Evolutionary Bioinformatics

ORIGINAL RESEARCH

Analysis of Protein Thermostability Enhancing Factors in Industrially Important *Thermus* Bacteria Species

Benjamin Kumwenda[1,*], Derek Litthauer[2,3], Özlem Tastan Bishop[4,5] and Oleg Reva[1]

[1]Bioinformatics and Computational Biology Unit, Department of Biochemistry, University of Pretoria, South Africa. [2]Department of Microbial Biochemical and Food Biotechnology, University of the Free State, South Africa. [3]National Control Laboratory for Biological Products, University of the Free State, South Africa. [4]Rhodes University Bioinformatics (RUBi), Department of Biochemistry, Microbiology and Biotechnology, Rhodes University, Grahamstown 6140, South Africa. [5]Biological Sciences and Bioengineering, Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey. *Corresponding author email: benjkum@gmail.com

Abstract: Elucidation of evolutionary factors that enhance protein thermostability is a critical problem and was the focus of this work on *Thermus* species. Pairs of orthologous sequences of *T. scotoductus* SA-01 and *T. thermophilus* HB27, with the largest negative minimum folding energy (MFE) as predicted by the UNAFold algorithm, were statistically analyzed. Favored substitutions of amino acids residues and their properties were determined. Substitutions were analyzed in modeled protein structures to determine their locations and contribution to energy differences using PyMOL and FoldX programs respectively. Dominant trends in amino acid substitutions consistent with differences in thermostability between orthologous sequences were observed. *T. thermophilus* thermophilic proteins showed an increase in non-polar, tiny, and charged amino acids. An abundance of alanine substituted by serine and threonine, as well as arginine substituted by glutamine and lysine was observed in *T. thermophilus* HB27. Structural comparison showed that stabilizing mutations occurred on surfaces and loops in protein structures.

Keywords: biotechnology, enzyme, evolution, folding energy, thermostability, 3D structures

Evolutionary Bioinformatics 2013:9 327–342

doi: 10.4137/EBO.S12539

This article is available from http://www.la-press.com.

Background

Thermostability is the resistance to irreversibility of chemical or physical changes of a substance due to elevation in temperature. Protein thermostability is, therefore, the preservation of the unique structure and chemical properties of polypeptide chains under extreme temperatures.[1] Thermostability is a very important property of enzymes because it enhances productivity in industry, as enzymes operate at higher temperatures where more reagents and compounds are available.[2,3]

Reactions at higher temperatures using thermostable enzymes are more efficient due to increased solubility. With no need for frequent cooling, which slows down production, as is the case with mesophilic and some synthetic enzymes, overhead costs are lowered and consequently profits increase.[4–6]

Thermophilic organisms provide a natural source of thermostable enzymes for industrial applications. Thermostable enzymes are either extracted from cultured thermophilic organisms or are genetically engineered based on previously elucidated properties.[7] Furthermore, organisms can be directly inoculated in chemical processes to achieve desired byproducts as a result of biochemical processes such as respiration, oxidation, and degradation.[8]

Examples of the application of thermostable proteins, particularly those produced by species of the *Thermus* genus, include, but are not limited to, bioremediation for eradication of heavy metal pollution, waste and contaminated water treatment,[9] clearing clogged pipes,[10,11] bio-mining,[12–14] biofuel production,[15] controlling global methane fluxes into the atmosphere,[16,17] and reduction in toxicity in food.[18]

The ability of *Thermus* species to live in higher temperature environments, such as hot springs, hot domestic water, deep mines, compost manure,[19–21] and many others, do not only fascinate biologists, but also has extreme importance in biotechnological applications.[3]

Sustaining life at high temperatures suggests an evolved metabolic network system and use of thermostable proteins and enzymes which facilitate cellular biochemical processes necessary for survival. Such enzymes are both thermostable and resistant to chemical reagents, salt concentrations, high pressure, and acidic and alkaline conditions, making them even more suitable for biotechnological application.[22]

Several factors are implicated in thermal stabilization of proteins and enzymes in thermophilic organisms. These include amino acid preference, ratio of charged versus uncharged amino acids, ionic interactions, codon usage, hydrophobicity, and protein surface area.[1,23,24]

Thermal stability, however, is a consequence of the combination of several factors acting in synergistic manner and not due to a single specific major factor.[3,25] Moreover, stabilizing factors differ from taxon to taxon, organism to organism, and even within the same organism from protein to protein, rendering their general elucidation extremely complex.[26] It comes as no surprise, therefore, that studies have produced conflicting results on factors and evolutionary mechanisms that generally influence protein thermostability. Russell et al[24] found protein compactness as a contributing factor. Haney et al,[27] Zhou et al,[1] and Dill[28] reported hydrophobicity as a dominant force influencing thermostability. However, Kumar et al,[23] having compared proteins from mesophiles and thermophiles, found no impact on the level of thermostability due to compactness and hydrophobicity of proteins. However, these authors and Yip et al[29] reported salt bridges and side chain hydrogen bonds to positively influenced protein thermostability. In accordance with this, Kumar et al[23] did not observe any correlation between the thermostability and compactness or protein lengths. Contrarily, Russell et al[24] noted an increase in thermostability due to the shortening of proteins and loop deletions in secondary protein structures. These contradictions are partly due to insufficient data that is used in the analysis, which is usually in the author's interests, and different methods used to measure protein thermostability. Although thermostability-enhancing factors have been identified in other organisms, there was need to determine them in species of the *Thermus* genus.

Among several other approaches, protein thermostability is measured by optimum growth temperature,[30] protein melting temperature (T_m),[23] minimum folding energy (MFE) of RNA secondary structures,[31] and the (Glu+Lys)/(Gln+His) amino acid ratio.[32] Since the same organisms can inhabit different environments with different levels of environmental temperatures depending on salinity, acidity, and other factors, optimum growth temperature cannot precisely measure thermostability of individual

proteins to be generalized to entire species. In addition, thermostability differs from protein to protein within the same organism at the same environmental temperature. The conventional approach has been to crudely consider all proteins from thermophilic organisms as thermostable. In this study, however, the intention was to determine factors that can be used to exactly identify individual thermostable enzymes within species of the *Thermus* genus.

Melting temperature is the point where proteins lose conformity and denature.[23] It is an approach that is used to determine protein thermostability by heating proteins in the laboratory. Its limitation, however, is that it requires advanced expertise, equipment, and other resources that are both laborious and costly. There is also insufficient protein melting data that could be used for large-scale analysis of this nature. Melting temperatures cannot be derived from orthologous sequences because slight differences may be the source of the difference in stabilization. In silico approaches based on codon and amino acid statistics or estimation of the folding energy of mRNA molecules appear to be a more promising approach for a large-scale thermostability analysis. MFE computations combine estimation of canonical base pair energies of interior and exterior edges, resulting from the formation of stacking, hairpins, bulges, interior, and multi-branched loops.[33] Thermostability prediction is based on the assumption that structures that fold with minimal energy are more stable and less affected by external factors. The UNAFold algorithm[31] performs thermodynamic optimization of all possible structures of RNA molecules to obtain the one with the lowest free energy normalized by the sequence length. Free energy is determined by adding energy contributions of all base pairs in loops and hairpins.[34]

This study applied MFE to predict thermostability of *Thermus* proteins under the hypothesis that there is correlation between thermostability of mRNA molecules and encoded proteins in related organisms with similar genomic GC-content. The amount of free energy released during base pair formation defines the thermostability of mRNA secondary structures. According to thermodynamics rules, the more negative free energy that is released, the more stable the resulting structure.[35] It was experimentally proven that structural RNA molecules had lower folding energy than random RNA of the same nucleotide frequency.[36]

Having analyzed pairs of orthologous proteins in *T. thermophilus* HB27 and *T. scotoductus* SA-01, it was observed that there were differences in the predicted levels of thermostability due to the effect of selection and counter-selection of specific amino acids and changes in proteins structures. This study analyzed factors contributing to thermostability within thermophilic *Thermus* species by comparing proteins of the extreme thermophiles *T. thermophilus* HB27[37] against thermotolerant *T. scotoductus* SA-01[21] to elucidate evolutionary factors that enhance protein thermostability.

