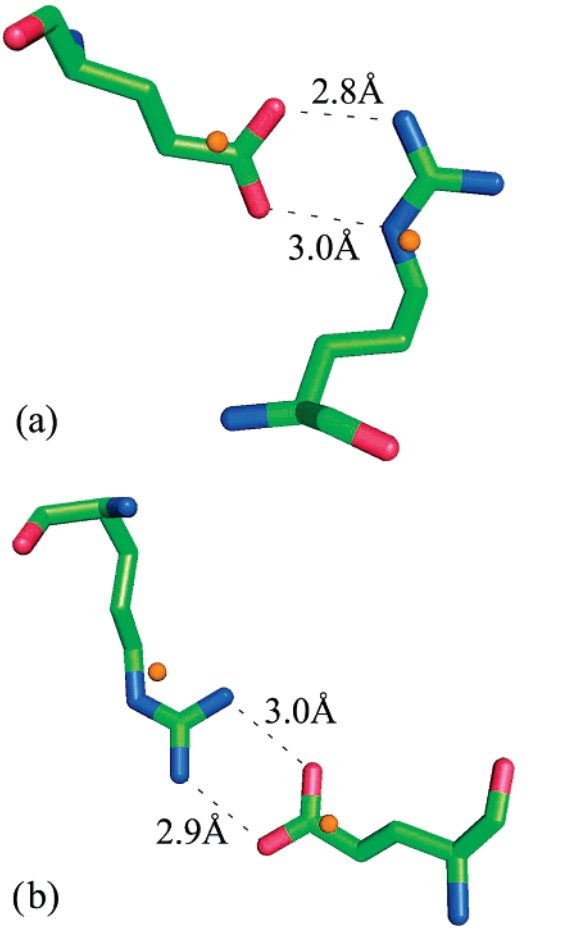

Figure 3. Representatives of the two classes of Glu−Arg salt bridge geometries described in Table 3. The side-chain centroid C_μ of each residue is represented as an orange sphere. ((a) Fork−stick geometry (inter-C_μ distance = 4.6 Å) and (b) fork−fork geometry (inter-C_μ distance = 5.9 Å).) The structures are represented using PyMol.[42]

temperature-resistant proteins.[13,14] However, this feature is less pronounced at the fork−stick minimum, and the potentials corresponding to hydrophobic interactions do not present it at all. This suggests that different amino acid pairs may contribute differently to the increase in compactness.

Note that the distance between the charge-carrying atoms remains constant in the fork−fork geometry and is not responsible for the increased compactness in thermostable proteins. Instead, the smaller inter-C_μ distances are due to slightly more-compact conformations of the side chains.

DISCUSSION

The main result of our analysis is that, when the temperature increases, the stabilizing contribution of a single hydrophobic interaction remains unchanged, compared to other interactions, whereas the salt bridges that involve Arg seem relatively more favorable. This result may be considered to be statistically significant, because only 0.1% of the random sets behave similarly to salt bridges that involve Arg; as for the hydrophobic interactions, most random sets exhibit larger temperature dependencies than those observed in the T_m-dependent sets.

This conclusion must be slightly modulated, because we could not reach firm conclusions for the Ile−Ile interactions and the salt bridges that involve Lys, because of the fact that the Ile and Lys contents in our sets of psychrostable and thermostable proteins are quite unusual and contradict previous analyses. This point will have to be re-examined when larger datasets of proteins with known T_m values become available. Note that, here, we have assumed implicitly that the only pressure on the evolutionary adaptation of the amino acid content is the thermal stability. This is clearly a simplification, because all molecules present in the cell must be resistant to the environmental temperature (in particular, RNA, DNA, and the interacting proteins). This is likely to impose new constraints on the amino acid content, which might explain some unexpected frequencies (for example, those of Ile or Lys).

Moreover, we found that two geometries must be distinguished for the Arg-containing salt bridges: the fork−stick geometry, in which the N_ϵ atom of Arg is part of the interaction, and the fork−fork geometry, where the side chains point toward each other. Both geometries are favorable, and they are more favorable in thermostable proteins than in psychrostable proteins, although to different extents. Note that, more generally, the different types of salt bridge geometries, as well as the salt bridges that involve Arg and Lys, are not necessarily equivalent, as far as thermostability is concerned, and distinguishing them should be helpful in analyzing and predicting thermal resistance. The rationale behind considering Lys and Arg separately is that the positive charge is localized in Lys and delocalized in Arg, hence implying different energetic contributions and thus possibly different temperature dependences. In contrast, Glu and Asp are expected to behave similarly, because of their similar

charge distribution, except perhaps for a slightly higher conformational entropy in Glu, which is due to its longer side chain.

Furthermore, the minima of the potentials that describe Glu−Arg and Asp−Arg salt bridges are shifted toward smaller distances in thermostable proteins, especially for the fork−fork geometry. In contrast, no such shift is observed in the potentials corresponding to hydrophobic interactions. These results suggest the preference for more-compact salt bridge geometries in thermostable proteins and the maintenance of similar hydrophobic packing.

