PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

How do thermophilic proteins resist aggregation?

Anthony Mary Thangakani,[1] Sandeep Kumar,[2] Devadasan Velmurugan,[1] and Maria Siluvay Michael Gromiha[3]*

[1] Department of Crystallography and Biophysics, University of Madras, Chennai, Tamil Nadu 600025, India
[2] Pharmaceutical Research and Development, Biotherapeutics Pharmaceutical Sciences, Pfizer, Inc., St. Louis, MO, USA
[3] Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India

ABSTRACT

Aggregation is an ancient threat that must be overcome by proteins from all organisms to maintain their native functional states. This is essential for the maintenance of metabolic flux and viability of their cellular machineries. Here, we compare the aggregation-resistance strategies adapted by the thermophilic proteins and their mesophilic homologs using a dataset of 373 protein families. Like their mesophilic homologs, the thermophilic protein sequences also contain potential aggregation prone regions (APRs), capable of forming cross-β motif and amyloid-like fibrils. Tetrapeptide and hexapeptide amyloid-like fibril forming sequence patterns and experimentally proven amyloid-like fibril forming peptide sequences were also detected in the thermophilic proteins. Both the thermophilic and the mesophilic proteins use similar strategies to resist aggregation. However, the thermophilic proteins show superior utilization of these strategies. The thermophilic protein monomers show greater ability to "stow away" the APRs in the hydrophobic cores to protect them from solvent exposure. The thermophilic proteins are also better at gatekeeping the APRs by surrounding them with charged residues (Asp, Glu, Lys, and Arg) and Pro to a greater extent. While thermophilic and mesophilic proteins in our dataset are highly homologous and show strong overall sequence conservation, the APRs are not conserved between the homologs. These findings indicate that evolution is working to avoid amyloidogenic regions in proteins. Our results are also consistent with the observation that thermophilic cells often accumulate small molecule osmolytes capable of stabilizing their proteins and other macromolecules. This study has important implications for rational design and formulation of therapeutic proteins and antibodies.

Proteins 2012; 80:1003–1015.

Key words: aggregation; proteins; thermophiles; mesophiles; biotherapeutics.

INTRODUCTION

Thermophilic organisms can produce proteins with extreme thermal stability. Some of the thermophilic proteins withstand temperatures up to 120°C.[1] When compared with their mesophilic homologs, the thermophilic proteins show greater temperature resistance (higher melting temperature) and greater maximal stabilities, which is the free energy difference between the native and denatured state at the temperature of maximal stability.[2] This observation is also supported by the thermodynamic simulations of reversible two-state proteins.[3] Further, the thermodynamic stability of proteins showed a good positive correlation with their average environmental growth temperature.[4,5]

The stability of thermophilic proteins has been elucidated with various experiments as well as using computational methods.[6–11] Gromiha *et al.*[4] showed that the increase in Gibbs free energy change of hydration ($-G_{hN}$) and shape enhanced the stability of thermophilic proteins. This has been supported by the experimental work of Hasegawa *et al.*[12] They increased the stability of mesophilic cytochrome c through five substitutions and observed that the $-G_{hN}$ contribute to the stability.[12] Furthermore, it has been reported that increase in number of salt bridges and side chain–side chain interactions, aromatic clusters, contacts between the residues of hydrogen bond forming capability, ion pairs, cation–π interactions, noncanonical interactions, electrostatic interactions of charged residues and the dielectric response, amino acid coupling patterns, hydrophobic residues at protein surface, and main-chain hydrophobic free energy in thermophilic proteins can enhance their thermodynamic stability.[13–26]

Aggregation is an ancient threat to productive protein folding, and cellular machineries spend considerable efforts to minimize it.[27–29] In humans and animals, the aggregation of endogenous proteins often causes diseases, unless the aggregates have functional utility such as hormone storage.[30] In case of biotherapeutics, the fastest growing class among the modern pharmaceuticals,

Additional Supporting Information may be found in the online version of this article.
*Correspondence to: M. Michael Gromiha, Department of Biotechnology, Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India. E-mail: gromiha@iitm.ac.in
Received 20 October 2011; Revised 18 November 2011; Accepted 21 November 2011
Published online 5 December 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24002

aggregation remains the most common and least understood obstacle in their successful development and manufacturing.[31,32] Perhaps, due to these reasons, aggregation has emerged as one of the most important research topics in protein science. It has been reported that specific regions of protein sequences, especially the ones with amyloidogenic properties, tend to drive aggregation. These short regions, called aggregation prone regions (APRs), follow specific sequence and structural motifs.[33,34] Especially, the cross-β motif has been found in aggregates of more than 100 different proteins including bacterial inclusion bodies.[31,35] The APRs have unique features with respect to charge, hydrophobicity, aromaticity, and secondary structural preference, and several computational approaches have been developed to predict potential APRs in proteins.[31,32] Most of these prediction methods use only the protein sequences as input to identify short APRs of 5–9 residues capable of forming amyloid-like fibrils.[36] Other methods are based on pattern recognition, three-dimensional profiles, and molecular simulations.[37–40]

Aggregation is a universal phenomenon, and all proteins including those from thermophilic organisms can aggregate. In fact, the threat posed by protein aggregation to maintenance of metabolic flux, and thus, viability of the thermophilic bacterial cells cannot be under-estimated, especially at their ambient temperatures. At the elevated temperatures, the thermophilic microbial cells must struggle to preserve the native structures of their proteins. Even slight structural perturbations can quickly shift the equilibrium in favor of misfolding and aggregation. Consistently, the thermophilic cells also use osmolytes to stabilize their proteins and other cellular components. The osmolytes stabilize proteins by crowding them and, thus, altering their solvation via changes in water structure around the proteins.[41–43] The osmolytes can be of several types including sugars, polyols, amino acid derivatives, and salts. For example, the hyperthermophiles *Pyrobaculum aerophilum* and *Thermoproteus tenax* accumulate the disaccharide trehalose in large amounts. Trehalose is also used as an excipient in formulation of biopharmaceuticals.

The increased thermodynamic stability of thermophilic proteins may not necessarily equip them with lower susceptibility toward aggregation. Several mechanisms have been proposed for the process of irreversible aggregation, which is initiated by the formation of nascent seeds or aggregation nuclei. Subsequent propagation steps involve growth of aggregation nuclei to soluble aggregates to visible particulates to precipitates or fibrils, fibers, and plaques (see Buck *et al.*[44] for a review). Almost all these mechanisms postulate destabilization of native state, N, to the aggregation competent state, N*. It follows that the heights of energy barriers that separate N and N* states must be crucial to a protein's ability to resist aggregation. The N* state is different from the unfolded state, U. The energy barriers that separate the N state from the U state may have nothing to do with those that separate the N state from the N* state.[44] The emerging experimental data suggest that there is no direct correlation between thermodynamic stability of proteins and their aggregation proneness.[45,46] Even if we grant that thermophilic proteins are able to better preserve their native structures due to their greater thermodynamic stability and their transition from N to N* state may be slower, the question remains what happens when some of these proteins reach N* state. Hence, here, we asked the following questions: Do thermophilic proteins also contain APRs like their mesophilic homologs? After a thermophilic protein reaches the N* state, what molecular sequence and structural features assist/hinder the protein to form aggregates? Are there any differences among the strategies used by thermophiles and mesophiles to counter aggregation? Are the APRs conserved between thermophilic and mesophilic proteins? To our knowledge, this is the first attempt to systematically compare the strategies to resist aggregation by the proteins from thermophiles and mesophiles.

Several lessons were learnt from this data analysis study. Both thermophilic and mesophilic proteins use common strategies to resist aggregation. However, these features are stronger in the thermophilic proteins. Like their mesophilic homologs, the thermophilic protein sequences also contain several APRs capable of forming cross-β motifs and amyloid-like fibrils as detected by Tango/PAGE combination and Waltz and amyloidogenic tetrapeptide and hexapeptide patterns. Three experimentally proven amyloid-like fibril forming peptide sequences were also detected in the thermophilic proteins. These are GYE from amyloid β peptide,[47] designed tetrapeptide KVVE,[48] and a hexpeptide VSFEIV from the Waltz training set.[30] However, the thermophilic protein monomers have greater ability to "stow away" the APRs in hydrophobic cores where they are protected from solvent exposure. The thermophilic proteins are also better at gatekeeping the APRs by surrounding them with charged residues (Asp, Glu, Lys, and Arg) and Pro, which is incompatible with β-strand conformation. While thermophilic and mesophilic proteins in our dataset are highly homologous and show strong overall sequence conservation, the APRs are not conserved between the homologs. This observation further validates the notion of evolutionary bias against amyloidogenic regions in the protein sequences. Our results are consistent with the finding that the cells from thermophiles rely on small molecule organic cosolvents, osmolytes capable of stabilizing their proteins.[49]

MATERIALS AND METHODS

Dataset

We have used a dataset of 373 pairs of thermophilic and mesophilic proteins compiled by Glyakina *et al.*[23] in this study. The dataset has the following features: (i)

multidomain proteins were divided into separate single domains, (ii) a domain has not more than 400 residues, (iii) if one partner of the pair had longer sequences at the N or C termini, the extended segment of residues were truncated, (iv) the difference in the length between the proteins in a pair was not more than 10%, (v) number of residues that lack 3D coordinates were not more than 10%, and (vi) the structural alignment score computed with Maxsub was more than 70%.