Methods

Calculation of minimum folding energy of mRNA secondary structures

This work aimed at determining factors that enhance protein thermostability in *Thermus* species. A measure of thermostability was required to categorize protein sequences based on which factors could be assessed. To accomplish this, UNAFold algorithm[31] was applied to DNA sequences to calculate MFE (kcal/mol) of predicted mRNA secondary structures. The algorithm was extended in this work using an in-house Python script. The script extracted all coding sequences in a given genome in GenBank format and computed folding energies for all sequences at one run. This was an improvement from its original implementation which computed one sequence at a time and then used another program on the output to extract folding energy values.

Thermostability was analyzed in moderate thermophiles (*Meiothermus silvanus* DSM 9946 [NC_014212] and *M. ruber* DSM 1279 [NC_013946]; thermotolerant *T. scotoductus* SA-01 [NC_014974]); thermophilic *T. thermophilus* HB8 [NC_006461] and HB27 [NC_005835]). All these organisms belong to the Thermaceae family and the GC-content of their genomes is in the range from 62% to 69%. Their genome sequences are available in the NCBI database under the given accession numbers. The analysis was narrowed and focused on coding sequences of *T. thermophilus* HB27 and *T. scotoductus* SA-01, in order to be generalized to *Thermus* species. Although there are over fifty known *Thermus* species, only three strains were completely sequenced at the commencement of this work: *T. thermophilus* HB8 and HB27, and *T. scotoductus* SA-01.

Identification and analysis of orthologous sequences

Thermostability enhancing factors were analyzed between pairs of orthologous sequences of *T. thermophilus* HB27 and *T. scotoductus* SA-01. Orthologous sequences with largest negative difference in MFE as computed by UNAFold algorithm were analyzed. Mutations of dominant amino acids and their properties were analyzed between orthologous sequences that met the selection criteria. Orthologous sequences were identified using an in-house Python script that implemented BLASTp[38] and MUSCLE[39] alignments. An assumption was made that since the analyzed organisms were closely related, protein sequences were orthologs if they showed the best BLASTp hit results with the e-value below 0.0001 and scored above 100 with a 75% alignment. For each MUSCLE codon alignment, the alignment length and numbers of nucleotide and amino acid substitutions were calculated.

Composition and substitution of amino acid properties

The goal was to identify preferred amino acid residues and properties that enhance thermostability in *Thermus* species. Orthologs for *T. scotoductus* SA-01 were identified in *T. thermophilus* HB27 using BLASTp as previously discussed. Composition of amino acid residues and properties were analyzed in 500 orthologous sequences with the largest negative difference in MFE on normalized values. Differences in energy values after 500 pairs of sequences became insignificantly smaller. The analysis was done only between *T. thermophilus* HB27 and *T. scotoductus* SA-01, which are closely related, in order to eliminate influence of other evolutionary factors that occur in distantly related organisms. *T. thermophilus* HB27 was chosen over *T. thermophilus* HB8 as it shared more homologous sequences with higher sequence identity with *T. scotoductus* SA-01. Analysis of proteins for closely related species allowed a possible conclusion that differences in the amino acid substitution and composition were more likely to be due to differences in levels of thermostability rather than to other evolutionary factors.

The occurrence of each amino acid residues in coding sequences was counted and expressed as a percentage of the sequence length. Average values were computed for each amino acid residue as it occurred in all coding sequences in the genome. In a similar manner, the average distribution of each amino acid property in each genome was calculated. Each property was calculated in a sequence, expressed as a percentage, and averaged over the entire genome. First, normality of the amino acid residues and property distribution was tested using the Shapiro-Wilk normality test.[40] Then, based on whether they were normally distributed or not, a parametric *t*-test[41] or nonparametric Wilcoxon *t*-test[41] was applied to determine statistical significance in the difference in distribution of amino acid residue and properties at a 95% level of confidence. The distribution of properties was generally not normal and therefore, the Wilcoxon *t*-test was mostly applied.

To determine the direction of amino acid substitutions between high and less thermophilic proteins as measured by the MFE, an amino acid substitution table was computed for all residues between the sampled sequences of *T. scotoductus* SA-01 and their orthologs in *T. thermophilus* HB27. The expectation was that the absolute values of amino acid substitutions should be proportional to the frequencies of these amino acids in the sampled proteins. The frequency values for each amino acid were computed and normalized as shown in the formula below:

$$X_{ij}^{norm} = \frac{N \times N_{ij}}{f_i f_j} \quad (1)$$

where X_{ij} is the computed value in the table; f_i and f_j are the corresponding frequencies of amino acids; and N is the total number of residues in the whole alignment of five-hundred (500) selected proteins. The difference between the direct and reverse substitutions for pairs of amino acids calculated on the distribution of normalized values that deviated beyond the 1.96σ threshold were considered statistically reliable. Amino acid properties that changed between high and less thermophilic sequences were deduced from the table.

Correlation coefficients were calculated between the amount of MFE changes and frequencies of substitutions in nucleotide and amino acid sequences of orthologous proteins. Statistical significance of the calculated correlation coefficients was confirmed by

checking that the standard *t-Student* parameter (1.96 for $P = 0.05$) is significantly smaller than the estimated ones calculated by the following equation:

$$t_{st} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2)$$

where r is a correlation coefficient and n is the number of orthologous genes.

Analysis of protein structures

After determining mutations of dominant amino acids and their properties that enhance thermostability, it was important to investigate their locations in protein structures and their energy contribution to stabilization. This was to identify locations and structural features that contribute to stabilization of protein structures. To achieve this, protein structures were modeled using target orthologous sequences which satisfied the selection criterion. Target sequences for searching template structures to build homology models were orthologous pairs from *T. thermophilus* HB27 and *T. scotoductus* SA-01 that were less than 50 nucleotides difference in length, sequence identity of greater than 50%, and largest MFE difference on normalized MFE values. Orthologous sequences that satisfied the selection criterion were used to search for 3D protein structure templates using the HHpred program[42] in the PDB database.[43] The process of constructing 3D protein model structures comprised the following steps: (i) target identification and selection of template structures; (ii) sequence alignment of the target sequence to the identified template; (iii) construction of the model based on the alignment with the template; and (iv) evaluation and refining the model structure.[44] Template structures with a sequence identity of greater than 30% and resolution of less than 3Å with the highest quality as validated by Procheck, Anolea, and Qmean6 programs within the Swiss-model suite[45] were used. The PDB coordinate files for protein structures were edited so that the start and stop positions of atoms corresponded to positions in the target sequence. Templates were verified using PyMOL[46] for loop breakages. Preferably, chain A was used for modeling. For each target sequence, other homologous sequences were searched in the NCBI public database using BLASTp for multiple sequence alignment. Homologous sequences with query coverage greater than 80% and E-values lower than 0.0001 qualified for multiple sequence alignment. These sequences, together with the sequence of the selected template structure, were subjected to multiple sequence alignment using different programs such as MAFFT,[47] Muscle,[39] and Promals3D[48] to establish the best alignment that could result in building accurate model structures. Multiple sequence alignments were assessed in two ways, by comparing against the HHpred alignment and also by comparing the quality of the resulting models in PyMOL.[44] Quality alignment had fewer gaps in helices and beta-sheets and with the most biologically meaningful substitutions. Such an alignment was refined, if necessary, before being used to build models using MODELLERv9.11.[49] For each protein, not less than one hundred models were constructed and the best three models were selected based on the lowest DOPE Z score.[50] The qualities of the three best models for each target protein were further validated using Procheck, Anolea, and Qmeans6 programs found in the Swiss-model suite. The best model out of the top three were refined, if necessary, by further adjusting the alignment and refining loops before performing structural comparisons.

