THE PROTEIN SOCIETY

Vibrational entropy differences between mesophile and thermophile proteins and their use in protein engineering

Vincent Frappier and Rafael Najmanovich*

Department of Biochemistry, Faculty of Medicine and Health Sciences, University of Sherbrooke, J1H 5N4, Quebec, Canada

Received 31 August 2014; Revised 16 October 2014; Accepted 17 October 2014
DOI: 10.1002/pro.2592
Published online 00 Month 2014 proteinscience.org

Abstract: We recently introduced ENCoM, an elastic network atomic contact model, as the first coarse-grained normal mode analysis method that accounts for the nature of amino acids and can predict the effect of mutations on thermostability based on changes vibrational entropy. In this proof-of-concept article, we use pairs of mesophile and thermophile homolog proteins with identical structures to determine if a measure of vibrational entropy based on normal mode analysis can discriminate thermophile from mesophile proteins. We observe that in around 60% of cases, thermophile proteins are more rigid at equivalent temperatures than their mesophile counterpart and this difference can guide the design of proteins to increase their thermostability through series of mutations. We observe that mutations separating thermophile proteins from their mesophile orthologs contribute independently to a decrease in vibrational entropy and discuss the application and implications of this methodology to protein engineering.

Keywords: mesophiles; thermophiles; thermostability; protein engineering; normal mode analysis; vibrational entropy; flexibility

Introduction

Increasing the thermostability of proteins is an important component of protein engineering.[1,2] In a number of industrial applications, it is more desirable or necessary to work at higher temperatures. Additionally, thermostable proteins have longer shelf lives. Perhaps the most widespread use in research of an enzyme with higher thermal stability is the DNA polymerase from *Thermus aquosus* or Taq polymerase for short[3] that replaced the earlier use of *E. coli* DNA polymerase in PCR. Increasing protein stability is also used as a general strategy in protein engineering to build in a stability buffer to offset unwanted changes in stability caused by the introduction of other alterations that are necessary to achieve the new function that is the objective of the optimization. The above design considerations affect also the development of biologics (therapeutic proteins).[4,5]

Thermophile and hyperthermophile organisms, thought to be among the earliest life forms,[6] provide us with some of the most thermostable proteins.[7] Currently, the record holder is *Geogemma barossii*, an obligatory lithoautotroph isolated from the active Finn black-smoker hydrothermal-vent at a depth of 2280 m.[8] *G. barossii* also known as strain 121, does not grow below 85°C, can grow at 121°C (106°C optimal) and survives up to two hours at 130°C.[9,10] While little is know about the characteristics of *G. barossii* proteins, the question of what are the structural factors that lead to the higher thermal stability of thermophile proteins has been addressed numerous times.[11–25]

Perhaps the earliest attempt at understanding the differences between thermophile and mesophile proteins was performed by Perutz *et al.*[11] who studied a number of ferrodoxins and hemoglobins. By

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Rafael Najmanovich; Department of Biochemistry, FMSS, Université de Sherbrooke, 3001, 12e Avenue Nord, Sherbrooke J1H 5N4, Sherbrooke, Quebec, Canada. E-mail: rafael.najmanovich@usherbrooke.ca

analyzing the potential effect of amino acid substitutions based on their positions in the known structures, the authors suggested that the amino acid substitutions observed in the thermophile proteins may lead to higher thermal stability through the creation of salt bridges and hydrogen bonds. Argos *et al*.[26] used a set of criteria based on α-helix and β-sheet preferences, hydrophobicity, bulkiness, polarity, and least required number of codon alterations to define a subset of 24 pairwise amino acid exchanges most likely to increase protein stability. Of these, the authors observed 9 among the most frequently occurring exchanges using a set of 15 sequences representing three proteins from different organisms. More recently, Sadeghi *et al*.[19] used a dataset of 60 mesophile/thermophile homolog pairs and obtained the percentage of amino acid exchanges from mesophiles to thermophiles. Using a cutoff of 5%, the most frequently seen exchanges confirm the earlier results of Argos *et al*.[26] as well as those of Perutz *et al*.[11] showing a slight increase in the number of salt bridges and hydrogen bonds. For a long time, it has been suggested that thermophile proteins show improved packing,[12–14] a result that was confirmed by Robinon-Rechavi *et al*.[18] studying the wealth of *Thermatoga maritima* proteins structures elucidated by the Joint Center for Structural Genomics. The authors showed that there are small but statistically significant differences in compactness as measured through contact order[27] between *T. maritima* proteins and their mesophile homologs. Although the increased thermostability of thermophiles appear to come from a number of different mechanisms,[16] amino acids preferences can be used as a guiding principle in the manual design of thermostable proteins.

One of the factors differentiating thermo- from mesophile proteins is increased compactness,[12–14,18] leading to the suggestion that thermophile proteins are more rigid. It is important to keep in mind though that this is based on studying protein structures at nonthermophile temperatures.[15] Excluding differences between mesophile and thermophile proteins that reflect their divergent evolution to account for factors other than the temperature differences, the properties of highly similar mesophile and thermophile proteins ought to be similar. In that respect, the dynamics of mesophile and thermophile proteins has been studied via molecular dynamics[28] where it was found that at room temperature thermophile proteins are more rigid but equally flexible at higher temperatures. This equivalence is known as the hypothesis of equivalent states.[13,14]