The present results also highlight the importance of selecting protein datasets based on the melting temperature rather than on the living temperature. Indeed, the thermostabilities of proteins and the thermophilicities of their host organisms are related but not sufficiently correlated to ensure the similarity of the properties derived from both types of datasets.

Note that our results neither confirm nor infirm the suggestion that different organisms use different strategies for thermal adaptation, and, in particular, that proteins from hyperthermophilic *archaea* are more compact, whereas those from hyperthermophilic bacteria have thermally optimized interactions.[28] Indeed, very few proteins of our dataset come from *archaea*. Most of them come from bacteria and eukaryotes.

Overall, our study is complementary to previous statistical analysis of thermostability conducted within families of homologous proteins.[5,9−11,13,24] Our results are in good agreement with those, as stronger electrostatic interactions and more-compact structures have repeatedly been pinpointed as characteristics of temperature-resistant proteins, and hydrophobic interactions have often been represented as not being crucial in promoting thermostability. However, the statistical mechanics framework of mean force potentials provides a different and valuable point of view, which has allowed us to also report more-specific features, namely the fact that distinct salt bridges may be affected differently by temperature, and that the influence on compactness may be dissimilar for different types of interactions.

Note also that ΔW_{T_m} is a novel type of potential, which describes not only the correlation between sequence and structure but also their relationship with the melting temperature. This opens the way toward analyzing, in a systematic and general way, the sequence and structure adaptation to other environmental characteristics, which occurs, for instance, in acidophilic, halophilic, and barophilic organisms.

Finally, an important advantage of our approach is that the potentials developed in this paper give the possibility to evaluate quantitatively the effective impact of temperature on the different types of interactions. Although the limited size of our database is currently still an issue, such quantitative assessments will be necessary for any predictive application. For example, a direct application of our potentials derived from protein sets with different average melting temperatures (T_m) would be the prediction of residue substitutions likely to modify the thermostability of a protein.

ACKNOWLEDGMENT

We thank Christophe Biot, Dimitri Gilis, and Jean Marc Kwasigroch for interesting discussions. We acknowledge support from the Communauté Française de Belgique (through the Action de Recherche Concertée 02/07-289), the Belgian State Science Policy Office (through an Interuniversity Attraction Poles Programme (DYSCO)), the Belgian Fund for Scientific Research (FRS) (through an FRFC project), and the BioXpr Bioinformatics Company. B.F. has benefited from a FRIA grant from FRS, and Y.D. has benefited from a First-Postdoc grant of the Walloon region (PROMeTHe). M.R. is a Research Director at FRS.

Supporting Information Available: Table S1 lists the 127 proteins used in this manuscript; Table S2 describes the content of our four protein groups. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Demirjian, D. C.; Moris-Varas, F.; Cassidy, C. S. Enzymes from extremophiles. *Curr. Opin. Chem. Biol.* **2001**, *5*, 144−155.
(2) Bruins, M. E.; Janssen, A. E. M.; Boom, R. M. Thermozymes and their applications. *Appl. Biochem. Biotechnol.* **2001**, *90*, 155−186.
(3) Haki, G. D.; Rakshit, S. K. Developments in industrially important thermostable enzymes: a review. *Bioresour. Technol.* **2003**, *89*, 17−34.
(4) Gerday, C.; Aittaleb, M.; Bentahir, M.; Chessa, J.-P.; Claverie, P.; Collins, T.; D'Amico, S.; Dumont, J.; Garsoux, G.; Georlette, D.; Hoyoux, A.; Lonhienne, T.; Meuwis, M.-A.; Feller, G. Cold-adapted enzymes: from fundamentals to biotechnology. *Trends Biotechnol.* **2000**, *18*, 103−107.
(5) Vogt, G.; Woell, S.; Argos, P. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.* **1997**, *269*, 631−643.
(6) Haney, P. J.; Stees, M.; Konisky, J. Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the genus *Methanococcus*. *J. Biol. Chem.* **1999**, *274*, 28453−28458.
(7) Cambillau, C.; Claverie, J.-M. Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.* **2000**, *275*, 32383−32386.
(8) Kannan, N.; Vishveshwara, S. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. *Protein Eng.* **2000**, *13*, 753−761.
(9) Kumar, S.; Tsai, C.-J.; Nussinov, R. Factors enhancing protein thermostability. *Protein Eng.* **2000**, *13*, 179−191.
(10) Szilágyi, A.; Závodszky, P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **2000**, *8*, 493−504.
(11) Gianese, G.; Argos, P.; Pascarella, S. Structural adaptation of enzymes to low temperatures. *Protein Eng.* **2001**, *14*, 141−148.
(12) Chakravarty, S.; Varadarajan, R. Elucidation of factors responsible for enhanced thermal stability of proteins: A structural genomics based study. *Biochemistry* **2002**, *41*, 8152−8161.
(13) Gianese, G.; Bossa, F.; Pascarella, S. Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. *Proteins* **2002**, *47*, 236−249.
(14) Vieille, C.; Zeikus, G. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 1−43.
(15) Perutz, M. F.; Raidt, H. Stereochemical basis of heat stability in bacterial ferrodoxins and in haemoglobin A2. *Nature* **1975**, *255*, 256−259.
(16) Hendsch, Z. S.; Tidor, B. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci.* **1994**, *3*, 211−226.
(17) Elcock, A. H. The stability of salt bridges at high temperatures: implications of hyperthermophilic proteins. *J. Mol. Biol.* **1998**, *284*, 489−502.
(18) Kumar, S.; Nussinov, R. Close-range electrostatic interactions in proteins. *ChemBioChem* **2002**, *3*, 604−617.
(19) Kumar, S.; Nussinov, R. Salt bridge stability in monomeric proteins. *J. Mol. Biol.* **1999**, *293*, 1241−1255.
(20) Strickler, S. S.; Gribenko, A. V.; Gribenko, A. V.; Keiffer, T. R.; Tomlinson, J.; Reihle, T.; Loladze, V. V.; Makhatadze, G. I. Protein stability and surface electrostatics: a charged relationship. *Biochemistry* **2006**, *45*, 2761−2766.
(21) Musafia, B.; Buchner, V.; Arad, D. Complex salt bridges in proteins: statistical analysis of structure and function. *J. Mol. Biol.* **1995**, *254*, 761−770.
(22) Xiao, L.; Honig, B. Electrostatic contributions to the stability of hyperthermophilic proteins. *J. Mol. Biol.* **1999**, *289*, 1435−1444.
(23) Yano, J. K.; Blasco, F.; Li, H.; Schmid, R. D.; Henne, A.; Poulos, T. L. Preliminary characterization and crystal structure of a thermostable