Identification of APRs and computation of aggregation score

Potential APRs in mesophilic and thermophilic protein sequences were identified using a combination of Tango[50] and PAGE[51] as described in an earlier publication.[52] Tango is based on the physicochemical principles of β-sheet formation, extended by the assumption that the core regions of an aggregate are fully buried. PAGE is based on physicochemical properties and computational design of β-aggregating peptide sequences. To complement the predictions from Tango and PAGE, we have also used a relatively new aggregation prediction program, Waltz. It uses position specific scoring matrices and does a better job in recognizing polar APRs.[30] These programs provide aggregation scores for the entire sequence and/or APRs. Recent study showed that the currently available computational tools for prediction of APRs in proteins are more than 80% accurate.[31]

Identification of amyloidogenic patterns and experimentally validated aggregating peptide sequences

Lopez de la Paz and Serrano[33] have studied the link between amino acid sequence and amyloid fibril formation. They have used a *de novo* designed amyloid hexapeptide STVIIE and mutated each of the six positions with all the possible 19 natural amino acids and studied amyloid fibril formation. The authors have described two amyloidogenic sequence patterns stated below:

1. $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQE]_3$-$[ILTYWFNE]_4$-$[FIY]_5$-$\{PKRH\}_6$ for acidic pH.
2. $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQ]_3$-$[ILTYWFN]_4$-$[FIY]_5$-$\{PKRH\}_6$ for neutral pH.

These sequence patterns are written in PROSITE format. The numbers 1–6 represent positions in the hexapeptide. The curly ({ }) and the straight ([ ]) brackets indicate disallowed and allowed residues at a given position, respectively.

A third sequence pattern has been described by Tjernberg *et al.*[48]:

3. [KE]1-[FV]2-[FV]3-[EK]4 where the residue at position 1 is not the same as the one at position 4.

In addition to these patterns, we also collected 517 peptide sequences, which are known to aggregate via cross-β motif and form amyloid-like fibrils. The studies that report positive results in at least two of the following essential experiments were taken from the literature: The binding curves and/or stain pictures, Atomic Force Microscopy/Transmission Electron Microscopy (AFM/TEM) fibril pictures, powder X-ray diffraction patterns, Congo red staining pictures and/or absorbance curves, and turbidity curves. Additional pieces of evidence such as enrichment of β strand conformation by Circular Dichroism/Fourier Transform-Infra Red Microscopy (CD/FT-IR) and resistance to proteolysis were also considered. These 517 nonidentical short peptide sequences belong to ~110 different proteins and numerous synthetic peptides. Table S1 in Supporting Information lists these sequences. We searched for occurrences of the above-described library and the sequence patterns in all thermophilic and mesophilic proteins.

Amino acid composition of APRs in mesophilic and thermophilic proteins

The amino acid composition of APRs is computed using the ratio between the number of specific residues and total number of residues in APRs. It is given by

$$\mathrm{Comp}(i) = n(i)/N \tag{1}$$

where i stands for the 20 amino acid residues, and N is the total number of residues.

Hamming and Euclidean distance

Hamming and Euclidian distances were computed as described by Kumar and Bansal.[53] The Hamming distance of the amino acid composition in a mesophile–thermophile pair is given by,

$$D^{\mathrm{H}} = \Sigma\left|\mathrm{Comp(meso)}_i - \mathrm{comp(thermo)}_i\right| \tag{2}$$

where i stands for a specific amino acid residue and summation is for all the 20 amino acid residues.

The Euclidean distance is computed using the formula:

$$D^{\mathrm{E}} = \{\Sigma[\mathrm{Comp(meso)}_i - \mathrm{comp(thermo)}_i]^2\}^{1/2} \tag{3}$$

Computation of surrounding hydrophobicity

The amino acid residues in a protein molecule are represented by their α-carbon atoms, and each residue is assigned with the hydrophobicity index obtained from thermodynamic transfer experiments.[54,55] The surrounding hydrophobicity (H_{p}) of a given residue is defined as the sum of hydrophobic indices of various residues, which appear within 8 Å radius limit from it.[56]

$$H_{\mathrm{p}}(i) = \sum_{j=1}^{20} n_{ij} h_j \tag{4}$$

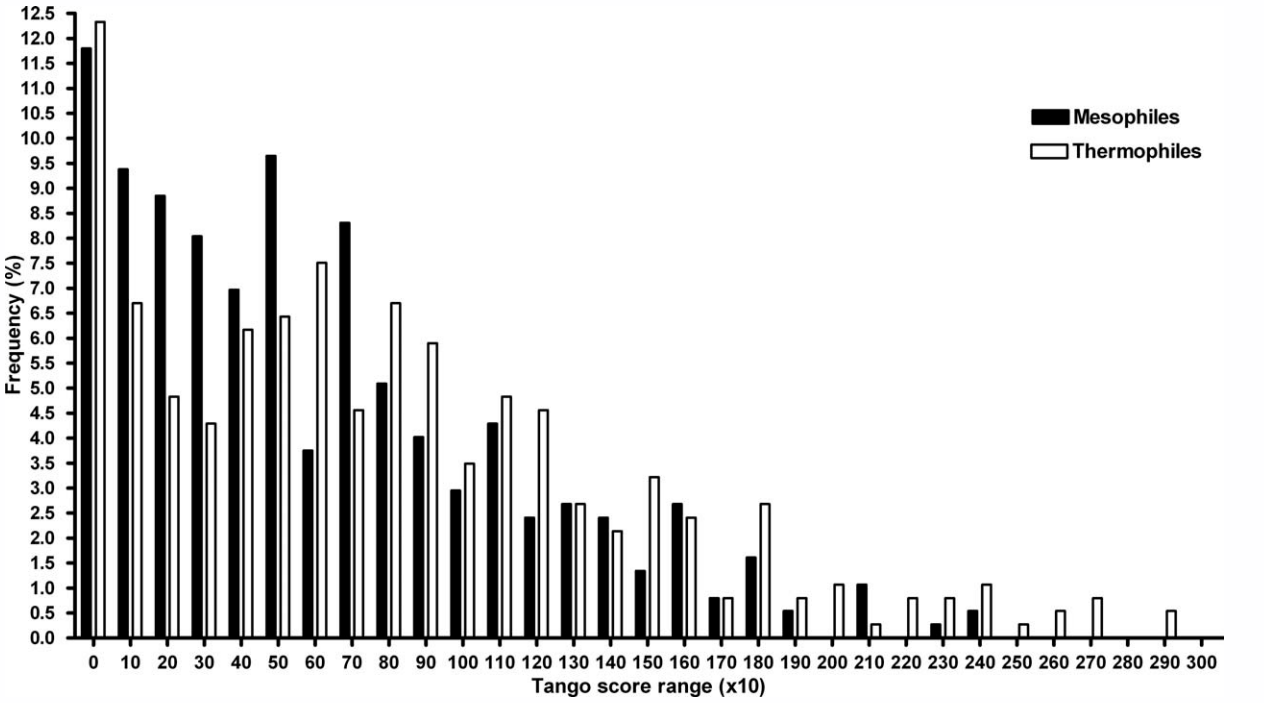

Figure 1

Frequency of occurrence of mesophilic and thermophilic proteins at various ranges of Tango score.

where n_{ij} is the total number of surrounding residues of type j around ith residue of the protein and h_j is the experimental hydrophobic index of residue type j in kcal/mol.[54,55] The average surrounding hydrophobicity of a protein is the sum of the H_p values of all the residues normalized by the total number of residues.