Radestock *et al*.[22] compared 19 mesophile/thermophile homolog protein pairs using constraint network analysis.[29–31] The method uses the FIRST approach[31] based on the detection of regions where flexibility is affected by existing energetic constraints from covalent bonds and nonbonded interactions (primarily hydrogen bonds using a simplified distance and angle dependent potential[32]). The increase in temperature is simulated through the removal of hydrogen bonds from the list of constraints according to their calculated entropy, producing a phase transition as a function of the mean coordination of residues[33] akin to an artificial temperature. This artificial melting temperature is observed to be higher for the thermophile protein compared to its mesophile homolog in 13 out of 19 cases.[22]

Normal mode analysis[34–37] of proteins has almost as long a history as molecular dynamics[38,39] in the study of proteins. Like molecular dynamics, the number of atoms being studies is a limiting factor given the complexity of the calculations involved. However, molecular dynamics and normal mode analysis are not equivalent methods, they are complementary to each other. Both methods can use the same potential energy functions. Setting aside the caveat of how well such functions represent reality, there is a fundamental difference between molecular dynamics and normal mode analysis. Whereas molecular dynamics generates a trajectory in conformational space starting from an equilibrium structure, normal mode analysis determines a basis set of modes of movements that when combined with the appropriate choice of amplitudes, can generate any other configuration of the system. In other words, while molecular dynamics generates actual movements, normal mode analysis generates possible movements. The eigenvalues obtained by normal mode analysis can be directly used to calculate vibrational entropy differences.[40]

Over the years a number of strategies have been devised to simplify the calculations allowing the two techniques to treat larger systems, or in the case of molecular dynamics longer time scales, to address biologically relevant processes. For molecular dynamics for example among other strategies,[41] steered molecular dynamics applies an external force to drive the dynamics.[42] For normal mode analysis, a number of simplifications in the representation of the protein have allowed to tackle larger systems or perform large-scale comparisons due to the lower requirements in terms of computational time. The vast majority of simplifications for normal mode analysis involve representing amino acids through their C_α atoms[43,44] or using block representations.[45]

Our group recently introduced ENCoM as the first coarse-grained normal mode analysis method that accounts for the nature of amino acids present in the protein and not just the structure.[46] Prior to ENCoM, coarse-grained normal mode analysis methods ignored the nature of the amino acid sequence of the protein such that two proteins with the same

structure would produce equivalent results. In particular, such methods could not account for the effect of mutations. ENCoM modulates the spring constants between amino acids via the surface area in contact[47] between atoms weighted by an atom-type pairwise potential. With this more realistic representation of side-chain interactions, ENCoM performs better than ANM[48] and other methods in terms of the prediction of conformational changes.[46] Furthermore, ENCoM was tested on a large nonredundant subset of the ProTherm database with 303 cases of point mutations with experimentally calculated $\Delta\Delta G$. We used vibrational entropy differences calculated with ENCoM as an approximation for $\Delta\Delta G$ to predict experimental $\Delta\Delta G$ values for single mutants and compared ENCoM against eight existing methods to predict the effect of mutations on thermal stability. We showed that ENCoM is less biased than existing methods and particularly good at predicting stabilizing mutations.[46] To date except for ENCoM, all existing methods for the prediction of the effect of mutations on thermostability are based on machine learning approaches and enthalpy calculations. ENCoM represent an entirely new way to predict thermostability based on vibrational entropy.

In this article, we use a curated extensive nonredundant dataset of mesophile/thermophile homolog pairs[24] to study the extent to which vibrational entropy variations calculated using normal mode frequencies can differentiate thermophile proteins from their mesophile homologs. We then proceed to study how different types of mutations affect vibrational entropy differences and how such entropy changes vary with the order in which mutations leading from mesophile to the thermophile are introduced in rubredoxin. Lastly, we perform all possible mutations in all positions for rubredoxin from *D. vulgaris* and rank the mutations observed in the thermophile homolog to determine if the methodology can be used to guide the selection of mutations for added thermal stability.

Results

Dataset

Out of the initial 373 proteins pairs,[24] 314 pairs were kept based on the criteria described in the method section. The mean RMSD between pairs is 1.32 +/− 0.49 Å, the average percentage of sequence identity is 42 +/− 16 and the average sequence length is 165 +/− 80 amino acids.

Vibrational entropy

Normal mode eigenvectors and eigenvalues are used to estimate vibrational entropy differences (ΔS_{vib}) as described in the methods section.[46] ENCoM predict that thermophilic proteins have statistically significant (P-value = 0.03) smaller vibrational entropy in 186 thermophile proteins relative to their mesophilic counterparts. There is no correlation between the predicted difference in vibrational entropy of ENCoM with the RMSD between the pair ($r = 0.04$, P-value = 0.49), the sequence identity ($r = -0.04$, P-value = 0.51) or the difference in sequence length ($r = 0.03$, P-value = 0.61). The P-values above represent the statistical significance of the alternative hypothesis that the true correlation is different from 0, therefore the high P-values obtained mean that the correlation cannot be said to be different from 0, showing the statistical significance of the near null found correlations.

