Elucidation of Factors Responsible for Enhanced Thermal Stability of Proteins: A Structural Genomics Based Study†

Suvobrata Chakravarty‡ and Raghavan Varadarajan*,‡,§

Molecular Biophysics Unit, Indian Institute of Science, Bangalore-560012, India and Chemical Biology Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Jakkur, Bangalore-560012, India

Received January 8, 2002; Revised Manuscript Received April 25, 2002

ABSTRACT: Understanding the molecular basis for the enhanced stability of proteins from thermophiles has been hindered by a lack of structural data for homologous pairs of proteins from thermophiles and mesophiles. To overcome this difficulty, complete genome sequences from 9 thermophilic and 21 mesophilic bacterial genomes were aligned with protein sequences with known structures from the protein data bank. Sequences with high homology to proteins with known structures were chosen for further analysis. High quality models of these chosen sequences were obtained using homology modeling. The current study is based on a data set of models of 900 mesophilic and 300 thermophilic protein single chains and also includes 178 templates of known structure. Structural comparisons of models of homologous proteins allowed several factors responsible for enhanced thermostability to be identified. Several statistically signficant, specific amino acid substitutions that occur going from mesophiles to thermophiles are identified. Most of these are at solvent-exposed sites. Salt bridges occur significantly more often in thermophiles. The additional salt bridges in thermophiles are almost exclusively in solvent-exposed regions, and 35% are in the same element of secondary structure. Helices in thermophiles are stabilized by intrahelical salt bridges and by an increase in negative charge at the N-terminus. There is an approximate decrease of 1% in the overall loop content and a corresponding increase in helical content in thermophiles. Previously overlooked cation$-\pi$ interactions, estimated to be twice as strong as ion-pairs, are significantly enriched in thermophiles. At buried sites, statistically significant hydrophobic amino acid substitutions are typically consistent with decreased side chain conformational entropy.

Hyperthermophilic organisms grow optimally at temperatures between 80 and 110 °C (*1*−*3*). Proteins of these organisms have evolved to function optimally at temperatures above 70 °C. Some enzymes from these organisms are active at temperatures as high as 110 °C and above (*4*, *5*). Thermophilic organisms grow optimally between 50 and 80 °C with their proteins showing optimal activity above 60 °C (*3*). Despite intense efforts, a detailed understanding of the factors responsible for enhanced thermal stability of protein from thermophiles/hyperthermophiles remains elusive. Thermophilic proteins and their mesophilic homologues typically share 40 to 85% sequence similarity, their three-dimensional structures are superimposable, and they have the same catalytic mechanism (*6*−*8*). In spite this close similarity, thermophilic and hyperthermophilic proteins are intrinsically more stable compared to their mesophilic homologues. In fact, the majority of thermophilic proteins studied to date, when expressed in *Escherichia coli*, retain all of the native protein's biochemical properties, proper folding, themostability, and optimal activity at high temperatures (*6*, *9*, *10*). Thus, most thermophilic proteins are not only intrinsically more stable but can also fold properly even at temperatures as much as 60 °C below their physiological temperature. However, there are reports of a few thermophilic proteins requiring extrinsic factors (e.g., salts or polyamines) or posttranslational modifications such as a glycosylation for thermostability (*3*, *11*). There is obvious biotechnological interest in engineering proteins with enhanced thermal stability and altered activity (*12*−*16*). This requires a comprehensive understanding of the factors responsible for enhancing thermal stability (*17*, *18*). Although understanding of the molecular mechanisms of thermal adaptation of proteins have been the focus of many studies for several decades, it has so far been difficult to pin down any single factor as being primarily responsible for enhancing thermal stability (*19*−*26*). This is probably because protein stability is determined by a multitude of both local and long-range interactions, and there is a fine balance between several contributing factors (*27*, *28*). It has been shown that pronounced thermal stability was achieved when several changes each with a relatively small effect were combined (*29*−*32*). This emphasizes the fact that in many cases the thermostabilization effects of individual changes are independent and nearly additive.

Studies of thermal stability carried out over the past several years can be broadly divided in to three categories: (i) those involving comparison of atomic structure of a single ther-

† This work was supported by grants from Department of Biotechnology and Department of Science and Technology, Government of India, to R.V.

* To whom correspondence should be addressed. Phone: 91-80-3092612. Fax: 91-80-3600683 or 91-80-3600535. E-mail: varadar@mbu.iisc.ernet.in.

‡ Indian Institute of Science.

§ Jawaharlal Nehru Centre for Advanced Scientific Research.

10.1021/bi025523t CCC: $22.00
Published on Web 05/25/2002

mophilic protein with one or more mesophilic homologues (*33*−*37*); (ii) systematic study involving analyses based on sequence and structural information for a group of proteins to reach general conclusions (*20*, *26*, *38*−*41*); and (iii) large scale comparison between thermophilic and mesophilic genomic sequences (*42*−*46*). This is now possible due to recent progress in genome sequence projects.

The main focus of the systematic studies has been to correlate specific sequence and structural features with enhanced thermal stability (*24*, *37*, *38*, *47*−*50*). These features have included amino acid composition, helix stabilization, hydrophobic and electrostatic optimization, increased hydrogen bonding, and enhanced secondary structure content. A recent systematic study on 25 protein families consisting of 64 mesophilic and 19 thermophilic proteins (the largest so far) concluded that different protein families adapt to higher temperatures utilizing different sets of structural devices (*26*). The only generally observed trend of an increase in the number of ion pairs with increasing growth temperature was also confirmed in the above study. However, the number of hydrogen bonds and polarity of buried surface did not show any clear-cut correlation with growth temperature.

The difficulty in reaching a general conclusion about the role of specific structural features for enhancing thermal stability is largely due to the limited amount of sequence and structural data available for proteins from hyperthermophiles and thermophiles. Recent progress in genome projects has made it possible to overcome this difficulty by performing statistical analyses on protein sequences from entire genomes to bring out differences between proteins from thermophiles and mesophiles (*43*−*46*). These studies have focused on statistics derived from sequence alone or from predicted secondary structures (*43*, *44*) and do not make any reference about tertiary structural features that contribute to thermal stability. In the present study, we present a systematic study on a large set of modeled proteins of mesophilic and thermophilic origin, which helps to overcome this lacuna. Complete genome sequences from 9 thermophilic and 21 mesophilic bacterial genomes were aligned with protein sequences with known structures from the RCSB[1] (*51*). Sequences with high homology to proteins with known structures were chosen for further analysis. High quality models of these chosen sequences were obtained from MODBASE (*52*). The current study is based on a data set of models of 900 mesophilic and 300 thermophilic protein single chains. This data set has been classified into 125 groups such that members within a group share more than 40% homology to each other and that every group contains at least one thermophilic, one mesophilic, and one protein of known structure from the RCSB. Systematic statistical analysis of various potential stabilizing factors in these 125 groups was carried out to obtain insight into structural features responsible for increased thermal stability.

[1] Abbreviations: RCSB, research collaboratory of structural bioinformatics; ORF, open reading frame; BLAST, basic local alignment search tool; ASA, accessible surface area; DSSP, dictionary of secondary structure of proteins; SCOP, structural classification of protein; PDB, protein data bank; TIM, triose phosphate isomerase; NADP, nicotinamide adenine dinucleotide phosphate.

MATERIALS AND METHODS

Construction of the Data Set. The translated open reading frames (ORFs) of the 30 completed microbial (only bacterial and archaeal) genomes were obtained from the web-page http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html. The list of PDB templates was taken from the 1999 version of the PDBselect set of 95% homologous proteins (http://www.sander.embl-heidelberg.de/pdbsel) (*53*). The coordinates of all the templates were obtained from RCSB (*51*). The sequences of the templates were taken from the "SEQRES" information provided in the header of each coordinate file. All the pairwise alignments between ORFs and templates were performed using the gapped version of the BLAST sequence alignment algorithm (*54*). The multiple alignments used in this analysis were carried out using CLUSTALW (*55*). All the 3D models used in this analysis were obtained from MODBASE (http://guitar.rockefeller.edu/modbase) (*52*). MODBASE is a database of annotated comparative protein structure models generated by the program MODELLER (*56*, *57*). The models consist of coordinates for all non-hydrogen atoms in the modeled part of the protein. Models are generated entirely by an automated procedure (*52*, *58*). MODELLER, a homology modeling program, uses pairwise sequence−structure alignment to build 3D models for the matched part of the target sequence based on the template structure (*56*, *57*).