The limit of 8 Å is sufficient to characterize the hydrophobic behavior of amino acid residues[57] and accommodate both the local and the nonlocal interactions.[58,59] Further, 8 Å limit has been used in several studies, such as, to understand the folding rate of two-state proteins,[60,61] protein stability on mutations[62] and to determine the transition state structures of proteins.[63]

Solvent accessibility

The solvent accessibility of each residue in mesophilic and thermophilic proteins is obtained with the program, DSSP.[64]

Computation of inter-residue interaction energy

We have calculated the interaction energy between atoms in protein structures using AMBER potential,[65] which is widely used in protein folding and stability analysis. It is given by:

$$E_{\text{inter}} = \sum [(A_{ij}/r_{ij}^{12} - B_{ij}/r_{ij}^{6}) + q_i q_j/\varepsilon r_{ij}] \qquad (5)$$

where $A_{ij} = \varepsilon_{ij}^{*} (R_{ij}^{*})^{12}$, $B_{ij} = 2\varepsilon_{ij}^{*} (R_{ij}^{*})^{6}$, $R_{ij}^{*} = (R_i^{*} + R_j^{*})$, and $\varepsilon_{ij}^{*} = (\varepsilon_i^{*} \varepsilon_j^{*})^{1/2}$; R^{*} and ε^{*} are, respectively, the van der Waals radius and well depth, and these parameters are obtained from Cornell *et al.*[65]; q_i and q_j are, respectively, the charges for the atoms i and j, and r_{ij} is the distance between them. We have used the distant dependent dielectric constant ($\varepsilon = r_{ij}$) to take account of the dielectric damping effect of the Coulomb interactions, as used in other studies.[66,67]

RESULTS

Aggregation score for mesophilic and thermophilic proteins

We have computed the aggregation score for all mesophilic and thermophilic proteins and their difference using Tango. The results were grouped into three different clusters: (i) both mesophiles and thermophiles have similar aggregation scores, (ii) mesophiles with higher aggregation scores, and (iii) high scores for thermophilic proteins. The variations of Tango aggregation scores for mesophilic and thermophilic proteins at different intervals are shown in Figure 1. We noticed that mesophilic proteins tend to have lower aggregation scores. The mesophilic proteins (~55%) have the aggregation score of less than 600. The higher aggregation scores (>800) are more frequent for the thermophilic proteins. This indicates that there may be amino acid compositional differ-

Table I
Variation of Aggregation Score Obtained with Tango with H_p and ASA

Tango score	H_p	%	ASA	%
Mesophiles				
Low (1–300)	Low (<12.3)	50.0 (56)	Low (<50)	14.3 (16)
	Medium (12.3–13.2)	28.6 (32)	Medium (50–58)	25.0 (28)
	High (>13.2)	21.4 (24)	High (>58)	60.7 (68)
Medium (301–800)	Low	36.5 (50)	Low	27.0 (37)
	Medium	31.4 (43)	Medium	35.8 (49)
	High	32.1 (44)	High	37.2 (51)
High (>800)	Low	9.7 (12)	Low	57.3 (71)
	Medium	40.3 (50)	Medium	33.1 (41)
	High	50.0 (62)	High	9.7 (12)
Thermophiles				
Low (1–300)	Low (<12.3)	56.2 (59)	Low (<50)	9.5 (10)
	Medium (12.3–13.2)	26.7 (28)	Medium (50–58)	33.3 (35)
	High (>13.2)	17.1 (18)	High (>58)	57.1 (60)
Medium (301–800)	Low	33.8 (47)	Low	33.1 (46)
	Medium	41.7 (58)	Medium	30.2 (42)
	High	24.5 (34)	High	36.7 (51)
High (>800)	Low	15.5 (20)	Low	50.4 (65)
	Medium	39.5 (51)	Medium	35.7 (46)
	High	45.0 (58)	High	14.0 (18)

The values in parentheses show the number of proteins.

ences among the APRs from thermophilic and mesophilic proteins.

Relationship between aggregation score difference and Hamming/Euclidian distance

We have computed the Hamming and Euclidian distance using the amino acid compositions of all the mesophilic and thermophilic proteins as described in "Materials and Methods" section. The computed distances have been compared with the aggregation score difference. The data are scattered among the thermophile–mesophile homologs, indicating that the amino acid composition differences among the homologous protein sequences do not correlate with differences in their aggregation tendencies. This result can be rationalized as follows: The protein sequences determine native structure and thermodynamic stabilities of the proteins. The overall thermodynamic stabilities of the proteins may not always be correlated with their aggregation proneness.

Aggregation score versus surrounding hydrophobicity and solvent accessibility in mesophilic and thermophilic proteins

We have analyzed the variation of aggregation score with surrounding hydrophobicity (H_p) and solvent accessibility [accessible surface area (ASA)]. Table I compares these at low, medium, and high aggregation scores. We observed that both mesophiles and thermophiles show similar tendency with respect to surrounding hydrophobicity and solvent accessibility. The proteins with low aggregation score prefer to have lower surrounding hydrophobicity and are more exposed to solvent. On the other hand, proteins with high aggregation score (>800) have high surrounding hydrophobicity and low solvent accessibility. Furthermore, we have evaluated the relationship between aggregation score and surrounding hydrophobicity/solvent accessibility (Fig. 2). For the thermophilic proteins, aggregation score showed a correlation of 0.38 with surrounding hydrophobicity [Fig. 2(a)] and −0.45 with solvent accessibility [Fig. (2b)]. The corresponding values for the mesophilic proteins are 0.37 and −0.47, respectively [Fig. 2(c,d)]. These values are significant, as the r^2 is higher than $2/(n)^{0.5}$, where n is the total number of proteins. Taken together, these results indicate that regions with high aggregation scores are likely to be buried in the protein cores and protected from solvent. Moreover, this aggregation-resistance strategy appears to be common between thermophilic and mesophilic proteins. To get more details, it is essential to analyze the APRs and their characteristic features (described later).

Amyloid-like fibril forming tetrapeptide and hexapeptide patterns and predicted APRs in thermophile—mesophile homologs

We have evaluated the existence of one tetrapeptide and two hexapeptide amyloid-like fibril forming patterns (acidic pH and neutral pH) proposed from experimental studies by Tjernberg *et al.*[48] and Lopez de la Paz and Serrano[33] in mesophilic and thermophilic proteins. These patterns were detected by using ScanProsite (http://prosite.expasy.org/scanprosite/) and pattern matching tools available at http://www.bioinformatics.org/sms2/protein_pattern.html. These patterns are listed in "Materials and Methods" section. The tetrapeptide pattern (Pattern 3) showed 18 hits in 17 mesophilic sequences; whereas, it showed 40 hits in 38 thermophilic

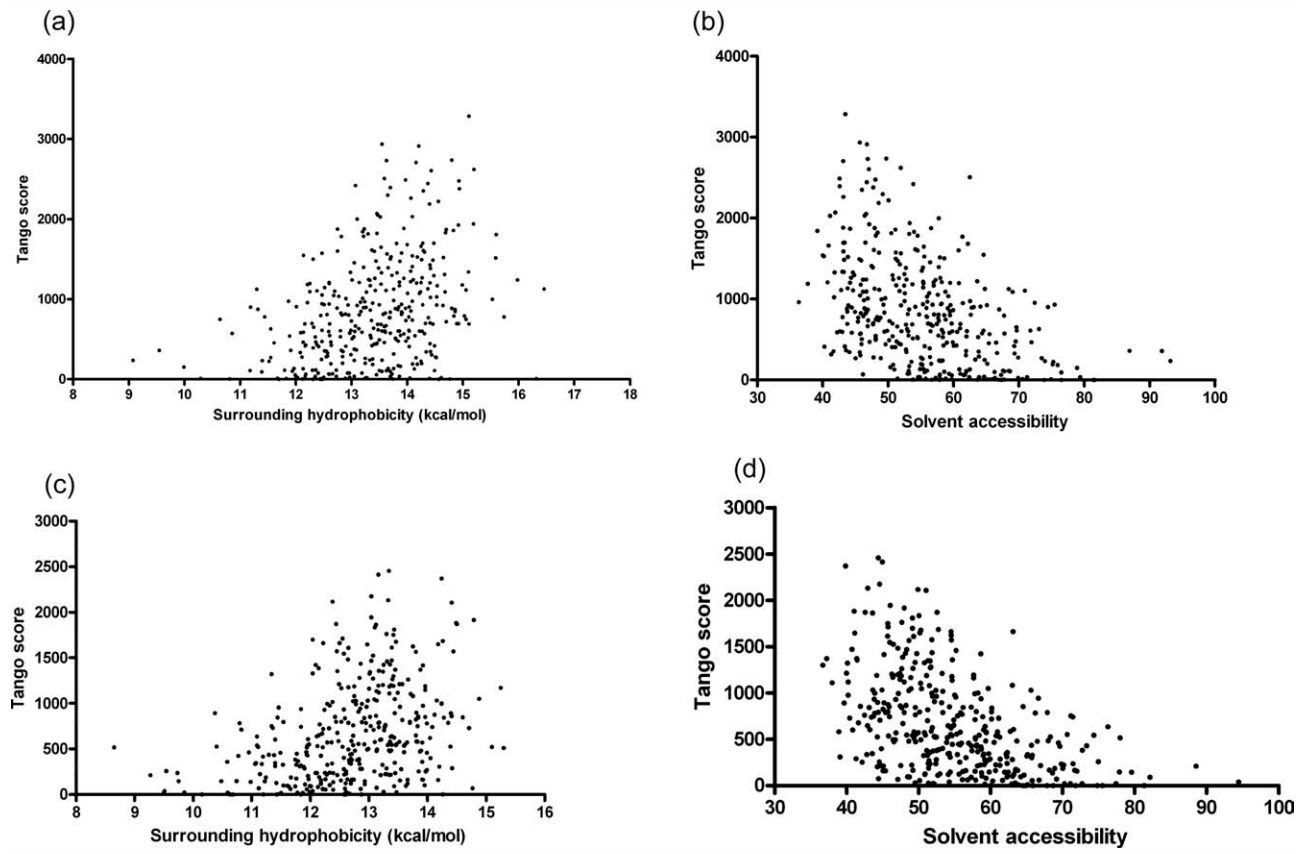

Figure 2

Relationship between surrounding hydrophobicity/solvent accessibility and Tango score in thermophilic and mesophilic proteins: (**a**) Surrounding hydrophobicity and Tango score in thermophilic proteins. (**b**) Solvent accessibility and Tango score in thermophilic proteins. (**c**) Surrounding hydrophobicity and Tango score in mesophilic proteins. (**d**) Solvent accessibility and Tango score in mesophilic proteins.