To assess the robustness of the predictions above and to what extent these are attributable to small structural differences, we generated for every PDB structure 50 models by homology modeling using Modeller with a flexible backbone and compared the vibrational entropy for each pair of ensembles using a Student t-test for each case in the database. ENCoM predicts that thermophilic proteins have less vibrational entropy on average than their mesophilic counterparts ($\Delta S_{vib} > 0$) in 195 cases, 104 with $\Delta S_{vib} < 0$ and 15 with no statistically significant differences (P-value 0.05). We obtain a Pearson correlation coefficient of 0.33 between the average ΔS_{vib} for the models and the ΔS_{vib} of the crystal structure only. In the 104 cases where $\Delta S_{vib} < 0$, it is likely that factors other than vibrational entropy contribute for the higher stability of the thermophile protein as previously observed.[49]

The above results for a dataset of 314 mesophile and thermophile homolog protein pairs were performed using a single mesophile homolog. It is possible however to perform such ΔS_{vib} calculations in cases were there is more than one mesophile homolog. We used a similar approach as above generating 100 models for each of 4 structures for rubredoxin from three mesophile species: *C. pasteurianum* (PDB ID 1IRO), *D. vulgaris* (PDB ID 8RXN) and *D. gigas* (PDB ID 1RDG) and one thermophile *P. furiosus* (PDB ID 1CAA). Rubredoxin is a small protein (51 amino acids) with an iron-sulphur cluster that is present in all homologs. We observe that the thermophile vibrational entropy is higher on average for the thermophile than any of its mesophile counterparts (Fig. 1) suggesting that the positive ΔS_{vib} differences observed in the large database are not the result of the particular mesophile species selected but a property conserved across mesophile homologs with respect to a thermophile.

Mutations affecting ΔS_{vib}

In this section, we are interested in what amino acid changes on average are more likely to lead to positive entropy changes ($\Delta S_{vib} > 0$) leading to a more rigid thermophile protein than its mesophile counterpart. For that purpose, we focus only on the 186

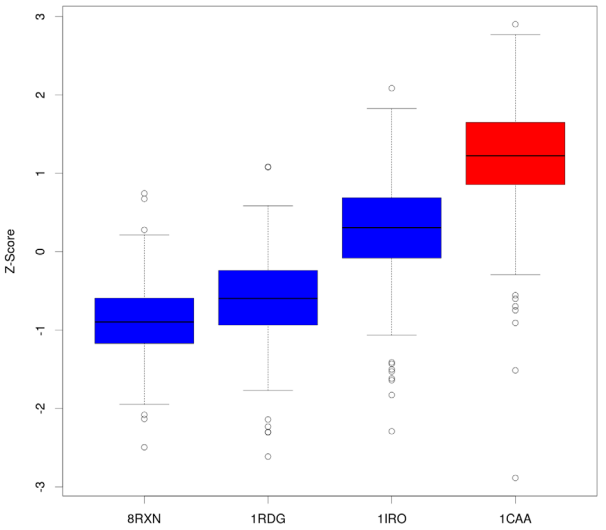

Figure 1. Average ΔS_{vib} Z-scores for rubredoxin homologs. The mesophile homologs from *C. pasteurianum* (PDB ID 1IRO), *D. vulgaris* (PDB ID 8RXN), and *D. gigas* (PDB ID 1RDG) in blue are compared with that of the thermophile *P. furiosus* (PDB ID 1CAA) in red. Bootstrapped differences are statistically significant with *P*-value < 0.001. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

cases where $\Delta S_{vib} > 0$. Furthermore, we remove the 10% outliers at both extremes of ΔS_{vib} variation (see the methods section for justification) to obtain a dataset of 14,884 mutations comprising 99 mutations types (out of the 380 possible combinations) that produce statistically significant (P-value < 0.01) ΔS_{vib} variations. We analyzed if their mean ΔS_{vib} correlates with the rate at which the given residue is observed to mutate from mesophile to thermophile proteins. In other words, are mutations more frequently seen to occur between mesophiles and thermophiles more likely to affect flexibility? We observe a correlation of 0.54 when looking at the average effect of residues mutating to any residue and −0.56 when looking at any residue mutated to a given residue (Table I). For example, mutating an alanine to any residue was observed 1559 times while the inverse, any residue to alanine 1236 times (−20.6%). Replacing an alanine for other residues increases rigidity as measured by an average ΔS_{vib} of 0.35 for the 1559 single mutations, representing an increase in stability. Other amino acids, such as the aromatic amino acids tyrosine and tryptophan or the charged amino acid arginine produce the opposite effect, are more abundant in thermophiles, lead to increased rigidity as measured by ΔS_{vib} and consequently contribute to the increase in stability of the thermophile protein. Overall, this positive correlation between abundance and ΔS_{vib} means that residues that increase rigidity are found more often in the thermophile and residues that are known to increase flexibility are more often found in the mesophile proteins.

Not just particular amino acids are preferred when changing from mesophile to thermophile, but specific preferences in amino acid pairwise replacements can be observed. For example, replacing an alanine in the mesophile for any residue other than glycine increases ΔS_{vib} (therefore stability). For alanine, the highest ΔS_{vib} is obtained when mutating to a tryptophan. The heatmap in Figure 2 shows clear

Table I. *Average Effect of Single Point Mutations*

	Number of cases				ΔS_{vib}	
Amino acid	From	To	Difference[a]		From	To
A	1559	1236	−323	(−20.72)	0.35	−0.38
C	223	112	−111	(−49.78)	NS[b]	NS
D	852	734	−118	(−13.85	NS	−0.17
E	916	1418	502	(54.80)	−0.09	NS
F	499	593	94	(18.84)	−0.23	0.56
G	730	667	−63	(−8.63)	0.20	−0.39
H	344	282	−62	(−18.02)	NS	0.24
I	951	1093	142	(14.93)	NS	0.17
K	859	1211	352	(40.98)	NS	0.12
L	1239	1314	75	(6.05)	−0.07	0.23
M	397	364	−33	(−8.31)	−0.17	0.13
N	634	531	−103	(−16.25)	0.12	−0.12
P	505	532	27	(5.35)	NS	−0.17
Q	718	432	−286	(−39.83)	NS	NS
R	729	899	170	(23.32)	−0.22	0.21
S	1001	776	−225	(−22.48)	NS	NS
T	990	740	−250	(−25.25)	NS	−0.14
V	1198	1285	87	(7.26)	0.14	−0.10
W	137	140	3	(2.19)	−0.27	0.68
Y	403	525	122	(30.27)	−0.21	0.55

[a] The number in parenthesis represents percent of change, for example, −323/1559 for A.
[b] Statistically nonsignificant value.