Analysis of amino acid substitutions on protein structural stability by FoldX

In order to determine the stabilizing effect of each mutation, FoldX plugin for Yasara program version 1.4.21[51,52] was used to calculate the energy difference between the structure (wild type) and the mutant. The original structures that matched with *T. thermophilus* HB27 were used in this analysis. As shown in Table 2, these had high sequence identity: 100% for 2DP9 and 2CWY; 99% for 2EBJ; 98% for 2FK5; and 95% for IV8D. In order to be precise, only structures with sequence identity greater than 95% were used in the stability change analysis. Only predominant substitutions identified in the study were analyzed for stabilizing effect in protein structures. Mutations were introduced into structures and energy changes were observed for each mutation. Negative energy values in energy change indicated stabilization while positive values implied protein destabilization. FoldX has an error margin of ±0.5 kcal/mol, hence energy values above the error margin were considered to have

significant stabilizing effect. PyMOL was used to locate mutations in 3D protein structures.

Results

Choice of MFE for prediction of thermostability in *Thermus* proteins

Optimum growth temperature of an organism and protein melting temperature were considered to be impractical measures of thermostability in this work for reasons previously given. Therefore, the Fariás-Bonato ratio and the MFE were the two main in silico approaches that could be applied. Before reverting to using MFE, the Fariás-Bonato ratio was examined to determine if it could be applicable in *Thermus* species. The ratio determines levels of thermostability based on amino acid composition in protein sequences. Particularly, the increased number of charged residues (Glu+Lys) against the decreased number of polar residues (Gln+His) to create a ratio (Glu+Lys)/(Gln+His) which is used to identify thermostable proteins.[32] Whilst the Fariás-Bonato ratio is applicable in other organisms, it was found not to be suitable for determining thermostability of individual proteins in *Thermus* species. Calculated for *T. scotoductus* SA-01, the ratio was below the defined range of 3.2 to 4.5 that characterized mesophilic organisms.[32] In addition, analysis showed that the preferred amino acid usage also did not comply with the prescribed distribution based on which the ratio was developed. Bacteria of the genus *Thermus* therefore, exploited alternative mechanisms of increasing thermostability. The Fariás-Bonato ratio is ideal for computing thermostability of proteins of distantly related organisms, but not for closely related ones as in the case of this study.

The working hypothesis was that there is a strong correlation between the thermostability of mRNA molecules and the encoded proteins, and that thermal stability of mRNA molecules directly affects the functionality of encoded proteins. This hypothesis was investigated by analyzing MFE of all coding sequences from genomes of different *Thermus* bacteria.

MFE distribution curves and protein thermostability

The UNAFold algorithm was applied to determine thermostability of individual protein sequences. It was expected that in extreme thermophiles, these values would be higher than in thermotolerant Thermaceae. Calculated MFE values were normalized by the length of mRNA sequences to avoid bias. Figure 1 shows the distribution of MFE values calculated for all predicted genes in five Thermaceae genomes. Genes were ordered by MFE values and ranked into groups of approximately 0.5% of the total number of the genes in each genome for better presentation.

MFE distribution curves calculated for all organisms were similar, but shifted to lower MFE in both thermophilic *T. thermophilus* strains indicating higher thermostability for all their mRNAs. These were followed by *T. scotoductus* SA-01, *M. ruber* DSM 1279, and *M. Silvanus* DSM 9946, respectively. This observation was consistent with phylogenetic classification of micro-organisms based on 16s rRNA

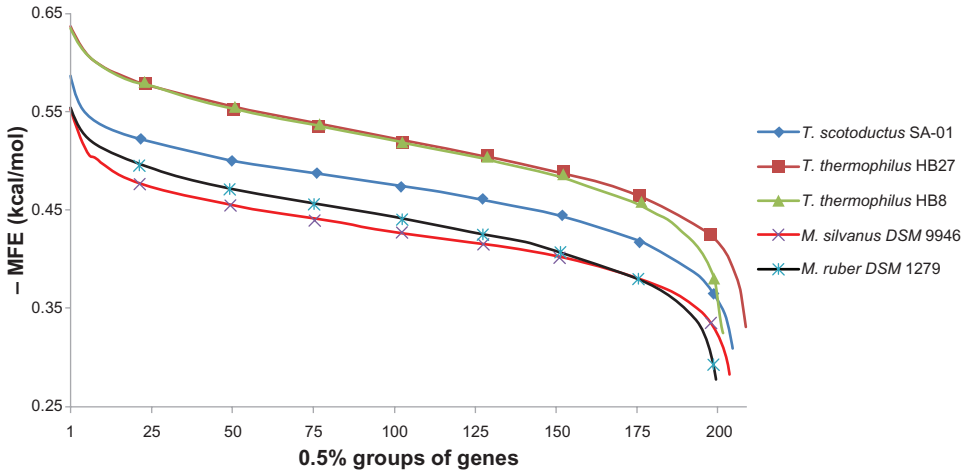

Figure 1. Distribution of MFE calculated for genes of five bacterial genomes. Negative MFE values are ordered along the axis Y from bigger to smaller.

sequences and their optimum environmental growth temperature. Interestingly, the top 100 thermostable mRNAs in *T. scotoductus* SA-01 were not the same as the top 100 ones in *T. thermophilus,* as shown in Supplementary Tables 1 and 2. These groups overlapped only by 35% and 30% for *T. thermophilus* HB27 and *T. thermophilus* HB8, respectively. Thus, some mRNAs acquired more thermostability than others. In an attempt to identify the mechanisms of enhancing of thermostability in *Thermus*, we computed correlations between differences in MFE values calculated for orthologous proteins in *T. thermophilus* HB27 and HB8, and *T. scotoductus* SA-01, and the rates of nucleotide and amino acid substitutions and deletions. The increase in MFE values in *T. thermophilus* HB27 and HB8 calculated for predicted mRNA secondary structures correlated with both the rate of nucleotide substitutions ($r = 0.38$) and amino acid substitutions ($r = 0.30$) in encoded proteins (Supplementary Tables 1 and 2). These correlations were statistically significant as confirmed by a *t*-test. It was assumed that the correlation between changes in MFE and frequencies of nucleotide and amino acid substitutions between orthologous protein coding genes implies a parallel adaptation of mRNA and proteins to higher temperatures. A statistically reliable correlation of 0.18 was also found between changes in MFE and dN/dS non-synonymous/synonymous nucleotide substitution rate ratios (Supplementary Tables 1 and 2). An increased frequency of dN substitutions over dS indicate a positive selection.[51] Thus, it can be concluded that significant changes in MFE and possible higher thermostability of mRNA molecules and encoded proteins are under positive evolutionary selection in these microorganisms.