Secondary Structure. The secondary structure definition of residues in both modeled and template PDB structures were obtained using DSSP (*59*). We merged different DSSP secondary structure elements into three categories: helix, strand, and other. Residues having a letter H or G in the DSSP output were considered to be in helices, those having E or B were considered to be in strands, and the remaining residues were considered to be in irregular regions (*26*).

Residue Depth. Residue depth (*60*) was used to distinguish buried residues from solvent exposed ones. Residues with depth greater than 5.5 Å were considered buried. A pair of residues, such as those involved in a salt bridge, was considered buried if the average depth of the pair was greater than 5.5 Å.

Salt Bridge/Ion Pair. Salt bridge/ion pairs were defined using a simple distance based criterion: two oppositely charged residues were considered to form a salt bridge if both the distance between the centroids of the two side chains as well as the shortest distance between the oppositely charged atoms in the salt bridge were less than 6 Å. Arg, Lys, His, Asp, and Glu residues were considered for salt bridge calculations.

Residue Contacts. We define residue contacts as the number of neighoring residues in contact with a particular residue within 6.5 Å. The distance in this case is measured between the side chain centroids of residues, and for Gly it is calculated from the centroid of the peptide unit. This parameter is used to compare the residue pair proportions in the two populations (thermophiles and mesophiles). Two residues are paired if the side chain centroids of the residues are within 6.5 Å of each other.

Hydrogen Bonding. Hydrogen bonds were determined by the program Hbplus (*61*). The number of unsatisfied hydrogen bonds was determined by counting the number of buried

(depth > 6 Å) hydrogen bond donors and acceptors that did not have an opposing acceptor/donor partner.

Helix Propensity. The preference of a particular amino acid to be in helices, is defined as $(n_i/N_i)/h$, where n_i is the number of times residue *i* occurs in helices, N_i is the total number of residue *i* in the population, and *h* is the fraction of helical residues in the population (*62*, *63*). Helix propensity of each of 20 amino acids were determined for thermophilic and mesophilic populations.

Helix Stabilization. The helical residues were classified according to Richardson and Richardson (1988) (*64*) as follows:

-N″−N′−Ncap−N1−N2−N3-...
C3−C2−C1−Ccap−C′−C″-

where Ncap and Ccap are the N- and C-terminal residues of the helices, respectively. The occurrence of the following factors known to affect helix stability was examined.

(i) β-branched residues: β-branched residues are found to destabilize helices (*65*, *66*). Rotameric conformations of C-β branched side chains are constrained by the helical conformation of the backbone (*67*). The number of C-β branched amino acids, i.e., Val, Ile, and Thr was calculated for residues in the N1-...-C1 segment.

(ii) Salt bridges: Oppositely charged residue pair interactions, i.e., Glu−Lys, Glu−Arg, Asp−Lys, Asp−Arg at positions (i, $i \pm 3$) and (i, $i \pm 4$) are stabilizing (*68*).

(iii) Charge−dipole interaction: Ptitsyn (1969) (*69*) pointed out that negatively (Asp and Glu) and positively (Arg, Lys, and His) charged residues are preferentially found at the N- and C-terminal end of helices. This suggested that the helix dipole is stabilized by positively charged residues at the C-terminus and negatively charged residues at the N-terminus (*70*). Charged dipole interactions were evaluated in terms of the number of Asp and Glu at Ncap, N1, N2, and N3 positions, and the number of Arg, Lys, and His at the C3, C2, C1, and Ccap positions. Only helices longer than eight residues were considered to avoid the overlap of N- and C-terminal regions. The net charge (number of Arg, Lys, and His minus the number of Asp and Glu) for each set of four terminal helix positions was determined.

(iv) N-capping box: In an N-capping box, the side chain of the residue at N3 is hydrogen bonded to the amide group of Ncap and the side chain of Ncap is hydrogen bonded to the amide group of N3 (*71*). Potential N-capping boxes are characterized by the simultaneous presence of Ser, Thr, Asp, Asn, His, Glu, or Gln at both the Ncap and N3 positions (*71*).

(v) Schellman motif: This typical C-terminal helix termination signature is defined by the presence of Gly at C′, hydrogen bonds between C2 and C3 carbonyl groups and C′ and C″ amides, and the presence of a hydrophobic interaction between C3 and C″ (*72*).

(vi) Hydrophobic staple: The interaction between hydrophobic residues at N′ and N4 positions was observed in several proteins, and considered an expansion of the N-capping box (*73*) or as a specific motif, called the "Hydrophobic-staple motif" (*74*). The presence of this motif has been highlighted by the simultaneous presence of Leu, Ile, Val, Met, or Phe at both the N′ and N4 positions (*74*).

(vii) Phe-Cys/Met interaction: The interaction between Phe and Cys/Met side chains at positions (i, $i \pm 4$) can stabilize helices through interaction between aromatic electrons and sulfur atoms (*75*).

Statistical Test. We have carried out the Z test and the paired *t* test to evaluate the statistical significance of the difference in a property/feature between the thermophilic and mesophilic proteins for the entire data set.

(i) *Z test*: The Z test was used for comparisons of proportions, e.g., fraction of amino acids or fraction of residues in different secondary structures, etc. Let p_1 ($= x_1/N_1$) and p_2 ($= x_2/N_2$) be the proportions of a quantity (e.g., fraction of Ala) in the thermophilic and mesophilic populations, respectively. x_1 and x_2 are total number of occurrences of that quantity, and N_1 and N_2 are the total number of residues in the thermophilic and mesophilic populations, respectively. $p = (x_1 + x_2)/(N_1 + N_2)$; $q = 1 - p$; $D = pq(1/N_1 + 1/N_2)$; $Z = (p_1 - p_2)/(\sqrt{D})$.

A positive value of *Z* indicates that the proportion of the quantity is higher in the thermophilic population. The 99.9% confidence interval is $Z > 3.0$ or $Z < -3.0$.

(ii) *t Test*: The *t* test was carried out for systematic comparison of various traits in the 125 groups of homologous proteins used in this study. For a particular trait, e.g., number of salt bridges, we calculated the average number of the trait per protein for mesophilic and thermophilic members in a particular group. This gives 125 pairs of numbers for a particular trait, one from each group. A two-tailed paired *t* test of these 125 pairs of numbers was carried out for the statistical significance with 124 degrees of freedom. A positive value of *t* would indicate the property/feature to be higher in the thermophilic population. The 99.9% confidence interval is $t > 2.6$ or $t < -2.6$.

RESULTS AND DISCUSSIONS

The Construction of Data Set for Structural Comparison. The completed proteomes of 21 mesophilic and 9 thermophilic organisms used in this work along with their respective ORFs and optimal growth temperatures are listed in Table 1. All the ORFs (targets) from all 30 proteomes (17 275 thermophilic and 47 105 meophilic targets) were aligned to 3084 chains (templates) from the 1999 version of the PDBselect set of 95% homologous representative proteins (*53*) using the gapped version of the BLAST sequence alignment algorithm (*54*). Only those alignments with an *E*-value $< 10^{-5}$ were chosen for this analysis. From this set of alignments, we only retained those where the sequence identity of the targets was > 40% over at least 90% of the template length. The 40% sequence identity cut off was chosen because above this cutoff, the median overlap between a model and the corresponding experimentally determined structures is > 90% (*58*). The total number of target−template pair (alignments) that satisfied these criteria were 6750, and the individual numbers of targets and templates were, respectively, 3652 and 666. The sequences of these 666 templates were clustered into 394 groups such that members within a group had > 40% sequence identity to each other. Each target sequence was assigned to the group containing the corresponding template. A total of 125 of these 394 groups contained at least one each of thermophilic and mesophilic target sequences; twenty contained only thermo-