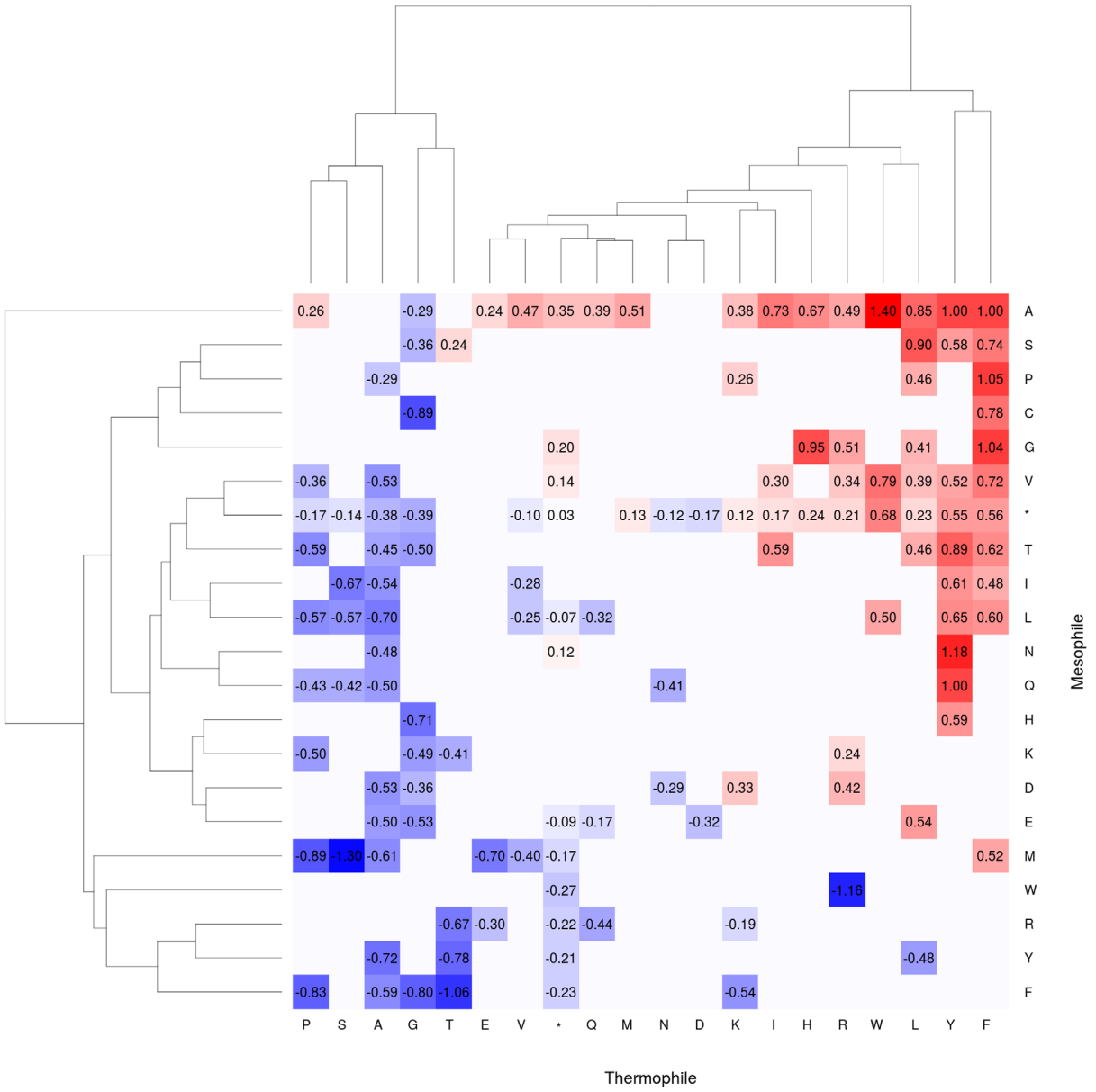

Figure 2. Heatmap of average ΔS_{vib} for pair-wise amino acid substitutions from mesophile to thermophile proteins. The top right half matrix around the inverse diagonal represent for the most part mutations that increase the stability of the thermophile protein. Missing values represent pairwise amino acid substitutions without statistically significant results. ΔS_{vib} values are scaled by 10^3 for visualization purposes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

preferences for particular amino acids with essentially the upper half triangle around the inverse diagonal discriminating between mutations that increase ΔS_{vib} and pairwise mutations in the lower right half triangle decreasing ΔS_{vib}, thus diminishing stability. Only pairwise combinations with statistically significant results are shown in the heatmap. The heatmap also includes a wildcard row (marked by *) that represents the data in Table I, that is, any amino acid in the mesophile mutated to some particular amino acid in the thermophile and a wildcard column, for mutations of particular amino acids in the mesophile to any amino acid in the thermophile. Despite not having sufficient data or no clear trend to assign average ΔS_{vib} values for every pair of amino acids, the data in Figure 2 can be used as a guide when seeking to introduce mutations into a protein in order to affect its rigidity and stability.