Distribution and substitution of amino acids

In a comparative analysis of genomes of *T. thermophilus* HB27 and *T. scotoductus* SA-01, 1526 orthologous proteins were identified. Out of these, 500 orthologs with largest MFE difference were analyzed (Supplementary Table 3A and B). Figure 2 shows amino acid frequency distribution normalized by length between *T. thermophilus* HB27 and *T. scotoductus* SA-01. It was hypothesized that there was no difference in amino acid distribution between highly thermophilic and less thermophilic sequences in *T. thermophilus* HB27 and *T. scotoductus* SA-01, respectively. The distribution of each amino acid was independently analyzed between the two species to determine if it was statistically significant. The hypothesis of preferable use of amino acids in thermostable proteins, were rejected at a significance level of 0.05 (Supplementary Table 4). High occurrence of Ile, Ser, Thr, Asn, Gln, and Lys was observed in less thermophilic *T. scotoductus* SA-01 sequences while occurrence of Pro, Gly, Arg, and Ala were higher in *T. thermophilus* HB27. This hypothesis was also analyzed for amino acid properties. The differences were statistically significant in distribution of polar, sulfur containing, acidic, positively charged, tiny, small, and basic amino acids (Supplementary Table 4). Polar, sulfur containing, and acidic were dominant characteristics in *T. scotoductus* SA-01, while *T. thermophilus* HB27 had high occurrence of non-polar, positively

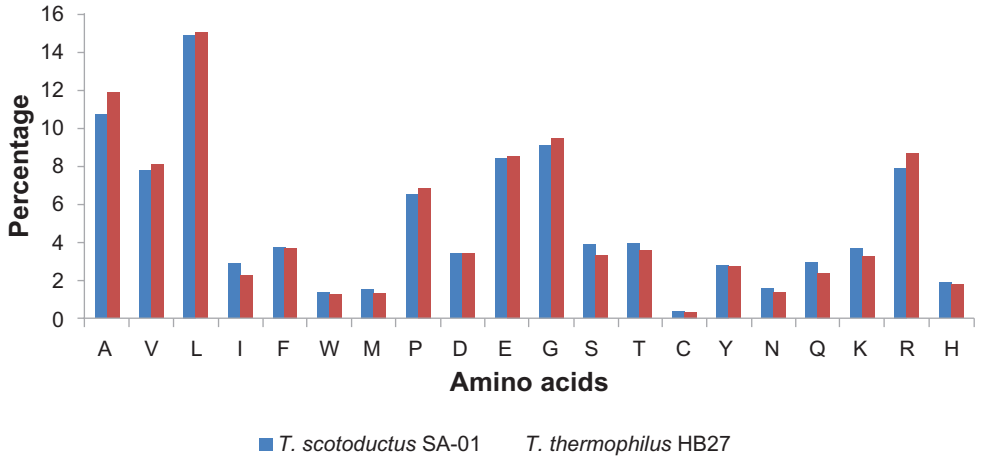

Figure 2. Composition of amino acids in sequences of *T. thermophilus* HB27 and *T. scotoductus* SA-01 sequences.

charged, tiny, small, and basic residues. These results agree with frequencies of amino acid substitutions in Table 1.

Pairwise frequencies of amino acid substitutions were calculated for orthologous sequences. Table 1 shows differences between direct substitutions from amino acids shown in row titles to amino acids in column titles in the pairs of orthologous proteins of *T. thermophilus* HB27 and *T. scotoductus* SA-01. Positive values in Table 1 meant that amino acids shown in columns were accumulated in the proteins of thermophilic *T. thermophilus* HB27. An overall increase or decrease of the amino acids in the sampled proteins of *T. thermophilus* is indicated in the row difference. The rows 'Increased' and 'Decreased' indicate which amino acids underwent this overall increase or decrease of the amino acids shown in column titles.

The expectation was that the absolute values of amino acid substitutions would be proportional to the frequencies of these amino acids in the sampled proteins. The frequency values for each amino acid are shown in Table 1 in the column AMC. Pairs of amino acids for which the difference between the direct and reverse substitutions deviated beyond the 1.96σ threshold calculated on the distribution of normalized values, and thus statistically reliable, are highlighted in the table.

Protein homology modeling

The objective was to determine locations of amino acid substitutions in protein structures to indicate areas useful for protein stabilization in *Thermus* species. Predominant amino acid substitutions were located in 3D protein structure models (Fig. 3). PyMOL was used to visualize the structure in order to determine and contrast the localization of predominant substitutions. Template structures for building 3D models were chosen based on resolution and sequence identity as indicated in Table 2. For each orthologous pair of proteins, one common template with the highest scores was selected. Based on these stringent template selection criterions, five templates

Table 1. Pairwise substitutions of amino acids in orthologous proteins of thermotolerant *T. scotoductus* SA-01 and thermophilic *T. thermophilus* HB27.

	Ala	**Cys**	**Asp**	**Glu**	**Phe**	**Gly**	**His**	**Ile**	**Lys**	**Leu**
Difference	1836	–2	63	128	–184	415	–57	–986	–795	143
Increased								Val	Arg	
Decreased	Ser, Thr		Asn	Gln			Tyr			
Ala	0									
Cys	0	0								
Asp	20	0	0							
Glu	290	–1	13	0						
Phe	23	–1	0	–3	0					
Gly	10	–1	–47	–39	–15	0				
His	22	0	8	20	–13	8	0			
Ile	46	0	0	9	9	1	4	0		
Lys	71	0	11	110	1	47	–5	–1	0	
Leu	83	4	5	6	–112	19	–14	–227	–9	0
Met	40	1	0	20	4	9	0	2	0	44
Asn	52	–1	**56**	25	4	40	10	0	3	5
Pro	19	0	–5	22	–1	–6	–3	–4	–30	–32
Gln	108	1	12	**249**	0	35	16	0	16	46
Arg	47	–5	–4	–92	–16	19	–110	–11	**–541**	–71
Ser	**416**	2	21	51	5	114	12	–6	7	–1
Thr	**349**	1	14	41	–5	33	0	–9	2	13
Val	235	–1	0	2	–22	2	1	**–656**	–7	–110
Trp	–3	1	1	11	16	3	7	–2	0	0
Tyr	8	–2	–2	–2	–20	–1	**70**	–3	–2	4

were identified which matched five pairs of orthologous proteins (Table 2).

Among them, were two proteins of *T. thermophilus* HB27 (TTC1891 and TTC1937), which were previously crystallized. They were used to model the structures of *T. scotoductus* SA-01 proteins TSC_C20070 and TSC_C00450. The template 2DP9 was used to build homology models of two ASCH domain containing proteins TSC_C20070 (*T. scotoductus* SA-01) and TTC1891 (*T. thermophilus* HB27), shown in Figure 3A. Deletion of a short helix in *T. thermophilus* HB27 was observed, with Pro in *T. scotoductus* SA-01 mutated to a polar Glu residue in *T. thermophilus* HB27 (indicated in Figure 3A by an arrow *a*). In the same orthologous sequences, a relatively shorter beta-sheet was noticed in the structure of the thermophilic protein of *T. thermophilus* HB27, with no mutational differences observed (arrow *b*). Protein alignment analysis of TTC1891 against TSC_C20070 revealed one (1) Asn → Asp, two (2) Gln → Glu, two (2) Ile → Val and three (3) Lys → Arg substitutions located predominantly in surface oriented loops and beta-sheets.

The template 2EBJ was used for 3D models of two (2) peptidases TSC_C13250 (*T. scotoductus* SA-01) and TTC0531 (*T. thermophilus* HB27), shown in Figures 3A and 3B. No differences were observed in the structures (Fig. 3B). Protein alignment analysis of TTC0531 against TSC_C13250 revealed one (1) Ser → Ala, one (1) Thr → Ala, two (2) Ile → Val, and one (1) Lys → Arg substitutions located in surface oriented loops and domains.

The template 2CWY was used to build 3D models of two hypothetical proteins TSC_C00450 (*T. scotoductus* SA-01) and TTC1937 (*T. thermophilus* HB27), shown in Figure 3C. No differences were observed in the structures. Protein alignment analysis of TTC1937 against TSC_C00450 revealed three (3) Lys → Arg and one (1) Glu → Arg substitutions located in surface oriented alpha-helices.