Table 1: List of the Complete Genomes

organism	type	ORF	optimal growth temp (°C)[a]
Hyperthermophile/thermophile			
Aquifex aeolicus	bacteria	1522	80
Thermotoga maritime	bacteria	1864	80
Archeoglobus fulgidus	archaea	2409	83
Aeropyrum pernix	archaea	2694	90
Methanococcus jannaschii	archaea	1771	83
Methanobacterium thermoautotrophicum	archaea	1871	65
Pyrococcus abyssi	achaea	1765	98
Pyrococcus horikoshii	archaea	2061	98
Thermoplasma acidophilum	archaea	1478	66
Mesophile			
Borrelia burgdorferi	bacteria	1638	37
Bacillus halodurans	bacteria	4066	30
Bacillus subtilis	bacteria	4100	30
Campylobacter jejuni	bacteria	1634	37
Chlamydia pneumoniae	bacteria	1052	37
Chlamydia trachomatis	bacteria	894	37
Escherichia coli	bacteria	4290	37
Helicobacter sp	bacteria	2058	37
Helicobacter pylori	bacteria	1577	37
Haemophilus influenzae	bacteria	1707	37
Mycoplasma genitalium	bacteria	479	37
Mycoplasma pneumonia	bacteria	672	37
Mycobacterium tuberculosis	bacteria	3924	37
Neisseria meningitidis	bacteria	2025	37
Pseudomonas aeruginosa	bacteria	5565	37
Rickettsia prowazekii	bacteria	837	37
Synechocystis sp strain PCC6803	bacteria	3168	25
Treponema pallidum	bacteria	1030	37
Ureaplasma urealyticum	bacteria	611	37
Vibrio cholerae	bacteria	3828	28
Xylella fastidiosa	bacteria	2766	26

[a] The optimal growth temperatures of the organisms were obtained from the German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de).

philic target sequences, and 249 contained only mesophilic sequences. Our current study is based on this set of 125 groups of protein sequences comprising a total number of 300 thermophilic and 900 mesophilic targets and 178 templates such that every group has at least one thermophilic member, one mesophilic member, and one template. The sequences were clustered into groups because we chose to perform a systematic study on homologous proteins of thermophilic and mesophilic origin. Using BLAST sequence alignments, we checked that no particular stretch of a sequence of a particular group was aligned with members of any other group with an *E*-value less than 10^{-5}. The total number of members in a group varied from 2 to 25 and the fraction of thermophilic sequences in a group varied from 10 to 50%. Models of these 1200 proteins were obtained from MODBASE (http://guitar.rockefeller.edu/modbase) (*52*). In cases where a particular target had multiple models, we accepted only that model that had the highest sequence homology to the cognate template or had a better model quality score as described by Sanchez and Sali (*58*). We also verified that the template for a particular group we picked was identical to that used by MODBASE. In cases where the PDB code for the template differed from that of ours, we incorporated the one used by MODBASE into our data set. The reason for the difference was mainly due to our use of the 1999 PDBselect set, whereas MODBASE uses the most recent version of the PDB. We then checked the distribution of structural classes and folds of domains in our data set using the SCOP database (*76*). Our data set contained 175 domains belonging to the following SCOP classes: 50 of $\alpha+\beta$ (28%), 71 of α/β (40%), 21 of all α (12%), 23 of all β (13%), 9 of multidomain of α & β (7%), and 1 small protein. This indicates that the data set was not biased to any particular structural class. The most frequent folds of the $\alpha+\beta$ class were the ferredoxin-like fold (10%) and the class II aa RS and biotin synthase (8%) and that of the α/β class were TIM β/α barrel (20%), NAD(P) binding Rossmann fold (10%), and P-loop containing nucleotide triphosphate hydrolase (15%). On the basis of biochemical functions, the most common groups were ribosomal proteins (S5, S8, S15, S7, S17, L1, L6, L7/L12, L9, L11, L14, and L30), tRNA aminoacyl synthetases, and transcription and translation factors (IF1, IF3, TFII3, Ef_Tu, and EF-G).

Accuracy of Models. Earlier studies indicated factors such as salt bridges/ion pairs, secondary structure content, and helix stability as being likely contributors to the higher stability of proteins of thermophilic origin (*24*). To study these factors from modeled structures, we verified the reliability of the models for our study. An earlier study by Sanchez and Sali (*58*) quantified the accuracy of the models by measuring the overlap between the model and the actual structure using 1085 models of protein chains of known structure (PDB). The overlap was defined as the fraction of residues whose C_α atoms are within 3.5 Å of each other in the globally superposed target−template pair of structures (*58*). "Good" models are those that have > 30% of their residues overlapped with the corresponding actual structures (*58*). We selected 625 "good" models of 267 PDB templates from the earlier work of Sanchez and Sali (1998) whose chain lengths were > 50 residues and target−template sequence identity was > 40% over 90% of the template length. A total of 267 PDB structures (actual structures) were compared with their corresponding models from the set of 625 "good" models for the following features: secondary structure, salt bridge/ion pair, and residue pair contacts. Figure 1a,b shows the histograms of the percentage of models as a function of the fraction of residues with correctly predicted and overpredicted secondary structures, respectively. The average accuracy of correctly predicted secondary structure (ratio between number of correctly predicted and total number of residues) is 83%, and there is 16% overprediction. Figure 1c,d shows histograms of the percentage of models as a function of the fraction of residues with correctly predicted and overpredicted salt bridges. On average, the accuracy of correctly predicting a salt bridge is 64%, and the overprediction is about 28%. Figure 1, panels e and f, respectively, show histograms for correct and overprediction of residue contacts. The accuracy of prediction is 72%, and the overprediction is 22%.

These prediction accuracies are high enough to permit valid comparisons of differences in these features between modeled structures from thermophiles and mesophiles. It should be noted that many of the comparisons made here depend primarily on the quality of the target/template alignment and the overall similarity in the three-dimensional structures. The high level of identity in the alignments used in this study suggests that the alignments are very likely to be correct and that the true structures of the target will typically be similar to that of the template.

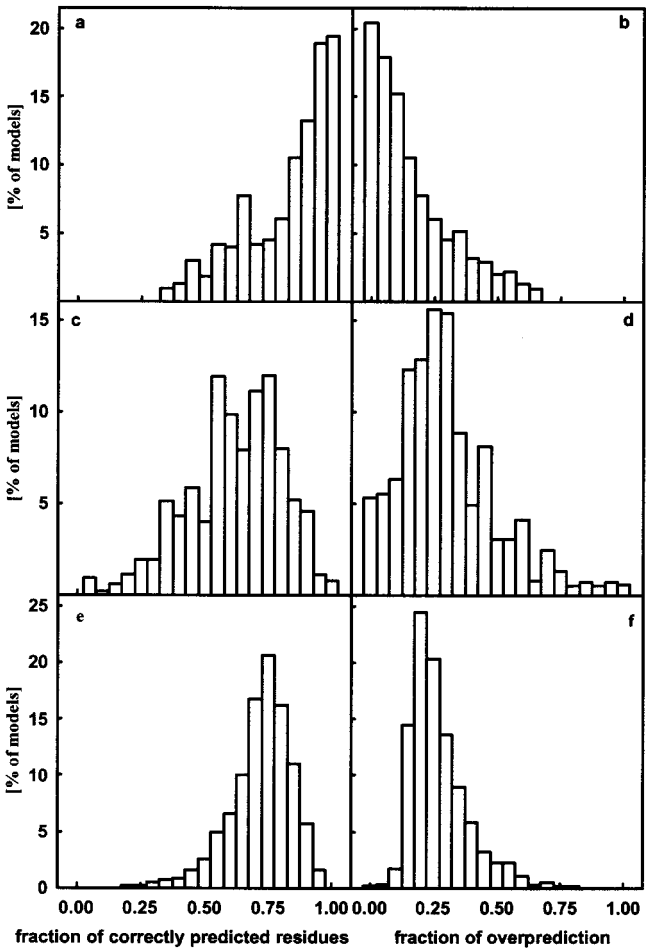


FIGURE 1: Evaluation of model accuracy. Models of known PDB structures were obtained to check the model accuracy as described in the text. The distribution of the models (in percentage) as function of fraction (a) of correctly predicted secondary structures, (b) of overpredicted secondary structures (c) of correctly predicted salt bridges, (d) of overpredicted salt bridges, (e) of correctly predicted residue contacts, and (f) overpredicted residue contacts. Bin size is 0.05.