sequences. Similarly, acidic pH hexapeptide pattern (Pattern 1) showed 811 and 894 hits in mesophilic and thermophilic proteins, respectively. The neutral pH hexapeptide pattern (Pattern 2) is a subset of acidic pH pattern (Pattern 1). This pattern has 538 and 529 hits in the mesophilic and thermophilic proteins, respectively. We have also evaluated the overlap of these patterns with APRs detected by Tango/PAGE and Waltz programs. The results are shown in Table II. Pattern 2 showed the greatest number of overlaps with APRs. Furthermore, the number of matches between patterns and predicted APRs was greater for thermophilic proteins than mesophilic proteins. Interestingly, the incidence of these patterns is very similar between the thermophilic and mesophilic proteins.

We have also compared the existence of experimentally validated aggregating peptide sequences in mesophilic and thermophilic proteins by scanning their sequences against the library of 517 experimentally validated peptide sequences found in more than 110 proteins and numerous synthetic peptides. We found that mesophilic and thermophilic proteins have 19 and 22 aggregating peptide sequences, respectively. These proteins and the peptide sequences are listed in Table III. We noticed that the peptide sequence, GYE, found at positions 9–11 in amyloid β_{1-42} occurs in several thermophilic proteins and their mesophilic homologs. The tripeptide GYE from $A\beta_{1-42}$ is the smallest peptide known to form amyloid fibrils (Naskar *et al.*[47]). Tetrapeptide KVVE designed to

Table II

Number of Matches Between the APRs Obtained with Three Different Patterns and Predicted APRs in Mesophilic and Thermophilic Proteins Using Tango/PAGE and Waltz

Proteins	Pattern 1	Pattern 2	Pattern 3
Tango/PAGE			
Mesophilic	115	110	0
Thermophilic	142	133	2
Waltz			
Mesophilic	308	246	1
Thermophilic	348	271	1

Pattern 1: $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQE]_3$-$[ILTYWFNE]_4$-$[FIY]_5$-$\{PKRH\}_6$; Pattern 2: $\{P\}_1$-$\{PKRHW\}_2$-$[VLS(C)WFNQ]_3$-$[ILTYWFN]_4$-$[FIY]_5$-$\{PKRH\}_6$; Pattern 3: $[KE]_1$-$[FV]_2$-$[FV]_3$-$[EK]_4$.

Table III
Occurrence of Experimentally Validated Aggregating Peptides and Their Conservation in Mesophilic and Thermophilic Proteins

	Aggregating peptide			
PDB code 1	Name	Sequence	PDB code 2	Conservation
Mesophilic proteins (19)				
1u1ha	Designed peptide	KVVE	*1t7la*	K+−E
1g6ha	Designed peptide	KVVE	*1g9xa*	KVVE
1td9a	a-Synuclein	KTKEGV	*1qzta*	KT−+G−
1piea	Ab 9–11	GYE	*1s4ea*	GY+
1dv1c	Designed peptide	KVVE	*1ulzc*	K+VE
1uaae	Ab 9–11	GYE	*2pjre*	−−−
1lvlac	Ab 9–11	GYE	*1ebda*	G−E
1ou5a	Ab 9–11	GYE	*1miwa*	G−E
1yx2a	Ab 9–11	GYE	*1wosa*	G+E
1yx2c	Ab 9–11	GYE	*1wosc*	GYE
1on3a	Designed peptide	KVVE	*1vrga*	K+V+
1vbja	Ab 9–11	GYE	*1vp5a*	GYE
2gyia	Ab 9–11	GYE	*1bxba*	GY+
1pcqc	GroES (61–74)	VGDIVIFNDGYGVK	*1we3c*	−GDIV+F−−−YG−−
	GroES (58–67)	DVKVGDIVIF		+VK−GDIV+F
	GroES core	VGDIVIF		−GDIV+F
1d8la	Ab 9–11	GYE	*1ixra*	G+−
1oena	Ab 9–11	GYE	*1j3ba*	G−−
1ox6a	Ab 9–11	GYE	*1ka9a*	G+−
Thermophilic proteins (22)				
1hjza	Designed peptide	KVVE	1spva	−−−E
1gtda	Ab 9–11	GYE	1t4aa	−−E
1g9xa	Designed peptide	KVVE	1g6ha	KVVE
1e19a	Ab 9–11	GYE	1b7ba	G+−
1wlsa	Ab 9–11	GYE	1nnsa	G−E
1xtta	Ab 9–11	GYE	1bd3a	−−E
1ny5c	Ab 9–11	GYE	1ojla	G+E
1c3ra	Ab 9–11	GYE	1t64a	−Y+
1i5fa	Waltz training set	VSFEIV	1mjca	VSF−I−
3pvaa	Ab 9–11	GYE	2bjfa	−++
1b04a	Ab 9–11	GYE	1ta8a	−+E
1phpa	Designed peptide	KVVE	1hdia	−+−+
2bm3a	Ab 9–11	GYE	1qzna	−−−
1wosc	Ab 9–11	GYE	1yx2	GYE
1pvta	Ab 9–11	GYE	1gt7a	−−−
1vkna	Ab 9–11	GYE	1xyga	−−E
1cz3a	Ab 9–11	GYE	1qzfa	−−E
1o0wa	Ab 9–11	GYE	2a11a	G−E
1zh8a	Designed peptide	KVVE	1h6da	++++
1vp5a	Ab 9–11	GYE	1vbja	GYE
1ve1a	Designed peptide	KVVE	1y7la	−−−−
1j33a	Ab 9–11	GYE	1l50a	−−−

PDB codes 1 and 2 indicate the pairs of mesophilic and thermophilic proteins. The PDB codes of thermophilic proteins are shown in italics; +: similar: −: not conserved.

from amyloid-like fibrils by Tjernberg *et al.*[48] was detected in five thermophilic proteins. These proteins are 3-phosphoglycerate kinase from *Bacillus stearothermophilus* (PDB code: 1PHP.A), Histone AF1521 from *Archaeoglobus fulgidus* (1HJZ.A), CobT from *Thermus thermophilus* (1J33.A), ATP binding protein from *Methanocaldococcus jannaschii* (1G9X.A), and Oxidoreductase from *Thermotoga maritima* (1ZH8.A). An amyloidogenic hexapeptide, which was part of Waltz training set (VSFEIV),[30] was detected in one thermophilic protein (cold shock protein from *Bacillus caldolyticus*, 1I5F.A). Taken together, the above observations indicate that the thermophilic proteins may also aggregate and form amyloid-like fibrils in a manner similar to their mesophilic homologs, because sequence features that facilitate cross-β motif were observed in the thermophilic proteins as well. Ala variants of cytochrome c552 from *Hydogenobacter thermophilus* and the protein S6 from *T. thermophilus* have been shown to form amyloid-like fibrils.[68,69]

Conservation of APRs between thermophilic and mesophilic proteins

We have computed the conservation of the above experimentally validated aggregating peptide sequences, and the results are included in Table III. Of the six

Table IV
Multiple Sequence Alignment for Cold Shock Protein from *Bacillus caldolyticus* Showing the Conservation of the Hexapeptide VSFEIV[a]