The template 2FK5 was used for creation of 3D structures of two (2) fuculose phosphate aldolases

(**Table 1.** continued from page 8)

Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr	AMC
–185	–263	234	–823	1297	–816	–560	690	–60	–75	
	Asp		Arg, Glu		Ala	Ala			His	
				Gln, Lys			Ile			
										32594
										936
										9433
										25858
										10072
										25751
										5125
										6434
										9729
										43451
0										3299
–1	0									3745
–3	–7	0								18184
2	–2	13	0							7155
–13	–41	26	**–304**	0						24379
–4	–26	89	–5	92	0					9364
4	11	30	–6	42	–34	0				9772
–52	–5	20	–11	10	–13	–79	0			22490
2	0	–2	1	19	1	3	–2	0		3126
0	0	8	–2	18	–3	2	6	–4	0	7595

Notes: Positive values mean that amino acids shown in columns were accumulated in proteins of thermophilic *T. thermophilus* HB27. An overall increase or decrease of the amino acids is indicated in the row difference. The rows 'Increased' and 'Decreased,' indicate amino acids that underwent overall increase or decrease as shown in column titles.

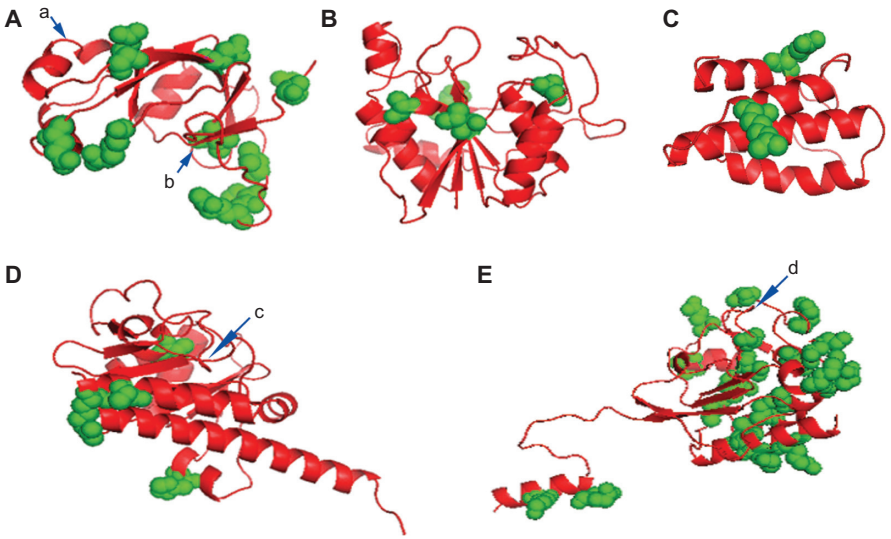

Figure 3. Protein structural models based on templates: (**A**) 2DP9; (**B**) 2EBJ; (**C**) 2CWY; (**D**) 2FK5; and (**E**) 1V8D. Positions of minor structural changes between the orthologous proteins of *T. scotoductus* SA-01 and *T. thermophilus* HB27, are depicted by arrows.

TSC_C04250 (*T. scotoductus* SA-01) and TTC1459 (*T. thermophilus* HB27), shown in Figure 3D. TSC_C04250 contains one (1) additional surface alpha-helix that is missing in the structure of more thermostable TTC1459 (arrow c in Fig. 3D), although there were no mutational differences observed within this region. Protein alignment analysis of TTC1459 against TSC_C04250 revealed one (1) Ser → Ala, one (1) Tyr → His, two (2) Ile → Val, and two (2) Lys → Arg substitutions located mostly in surface oriented beta-sheets.

Table 2. PDB Templates and target pairs of orthologous proteins used for constructing 3D protein structure models.

Templates (PDB ID)	Annotation	MFE (kcal/mol)	Sequence identity (%)	Coverage (%)	Coverage range	E-value	DOPE Z score
2DP9	TSC_C20070 ASCH domain superfamily protein (*T. scotoductus* SA-01)	–0.39	82	96.77	3–124	1.4E-49	–1.56
	TTC1891 hypothetical protein (*T. thermophilus* HB27)	–0.50	100	82.55	3–124	9.9E-41	–0.65
2EBJ	TSC_C13250 peptidase (*T. scotoductus* SA-01)	–0.36	69	99.48	1–192	2.7E-62	–2.42
	TTC0531 peptidase (*T. thermophilus* HB27)	–0.49	99	99.48	1–192	5E-63	–2.42
2CWY	TSC_C00450 conserved hypothetical protein (*T. scotoductus* SA-01)	–0.44	60	93.62	5–94	5.9E-35	–2.19
	TTC1937 hypothetical protein (*T. thermophilus* HB27)	–0.54	100	100	1–94	3.7E-38	–2.78
2FK5	TSC_C04250 fuculose-1-phosphate aldolases (*T. scotoductus* SA-01)	–0.48	83	94.5	1–190	1.4E-51	–1.78
	TTC1459 L-fuculose phosphate aldolases (*T. thermophilus* HB27)	–0.57	98	99.5	1–200	1E-54	–1.73
1V8D	TSC_C14180 conserved hypothetical protein (*T. scotoductus* SA-01)	–0.46	80	82.55	40–235	5E-104	–0.84
	TTC0214 transcriptional regulator (*T. scotoductus* SA-01)	–0.55	95	82.55	40–235	5E-107	–1.14

The template 1V8D was used for modeling 3D structures of the conserved hypothetical protein TSC_C14180 from *T. scotoductus* SA-01 and the orthologous transcriptional regulator TTC0214 from *T. thermophilus* HB27 shown in Figure 3E. TSC_C14180 contains one (1) additional surface alpha-helix that is missing in the structure of more thermostable TTC0214 (arrow d) without any mutational differences observed at sequence level. Several mutations have been observed located on the surface and loops of the structure. The mutations were three (3) Thr → Ala, four Ile → Val, four (4) Lys → Arg, and one (1) Glu → Arg.

Table 3 indicates the effects of individual amino acid substitutions, which varied and were often below the FoldX sensitivity. Protein stabilization by increasing the folding energy is just one of the factors for adaptation in higher temperature environments. Redressing of surface amino acids from polar to non-polar and from uncharged to charged, as well as adjustment of protein-protein interactions, are other means exploited for higher temperature adaptation. In terms of affecting the protein folding energy, similar substitutions have different effects, depending on their localization in the protein structure.

Discussion

This study aimed to determine factors that enhance thermostability in species of the *Thermus* genus in the Thermaceae family. Although there are over fifty known *Thermus* species, only three strains (*T. thermophilus* HB8 and HB27, and *T. scotoductus* SA-01) were sequenced at the commencement of this study. Contrary to prior studies, which analyzed thermostability enhancing factors between distantly related species such as mesophiles and thermophiles, this work analyzed closely related *Thermus* species in the family of Thermaceae, specifically thermotolerant *T. scotoductus* SA-01 against thermophilic *T. thermophilus* HB27. This was to eliminate other evolutionary factors that exist in distantly related organisms, which have no effect on thermostability.