Comparison of Various Features between Thermophilic and Mesophilic Proteins. (a) Amino Acid Composition. The proportions of 20 amino acids (in percentage) in the thermophilic and mesophilic populations are shown in Table 2. The Z test was performed with an assumption that the frequency of any particular amino acid is independent of other amino acids and the changes in frequency seen are solely due to thermal adaptation. A boxed Z score indicated that the difference in composition between the two populations is significant at the 99.9% confidence level. Compositional differences for buried and surface-exposed residues are shown separately. Residues are classified as charged (Arg, Lys, His, Asp, and Glu), polar (Asn, Gln, Ser, Thr, Cys), and hydrophobic (Ala, Val, Ile, Leu, Met, Phe, Tyr, Trp, Pro).

Compositional differences are much more pronounced at exposed positions. At exposed sites, there is a large increase in charged residues and decrease in polar residues. Unlike other charged residues, His and Asp are decreased in thermophiles. There is a large decrease in Ala content at exposed sites in thermophiles, but there is no change in Ala content at buried sites. Hydrophobic amino acids are increased at exposed sites due to increases in Val, Pro, Tyr, and Trp. Leu is decreased whereas Val is increased at buried sites in thermophiles. Cys is decreased at both buried and exposed sites in thermophiles. The amino acid frequencies presented here are obtained from a set of homologous proteins and need to be compared to frequencies obtained from entire genomes. Our previous analysis of amino acid compositions based on putative soluble proteins from 20 genomes had shown Val, Glu, and total charged residue content to be significantly higher and Gln, Asn, Ser, Thr, His contents to be significantly lower in thermophilic genomes (*43*). This is in agreement with the current result, although changes in the proportions of other charged residues (Arg, Lys, Asp), aromatic residues, Pro and Ala could not be concluded to be significantly different from the previous study. Cambillau and Claverie have recently compared the amino acid compositions of proteins from 30 genomes (22 mesophiles, 1 thermophile, and 7 hyperthermophiles) and observed statistically significant increase in the proportions of (Lys, Arg, Asp, Glu) and decrease in the proportions of (Asn, Gln, Ser, Thr) in thermophiles (*45*). Computing water-accessible surface area for amino acids in 131 mesophilic and 58 thermophilic proteins, they concluded that proportions of charged residues are higher at the surface for thermophilic proteins.

A related study by Haney et al. based on comparison of sequences of 115 proteins from hyperthermophilic archaeon *Methanococcus jannaschii* with their homologues from mesophilic *Methanococcus* species arrived at similar conclusions (*37*). In this work, the number of gains and losses of a particular amino acid replacement in the direction of mesophile → thermophile was calculated based on the pairwise alignment. A two-tailed binomial distribution was used to calculate the probability of such replacements. The

Table 2: Amino Acid Composition (%) in Thermophiles and Mesophiles

	Overall (%)			Exposed (%)			Buried (%)		
Amino Acid	TH[a]	ME[b]	Z[c]	TH[a]	ME[b]	Z[c]	TH[a]	ME[b]	Z[c]
G	8.3	8.3	0.0	8.9	8.9	-0.2	6.8	7.0	-0.8
A	8.0	9.2	-10.6	6.5	8.3	-14.0	13.3	12.9	1.3
V	8.3	7.6	7.2	6.5	5.9	5.2	14.7	13.8	3.1
I	7.2	7.0	2.0	5.4	5.2	1.9	12.8	13.1	-1.1
L	8.6	8.9	-2.8	7.0	7.1	-0.8	13.7	14.8	-3.6
M	2.5	2.4	1.1	1.9	1.9	1.5	3.5	3.6	-0.5
P	4.4	3.9	7.6	5.0	4.3	7.6	2.7	2.7	0.2
F	3.7	3.5	1.7	2.9	2.9	0.6	5.8	5.6	1.1
Y	3.3	2.6	11.4	3.5	2.7	10.1	3.1	2.7	3.4
W	0.8	0.6	4.6	0.8	0.6	5.0	0.8	0.8	-0.6
R	5.6	4.8	9.4	6.8	5.8	9.3	1.7	1.6	0.9
K	7.9	6.8	11.2	9.9	8.4	10.9	1.5	1.2	2.6
H	1.9	2.0	-3.3	2.0	2.3	-4.6	1.7	1.6	0.9
D	5.3	5.8	-5.6	6.5	6.9	-3.6	2.0	2.2	-1.7
E	9.4	7.3	20.3	11.6	8.8	20.2	2.4	2.0	2.7
S	4.1	5.5	-17.2	4.1	5.8	-15.3	3.6	4.3	-4.7
T	4.7	5.3	-8.0	4.4	5.4	-9.3	5.3	5.1	1.2
C	0.8	0.9	-4.9	0.5	0.6	-3.6	1.5	2.0	-4.1
N	3.2	4.1	-12.4	3.5	4.6	-12.1	1.9	1.8	1.1
Q	2.3	3.5	-17.0	2.5	3.9	-16.5	1.3	1.4	-0.2
Charged	30.0	26.7	19.3	36.7	32.1	20.5	9.2	8.6	2.5
Polar	15.0	19.4	-29.6	15.0	20.3	-28.8	13.6	14.5	-3.1
Hydrophobic	46.8	45.7	5.7	39.5	38.7	3.4	70.4	70.0	1.2

[a] Thermophile. [b] Mesophile. [c] Z score in the box indicates the difference between TH and ME is significant at the 99.9% confidence level.

gain in Arg, Lys, Glu, Ile, Pro, Tyr in thermophiles was significantly higher than their losses, while the reverse was observed for Ser, Asn, Thr, Gln, Gly (Table 1 of Haney et. al. 1999). Arg showed the maximum gain of 17% and Ser showed a maximum loss of 32%.

In the present dataset, the percent change in the amino acid composition was also computed: $(T_i - M_i)/M_i$ where T_i and M_i are the proportion of a particular amino acid in the thermophile and mesophile in the entire dataset. In the present case, percent changes for Arg and Ser are +17 and −26%, respectively, similar to that observed earlier. However, we find changes in several other amino acids not observed in the previous study such as Ala, Val, Cys, Asp, and His. In addition, we show that the amount and the nature of protein compositional variation differ at buried and exposed locations.

The overall compositional difference between the mesophilic and thermophilic population is largely due to differences in the surface-exposed regions of proteins. This is in agreement with a recent study by Fukushi and Nishikawa (*46*). These results are also in agreement with recent mutational studies (*77*, *78*) that highlight the important role of surface residues in protein thermal stability. Changes in densely packed protein cores (*79*) often create packing defects and are thus destabilizing (*80*). Residues at the surface tend to be flexible and show few intra-protein interactions, and thus their contribution to stabililty are often locally confined and additive (*81*). A number of recent studies have shown that surface salt bridges make a significant contribution to protein stability (*82*−*86*). Thermophiles appear to have an increased proportion of charged residues on the protein surface enabling them to form surface salt bridges. Karshikoff and Ladenstein have very recently demonstrated that optimization of electrostatic interaction by increasing the number of salt bridges is the driving force for the enhancement of thermotolerence of proteins from thermophilic organisms (*87*). The dramatic decrease in noncharged polar residues at exposed sites in thermophiles is probably because Asn and Gln side chain deamidation is pronounced at high temperatures (*37*, *88*), and Ser and Thr are known to facilitate this deamidation process.

It has been shown that proteins can be stabilized by decreasing the conformational entropy of the unfolded state (*89*). In the unfolded state, Gly and Pro are residues having the highest and lowest contribution to conformational entropy, respectively. Thus, the mutations Gly → Xaa or Xaa → Pro should decrease the conformational entropy of a protein's unfolded state and result in protein stabilization. We do not find a difference in Gly content, but the Pro content is higher at exposed sites in thermophiles indicating a Xaa → Pro mechanism of protein stabilization. Above 100 °C, the thermal stabilities of amino acids are (Val, Leu) > Ile > Tyr > Lys > His > Met > Thr > Ser > Trp > (Asp, Glu, Arg, Cys) (*25*). The observed preference of branched chain amino acids in thermophiles may be associated with the stability of amino acids itself apart from the conformational entropy factor.