Organism	Sequence
Bacillus caldolyticus	MQEGKVKWFNNEKGYGFIEVEGGSDVFVHFTAIQGEGFKT
Bacillus subtilis	MQNGKVKWFNNEKGFGFIEVEGGDDVFVHFTAIEGDGYKS
Bacillus cereus	MQTGKVKWFNSEKGFGFIEVEGGDDVFVHFSAIQGDGFKT
Bacillus coagulans	MEQGKVKWFNSEKGYGFIEREGGSDVFVHFSAIQGEGFKT
Listeria grayi	MQNGKVKWFNNEKGYGFIETENGEDIFAHFTAIQGDGYKT
Bacillus thuringiensis	MQNGKVKWFNSEKGFGFIEVEGGEDVFVHFSAIQGEGFKT
Paenibacillus curdlanolyticus	MQQGTVKWFNAEKGFGFIEVEGGSDVFVHFSAIQGDGFKT
Bacillus cellulosilyticus	M– –GKVKWFNSEKGFGFIEREGGDDVFVHFSAIQGEGFKT
Geobacillus sp.	MNKGKVKWFNAEKGYGFIEMDGGNDVFVHFTAIQGEGFKT
Lysinibacillus sphaericus	MQQGIVKWFNNEKGYGFIECDDGEDVFVHFTGIQEEGFRT
Oceanobacillus iheyensis	MNTGSVKWFNAEKGFGFIEVEGGDDVFVHFSAIQGEGFKT
Turicibacter sanguinis	MTTGTVKWFNAEKGFGFITVDGGNDVFAHFSAITGEGFKT
Listeria grayi	MQTGTVKWFNSEKGFGFIEVEGGDDVFVHFSAIQGDGFKT
Bacillus caldolyticus	LEEGQEVSFEIVQGNRGPQAANVVKL
Bacillus subtilis	LEEGQEVSFEIVEGNRGPQASNVVKL
Bacillus cereus	LEEGQEVSFEIVEGNRGPQAANVTKN
Bacillus coagulans	LEEGQSVSFDIEEGNRGPQAANVSKL
Listeria grayi	LEEGQSVSFEVVEGNRGPQAANIEKLS
Bacillus thuringiensis	LEEGQEVTFEVEQGNRGPQATNVNKK
Paenibacillus curdlanolyticus	LDEGQRVEFTIAQGNRGPQAENVVKL
Bacillus cellulosilyticus	LEEGQEVEFEIVDGDRGPQAANVVKL
Geobacillus sp.	LEEGQTVMFDIIDGNRGPQAANVQKA
Lysinibacillus sphaericus	LEEGQKVSFDVVEGNRGPQASNVVKL
Oceanobacillus iheyensis	LEEGQSVSFDIEEGNRGPQAANVVKN
Turicibacter sanguinis	LEEGQKVSFEIVEGNRGPQAANIVKL
Listeria grayi	LDEGQSVEFEIVEGQRGPQAENVVKL

The hexapeptide is shown in yellow background. The conservation of V and F are shown in boxes.
[a]Only partial sequences are shown.

residue in this hexapeptide (VSFEIV), only four residues (V, S, F, and I) are conserved in homologous mesophilic proteins. When we compared with the homologous sequences of *B. caldolyticus* cold shock protein (1I5F.A), two residues (V and F) are conserved, two of them are variable (S and E), and two residues (I and V) have moderate conservation score. The analysis was performed using Consurf server,[70] and it indicates that the hexapeptide amyloidogenic sequence, VSFEIV, has been disrupted in the homologs of *B. caldolyticus* cold shock protein (Table IV). The tetrapeptide sequence (KVVE), although found in five thermophilic proteins, is conserved in only one thermophilic–mesophilic protein pair (1G9X.A–1G6H.A). This tetrapeptide is not conserved in other three proteins and is partially conserved in 1HJZ.A. Further, only three peptides among 19 mesophilic and 22 thermophilic proteins are conserved (Table III). These results show that the experimentally proven amyloid-like fibril forming peptide sequences are not conserved among homologous thermophilic and mesophilic proteins.

We have analyzed the conservation of APRs, predicted by Waltz, in thermophilic proteins by comparing the alignments of them with their mesophilic counterparts. Of the 598 APRs identified in thermophilic proteins, only 55 (9.2%) APRs are conserved in their respective mesophilic homologs. In most cases, only one or two residue mutations caused the APR disruption. In mesophilic proteins, Waltz identified 515 APRs, and the alignment of mesophilic and their respective thermophilic proteins showed that only 56 (10.9%) APRs are conserved between them.

Taken together, the observation presented in this section indicates that APRs are not conserved among thermophilic–mesophilic pairs.

Amino acid composition in APRs of mesophilic and thermophilic proteins

We have analyzed the amino acid composition of identified APRs using Tango/PAGE and Waltz in mesophilic (Table S2 in the Supporting Information) and thermophilic (Table S3 in the Supporting Information) proteins. The comparison of amino acid composition of Tango/PAGE predicted APRs with that of the complete mesophilic/thermophilic protein sequences are shown in Figure 3(a,b). For the mesophilic proteins [Fig. 3(a)], the Tango/PAGE predicted APRs favor aliphatic residues Ala, Gly, Leu, and Val and disfavor charged/polar residues, Asp, Glu, Asn, Gln, and Tyr when compared with the overall sequence compositions of the mesophilic proteins. In case of the thermophilic proteins [Fig. 3(b)], the Tango/PAGE predicted APRs favor β-branched residues Leu and Val and disfavor all the polar residues when compared with the overall sequence compositions of the thermophilic proteins. Hence, the Tango/PAGE identified APRs in the thermophilic proteins tend to be more selective in the nonpolar residues. Similar calculations were repeated for the APRs predicted by Waltz program[30] (Tables S4 and S5 in the Supporting Information). The Waltz predicted APR sequence compositions favor Glu, Gly, Lys, Pro, and Arg when compared with the overall sequence compositions of mesophilic proteins. For Waltz predicted APRs in thermophilic proteins, Glu, Gly, Lys, Pro, and Arg are favored when compared with the overall sequence compositions of the thermophilic proteins. These observations are consistent with the differences in the Tango/PAGE and Waltz algorithms. Waltz tends to identify more hydrophilic/polar APRs, whereas Tango and PAGE identify more nonpolar APRs.[30] We have also presented the APR amino acid composition differences between thermophiles and mesophiles predicted by Tango/PAGE and Waltz [Fig. 3(c)]. Tango/PAGE and Waltz predicted APRs in the thermophilic proteins are richer in Glu, Ile, Lys, Arg, Val, and Tyr (−ve change in occurrence) and poorer in Ala, Cys, Gly, Met, Asn, Gln, and Thr (+ve change in occurrence) as compared to the APRs in mesophilic proteins. Asp, Leu, and Ser show ambiguous trends. Tango/ PAGE and Waltz follow different algorithms, and amino acid compositional differences in the APRs predicted by these algorithms are expected. Despite this, change in occurrence of a majority of amino acid residues in the APRs of mesophilic versus thermophilic proteins is consistent between Tango/PAGE and Waltz predictions.

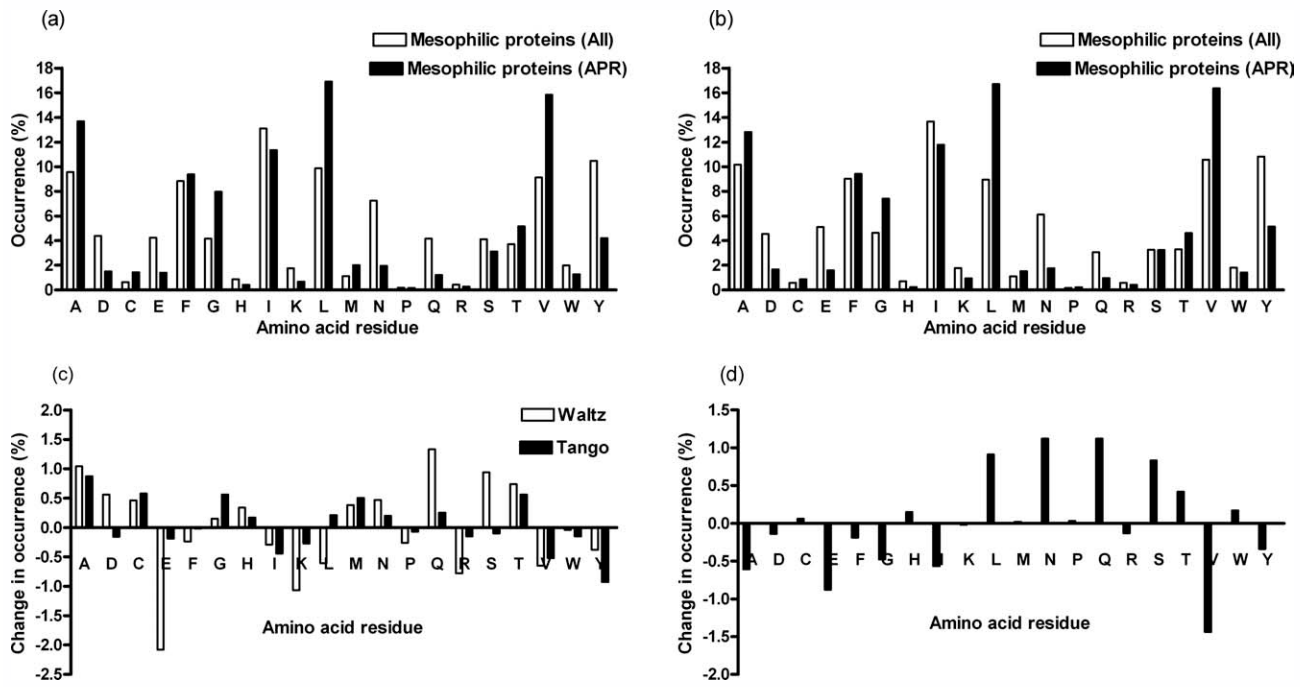

Figure 3

(**a**) Occurrence of amino acid residues in Tango/PAGE predicted APRs and all mesophilic proteins. (**b**) Occurrence of amino acid residues in Tango/PAGE predicted APRs and all thermophilic proteins. (**c**) Difference in amino acid occurrence in the APRs between mesophilic and thermophilic proteins. The data for the APRs obtained with both Tango/PAGE and Waltz are shown. (**d**) Difference in amino acid occurrence between mesophilic and thermophilic proteins.