Predominant accumulation of non-polar alanine and positively charged arginine was observed in thermostable proteins of *T. thermophilus* HB27. The accumulation of non-polar amino acids in proteins of thermophiles was previously reported to increase rigidity and hydrophobicity.[54] As it was found by these authors, alanine residues with methyl groups, occurred with lower frequency in exposed states and higher frequency in well-buried states in thermophiles. Alanine is also believed to be the best helix forming residue in thermophiles.[23,55] Alanine residues were mostly substituted by serine and threonine in the homologous *T. scotoductus* SA-01 proteins. Replacement of serine by alanine residues was reported as a factor that increases thermostability of proteins in *Methanococcus*, *Bacillus,*[27,56] and *Corynebacterium.*[57] Serine residues tend to impair hydrophobic interactions between beta-strands, whereas alanine is known to be effective at bridging up strands.[57] Substitution of serine residues by alanine in thermophiles most likely occurs in a series of steps with intermediate glycine residues[58,59] consistent with our observed increase of serine to glycine substitutions (Table 1). Accumulation of threonine in thermostable proteins was observed[59] in contradiction to our results, where alanine was predominantly substituted, a phenomenon that is rather specific to *Thermus* species. However, it is known that threonine, as well as serine, interacts with water molecules, which probably increase instability of proteins at higher temperatures.[60] Arginine and glutamate residues, which accumulate in *T. thermophilus* proteins (Table 1), stabilize exposed structures of thermophilic proteins. Arginine has reduced chemical reactivity and high tendency to participate in salt-bridges (ion pair) interactions.[23,54,61] Several studies have reported high incidence of salt-bridges and ion-pairing as a stabilizing factor in thermophilic proteins.[25,29,62]

In thermostable proteins of *Thermus* species, arginine residues predominantly substituted polar glutamine and lysine. Polar residues are found to be much lower in thermophilic proteins. Asparagine and glutamine residues undergo deamination at high temperatures.[63] Avoidance of lysine and replacement of isoleucine by valine were reported in proteins of some thermophiles. However, lysine and isoleucine are frequently found in proteins of thermophilic *Methanococcus*,[32,64] and thus their role in thermostability is not clear. The frequency of histidine is reduced in general in *T. thermophilus* HB27 in pairs with tyrosine. The former amino acid is preferable in proteins of thermophilic *T. thermophilus* HB27 than in their orthologs from thermotolerant *T. scotoductus* SA-01. This contradicts previous publications stating

Table 3. Energy difference impacts of predominant amino acid substitutions on protein stability as predicted by FoldX.

PDB ID	Mutation SA-01 => HB27	Property change	Position	Location in 3D	Energy change (kcal/mol)
2DP9	Lys => Arg	Conserved	3	S, L	–0.61*
	Lys => Arg	Conserved	29	S, L	+0.03
	Ile => Val	Conserved	41	B, BS	+0.61*
	Asn => Asp	To charged	51	S, BS	–0.09
	Gln => Glu	To charged	55	S, L	0.00
	Gln => Glu	To charged	56	S, BS	+0.10
	Ile => Val	Conserved	104	B, L	+0.52*
	Lys => Arg	Conserved	106	S, L	–0.48
Total energy change					**+0.08**
2EBJ	Thr => Ala	To non-polar	20	S, H	–1.50*
	Ile => Val	Conserved	38	S, BS	+0.48
	Ser => Ala	To non-polar	43	S, L	–0.86*
	Ile => Val	Conserved	80	S, BS	–0.49
	Lys => Arg	Conserved	129	S, BS	–1.20
Total energy change					**–3.57**
2CWY	Glu => Arg	To charged	9	S, H/L	–0.82*
	Lys => Arg	Conserved	49	S, H	–0.32
	Lys => Arg	Conserved	59	S, H	+0.19
	Lys => Arg	Conserved	64	S, H	+0.31
Total energy change					**–0.64**
2FK5	Lys => Arg	Conserved	4	S, H	+0.27
	Ser => Ala	To non-polar	7	B, H	–0.75*
	Ile => Val	Conserved	73	S, H	–0.14
	Tyr => His	To charged	113	S, H	+0.30
	Lys => Arg	Conserved	141	S, H	+0.25
	Ile => Val	Conserved	156	S, BS	+0.38
Total energy change					**+0.31**
1V8D	Lys => Arg	Conserved	6	L, B	N/A
	Lys => Arg	Conserved	10	H, S	N/A
	Ile => Val	Conserved	12	H, B	N/A
	Lys => Arg	Conserved	40	L, S	N/A
	Lys => Arg	Conserved	44	H, S	N/A
	Ile => Val	Conserved	53	H, B	N/A
	Ile => Val	Conserved	66	L, S	N/A
	Ile => Val	Conserved	68	BS, B	N/A
	Thr => Ala	To non-polar	89	H, S	N/A
	Thr => Ala	To non-polar	96	L, S	N/A
	Thr => Ala	To non-polar	106	B, H	N/A
	Glu => Arg	To charged	181	H, S	N/A

Notes: The location of substitutions was marked as S—surface; BS—beta sheet; H—helices and L—in loops. Asterisks mark values of energy change above the FoldX error margin of ± 0.5 kcal/mol. FoldX calculations were not applicable for the template 1V8D due to its low identity (95%) with the studied proteins.

that His/Tyr sites have histidine in mesophiles and tyrosine in thermophiles,[56] which once more demonstrates the taxonomic specificity of the preferable accumulation of amino acids in thermostable proteins. This specificity is associated with the fact that the high temperature of the environment is not the only selective factor to which bacteria adapt. Particularly, as previously discussed, proteins of *Thermus* species are characterized with increased resistance to many other adverse conditions such as extreme *pH* and high salt concentrations, making them even more suitable for biotechnology application.[22]

On the level of RNA and DNA sequences, adaptation was achieved by accumulation of base stacking energy rich oligomers in coding sequences. The increase in mRNA thermostability correlated

to the general increase in GC-content of the tested genomes from 62%–63% in *M. ruber* DSM 1279 and *M. silvanus* DSM 9946, to 69% in *T. thermophiles*. This could not, however, be explained by random substitutions of AT pairs of nucleotides to energy rich GC pairs. Comparisons of tetranucleotide patterns of concatenated coding sequences of *T. scotoductus* SA-01 against those of *T. thermophilus* HB27 showed a relative increase of energy rich and a decrease of energy poor oligomers in *T. thermophilus*. For example, in *T. thermophilus* the frequency of GAGG (−33.11 kcal/mol) increased to an average of 194 oligomers per 100 kbp from 126 oligos per 100 kbp in *T. scotoductus* SA-01, and the frequency of TTTC (−27.33 kcal/mol) decreased from 38 to 18 oligos per 100 kbp. Similar changes in frequencies of several other energy rich and poor oligonucleotides were also observed; however, the choice between preferable and discarded oligonucleotides was selective. For example, frequencies of low energy TTTG and TTGT motifs, and energy rich GCGC remained the same in coding sequences of both bacteria, but the frequency of the latter doubled in non-coding sequences of *T. thermophilus* HB27.

Orthologous coding sequences from *T. scotoductus* SA-01 and *T. thermophilus* HB27 that differed significantly by MFE were further analyzed. To elucidate possible adaptive reasons for predominant amino acid substitutions in *Thermus* organisms, we mapped the identified mismatches in sequences after alignment against 3D structure sequences. Localization of predominant amino acid substitutions highlighted in Table 1 were studied on 3D structural models of five *T. thermophilus* HB27 and *T. scotoductus* SA-01 proteins, for which appropriate templates were identified. A general trend of localization of substitutions was observed in surface areas, at the ends of conserved domains and in loops. Alteration of loops has been observed in several studies to affect protein stabilization.[24,65,66]

It is interesting to note that in all five structures predominant substitutions did not affect overall structures. The gain in surface and exposed tiny, small, and charged amino acid residues in *T. thermophilus* HB27 was observed, which contributed to the stabilization of proteins at high temperatures by reduced chemical reactivity and the high tendency to participate in salt-bridge interactions,[23,54,57,61] an increased rigidity and hydrophobicity,[54,60] and stabilization of helixes.[23,55] Amino acid substitutions also contribute to protein stabilization by increasing the overall protein folding energy. FoldX program was used to estimate the energy change of predominant substitutions (Table 3). The energy contribution of each substitution deferred depending on its location and neighbors in the sequence. The overall substitution analyses, showed that stabilizing effect (negative energy values) of mutations, were much higher, mostly above the error margin of ± 0.5 kcal/mol for FoldX as compared to destabilizing effect (positive energy values) on protein structures. This showed that abundance of such mutations could indeed lead to a more stable protein structure, as previous observed.