(b) Trends in Amino Acid Composition with Increasing Growth Temperature. As shown in Table 1, of the organisms that grow at high temperature, six are from hyperthermophiles with optimal growth temperatures ≥ 80 °C, while two are from thermophiles with growth temperatures around 65 °C. Since the bulk of the data are from hyperthermophiles most of the comparisons desrcribed in this work will relate to differences between mesophiles and hyperthermophiles rather than between mesophiles and thermophiles. To examine if there were detectable trends in the amino acid composition with growth temperature, amino acid compositions of mesophilic, thermophilic, and hyperthermophilic proteins were compared. In the vast majority of cases where Table 2 showed a statistically significant difference betweeen mesophiles and thermophile/hyperthermophile, a clear trend could be observed in the direction mesophile → thermophile → hyperthermophile. For example, relative percentages of Ala, Leu, Ser, Thr, Asn, and Gln all decrease going from mesophile → thermophile → hyperthermophile while Pro, Tyr, Arg, Glu show corresponding increases in the same direction.

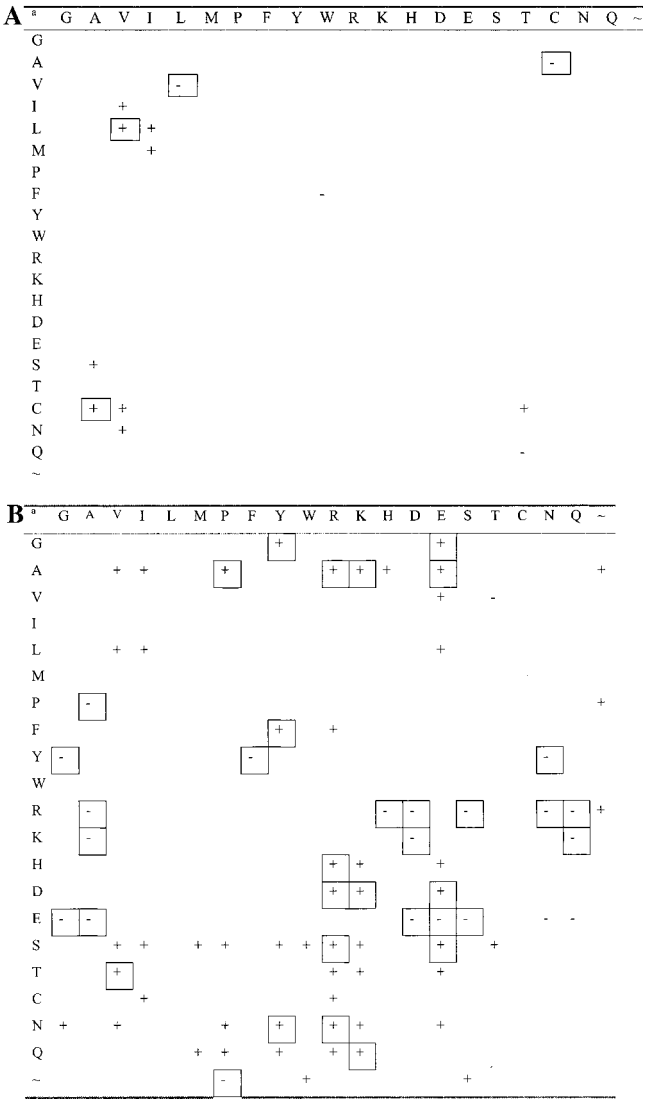
(c) ASA and Solvation. The role of protein−solvent interactions on protein stability has been extensively studied (*24*, *49*, *90*, *91*). Most of the earlier studies have used a simplification based on the accessible surface area of atom types to compute the free energy of interaction between protein surface residues and solvent molecules. Table 2 shows that surface residue compositions differ significantly between thermophiles and mesophiles suggesting that protein/solvent interactions may also differ. We have therefore compared the polar and the apolar accessible surfaces of mesophilic and thermophilic proteins. The accessible surface areas (ASA) were calculated using the implementation of the Lee and Richards algorithm (*92*) with a probe radius of 1.4 Å and a z-section of 0.05 Å. NE, CZ, NH1, NH2 atoms of Arg; NZ of Lys; HD1, NE2 of His; CG, OD1, OD2 of Asp; CD, OE1, OE2 of Glu; OG of Ser; OG1 of Ser; CG, OD1, ND2 of Asn; CD, OE1, NE2 of Gln, OH of Tyr; and C, N, O of backbone were considered polar atoms based on the partial charges assigned to atoms in the AMBER suite of programs (*93*). The remaining atoms were considered apolar. Polar and apolar areas of each protein were computed by summing the ASA of the respective polar and apolar atoms. The total polar surface area of thermophilic proteins are significantly different from that of corresponding mesophilic homologues ($t = 6.0$). The fractions of polar (polar ASA/total ASA) and apolar ASA (apolar ASA/total ASA) for each protein were also calculated. The fractions of polar ASA ($t = 5.97$) and apolar ASA ($t = -5.97$) of a thermophilic protein are significantly different from that of a homologous mesophilic protein. On average, the ASA of a thermophilic protein consists of 46.4 (± 2.1)% polar surface and 53.6 (± 2.1)% apolar surface, whereas that of a mesophilic protein consists of 45.4 (± 3.2)% polar surface and 54.6 (± 3.2)% apolar surface in the present dataset. This qualitatively emphasizes the fact that protein−solvent interactions may make a significant contribution to protein thermal stability.

(d) Residue Substitution. Residue substitution is based on the multiple alignment of homologous proteins. The frequencies of substitution of residue X → Y were computed in the three following backgrounds:

1. $(XY)_{MT}$ vs $(YX)_{MT}$ [forward vs backward for M → T]
2. $(XY)_{MT}$ vs $(XY)_{MM}$ [(M → T) vs (M → M)]
3. $(XY)_{MT}$ vs $(XY)_{TT}$ [(M → T) vs (T → T)]

For computing the Z scores, we first computed the fractions of $(XY)_{MT}$ and $(YX)_{MT}$ replacements from the total

Table 3: Substitution Matrix for (A) Buried Residues and (B) Exposed Residues[a]



[a] A positive sign indicates the substitution row → column is favored in the M → T direction. We report only those substitutions for which the Z scores for all three backgrounds (described in the text) are all either > 3.0 or < −3.0. Only the sign is shown. A box (□) indicates that this holds true for the YX pair with an inverted sign. A tilda (∼) indicates deletion.

of number of M → T replacements. Similarly, the fractions of $(XY)_{MM}$ and $(XY)_{TT}$ were determined from the total of number of M → M and T → T replacements, respectively. Z scores were then determined by treating these as simple comparison of proportions. In doing so, we assume that residue substitutions are independent events and that substitution of amino acids in the M → T direction is primarily due to thermal adaptation.

Table 3 (A, B) shows the most frequent substitutions in buried and exposed regions, respectively. We report only those substitutions that are statistically significant (i.e., $Z > 3$ or $Z < -3$) at all the above three backgrounds. We report the substitution bias in the (M → T) direction. Substitution bias is more prominent at an exposed site than a buried site in the M → T direction. At buried sites, all significant hydrophobic replacements are consistent with decreased side chain conformational entropy, e.g., Leu is preferentially substituted by β-branched residues Val and Ile, Met by Ile and Ile by the smaller and more rigid Val residue. At buried sites, the polar residues Asn and Ser are also preferentially substituted by approximately isosteric Val and Ala. Most prominent substitutions at exposed sites involve charged and noncharged polar residues. Noncharged polar (Thr, Ser, Asn, and Gln) residues are replaced either by rigid (Pro), branched nonpolar (Ile or Val), large aromatic (Tyr or Trp), or charged (Lys, Arg, Asp, or Glu) residues. Asp and Glu residues behave quite differently with respect to substitution. Glu is rarely substituted, whereas Asp is preferentially substituted with Arg, Lys, and Glu. This is consistent with the reduced Asp content in thermophiles. His is replaced by Arg, Lys, and Glu (consistent with reduced His content in thermophiles). Ala is repeatedly substituted by a variety of other residues consistent with the reduced Ala content in thermophiles. The overall features of the substitutions in this study are qualitatively similar to those observed by Haney et. al. (*37*). However, there are several differences. The substitutions considered in this study are from multiple alignments and not from pairwise alignments. The total number of X → Y replacements in the data set is substantially larger than in the previous study. The number of occurrences of over 70% of the X → Y replacements in this study exceeds 300. In general, because of the much larger sample size used in the present study, several additional substitution biases can be detected.