Taken together, these observations indicate that APRs in thermophilic proteins utilize different amino acids than their mesophilic counterparts. The complete amino acid sequences of the thermophilic and mesophilic proteins also show amino acid compositional differences. In our dataset, the thermophilic proteins are richer in Ala, Asp, Glu, Phe, Gly, Ile, Val, and Tyr [−ve change in occurrence in Fig. 3(d)] and poorer in His, Leu, Asn, Gln, Ser, Thr, and Trp [+ve change in occurrence in Fig. 3(d)] when compared with their mesophilic homologs. The amino acid compositional differences in the overall sequences of thermophilic and mesophilic proteins do not completely match with the analogous differences in APRs.

We have also evaluated the predicted APR sequences by computing the amino acid composition to see, if they are dominantly hydrophobic like in amyloid β peptide or dominantly hydrophilic like in Q/N rich sequences found in prions. An APR is classified as dominantly hydrophobic, if it contains more than 50% hydrophobic residues (aliphatic and aromatic). If more than 50% residues in an APR are either polar or charged, then it is dominantly hydrophilic. If an APR contains equal proportions of hydrophobic and hydrophilic residues, then it was classified as amphipathic. We observed that among 515 APRs identified by Waltz in the mesophilic proteins, 21 (4.1%) APRs are dominantly hydrophilic, 403 (78.2%) APRs are dominantly hydrophobic, and 91 (17.7%) APRs are amphipathic. In the thermophilic proteins, there are 598 Waltz predicted APRs. Out of these, eight (1.3%) APRs are dominantly hydrophilic, 492 (82.3%) APRs are dominantly hydrophobic, and 98 (16.4%) APRs are amphipathic. Thus, most Waltz predicted APRs in both thermophilic and mesophilic proteins are dominantly hydrophobic. Tango/PAGE predictions also showed that most of the APRs are dominantly hydrophobic [1297/1311 (98.9%) in mesophilic proteins and 1382/1396 (99%) in thermophilic proteins]. Our results show that most of APRs in thermophilic proteins are dominantly hydrophobic like in their mesophilic homologs. Hence, even at elevated temperatures, exposure of hydrophobic surface on the perturbations in a thermophilic protein's native structure could potentially facilitate its aggregation.

Overall behavior of APRs based on surrounding hydrophobicity, solvent accessibility, and interaction energy

We have computed the parameters, surrounding hydrophobicity (H_p), solvent accessibility (ASA), and interaction energy of all APRs as well as for all the residues in mesophilic and thermophilic proteins as described in Materials and Methods section. The average values are presented in Table V. We observed that the surrounding hydrophobicity is higher for APRs than all other non-APR parts in amino

Table V
Average Values for Structure Based Parameters in APRs and All Residues

Parameter	Meso (APR)	Meso (all)	Thermo (APR)	Thermo (all)
H_p	15.905 (5.15)	12.72 (5.21)	16.670 (5.35)	13.40 (5.28)
ASA	37.24 (41.7)	51.67 (48.6)	35.74 (40.9)	52.03 (49.6)
Interaction energy/residue	−19.14 (5.58)	−16.23 (5.45)	−19.31 (5.65)	−16.48 (5.49)

The deviations are given in parentheses.

acid sequences of both mesophilic and thermophilic proteins. When complete amino acid sequences are compared between the mesophilic and thermophilic proteins, the thermophilic ones have higher H_p. Similarly, the APRs have stronger interaction energy than other non-APR parts for both thermophilic and mesophilic proteins. Furthermore, APRs are more buried (lower ASA) than non-APR parts for both thermophiles and mesophiles.

However, there are subtle differences between thermophilic and mesophilic proteins. We have analyzed the frequencies of APRs in different intervals of surrounding hydrophobicity values in mesophilic and thermophilic proteins. Histograms of APR frequencies versus H_p values in mesophilic and thermophilic proteins are compared in Figure 4. For $H_p \leq 16$ kcal/mol, the APRs from mesophilic proteins are more frequent in each interval. For $H_p \geq 18$ kcal/mol, APRs from thermophilic proteins are more frequent. These trends indicate that thermophilic proteins are better at surrounding their APRs with other hydrophobic regions and making it harder for them to become solvent exposed on native structure perturbations.

Compositions of residues at gatekeeper positions

Gatekeeper residues (Asp, Glu, Lys, Pro, and Arg) can significantly affect the ability of an APR to initiate aggregation. Thus, punctuating the APR sequences with the gatekeeper residues is a strategy adapted by proteins to resist aggregation.[71] Three preceding and three succeeding positions have the greatest influence on an APR's potential to initiate aggregation.[72] The amino acid compositions at these gatekeeper positions showed specific trends for mesophilic and thermophilic proteins as seen in Table VI. The gatekeeper positions are dominated by Asp, Glu, Gly, and Lys in both mesophilic and thermophilic proteins. Frequency of each residue at these positions is >7%. Thermophilic proteins show greater incidence of Glu, Pro, and Arg at the gatekeeper positions. The frequencies of these residues at the gatekeeper positions increase by >1% in the thermophilic proteins as compared to their mesophilic homologs. These results are consistent in both Tango/PAGE and Waltz predictions (Table VI). We summed the incidence of the gatekeeper residues Asp, Glu, Lys, Arg, and Pro[71,72] at the gatekeeper positions surrounding the Tango/PAGE and Waltz predicted APRs in thermophilic and mesophilic proteins (Table VI). The incidence of the gatekeeper residues is greater for the APRs in the thermophilic proteins. These observations indicate that thermophilic proteins have stronger gatekeeper control on the APRs than their mesophilic homologs.

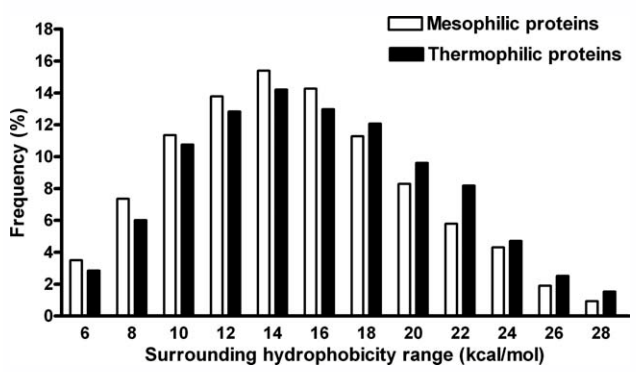

Figure 4
Histogram plot showing comparison of surrounding hydrophobicity (H_p) between thermophilic and mesophilic proteins.

Table VI
Amino Acid Composition of Gatekeeper Residues Predicted by Tango/PAGE and Waltz in Mesophilic and Thermophilic Proteins

	Tango			Waltz		
Residue	Meso	Thermo	Diff	Meso	Thermo	Diff
Ala	7.53	6.51	1.01	8.58	8.05	0.53
Asp	9.90	9.25	0.65	7.53	6.92	0.61
Cys	1.28	0.63	0.65	1.55	0.84	0.70
Glu	9.53	11.17	−1.64	7.33	9.20	−1.87
Phe	1.45	1.46	−0.01	2.53	2.50	0.03
Gly	9.63	9.89	−0.26	10.72	9.20	1.52
His	3.15	2.95	0.21	2.66	2.56	0.10
Ile	2.71	3.08	−0.37	4.11	4.73	−0.62
Lys	8.85	9.23	−0.38	7.04	8.53	−1.49
Leu	3.96	4.54	−0.58	6.74	6.81	−0.07
Met	1.60	1.49	0.11	1.87	1.63	0.24
Asn	5.66	5.26	0.40	4.87	4.45	0.42
Pro	5.88	6.98	−1.10	5.29	6.16	−0.87
Gln	4.35	2.92	1.43	4.18	2.48	1.70
Arg	6.68	7.72	−1.04	4.93	5.32	−0.39
Ser	6.62	5.60	1.02	6.12	5.54	0.57
Thr	5.02	4.76	0.26	5.59	5.77	−0.18
Val	3.48	3.73	−0.25	5.49	5.91	−0.42
Trp	0.55	0.82	−0.27	0.72	0.73	−0.01
Tyr	2.18	2.02	0.16	2.14	2.65	−0.51
Gate-keeping residues	40.84	44.35	−3.51	32.12	36.13	−4.01

Gate-keeping residues: Asp, Glu, Lys, Arg, and Pro. Difference (Diff) = Mesophilic (Meso) − Thermophilic (Thermo).