Conclusion

The UNAFold algorithm was extended in this work to efficiently handle large data sets to compute MFE for coding sequences in entire genomes. MFE values were higher in thermophiles *T. thermophilus*, which correlated with the general increase in the genomic GC-content. The prediction approach proved to be a useful for the identification of thermostable proteins in closely related organisms. Although this approach yielded consistent results when applied on *Escherichia coli* K-12 and *B. subtilis*, further testing of approaches on distantly related organisms with diverse GC-content is necessary. MFE values calculated for AT-rich sequences may vary, as they utilize different mechanisms of mRNA stabilization.[67–70]

The application of optimum growth temperature and protein melting temperature were considered unsuitable to achieve the objectives of this work. It was observed that the Fariás-Bonato ratio (Glu+Lys)/(Gln+His), recommended for identifying thermostable proteins of thermophilic and mesophilic bacteria,[32] was also not feasible for *Thermus* species because of the observed unexpected decrease of lysine residues and ulterior differences in frequencies of glutamate and histidine (Table 1). The calculated ratio for *T. scotoductus* SA-01 was found to be below the prescribed value for thermophilic bacteria. This observation indicated that bacteria of different taxa exploit different evolutionary adaptation mechanisms to bio-stresses. The ratio is not ideal for computing thermostability of proteins from closely related organisms.

Adaptation of proteins for higher thermostability was achieved by amino acid substitutions.

Comparison of orthologous protein sequences showed existence of dominant trends in amino acid substitutions and their properties consistent with the difference in thermostability between orthologous sequences. *T. thermophilus* HB27 proteins had an increased occurrence of non-polar, small, tiny, and charged amino acids. An abundance of alanine in thermophilic sequences of *T. thermophilus* HB27 was substituted by serine and threonine *T. scotoductus* SA-01. An abundant occurrence of arginine was observed in *T. thermophilus* HB27 substituted by glutamine and lysine.

Structural changes were detected in three out of five modeled structures (Fig. 3). In all cases, the conserved domains disappeared or became shorter in proteins of *T. thermophilus*, consistent with a published report stating that the shortening of loops and domains contribute to compactness of proteins, which is essential in achieving protein thermostability.[24] Structural changes were associated with Pro → Glu substitution in TSC_C20070 versus TTC1891 proteins (Fig. 3, substitution a). Structural comparison of 3D protein structures showed that stabilizing mutations occurred at surfaces. The energy contribution of identified mutations in sampled structures revealed a greater stabilizing effect as compared to destabilizing effect. It may be concluded that adaptation of proteins to high temperature environments is driven by the combination of multiple factors, which may act in synergistic manner as has been reported in other studies.[3,25]

Author Contributions

BK and OR contributed equally to this research and preparation of the manuscript. DL provided sequence and annotation data for *Thermus scotoductus* SA-01, design of the study and preparation of the manuscript. ÖTB guided modeling and analysis of protein structures and preparation of the manuscript. All authors reviewed and approved of the final manuscript.

Funding

This work was funded by SABINA Fellowship Grant, and partly by the South African National Research Foundation Grant 71261.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Zhou X-X, Wang Y-B, Pan Y-J, Li W-F. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*. 2008;34:25–33.
2. Leuschner C, Antranikian G. Heat-stable enzymes from extremely thermophilic and hyperthermophilic microorganisms. *World Journal of Microbiology and Biotechnology*. 1995;11:91–114.
3. Lasa I, Berenguer J. Thermophilic enzymes and their biotechnology potential. *Microbiologia*. 1993;9:77–89.
4. Becker P, Abu-Reesh I, Markossian S, Antranikian G, Márkl H. Determination of the kinetic parameters during continuous cultivation of the lipase-producing thermophile Bacillus sp. IHI-91 on olive oil. *Applied Microbiol Biotechnology*. 1997;48:184–90.
5. Haki GD, Rakshit SK. Developments in industrial important thermostable enzymes: a review. *Bioresource Technology*. 2003;89:17–34.
6. Lioliou EE, Pantazaki AA, Kyriakidis DA. *Thermus thermophilus* genome analysis: benefits and implications. *Microbial Cell Factories*. 2004;3(5):1723–7.
7. Hilvert D. Enzyme engineering. *Chimia*. 2001;55:867–9.
8. Seeliger D, de Groot BL. Protein thermostability calculations using alchemical free energy simulations. *Biophysical Journal*. 2010;98:2309–16.
9. Opperman DJ, van Heerden E. Aerobic Cr (VI) reduction by *Thermus scotoductus* strain SA-01. *Journal of Applied Microbiology*. 2007;103:1907–13.
10. Lovley DR, Baedecker MJ, Lonergan DJ, Cozzarelli IM, Phillips EJP, Siegel DI. Oxidation of aromatic contaminants coupled to microbial iron reduction. *Nature*. 1989;339:297–300.
11. Baedecker MJ, Back W. Modern marine sediments as a natural analog to the chemically stressed environment of a landfill. *Journal of Hydrology*. 1979;43:393–414.
12. Anold RG TJ, Di Christina TJ, Hoffmann MR. Reductive dissolution of Fe(III) oxides by *Pseudomonas* sp 200. *Biotechnology and Bioengineering*. 1998;32:1081–96.
13. Bell PE, Mills AL, Herman JS. Biogeochemical conditions favoring magnetite formation during anaerobic iron reduction. *Applied and Environmental Microbiology*. 1987;53(11):2610–6.
14. Mumford EM. New Iron Bacterium. *J. Chem. Soc*. 1913;103:645–50.
15. Turner P, Mamo G, Karlsson EN. Potential and utilisation of thermophiles and thermostable enzymes in biorefining. *Microbial Cell Factories*. 2007;6:9 doi:10.1186/1475–2859–6-9.
16. Cicerone RJ, Oremland RS. Biogeochemical aspects of atmospheric methane. *Global Biogeochemical Cycles*. 1988;2:299–327.
17. Lovley DR. Dissimilatory Fe(III) and Mn(IV) reduction. *Microbiological Reviews*. 1991;55(2):259–87.
18. Rowbotham AL, Levy SL, Shuker LK. Chromium in the environment: An evaluation of exposure of the UK general population and possible adverse health effects. *Journal of Toxicology and Environmental Health*. 2010;3(3):145–78.