Table 4: Residue Pair Proportions

buried		exposed					
pair	Z[a]	pair	Z[a]	pair	Z[a]	pair	Z[a]
VE	5.9	KE	10.2	AT	−3.0	DS	−5.3
ME	3.5	RE	7.0	NQ	−3.0	QQ	−5.5
RN	3.3	PY	5.3	HQ	−3.1	AS	−6.7
AV	3.0	DE	5.3	IQ	−3.1		
MY	3.0	VI	5.1	IN	−3.1		
FF	3.0	YK	5.0	KS	−3.2		
AH	−4.5	VK	4.9	KQ	−3.2		
YS	−3.3	PE	4.9	ST	−3.3		
SC	−3.3	EE	4.7	FT	−3.3		
VC	−3.2	YE	4.3	NN	−3.3		
IC	−3.2	YY	4.2	FQ	−3.4		
LC	−5.0	VP	4.1	AW	−3.5		
		PV	4.1	LS	−3.5		
		WK	3.9	IS	−3.6		
		PD	3.6	KN	−3.6		
		LY	3.6	VQ	−3.7		
		VY	3.6	TN	−4.1		
		KD	3.5	LN	−4.2		
		RD	3.5	GS	−4.3		
		ME	3.3	TQ	−4.3		
		YW	3.2	DQ	−4.3		
		RK	3.0	AA	−4.8		
		IK	3.0	AQ	−5.0		
		PK	3.0	SS	−5.0		
				AN	−5.1		

[a] Only residue pairs with Z scores > 3 or < −3 are reported here.

(e) Residue Pair Comparisons. We determined the residue pair propensity at buried and exposed sites (Table 4). Once again, the largest differences between thermophiles and mesophiles occur at exposed sites. Exposed residue pairs involving at least one of the polar uncharged residues (Ser, Thr, Ala, and Gln) show a significant drop in thermophiles. This is consistent with the observation that uncharged polar residues are drastically reduced at exposed sites on thermophilic proteins. Oppositely charged pairs (RE, RD, KE, and KD) on the surface are significantly higher in thermophiles

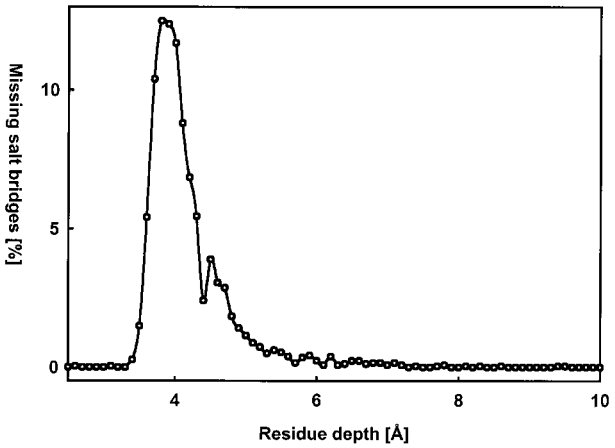


FIGURE 2: Depth distribution of the missing salt bridges. The distribution peaks at 4 Å. The depth of the salt bridge is the average of the depths of the partners.

consistent with increased salt bridge formation. Proportions of charged pairs of similar type (RK, DE, EE) on the surface are also higher in thermophiles, indicating that these residues take part in ion pair networks in thermophiles. There are also increases in aromatic−aromatic and cation−aromatic pairs at exposed sites in thermophiles. Residue pairs involving Pro are increased in thermophiles at exposed sites.

(f) Salt Bridges. The number of salt bridges for every modeled protein in a group was determined as well as the difference in the average number of salt bridges between a thermophile and mesophiles. The results of the paired *t* test are shown in Table 4. The number of salt bridges in a thermophilic protein is significantly higher than in the corresponding mesophilic homolog. Although this result is consistent with earlier observations (*41*), to minimize the errors resulting from the use of modeled structures, "missing salt bridges" were examined in further detail. The missing salt bridges in mesophiles are defined as follows: (i) A salt bridge that is conserved in all thermophilic members of a group and absent in at least one mesophile, or (ii) present in at least one thermophilic member of a group and absent in all mesophilic members. Since every group has a true structure (template) results from missing salt bridges would have low error associated with them. There are 1231 of type 1 and 2626 of type 2 missing salt bridges. The depth distribution of the missing salt bridges (Figure 2) shows that most missing salt bridges are on the surface. This is consistent with the high frequency of surface substitutions. There are controversies regarding the stabilizing effect of salt bridges (*94*). It has been suggested that the free energy of hydration of charged groups become less favorable at higher temperatures. Hence, the unfavorable desolvation penalty incurred on forming a salt bridge is reduced in magnitude at high temperatures (*95*). In the present study, most of the missing salt bridges are on the surface and are solvent exposed. For such exposed salt bridges, there is unlikely to be a large desolvation penalty at any temperature.

A total of 35% of these missing salt bridges are in the same unit of secondary structure (24% in helices, 10% in loops, and 1% in β-strands). Of the 24% in helices, 19% are separated by 3−4 residues, and the remaining 5% are < 3 residues apart. Of the remaining 65% of the missing salt bridges, 12, 10, and 4% are in adjacent helices, loops, and strands, respectively, and 26, 8, and 5% are in helix−loop, strand−loop and helix−strand, respectively. This analysis suggests that such exposed salt bridges significantly stabilize helices and that stabilization of individual helices can lead to increased protein thermal stability.

Table 5: Results of Paired *t* Test for Various Factors Potentially Linked to Protein Thermal Stabilization

factor	*t*-statistic[a]
no. of salt bridge/protein	3.4
cation−π interactions	4.2
difference in average helix length	3.0
difference in average loop length	−2.2
helix stabilization	
ion pair/helix	3.1
negative charge at N-terminus	3.0

[a] Other factors such as unsatisfied hydrogen bonds, Beta-branched residues/helix, positive charge at C-terminus, N-cap box, Schellman motif, hydrophobic staple, prolines/helix, Cys/Met−Phy interactions, and protein size do not show a significant difference.

Several other helix stabilizing factors were also analyzed (Table 5). In addition to an increase in intrahelical salt bridges, there is also an increase in the net negative charge close to the N-terminus of helices in thermophilic proteins. This was also observed previously by the whole genome comparison study (*43*). An earlier study (*48*) concluded that 69% of helices from 13 thermophilic proteins were more stable than their mesophilic homologues, and the enhanced stabilization was observed primarily due to the intrinsic helical propensities of amino acids present in helices from thermophiles and only minor effects were linked to side chain−side chain interaction, helix capping, and charge dipole effects (*48*). We do not find helix-capping, proportion of β-branched residues/helix, Cys/Met−Phe interactions, Schellman motif, hydrophobic staple, or proline/helix to be significantly different in mesophile and thermophile. Furthermore, the helical propensities of various amino acids were found to be identical in thermophiles and mesophiles, in contrast to the earlier study (*48*). This suggests that helix stabilization in thermophiles does not occur by using amino acids with higher propensity.