Engineering mutations

For each of the mesophile homologs of rubredoxin described above we calculated the best and worst and most probable order of mutations to reach the thermophile sequence as described in the methods section. In Figure 3, we show a sequence alignment of the four sequences with colors that indicate the

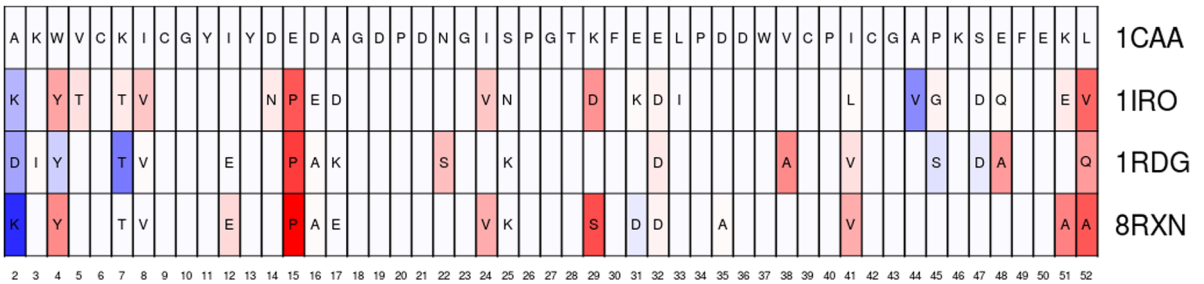

Figure 3. Alignment of rubredoxin homolog sequences showing ΔS_{vib} for single mutations. The color gradient is proportional to ΔS_{vib} with mutations predicted to increase the stability of the thermophile in red and those decreasing ΔS_{vib} in blue. Most mutations have equivalent effects across homologs. Unlabeled positions are unchanged with respect to the thermophile (1CAA). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

ΔS_{vib} of each mutation. We observe that for the most part the mutations in each position have equivalent effects across the different sequences. For each of the mesophile homologs, the best ΔS_{vib} path shows a steep increase followed by a peak/plateau and a decrease [Fig. 4(A–C)]. The single most contributing mutation according to ΔS_{vib} is P15E. Curiously, the mutation Y4W is stabilizing for two of the three homologs but destabilizing for 1RDG.

The most probable path bifurcates at times when more then one mutation of the same amino acid could be performed but in general, it also shows a steady increase. The ΔS_{vib} calculations here suggest that it may be possible to achieve the higher stability of the thermophile protein with less than the number of mutations observed. Additionally, the fact that different mutational pathways lead to approximately the same resulting the thermophile protein suggests that mutations have independent effects.

Lastly, for rubredoxin from *D. vulgaris* (PDB ID 8RXN) we performed every one of the 969 possible single point mutations (i.e., 19 mutations per position) and calculated ΔS_{vib}. All mutations were assigned an overall rank (out of 969) and a position-specific rank (out of 19). In Table II, we present the ranks of the 17 mutations observed for rubredoxin between *D. vulgaris* and *P. furiosus* (PDB ID 1CAA). The top 3 mutations in terms of ΔS_{vib} have an overall rank within the top 10%. Furthermore, for the 13 or 8 out of 16 positions that change in *P. furiosus*, the position-specific rank is among the top 10 or top 5, respectively (Table II). The results above suggest that it is possible to find among top ranking mutations those that were naturally selected. Thus, it is possible to use ΔS_{vib} with ENCoM to select mutations that increase rigidity and thermal stability in protein engineering.

Discussion

The number of factors contributing to the higher stability of thermophile proteins is as varied as the number of approaches used to understand them. From a thermodynamic point of view these led to three types of stability curves relative to a mesophile protein: an over increase in stability across the temperature range with associated increases in ΔS, a shift toward higher temperatures of the entire curve but with no change in the maximum or ΔS at a given temperature and a flattening of the curve with associated decrease in ΔS. In all cases, the Tm of the thermophile will be higher than that of the mesophile.[50] In this article, we explore the use of normal mode analysis as implemented in ENCoM to understand the differences between mesophile and thermophile proteins in terms of entropy changes. ENCoM is unique in that as a normal mode analysis method it permits to take in consideration the nature of amino acids of a protein in addition to its structure and as such account for the effect of mutations. Being a normal mode analysis method it is possible to calculate vibrational entropy differences that can be used to differentiate mesophile and thermophile proteins and being a coarse-grained method it allows us to perform a large scale analysis. As such, this work offers a new perspective on the problem. When comparing mesophile and thermophile homolog protein pairs, we observe that in around 60% of cases there is a decrease in entropy (increase in rigidity) of the thermophile protein relative to the mesophile homolog. As a result of evolution, different mechanisms to maintain the thermal stability of proteins at higher temperatures were selected.[49] The decrease in entropy as measured using ENCoM is one such factor that is statistically significant in around 60% of cases. Given the temperature dependence of the vibrational entropy,[51] the entropy difference between mesophile and thermophile homologs should decrease at the normal higher temperature of thermophile organisms, in what is known as the hypothesis of corresponding states.[20] Using crystal structures or models gave similar results in term ΔS_{vib} and no correlation was observed between this value and RMSD or the sequence identity, suggesting that ΔS_{vib} observed between the mesophile and thermophile proteins are independent of the conformation used.