19. Munster MJ, Munster AP, Woodrow JR, Sharp RJ. Isolation and preliminary taxonomic studies of *Thermus* strains isolated from Yellowstone National Park, USA. *International Journal of Systematics Bacteriology*. 1986;132:1677–83.
20. Williams RAD, Smith KE, Welch SG, MicallefSharp RJ. DNA relatedness of *Thermus* Strains, description of *Thermus brockianus* sp. nov., and proposal to re-establish *Thermophilus* (Oshima and Imahori). *International Journal of Systematics Bacteriology*. 1995;45(3): 495–9.
21. Gounder K, Brzuszkiewicz E, Liesegang H, et al. Sequence of the hyperplastic genome of the naturally competent *Thermus scotoductus* SA-01. *BMC Genomics*. 2011;12:577.
22. Niehaus F, Bertoldo C, Kahler M, Antranikian G. Extremophiles as a source of novel enzymes for industrial application. *Applied Mycrobiol Biotechnology*. 1999;51:711–29.
23. Kumar S, Ma B, Tsai C-J, Sinha N, Nussinov R. Folding and binding cascades: Dynamic landscapes and population shifts. *Protein Science*. 2000;9:10–9.
24. Russell RJM, Gerike U.M.J. Danson MJ, Hough DW, Taylor GL. Structural adaptations of the cold-active citrate synthase from an Antarctic bacterium. *Structure*. 1998;6:351–61.
25. Vogt G, Woell S, Argos P. Protein thermal stability, hydrogen bonds, and ion pairs. *Journal of Molecular Biology*. 1997;269:631–643.
26. Trivedi S, Gehlot H, Rao S. Protein thermostability in Archea and Bacteria. *Genetics and Molecular Research*. 2006;5(4):816–27.
27. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *PNAS*. 1999;96:3578–83.
28. Dill, K. Dominant Forces in Protein Folding. *Biochemistry*. 1990;29:7133–55.
29. Yip KSP, Stillman TJ, Britton KL, et al. The structure of Pyrococcus furiosus glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure*. 1995;3:1147–58.
30. Huang S-L, Wu L-C, Liang H-K, Pan K-T, Horng J-T, Ko M-T. PGTdb: a database providing growth temperatures of prokaryotes. *Bioinformatics*. 2004;20:276–8.
31. Markham N, Zuker M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*. 2008;453:3–31. doi: 10.1007/978–1-60327–429–6_1.
32. Farias ST, Bonato MCM. Preferred codons and amino acid couples in hyperthermophiles. *Genome Biology*. 2002;3(8):1–18.
33. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*. 1981;9(1):133–47.
34. Giese MR,Betschart K, Dale T, Riley CK, Rowan C, Sprouse KJ, Serra MJ. Stability of RNA Hairpins Closed by Wobble Base Pairs. *Biochemistry*. 1998;37:1094–1100.
35. Henne A, Brüggemann H, Raasch C, et al. The genome sequence of the extreme thermophile *Thermus thermophilus*. *Nat*. *Biotechnol*. 2004;22:547–53.
36. Mohsen A, Khader A, Ramachandram D, and Ghallab A. Predicting the minimum free energy RNA Secondary Structures using Harmony Search Algorithm. *International Journal of Biological and Life Sciences*. 2010;6 (3):157–63.
37. Clote P, Ferré F, Kranakis E, Krizanc D.Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*. 2005;11(5):578–91.
38. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J*. *Mol*. *Biol*. 1990;215:403–10.
39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004;32(5):1792–7.
40. D.J. Sheskin. Handbook of Parametric and nonparametric statistical procedures. Chapman and Hall/CRC, third edition, 2004.
41. Shapiro S, Wilk M. An analysis of variance test for normality. Biometrika. 1965;52:591
42. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*. 2005;33: doi:10.1093/nar/gki408. 43. Berman HM, Westbrook J, Feng Z, et al. The Protein data bank. *Nucleic Acids Research*. 2000;28(1):235–42.
44. Martín-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A. Comparative protein structure modelling of genes and genomes. *Annu*. *Rev*. *Biophys*. *Biomol*. *Struct*. 2000;29:291–325.
45. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006;22:195–201.
46. DeLano WL. The PyMOL Molecular Graphics System, 2002. DeLano Scientific, San Carlos, CA, USA.http://www.pymol.org. Accessed June 30th 2012.
47. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res* .2002;30:3059–66.
48. Pie J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple sequence and structure alignment. *Nucleic Acids Res*. 2008;36(7):2295–300.
49. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J*. *Mol*. *Biol*. 1993;234:779–815.
50. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci*. 2006;15:2507–24.
51. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucl*. *Acids Res*. 2005;33:W382–8.
52. Van Durme J, Delgado J, Stricher F, et al. A graphical interface for the FoldX force field. *Bioinformatics*. 2011;27(12):1711–2.
53. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: A structural genomics based study. *Biochemistry*. 2002;41:8152–61.
54. Pack SP, Yoo YJ. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic. *Journal of Bacteriology*. 2004;111:269–77.
55. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.*MolBiolEvol*. 1998;15(5):568–73.
56. McDonald JH. Temperature adaptation at homologous sites in proteins from nine thermophile-mesophile species pairs. *Genome Biol*. *Evol*. 2010;2:267–76.
57. Nishio Y, Nakamura Y, Kawarabayasi Y, et al. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. *Genome Res*. 2003;13(7):1572–9.
58. Argos P, Rossman MG, Grau UM, Zuber H, Frank G, Tratschin JD. Thermal stability and protein structure. *Biochemistry*. 1979;18:5698–703.
59. Garg SK, Johri BN. Thermostability. In Johri BN, Satyanarayana T, Olsen J, editor. *Thermophilic moulds in biotechnology*. Kluwer Academic Publishers. Netherlands, USA. 1999;354.
60. Mattos C. Protein-water interactions in dynamic world. *Trends in Biochemical Sciences*. 2002;27(4): 203–8.
61. Das S, Paul S, Bag SK, Dutta C. Analysis of *Nanoarchaeum equitans* genome and proteome composition: indications for hyperthermophilic and parasitic adaptation. *BMC Genomics*. 2006;7:186 doi:10.1186/1471–2164–7-186.
62. Scandurra R, Consalvi V, Chiaraluce R, Politi L, Engel PC. Protein thermostability in extremophiles. *Biochimie*. 1998;80:933–41.
63. Cantanzano FG, Capasso S, Barome G. Thermodynamic analysis of the effect of selective monodeamination at asparagine 67 in ribonucleases A. *Protein Science*. 1997;6:1682–93.
64. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Mol*. *Biol*. *Evol*. 1999;16:1785–970.
65. Nagi, AD, Anderson KS, Regan L. Using loop length variants to dissect the folding pathways of a four-helix-bundle protein. *Journal of Molecular Biology*. 1999;286:257–65.
66. Sánchez IE. Protein folding transition states probed by loop extension. *Protein Science*. 2008;17:183–6.
67. Erickson JW, Gross CA. Identification of the sigma E subunit of Escherichia coli RNA polymerase: a second alternate sigma factor involved in high-temperature gene expression. *Genes Dev*. 1989;3:1462–71.
68. Takayama K, Kjelleberg S. The role of RNA stability during bacterial stress responses and starvation. *Environ Microbiol*. 2000;2:355–65.
69. Garrett SC, Rosenthal JJ. A role for A-to-I RNA editing in temperature adaptation. *Physiology (Bethesda)*. 2012;27:362–9.
70. Alexandre A, Oliveira S. Response to temperature stress in rhizobia. *Critical Reviews in Microbiology*. 2012; Accepted for publication (PMID:22823534).

Supplementary Tables

Supplementary Table 1: MFE values and residue substitution statistics calculated for orthologous genes of *T. scotoductus* SA-01 and *T. thermophilus* HB27.

Supplementary Table 2: MFE values and residue substitution statistics calculated for orthologous genes of *T. scotoductus* SA-01 and *T. thermophilus* HB8.

Supplementary Table 3A: Amino acid distribution of orthologous sequences of *T. scotoductus* SA-01 with largest MFE difference expressed as percentage of the sequence length.

Supplementary Table 3B: Amino acid distribution of orthologous sequences of *T. thermophilus* HB27 with largest MFE difference expressed as percentage of the sequence length.

Supplementary Table 4: *P*-values for *t*-test and Wilcox on *t*-test(*) analysis of distribution of amino acid residues and properties of orthologous sequences of *T. scotoductus* SA-01 and *T. thermophilus* HB27.