DISCUSSION AND CONCLUSIONS

Aggregation is an ancient common threat to productive protein folding in all life forms. All organisms from

different ecological niches must be able to counter protein aggregation, so that they can maintain metabolic flux and viabilities of their cells. On the practical side, comparison of aggregation between extremophiles and mesophiles could suggest design and formulation strategies to mitigate aggregation in biotherapeutics. This is because several conditions encountered during manufacture, purification, packaging, storage, and shipping of these drug products are similar to the environmental conditions faced by the proteins from extremophiles.[73] Hence, the degradation issues that occur during the development of these drug products are similar to the challenges faced by the extremophilic proteins in the ambient environments of the source organisms.[73] The formulation scientists and extremophilic organisms, both ultimately aspire toward the same goal. Maintenance of physicochemical integrity of the macromolecular structure–function is required in the face of environmental perturbations in both cases. Among all the degradation routes for biotherapeutic drug products, aggregation is the most common yet the least understood one.

The goal of this investigation was to find out how thermophilic proteins resist aggregation. Do thermophilic proteins also contain sequence regions prone to cross-β motif and amyloid-fibril formation? We found that the thermophilic proteins do contain sequence regions capable of forming cross-β motif like their mesophilic homologs. In this study, the evidence for cross-β-type aggregation for thermophilic proteins comes from detections of APRs by aggregation prediction algorithms, Tango/PAGE and Waltz and by the matching tetrapeptide and hexapeptide amyloidogenic patterns. Moreover, experimentally proven aggregating peptide sequences could be found in the thermophilic proteins. Experimental studies on Ala variants of cytochrome c552[68] and S6[69] from thermophilic organisms have shown that the thermophilic proteins can also aggregate in the same way as their mesophilic counterparts. To our knowledge, this report is a first systematic analysis of potential APRs in thermophilic proteins. Hopefully, this study will spur greater experimental interest in studying aggregation of thermophilic proteins.

Are the APRs in thermophilic proteins different than those in their mesophilic homologs? Overall, APRs from thermophilic proteins show similar sequence features as those from the mesophilic proteins. However, some differences were observed [Fig. 3(c)], which could not be explained on the basis of amino acid composition differences among complete sequences of thermophilic and mesophilic proteins [Fig. 3(d)].

To initiate or propagate aggregation, an APR must be solvent exposed or become solvent exposed on perturbation of protein native structure. On solvent exposure, the APR must trigger protein self-association by seeking out its counterparts from other protein molecules in the vicinity and become solvent unexposed again. The protein sequences and structures must counter these two features of the APRs to resist aggregation and maintain the native state. Toward this goal, APRs are often found in the hydrophobic cores and/or in protein: protein interfaces. Consistently, we found that Tango aggregation scores are positively correlated with surrounding hydrophobicity (H_p) and negatively correlated with ASA. These correlations are slightly stronger for the thermophilic proteins. Thus, thermophilic proteins are able to stow away their APRs in more hydrophobic regions than the mesophilic proteins. This could help them better resist the solvent exposure of their APRs. Because our dataset consisted of homologous thermophilic and mesophilic protein single domains, we could not compare the incidence of APRs at the domain–domain (or protein–protein) interfaces in thermophiles and mesophiles. However, similar trends are expected.

APRs in protein sequences are often punctuated by the gatekeeper residues (Asp, Glu, Lys, Arg, and Pro), which either keep these regions solvated or disrupt β strand conformation in these regions.[70,71] The thermophilic proteins are also better at gatekeeping their APRs. The incidence of gatekeeper residues, especially Glu, Arg, and Pro at the gatekeeper positions is greater in the thermophilic proteins than their mesophilic homologs.

The observations in previous two paragraphs indicate that aggregation-resistance strategies used by the thermophilic proteins are also similar to the ones used by the mesophilic proteins. However, thermophilic proteins appear to be "more skilled" at utilizing the aggregation-resistance strategies than their mesophilic homologs.

A third strategy could be used to disrupt the APRs during evolution. More than 90% of the APRs in the thermophilic proteins are not conserved in their mesophilic homologs. In the specific example of *B. caldolyticus* cold shock protein homologs (Table IV), the several sequence positions in and around the experimentally proven amyloid-fibril forming hexapeptide (VSFEIV) are highly variable.

If thermophiles are more ancient than the mesophiles, can we trace some of the well-known aggregating peptide sequences to their thermophilic ancestors? By scanning a library of experimentally proven aggregating peptide sequences against the sequences of the thermophilic proteins, GYE, the smallest peptide shown to form amyloid-like fibrils[47] and found in Aβ peptide sequence, was detected in several thermophilic proteins from *Pyrococcus furiosus, Pyrococcus horikoshii, Methanobacterium thermoautotrophicum, Sulfolobus solfataricus, Aquifex aeolicus, Bacillus sphaericus, B. stearothermophilus, Clostridium thermocellum, T. maritima*, and *T. thermophilus*. In this study, we are limited by the available peptide sequences proven to form amyloid-like fibrils. As the library of such sequences grows, a greater number of such amyloidogenic sequences shall be discovered in their ancestral proteins.

Do the thermophilic proteins use any aggregation-resistance strategies that may be unique to their sequences and structures? We did not find any evidence for this. Analogous observations were made when stabilities of homologous thermophilic and mesophilic were compared.[4,14–23,74] Even the proteins from hyperthermophiles use the same set of stabilization strategies as the mesophilic ones. It has been suggested that thermophilic, especially hyperthermophilic, organisms evolved earlier than the mesophilic organisms.[5] It appears that the evolution of mesophilic proteins from their thermophilic ancestors was a smooth process that involved relatively small "fine-tuning" of the protein sequences according to the environmental needs. We are not aware of scientific evidence that shows large-scale remaking of the protein sequences and structures between thermophiles and mesophiles. Taking all these observations together, it appears that factors external to protein sequence–structural features, such as osmolytes, could be important contributors to the extra stabilization of the proteins and other cellular components in the thermophilic cells.[41–43,49]

ACKNOWLEDGMENTS

A.M.T. and D.V. thank the Department of Biotechnology, Government of India for the DBT-BIF facilities. S.K. acknowledges Drs. Ron Peeples, Sandeep Nema, and Kevin King for their support. Pfizer Research Informatics is acknowledged for computational facilities. Drs. Satish Singh, Patrick Buck, and Neeti Sinha are thanked for discussions on this topic. S.K. acknowledges Dr. Mark A. Mitchell for critical reading of this manuscript. M.M.G. thanks IIT Madras for computational facilities. The three anonymous referees are acknowledged for their useful suggestions and enthusiasm for this work.