cytochrome P450 from *Thermus thermophilus*. *J. Biol. Chem*. **2003**, *278*, 608−616.
(24) Gromiha, M. M. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophys. Chem.* **2001**, *91*, 71−77.
(25) Querol, E.; Perez-Pons, J. A.; Mozo-Villarias, A. Analysis of protein characteristics related to thermostability. *Protein Eng.* **1996**, *9*, 265−271.
(26) Haney, P.; Konisky, J.; Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. Structural basis for thermostability and identification of potential active residues for adenilate kinase from archaeal genus *Methanococcus*. *Proteins* **1997**, *28*, 117−130.
(27) Makhatadze, G. I.; Privalov, P. L. Energetics of protein structure. *Adv. Protein Chem*. **1995**, *47*, 307−325.
(28) Berezovsky, I., and Shakhnovich, E. I. Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci., USA* **2005**, *102*, 12742−12747.
(29) Di Giulio, M. The universal ancestor and the ancestor of bacteria were hyperthermophiles. *J. Mol. Evol.* **2003**, *57*, 721−730.
(30) Gromiha, M. M.; Oobatake, M.; Sarai, A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys. Chem*. **1999**, *82*, 51−67.
(31) Bava, K. A.; Gromiha, M. M.; Uedaira, H.; Kitajima, K.; Sarai, A. ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **2004**, *32*, D120−D121.
(32) Henrick, K.; Thornton, J. M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci*. **1998**, *23*, 358−361.
(33) Wang, G.; Dunbrack, R. L. PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.
(34) Sippl, M. J. 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*. **1990**, *213*, 859−883.
(35) Kocher, J.-P.; Rooman, M.; Wodak, S. Factors influencing the ability of knowledge based potentials to identify native sequence-structure matches. *J. Mol. Biol*. **1994**, *235*, 1598−1613.
(36) Dehouck, Y.; Gilis, D.; Rooman, M. A new generation of statistical potentials for proteins. *Biophys. J.* **2006**, *90*, 4010−4017.
(37) Barlow, D. J.; Thornton, J. J. Ion-pairs in proteins. *J. Mol. Biol*. **1983**, *168*, 867−885.
(38) McDonald, I. K.; Thornton, J. M. Satisfying bonding potential in proteins. *J. Mol. Biol*. **1994**, *238*, 777−793.
(39) Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *A34*, 827−828.
(40) Gilis, D.; Wintjens, R.; Rooman, M. Computed-aided methods of evaluating thermodynamic and thermal stability changes of proteins. *Res. Devel. Protein Eng.* **2001**, *1*, 277−290.
(41) Dehouck, Y.; Gilis, D.; Rooman, M. Database-derived potentials dependent on protein size for in silico folding and design. *Biophys. J.* **2004**, *87*, 171−181.
(42) DeLano, W. L. The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, 2002 (URL: http://www.pymol.org/).

CI700237G