(g) Cation−π Interaction. Aromatic rings of Phe, Tyr, and Trp are nonpolar as they do not have a net permanent dipole moment. However, they have quadrupole moments that are quite substantial in magnitude (*96*). The quadrupole of the aromatic ring system can be viewed as two opposing dipoles originating from either face of the ring. Due to this, cations interact very strongly with the aromatic ring system, and the strength of these interactions are estimated to be twice as strong as salt bridges due to a smaller desolvation penalty for a cation−π pair compared to an ion pair (*97*−*99*). It has been shown that over 70% of all Arg side chains are near aromatic side chains (*99*, *100*) and 26% of all Trps are involved in energetically significant cation−π interactions on the surface of proteins (*99*, *101*). An increased frequency of both exposed aromatic and positively charged residues in thermophiles suggests that cation−π interactions may have a significant stabilizing effect in enhancing thermal stability. We performed a simple distance based calculation to pick potential cation−π interacting partners. We considered a potential cation−π interaction when the CZ and CE atom of Arg and NZ and CE atom of Lys were within 6.5 Å of the centroid of the phenyl ring of Phe, Tyr, or indole ring of Trp. The accuracy of detection of such interactions in the

data set is 65%, and there is an over- and underprediction of 30 and 27%, respectively. We had shown earlier that incorporation of an uncertainty factor, arising from prediction errors of similar magnitude, in the statistical tests did not make any change in the results obtained with and without error incorporation (*43*) as the prediction errors are equal for thermophilic and mesophilic proteins. Paired *t* tests were carried out to evaluate statistical significance (Table 5). Here, we show that previously neglected cation$-\pi$ interactions may contribute very significantly in enhancing thermal stability.

(h) Secondary Structure Content. The proportion of residues in regular secondary structure were determined from the ratio of total number of helical, strand, and loop residues to the total number of residues in the thermophilc and mesophilic populations. Loops include residues in turns, bends, and irregular secondary structure. The proportion of residues in helical, strand, and loop regions are 38.5, 17.9, and 43.6% for thermophilic proteins as compared to 36.9, 18.2, and 44.6% percent in mesophilic proteins. The corresponding Z scores for the test of relative proportions of helix, strand and loop are 8.3, −4.1, and −5.2, respectively. The combined fraction of residues in helices, strands, turn (regular secondary structure) in thermophilic protein is 75.3% as compared to 74.0% in mesophiles ($Z = 7.2$). Although corresponding mesophilic and thermophilic homologues are modeled on the same template, there is a slight change in the proportion of residues in different secondary structures.

The fraction of residues in regular secondary structures is greater in thermophiles than in mesophiles due to an increase in helical content and decrease in loop content. There is no difference in the size of proteins from thermophiles relative to their mesophilic homologues (Table 5). Hence, to ascertain if lengthening of helix or shortening of loops was responsible for the observed change in secondary structural content, the following analysis was carried out. From the multiple alignments of sequences in every group, we defined the helical and loop regions based on the length of the longest helix and loop segment, respectively, in each sequence. There are 1454 helical and 2454 loop regions in the data set. We computed the difference in the average lengths of corresponding thermophilic and mesophilic helices. In 46% of cases, a given helix was longer in thermophiles than in mesophiles. In 35% of cases, the mesophilic helix was longer than the corresponding thermophilic one and 19% had identical average helix lengths. A paired *t* test on the length of helices ($t = 3.0$) confirmed that thermophilic helices were longer for the data set. In the case of loops, in 38% of cases thermophilic loops were shorter, in 29% of cases mesophilic loops were shorter and 33% had identical average length. From the result of paired *t* test ($t = -2.2$), we could not conclude that loops of thermophilic proteins are shorter at 0.1% level of significance, although the negative value of the *t*-statistic indicated a shortening of loops in thermophilic proteins. The average lengths of thermophilic and mesophilic proteins in every group were also compared (Table 5). However, in the present data set, thermophilic and mesophilic proteins do not differ in length ($t = -0.8$).

In conclusion, the present study has identified several features responsible for enhanced thermal stability of proteins. The data set used here is considerably larger than in any previous study, and this has yielded several fresh insights. The major differences between proteins from mesophiles and thermophiles occur at exposed sites. Surface salt bridges, cation$-\pi$ interactions, increased protein rigidity, and stabilization of secondary structure all appear to be important contributors to increased thermal stability. A list of statistically significant, preferred amino acid substitutions that occur in thermophiles has also been provided. Algorithms that suggest substitutions to enhance thermal stability of proteins primarily focus on choosing combinations of residues that would improve packing interactions of the hydrophobic core (*13*). The present work clearly shows the importance of optimizing interactions between surface residues. These findings can be exploited in experimental studies to design proteins with improved thermal stability.

ACKNOWLEDGMENT

We thank the Bioinformatics Centre and the Supercomputer Educational and Research Centre at the Indian Institute of Science for access to databases and computational facilities and Dr. N. V. Joshi for help with statistical analysis. We thank Dr. Ursula Piper, Dr. Eswar Narayanan, Dr. M. S. Madhusudhan, and Dr. Andrej Sali for providing the MODBASE models and for useful suggestions. We thank Dr. Akshay Bhinge for carrying out the surface area calculations. We thank the Sanchez lab for allowing S.C. use of their computational facilities. R.V. is a recipient of the Swarnajayanthi Fellowship, Government of India, and a Senior Research Fellowship from the Wellcome trust.

REFERENCES

1. Adams, M. W. (1993) *Annu. Rev. Microbiol. 47*, 627−58.
2. Stetter, K. O. (1999) *FEBS Lett. 452*, 22−5.
3. Vieille, C., and Zeikus, G. J. (2001) *Microbiol. Mol. Biol. Rev. 65*, 1−43.
4. Vieille, C., Burdette, D. S., and Zeikus, J. G. (1996) *Biotechnol. Annu. Rev. 2*, 1−83.
5. Zeikus, J. G., Vieille, C., and Savchenko, A. (1998) *Extremophiles 2*, 179−83.
6. Vieille, C., Hess, J. M., Kelly, R. M., and Zeikus, J. G. (1995) *Appl. Environ. Microbiol. 61*, 1867−75.
7. Russell, R. J., Ferguson, J. M., Hough, D. W., Danson, M. J., and Taylor, G. L. (1997) *Biochemistry 36*, 9983−94.
8. Bauer, M. W., and Kelly, R. M. (1998) *Biochemistry 37*, 17170−8.
9. Grattinger, M., Dankesreiter, A., Schurig, H., and Jaenicke, R. (1998) *J. Mol. Biol. 280*, 525−33.
10. Tomschy, A., Glockshuber, R., and Jaenicke, R. (1993) *Eur. J. Biochem. 214*, 43−50.
11. Wormald, M. R., and Dwek, R. A. (1999) *Struct. Fold Des. 7*, R155−60.
12. Russell, R. J., and Taylor, G. L. (1995) *Curr. Opin. Biotechnol. 6*, 370−4.
13. Malakauskas, S. M., and Mayo, S. L. (1998) *Nat. Struct. Biol. 5*, 470−5.
14. Strop, P., Marinescu, A. M., and Mayo, S. L. (2000) *Protein Sci. 9*, 1391−4.
15. Arnold, F. H. (2000) *Adv. Protein Chem. 55*, ix−xi.
16. Lehmann, M., and Wyss, M. (2001) *Curr. Opin. Biotechnol. 12*, 371−5.
17. Munoz, V., and Serrano, L. (1995) *Curr. Opin. Biotechnol. 6*, 382−6.
18. Munoz, V., and Serrano, L. (1996) *Fold Des. 1*, R71−7.
19. Perutz, M. F., and Raidt, H. (1975) *Nature 255*, 256−9.
20. Argos, P., Rossman, M. G., Grau, U. M., Zuber, H., Frank, G., and Tratschin, J. D. 1979 *Biochemistry 18*, 5698−703.
21. Jaenicke, R., and Zavodszky, P. (1990) *FEBS Lett. 268*, 344−9.
22. Jaenicke, R. (1991) *Eur. J. Biochem. 202*, 715−28.
23. Querol, E., Perez-Pons, J. A., and Mozo-Villarias, A. (1996) *Protein Eng. 9*, 265−71.
24. Vogt, G., Woell, S., and Argos, P. (1997) *J. Mol. Biol. 269*, 631−43.