The analysis of ΔS_{vib} for single mutations shows distinct preferences in terms of amino acid

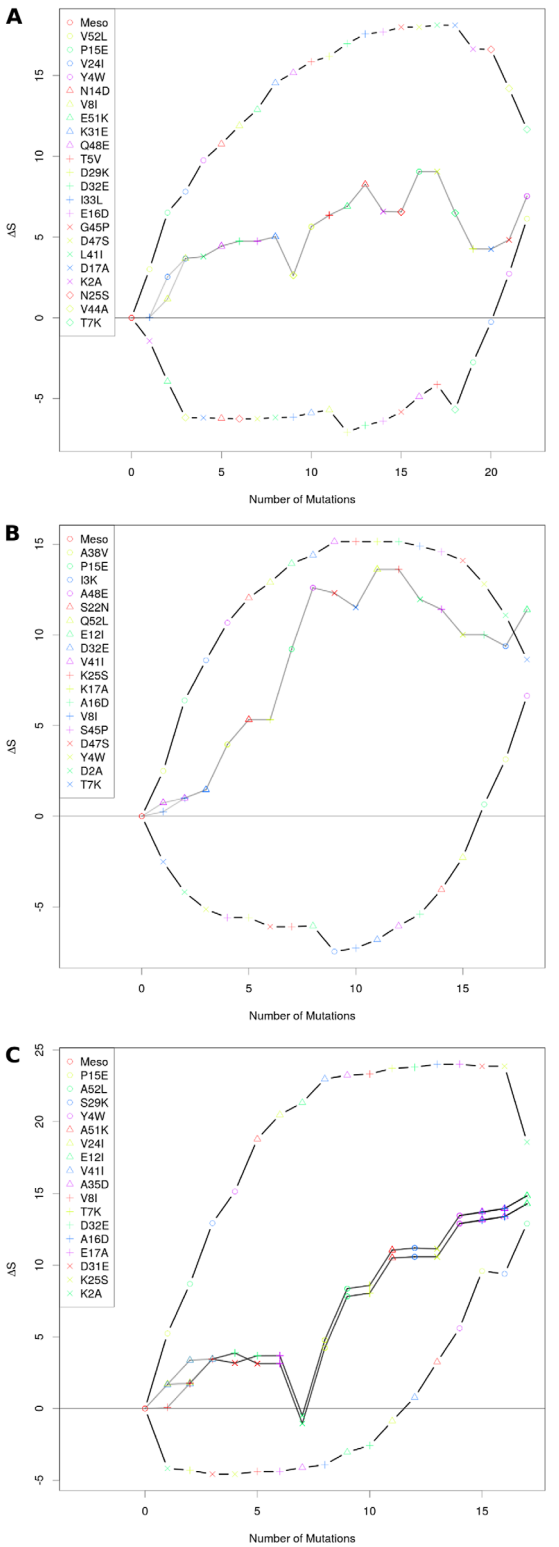

Figure 4. Mutational pathway for mesophile rubredoxin homologs to reach the thermophile sequence. Three pathways are presented, the best and worst pathways according to the ΔS_{vib} contributions of single mutations and the most probable (see methods). Each point from right to left represents a new structure with one extra mutation toward the thermophile protein. A: *C. pasteurianum* (PDB ID 1IRO), (B) *D. gigas* (PDB ID 1RDG), and (C) *D. vulgaris* (PDB ID 8RXN). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

substitutions that favor a decrease of vibrational entropy of the thermophile protein. Such substitutions can be used in protein engineering when trying to increase the thermal stability of proteins through modeling potential mutations and calculating ΔS_{vib} with ENCoM. When selecting mutations based on ΔS_{vib}, we observe a steady increase in ΔS_{vib} and a peak/plateau, suggesting that not all mutations affect the entropy and some of the thermostability effect may be obtained with a fraction of the mutations. We also observe no synergy from the order of the mutations. As different combinations of the order of performing mutations lead to the same final result, each mutation seems to contribute independently to the overall effect. The very good overall and position-specific rankings obtained for the mutations in rubredoxin from *D. vulgaris* are very promising and suggest that this strategy can be used in practice to increase the thermal stability of proteins.

Overall, the results shown here are encouraging from a protein-engineering standpoint, as it is possible to limit the search to a few top predicted residues.

Material and Methods

Database

We used the extensively curated dataset of 373 pairs of mesophile and thermophile homolog proteins of Glyakina *et al.*[24] We removed NMR structures (10 structures) as well as structure pairs with sequence lengths differing by more than 9 amino acids (as length has a direct effect on normal mode calculations). The structures were cropped on the basis of the structural alignment provided by the authors. RMSD and sequence identity were calculated for the cropped structures

Table II. *Rank of Observed Mutations for D. vulgaris Rubredoxin*

		Rank	
Mutation	ΔS_{vib}	Overall[a]	Position-specific[b]
K2A	−4.1781	914	19
Y4W	2.3298	137	1
T7K	−0.0181	620	8
V8I	0.0782	490	7
E12I	0.8392	295	6
P15E	5.2338	46	4
A16D	0.1984	449	5
E17A	0.0029	531	15.5
V24I	1.7099	190	7
S29K	3.5877	78	4
D31E	−0.2756	667	7
D32E	0.432	380	4
A35D	0.2632	425.5	3
V41I	1.6565	196	2
A51K	2.5215	125	2
A52L	3.4532	82	11

[a] Out of 969 performed single mutations (19 amino-acids in 51 positions).
[b] Out of 19 possibilities at the given position.

using TM-align in the I-TASSER software suite.[52] Heteroatoms and hydrogen atoms were removed from the structures, as the former would affect the normal mode calculations with ENCoM and the latter are not taken in account in ENCoM. The final dataset contains 314 pairs (Supporting Information File 1). The mean RMSD between pairs is 1.32 +/− 0.49 Å, the average percent sequence identity is 42 +/− 16 and the average sequence length is 165 +/− 80 amino acids.

All structures were rebuilt with Modeller using their own actual structure as template to generate a conformational ensemble for the given structure. A total of 50 models were generated for each protein.