REFERENCES

1. Huang SL, Wu LC, Liang HK, Pan KT, Horng JT, Ko MT. PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics 2004;20:276–278.
2. Gromiha MM. Protein bioinformatics: from sequence to function. New Delhi, India: Elsevier Publishers/Academic Press; 2010.
3. Kumar S, Nussinov R. Experiment-guided thermodynamic simulations on reversible two state proteins: implications for protein thermostability. Biophys Chem 2004;111:235–246.
4. Gromiha MM, Oobatake M, Sarai A. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 1999;82:51–67.
5. Gaucher EA, Govindarajan S, Ganesh OK. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 2008;451:704–707.
6. Ladenstein R, Antranikian G. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. Adv Biochem Eng Biotechnol 1998;61:37–85.
7. Jaenicke R, Bohm G. The stability of proteins in extreme environments. Curr Opin Struct Biol 1998;8:738–648.
8. Kumar S, Nussinov R. How do thermophilic proteins deal with heat? Cell Mol Life Sci 2001;58:1216–1233.
9. Yano JK, Poulos TL. New understandings of thermostable and peizostable enzymes. Curr Opin Biotechnol 2003;14:360–365.
10. Feller G. Protein stability and enzyme activity at extreme biological temperatures. J Phys Condens Matter 2010;22:323101.
11. O'Fágáin C. Engineering protein stability. Methods Mol Biol 2011;681:103–136.
12. Hasegawa J, Uchiyama S, Tanimoto Y, Mizutani M, Kobayashi Y, Sambongi Y, Igarashi Y. Selected mutations in a mesophilic cytochrome c confer the stability of a thermophilic counterpart. J Biol Chem 2000;275:37824–37828.
13. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000;13:179–191.
14. Kumar S, Tsai CJ, Nussinov R. Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 2001;40:14152–14165.
15. Gromiha MM. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. Biophys Chem 2001;91:71–77.
16. Gromiha MM, Thomas S, Santhosh C. Role of cation-pi interactions to the stability of thermophilic proteins. Prep Biochem Biotechnol 2002;32:355–362.
17. Ibrahim BS, Pattabhi V. Role of weak interactions in thermal stability of proteins. Biochem Biophys Res Commun 2004;325:1082–1089.
18. Xiao L, Honig B. Electrostatic contributions to the stability of hyperthermophilic proteins. J Mol Biol 1999;289:1435–1444.
19. Dominy BN, Minoux H, Brooks CL III., An electrostatic basis for the stability of thermophilic proteins. Proteins 2004;57:128–141.
20. Saraboji K, Gromiha MM, Ponnuswamy MN. Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins. Int J Biol Macromol 2005;35:211–220.
21. Zhang S, Zhang K, Chen X, Chu X, Sun F, Dong Z. Five mutations in N-terminus confer thermostability on mesophilic xylanase. Biochem Biophys Res Commun 2010;395:200–206.
22. Vijayabaskar MS, Vishveshwara S. Comparative analysis of thermophilic and mesophilic proteins using Protein Energy Networks. BMC Bioinform 2010;11(Suppl. 1):S49.
23. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics 2007;23:2231–2238.
24. Baldasseroni F, Pascarella S. Subunit interfaces of oligomeric hyperthermophilic enzymes display enhanced compactness. Int J Biol Macromol 2009;44:353–360.
25. Sawle L, Ghosh K. How do thermophilic proteins and proteomes withstand high temperature? Biophys J 2011;101:217–227.
26. Chan CH, Yu TH, Wong KB. Stabilizing salt-bridge enhances protein thermostability by reducing the heat capacity change of unfolding. PLoS One 2011;6:e21624.
27. Bruinsma IB, Bruggink KA, Kinast K, Versleijen AA, Segers-Nolten IM, Subramaniam V, Bea Kuiperij H, Boelens W, de Waal RM, Verbeek MM. Inhibition of α-synuclein aggregation by small heat shock proteins. Proteins 2011;79:2956–2967.
28. Pemberton S, Madiona K, Pieri L, Kabani M, Bousset L, Melki R. Hsc70 interaction with soluble and fibrillar α-Synuclein. J Biol Chem 2011;286:34690–34699.
29. Schulz EC, Ficner R. Knitting and snipping: chaperones in β-helix folding. Curr Opin Struct Biol 2011;21:232–239.
30. Maurer-Stroh S, Debulpaep M, Kuemmerer N, Lopez de la Paz M, Martins IC, Reumers J, Morris KL, Copland A, Serpell L, Serrano L, Schymkowitz JW, Rousseau F. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 2010;7:237–242.
31. Kumar S, Wang X, Singh SK. Identification and impact of aggregation prone regions in proteins and therapeutic mAbs. In: Wang W, Roberts CJ, editors. Aggregation of therapeutic proteins. Hoboken, NJ: Wiley; 2010. pp 103–118.
32. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL. Aggregation in protein-based biotherapeutics: computational studies and

tools to identify aggregation prone regions. J Pharm Sci 2011;100:5081–5095.
33. Lopez de la Paz M, Serrano L. Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci USA 2004;101:87–92.
34. Chiti F, Taddei N, Baroni F, Capanni C, Stefani M, Ramponi G, Dobson CM. Kinetic partitioning of protein folding and aggregation. Nat Struct Mol Biol 2002;9:137–143.
35. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D. Structure of the cross-β spine of amyloid-like fibrils. Nature 2005;435:773–778.
36. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M. Prediction of aggregation-prone regions in structured proteins. J Mol Biol 2008;380:425–436
37. Zhang Z, Chen H, Lai L. Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 2007;23:2218–2225.
38. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci USA 2006;103:4074–4078.
39. Chennamsetty N, Helk B, Voynov V, Kayser V, Trout BL. Aggregation-prone motifs in human immunoglobulin G. J Mol Biol 2009;391:404–413.
40. Ma B, Nussinov R. Simulations as analytical tools to understand protein aggregation and predict amyloid conformation. Curr Opin Chem Biol 2006;10:445–452.
41. Rösgen J. Molecular basis of osmolyte effects on protein and metabolites. Methods Enzymol 2007;428:459–486.
42. Rösgen J, Pettitt BM, Bolen DW. An analysis of the molecular origin of osmolyte—dependent protein stability. Protein Sci 2007;16:733–743.
43. Auton M, Bolen DW, Rösgen J. Structural thermodynamics of protein preferential solvation: osmolyte solvation of proteins, amino acids, and peptides. Proteins 2008;73:802–813.
44. Buck PM, Kumar S, Wang X, Agrawal NJ, Trout BL, Singh SK. Computational methods to predict of aggregation in therapeutic proteins. In: Voynov V, Caravella J, editors. Therapeutic proteins: methods & protocols, 2nd ed. Methods in molecular biology. USA: Humana Press, in press.
45. Colacino S, Tiana G, Broglia RA, Colombo G. The determinants of stability in the human prion protein: insights into folding and misfolding from the analysis of the change in the stabilization energy distribution in different conditions. Proteins 2006;62:698–707.
46. Brummitt RK, Nesta DP, Chang L, Chase SF, Laue TM, Roberts CJ. Nonnative aggregation of an IgG1 antibody in acidic conditions: part 1. Unfolding, colloidal interactions, and formation of high-molecular-weight aggregates. J Pharm Sci 2011;100:2087–2103
47. Naskar J, Drew MG, Deb I, Das S, Banerjee A. Water-soluble tripeptide Abeta (9–11) forms amyloid-like fibrils and exhibits neurotoxicity. Org Lett 2008;10:2625–2628.
48. Tjernberg L, Hosia W, Bark N, Thyberg J, Johansson J. Charge attraction and beta propensity are necessary for amyloid fibril formation from tetrapeptides. J Biol Chem 2002;277:43243–43246.
49. Lentzen G, Schwarz T. Extremolytes: natural compounds from extremophiles for versatile applications. Appl Microbiol Biotechnol 2006;72:623–634.
50. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 2004;22:1302–1306.
51. Tartaglia GG, Cavalli A, Pellarin R, Caflisch A. Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Sci 2005;14:2723–2734.
52. Wang X, Singh SK, Kumar S. Potential aggregation-prone regions in complementarity-determining regions of antibodies and their contribution towards antigen recognition: a computational analysis. Pharm Res 2010;27:1512–1529.
53. Kumar S, Bansal M. Dissecting alpha-helices: position-specific analysis of alpha-helices in globular proteins. Proteins 1998;31:460–476.
54. Nozaki Y, Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J Biol Chem 1971;246:2211–2217.
55. Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol 1975;50:167–183.
56. Manavalan P, Ponnuswamy PK. Hydrophobic character of amino acid residues in globular proteins. Nature 1978;275:673–674.
57. Manavalan P, Ponnuswamy PK. A study of the preferred environment of amino acid residues in globular proteins. Arch Biochem Biophys 1977;184:476–487.
58. Jiang Z, Zhang L, Chen J, Xia A, Zhao D. Effect of amino acid on forming residue–residue contacts in proteins. Polymer 2002;43:6037–6047.
59. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004;86:235–277.
60. Debe DA, Goddard WA. First principles prediction of protein folding rates. J Mol Biol 1999;294:619–625.
61. Gromiha MM, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rates of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol 2001;310:27–32.
62. Gromiha MM. Prediction of protein stability upon point mutations. Biochem Soc Trans 2007;35:1569–1573.
63. Gromiha MM. Influence of long-range contacts and surrounding residues on the transition state structures of proteins. Anal Biochem 2011;408:32–36.
64. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
65. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117: 5179–5197.
66. Pichierri F, Aida M, Gromiha MM, Sarai A. Free-energy maps of base-amino acid interactions for DNA–protein recognition. J Am Chem Soc 1999;121:6152–6157.
67. Gromiha MM, Yokota K, Fukui K. Energy based approach for understanding the recognition mechanism in protein–protein complexes. Mol Biosyst 2009;12:1779–1786.
68. Pertinhez TA, Bouchard M, Tomlinson EJ, Wain R, Ferguson SJ, Dobson CM, Smith LJ. Amyloid fibril formation by a helical cytochrome. FEBS Lett 2001;495:184–186.
69. Pedersen JS, Christensen G, Otzen DE. Modulation of S6 fibrillation by unfolding rates and gatekeeper residues. J Mol Biol 2004;341:575–588.
70. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–164.
71. Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F. Protein sequences encode safeguards against aggregation. Hum Mutat 2009;30:431–437.
72. Rousseau F, Schymkowitz J, Serrano L. Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol. 2006;16:118–126.
73. Kumar S, Singh SK, Gromiha MM. Temperature dependent molecular adaptations, microbial proteins. In: Flickinger MC, Editor. Encyclopedia of Industrial Biotechnology, Bioprocess, Bioseparation, and Cell Technology, New York: John Wiley & Sons, Inc.; Vol. 7; 2010. pp 4647–4761.
74. 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.