25. Jaenicke, R., and Bohm, G. (1998) *Curr. Opin. Struct. Biol. 8*, 738−48.
26. Szilagyi, A., and Zavodszky, P. (2000) *Struct. Fold Des. 8*, 493−504.
27. Dill, K. A. (1990) *Biochemistry 29*, 7133−55.
28. Pace, C. N., Shirley, B. A., McNutt, M., and Gajiwala, K. (1996) *FASEB J. 10*, 75−83.
29. Matsumura, M., Yasumura, S., and Aiba, S. (1986) *Nature 323*, 356−8.
30. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L., and Bryan, P. N. (1989) *Biochemistry 28*, 7205−13.
31. Serrano, L., Day, A. G., and Fersht, A. R. (1993) *J. Mol. Biol. 233*, 305−12.
32. Shih, P., and Kirsch, J. F. (1995) *Protein Sci. 4*, 2063−72.
33. Davies, G. J., Gamblin, S. J., Littlechild, J. A., and Watson, H. C. (1993) *Proteins 15*, 283−9.
34. Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquo, A., Chiaraluce, R., and Consalvi, V. (1995) *Structure 3*, 1147−58.
35. Aguilar, C. F., Sanderson, I., Moracci, M., Ciaramella, M., Nucci, R., Rossi, M., and Pearl, L. H. (1997) *J. Mol. Biol. 271*, 789−802.
36. Wallon, G., Kryger, G., Lovett, S. T., Oshima, T., Ringe, D., and Petsko, G. A. (1997) *J. Mol. Biol. 266*, 1016−31.
37. Haney, P. J., Badger, J. H., Buldak, G. L., Reich, C. I., Woese, C. R., and Olsen, G. J. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 3578−83.
38. Menendez-Arias, L., and Argos, P. (1989) *J. Mol. Biol. 206*, 397−406.
39. Haney, P. J., Stees, M., and Konisky, J. (1999) *J. Biol. Chem. 274*, 28453−8.
40. Gromiha, M. M., Oobatake, M., and Sarai, A. (1999) *Biophys. Chem. 82*, 51−67.
41. Kumar, S., Tsai, C. J., and Nussinov, R. (2000) *Protein Eng. 13*, 179−91.
42. Thompson, M. J., and Eisenberg, D. (1999) *J. Mol. Biol. 290*, 595−604.
43. Chakravarty, S., and Varadarajan, R. (2000) *FEBS Lett. 470*, 65−9.
44. Das, R., and Gerstein, M. (2000) *Funct. Integr. Genomics 1*, 76−88.
45. Cambillau, C., and Claverie, J. M. (2000) *J. Biol. Chem. 275*, 32383−6.
46. Fukuchi, S., and Nishikawa, K. (2001) *J. Mol. Biol. 309*, 835−43.
47. Warren, G. L., and Petsko, G. A. (1995) *Protein Eng. 8*, 905−13.
48. Facchiano, A. M., Colonna, G., and Ragone, R. (1998) *Protein Eng. 11*, 753−60.
49. Spassov, V. Z., Karshikoff, A. D., and Ladenstein, R. (1995) *Protein Sci. 4*, 1516−27.
50. Vogt, G., and Argos, P. (1997) *Fold Des. 2*, S40−6.
51. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) *Nucleic Acids Res. 28*, 235−42.
52. Sanchez, R., and Sali, A. (1999) *Bioinformatics 15*, 1060−1.
53. Hobohm, U., and Sander, C. (1994) *Protein Sci. 3*, 522−4.
54. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res. 25*, 3389−402.
55. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res. 22*, 4673−80.
56. Sali, A., and Blundell, T. L. (1993) *J. Mol. Biol. 234*, 779−815.
57. Sanchez, R., and Sali, A. (1997) *Proteins Suppl.* 50−8.
58. Sanchez, R., and Sali, A. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 13597−602.
59. Kabsch, W., and Sander, C. (1983) *Biopolymers 22*, 2577−637.
60. Chakravarty, S., and Varadarajan, R. (1999) *Struct. Fold Des. 7*, 723−32.
61. McDonald, I. K., and Thornton, J. M. (1994) *J. Mol. Biol. 238*, 777−93.
62. Chou, P. Y., and Fasman, G. D. (1974) *Biochemistry 13*, 211−22.
63. Levitt, M. (1978) *Biochemistry 17*, 4277−85.
64. Richardson, J. S., and Richardson, D. C. (1988) *Science 240*, 1648−52.
65. Piela, L., Nemethy, G., and Scheraga, H. A. (1987) *Biopolymers 26*, 1273−86.
66. Padmanabhan, S., Marqusee, S., Ridgeway, T., Laue, T. M., and Baldwin, R. L. (1990) *Nature 344*, 268−70.
67. Creamer, T. P., and Rose, G. D. (1994) *Proteins 19*, 85−97.
68. Scholtz, J. M., Qian, H., Robbins, V. H., and Baldwin, R. L. (1993) *Biochemistry 32*, 9668−76.
69. Ptitsyn, O. B. (1969) *J. Mol. Biol. 42*, 501−10.
70. Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M., and Baldwin, R. L. (1987) *Nature 326*, 563−7.
71. Harper, E. T., and Rose, G. D. (1993) *Biochemistry 32*, 7605−9.
72. Aurora, R., Srinivasan, R., and Rose, G. D. (1994) *Science 264*, 1126−30.
73. Seale, J. W., Srinivasan, R., and Rose, G. D. (1994) *Protein Sci. 3*, 1741−5.
74. Munoz, V., Blanco, F. J., and Serrano, L. (1995) *Nat. Struct. Biol. 2*, 380−5.
75. Viguera, A. R., and Serrano, L. (1995) *Biochemistry 34*, 8771−9.
76. Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995) *J. Mol. Biol. 247*, 536−40.
77. Martin, A., Sieber, V., and Schmid, F. X. (2001) *J. Mol. Biol. 309*, 717−26.
78. Maves, S. A., and Sligar, S. G. (2001) *Protein Sci. 10*, 161−8.
79. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng. 6*, 151−76.
80. Lim, W. A., and Sauer, R. T. (1991) *J. Mol. Biol. 219*, 359−76.
81. Reidhaar-Olson, J. F., and Sauer, R. T. (1990) *Proteins 7*, 306−16.
82. Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M., and Makhatadze, G. I. (1999) *Biochemistry 38*, 16419−23.
83. Takano, K., Tsuchimori, K., Yamagata, Y., and Yutani, K. (2000) *Biochemistry 39*, 12375−81.
84. Strop, P., and Mayo, S. L. (2000) *Biochemistry 39*, 1251−5.
85. Spector, S., Wang, M., Carp, S. A., Robblee, J., Hendsch, Z. S., Fairman, R., Tidor, B., and Raleigh, D. P. (2000) *Biochemistry 39*, 872−9.
86. Perl, D., Mueller, U., Heinemann, U., and Schmid, F. X. (2000) *Nat. Struct. Biol. 7*, 380−3.
87. Karshikoff, A., and Ladenstein, R. (2001) *Trends Biochem. Sci. 26*, 550−6.
88. Klibanov, A. M., and Mozhaev, V. V. (1978) *Biochem. Biophys. Res. Commun. 83*, 1012−7.
89. Matthews, B. W., Nicholson, H., and Becktel, W. J. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 6663−7.
90. Eisenberg, D., and McLachlan, A. D. (1986) *Nature 319*, 199−203.
91. Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D., and Argos, P. (1995) *Protein Sci. 4*, 2499−509.
92. Lee, B., and Richards, F. M. (1971) *J. Mol. Biol. 55*, 379−400.
93. Pearlman, D. A., Case, D. A., Caldwell, J. C., Seibel, G. L., Singh, U. C., Weiner, P., and Kollman, P. A. (1995) AMBER4.0, University of California, San Fransisco, USA.
94. Hendsch, Z. S., and Tidor, B. (1994) *Protein Sci. 3*, 211−26.
95. Elcock, A. H., and McCammon, J. A. (1997) *J. Phys. Chem. 101*, 9624−9634.
96. Dennis, G. R., and Ritchie, G. L. D. (1991) *J. Phys. Chem. 95*, 656−661.
97. Gallivan, J. P., and Dougherty, D. A. (2000) *J. Am. Chem. Soc. 122*, 870−874.
98. Dougherty, D. A. (1996) *Science 271*, 163−8.
99. Gallivan, J. P., and Dougherty, D. A. (1999) *Proc. Natl. Acad. Sci. U.S.A. 96*, 9459−64.
100. Singh, J., and Thornton, J. M. (1990) *J. Mol. Biol. 211*, 595−615.
101. Flocco, M. M., and Mowbray, S. L. (1994) *J. Mol. Biol. 235*, 709−17.

BI025523T