Preparation of rubredoxin structures

All four rubredoxin structures (PDB IDs 8XRN, 1RDG, 1IRO, and 1CAA) were structurally aligned and verified using PyMol. Terminal residues that were not structurally aligned were removed in order to achieve the same sequence length for every structure. The average RMSD between all structures is 0.51 +/− 0.05Å. In working with a small number of structures (as opposed to the entire dataset above), we increased the size of the conformational ensemble generated by Modeller to 100 models for each sequence using each of the four crystal structures as templates to see if the results would be affected by the number of models. The results represent the average of 400 models for each sequence.

Determination of vibrational entropy

Normal mode analysis methods are uniquely suited to calculate vibrational entropy differences.[40] All-atom normal mode calculations are computationally expensive and thus cannot be used in large scale. While coarse-grained normal mode analysis methods are computationally fast, with the exception of ENCoM,[46] such methods cannot by definition predict the effect of mutations when these do not change the backbone conformation of the protein. Therefore, the ENCoM method is particularly suited to perform scale analyses as required for the analysis of vibrational entropy changes both in terms of the number of proteins and number of mutations. We used the ENCoM method to calculate vibrational entropy differences (ΔS_{vib}) as described earlier[40,46] and normalized by the number of modes to account for the effect of varying sequence lengths. Specifically,

$$\Delta S_{Vib,A\rightarrow B} = \frac{N_B}{N_A} \ln \left(\frac{\prod_{n=7}^{3N_A} \lambda_{n,A}}{\prod_{n=7}^{3N_B} \lambda_{n,B}} \right) \quad (1)$$

where N_A and N_B represent the number of amino acids in proteins A and B and $\lambda_{n,i}$ represents the nth normal mode (the first 6 correspond to rotational and translational degrees of freedom) for protein i. The smaller ΔS_{vib}, the higher the flexibility of the thermophile relative to the mesophile protein. Likewise, the smaller ΔS_{vib}, the smaller the contribution to the stability of the thermophile protein relative to the mesophile. The ENCoM method is available for download or online use at http://bcb.med.usherbrooke.ca/encom.

Engineering protocol

Homolog pairs of mesophile and thermophile rubredoxin protein sequences were aligned using TM-align to identify the mutations that differentiate each pair. Starting from the mesophile protein, every single mutation was generated using the Modeller Mutated function to produce one structure per mutant. For one of the homologs (*D. vulgaris* PDB ID 8RXN), modeling with a flexible backbone lead to drastic conformational changes in the C-terminus of the protein that are not observed in the thermophile. This modeling artefact did not occur for the other homologs. Thus, for consistency we modelled all structures with a fixed backbone for all three homologs. Considering the number of mutations between homologs with minimal differences in the structures (average RMSD 0.51 Å) we feel that the restriction to use a fixed backbone in this experiment does not affect the results. The ΔS_{vib} relative to the mesophile was evaluated for every structure as described above. The mutations with highest ΔS_{vib} was selected and the process was repeated for the remaining mutations in turn until the thermophile sequence was reached. The same process was performed for the worst mutations in terms of ΔS_{vib}. This protocol produces an upper and lower bound on ΔS_{vib} and selects a particular order of performing the mutations that separates a thermophile protein from its mesophile counterpart. In addition, we also use the percent amino-acid replacements between mesophile and thermophile proteins of Sadeghi *et al.*[19] to choose at each step the most probable mutation to introduce. We call this mutational pathway as the most probable pathway.

Effect of mutations

We performed every mutation observed between mesophile and thermophile proteins in the 186 cases with $\Delta S_{\mathrm{vib}} > 0$ as an individual mutation using the Mutated function of Modeller, in this case with a flexible backbone. To be included in our analysis, a type of mutation (e.g., alanine to arginine) must appear in the dataset more than 5 times and have a consistent ΔS_{vib} effect (P-value < 0.01). A mutation type may have a large or small effect on entropy differences as a result of artefacts in the methodology (modeling errors) or due to its effect on a number of other molecular properties of the protein as discussed in the introduction. As such, we remove the

extremes (top and bottom) 5% ΔS_{vib} outlier mutations when trying to identify the effect on ΔS_{vib} most often associated to different types of mutations.

Acknowledgments

RJN is part of Centre de Recherche Clinique Étienne-Le Bel, the Institute of Pharmacology of Sherbrooke, PROTEO (the Québec network for research on protein function, structure and engineering) and GRASP (Groupe de Recherche Axé sur la Structure des Protéines).

References

1. Samish I, Macdermaid CM, Perez-Aguilar JM, Saven JG (2011) Theoretical and computational protein design. Annu Rev Phys Chem 62:129–149.
2. Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. Angew Chem Int Ed Engl 52:5700–5725.
3. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by compartmentalized self-replication. Proc Natl Acad Sci USA 98:4552–4557.
4. Cauchy M, D'Aoust S, Dawson B, Rode H, Hefford MA (2002) Thermal stability: a means to assure tertiary structure in therapeutic proteins. Biologicals 30:175–185.
5. Chennamsetty N, Voynov V, Kayser V, Helk B, Trout BL (2009) Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci USA 106:11937–11942.
6. Fujiwara S (2002) Extremophiles: developments of their special functions and potential resources. J Biosci Bioenegry 94:518–525.
7. Stetter KO (1999) Extremophiles and their adaptation to hot environments. FEBS Lett 452:22–25.
8. Delaney JR, Kelley DS, Mathez EA, Yoerger DR, Baross J, Schrenk MO, Tivey MK, Kaye J, Robigou V (2001) "Edifice Rex" Sulfide Recovery Project: Analysis of submarine hydrothermal, microbial habitat. Eos Trans Am Geophys Union 82:67–73.
9. Miroshnichenko ML, Bonch-Osmolovskaya EA (2006) Recent developments in the thermophilic microbiology of deep-sea hydrothermal vents. Extremophiles 10:85–96.
10. Kashefi K, Lovley DR (2003) Extending the upper temperature limit for life. Science 301:934–934.
11. Perutz MF, Raidt H (1975) Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. Nature 255:256–259.
12. Menéndez-Arias L, Argos P (1989) Engineering protein thermal stability. Sequence statistics point to residue substitutions in alpha-helices. J Mol Biol 206:397–406.
13. Jaenicke R, Závodszky P (1990) Proteins under extreme physical conditions. FEBS Lett 268:344–349.
14. Jaenicke R (1991) Protein stability and molecular adaptation to extreme conditions. Eur J Biochem 202: 715–728.
15. Scandurra R, Consalvi V, Chiaraluce R, Politi L, Engel P (1998) Protein thermostability in extremophiles. Biochimie 80:933–941.
16. Szilágyi A, Závodszky P (2000) Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure 8:493–504.
17. Demirjian D, Moris-Varas F, Cassidy C (2001) Enzymes from extremophiles. Curr Opin Chem Biol 5:144–151.
18. Robinson-Rechavi M, Godzik A (2005) Structural genomics of thermotoga maritima proteins shows that contact order is a major determinant of protein thermostability. Structure 13:857–860.
19. Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 119:256–270.
20. Radestock S, Gohlke H (2008) Exploiting the link between protein rigidity and thermostability for data-driven protein engineering. Eng Life Sci 8:507–522.
21. Mamonova TB, Glyakina AV, Kurnikova MG, Galzitskaya OV (2010) Flexibility and mobility in mesophilic and thermophilic homologous proteins from molecular dynamics and FoldUnfold method. J Bioinform Comput Biol 8:377–394.
22. Radestock S, Gohlke H (2011) Protein rigidity and thermophilic adaptation. Proteins 79:1089–1108.
23. Sterpone F, Melchionna S (2012) Thermophilic proteins: insight and perspective from in silico experiments. Chem Soc Rev 41:1665–1676.
24. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2007) Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics 23:2231–2238.
25. Taylor TJ, Vaisman II (2010) Discrimination of thermophilic and mesophilic proteins. BMC Struct Biol 10:S5.
26. Argos P, Rossmann MG, Grau UM, Zuber H, Frank G, Tratschin JD (1979) Thermal stability and protein structure. Biochemistry 18:5698–5703.
27. Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994.
28. Lazaridis T, Lee I, Karplus M (1997) Dynamics and unfolding pathways of a hyperthermophilic and a mesophilic rubredoxin. Protein Sci 6:2589–2605.
29. Pfleger C, Rathi PC, Klein DL, Radestock S, Gohlke H (2013) Constraint network analysis (CNA): a python software package for efficiently linking biomacromolecular structure, flexibility, (thermo-)stability, and function. J Chem Inf Model 53:1007–1015.
30. Krüger DM, Rathi PC, Pfleger C, Gohlke H (2013) CNA web server: rigidity theory-based thermal unfolding simulations of proteins for linking structure, (thermo-)stability, and function. Nucl Acids Res 41:gkt292–W348.
31. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF (2001) Protein flexibility predictions using graph theory. Proteins 44:150–165.
32. Dahiyat BI, Gordon DB, Mayo SL (1997) Automated design of the surface positions of protein helices. Protein Sci 6:1333–1337.
33. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: rigidity lost. Proc Natl Acad Sci USA 99:3540–3545.
34. Tasumi M, Takeuchi H, Ataka S, Dwivedi AM, Krimm S (1982) Normal vibrations of proteins: glucagon. Biopolymers 21:711–714.
35. Go N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci USA 80:3696–3700.
36. Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci USA 80: 6571–6575.
37. Levitt M, Sander C, Stern PS (1985) Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. J Mol Biol 181:423–447.

38. McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590.
39. Levitt M (1981) Molecular dynamics of hydrogen bonds in bovine pancreatic trypsin inhibitor protein. Nature 294:379–380.
40. Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332.
41. Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106:1589–1615.
42. Park S, Schulten K (2004) Calculating potentials of mean force from steered molecular dynamics simulations. J Chem Phys 120:5946–5961.
43. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I (2001) Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J 80:505–515.
44. Doruker P, Jernigan RL, Bahar I (2002) Dynamics of large proteins through hierarchical levels of coarse-grained structures. J Comput Chem 23:119–127.
45. Tama F, Gadea FX, Marques O, Sanejouand YH (2000) Building-block approach for determining low-frequency normal modes of macromolecules. Proteins 41:1–7.
46. Frappier V, Najmanovich RJ (2014) A coarse-grained elastic network atom contact model and its use in the simulation of protein dynamics and the prediction of the effect of mutations. PLoS Comput Biol 10: e1003569.
47. McConkey B, Sobolev V, Edelman M (2003) Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci USA 100:3215–3220.
48. Doruker P, Atilgan AR, Bahar I (2000) Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: application to alpha-amylase inhibitor. Proteins 40:512–524.
49. Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cell Mol Life Sci 58:1216–1233.
50. Beadle BM, Baase WA, Wilson DB, Gilkes NR, Shoichet BK (1999) Comparing the thermodynamic stabilities of a related thermophilic and mesophilic enzyme. Biochemistry 38:2570–2576.
51. McQuarrie DA (1976) Statistical Mechanics. Harper Collins, New York.
52. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309.