ORIGINAL PAPER

Structure-dependent relationships between growth temperature of prokaryotes and the amino acid frequency in their proteins

Gisle Sælensminde · Øyvind Halskau Jr · Ronny Helland · Nils-Peder Willassen · Inge Jonassen

Received: 30 November 2006 / Accepted: 27 February 2007 / Published online: 12 April 2007

Abstract We studied the amino acid frequency and substitution patterns between homologues of prokaryotic species adapted to temperatures in the range 0–102°C, and found a significant temperature-dependent difference in frequency for many of the amino acids. This was particularly clear when we analysed the surface and core residues separately. The difference between the surface and the core is getting more pronounced in proteins adapted to warmer environments, with a more hydrophobic core, and more charged and long-chained amino acids on the surface of the proteins. We also see that mesophiles have a more similar amino acid composition to psychrophiles than to thermophiles, and that archea appears to have a slightly different pattern of substitutions than bacteria.

Keywords Thermophiles · Psychrophiles · Cold-adaptation · Stability · Amino acid frequency · Protein structure

Abbreviations

OGT	Optimal growth temperature
SASA	Solvent accessible surface area

Electronic supplementary material The online version of this article (doi:10.1007/s00792-007-0072-3) contains supplementary material, which is available to authorized users.

Communicated by F. Robb.

G. Sælensminde (✉) · I. Jonassen
Computational Biology Unit (CBU), BCCS,
University of Bergen, Bergen, Norway
e-mail: gisle@cbu.uib.no

Ø. Halskau Jr
Department of Biomedicine, University of Bergen,
Bergen, Norway

R. Helland
Department of Chemistry, University of Tromsø,
Tromsø, Norway

R. Helland · N.-P. Willassen
Norwegian Structural Biology Centre,
University of Tromsø, Tromsø, Norway

N.-P. Willassen
Institute of Medical Biology, Faculty of Medicine,
University of Tromsø, Tromsø, Norway

G. Sælensminde · I. Jonassen
Department of Informatics, University of Bergen,
Bergen, Norway

Introduction

Micro-organisms are present in great variety on Earth and have found their way into almost all kinds of ecological niches imaginable. Many of these have adapted to extreme environmental conditions in terms of temperature, salinity and pressure, and are generally referred to as extremophiles. The extremophiles that have received most attention are those adapted to high temperatures (for reviews, see, e.g. Jaenicke and Bohm 1998; Li et al. 2005). Organisms are termed thermophiles if they live at temperatures above 55–60°C and hyperthermophiles if they live at temperatures above 80°C. Hyperthermophiles have been found in hot springs and hydrothermal vents on the ocean floor with temperatures up to 121°C (Kashefi and Lovley 2003). Organisms living in cold environments are known as psychrophiles. The exact definition of a psychrophile varies, but often organisms with optimal growth temperatures (OGTs) below 15–20°C are considered to be psychrophiles. These are generally found in the oceans and the polar and alpine regions of the Earth, and for example, some have been identified in pockets of salt brine in artic sea ice, where the temperature can drop to below –15°C (Marx et al. 2004). Organisms living in more temperate environments are called mesophiles.

Organisms living at extreme temperatures are faced with the challenge of overcoming altered physiochemical properties induced by the temperature. Thermophiles live at temperatures where most proteins from mesophilic species will unfold rapidly, while at the other end of the temperature scale, the psychrophiles have to cope with reduced molecular motion, lower membrane fluidity, reduced turnover rate of enzymes and increased viscosity. Still, both thermophiles and psychrophiles maintain metabolic processes on a level comparable to organisms living at moderate temperatures (for reviews, see, e.g. Deming 2002; Georlette et al. 2004). Although the vast majority of ecosystems on Earth are exposed to cold environments, like in Arctic, Antarctic and deep sea waters, and polar and alpine regions, most attention has, until recently, been given to proteins from organisms at the other end of the temperature scale. Several attempts have been made to try to identify mechanisms on the molecular level that relates temperature adaptation to nucleotide and/or amino acid composition, and some of the trends found have been reviewed by, e.g. Kumar and Nussinov (2001), Feller and Gerday (2003), Li et al. (2005) and Sadeghi et al. (2006).

One of the main issues for thermophilic organisms is to ensure stable and functional DNA, RNA and proteins that can withstand increased thermal motion caused by elevated temperatures. It has been suggested that thermophiles increase the stability of RNA and DNA by increasing the GC content since this nucleotide pair form three hydrogen bonds instead of two as in AT. Such a relationship has only been observed for ribosomal 16S RNA (Galtier and Lobry 1997), and the work of Hurst and Merchant (2001) indicate that increased CG content is not an adaptation strategy to higher temperatures, while Singer and Hickey (2003) observe that thermophiles are relative rich in purines and suggest that especially adenine (A) can have a stabilising effect on RNA. However, Nakashima et al. (2003) were able to predict reasonably well the OGT of bacteria using regression analysis of either the dinucleotide composition of DNA or the amino acid composition of the proteins. In agreement with this, Singer and Hickey (2000) observed that the amino acid composition in thermophiles and mesophiles differ significantly and Farias and Bonato (2003) observed that the ratio $(E + K)/(Q + H)$ increase with temperature, and is higher in thermophiles than in mesophiles, and yet higher in hyperthermophiles. Berezovsky and Shakhnovich (2005) suggest that there may be two main mechanisms for adaptation to high temperatures, where the proteins either have salt bridges that stabilise the protein tertiary fold or have a tightly packed hydrophobic core. They suggest that the former is more common in bacteria, while the latter is more common in archea. Analysis reported by Thompson and Eisenberg (1999) found that loops are systematically shortened in thermostable proteins compared to the mesophilic homologues on the genome scale.

In this work, we wanted to see which effect the temperature environment has on the amino acid composition and amino acid substitution rates in prokaryotes, and if possible, identify trends which may suggest how proteins from prokaryotic species have adapted be able to maintain metabolic, growth and reproduction rates at high and low temperatures. We also wanted to examine whether there are differences between how amino acid exposed to solvent and those buried in the core of the proteins adapts to extreme temperatures, and whether adaptation to cold and hot temperatures displays similar or different trends.

In the past, several analyses of amino acid substitution rates have been performed for thermophiles. Sadeghi et al. (2006) and Gianese et al. (2001) did a substitution analysis of psychrophiles versus a combination of mesophiles and thermophiles. They constructed multiple sequence alignments, each consisting of many related sequences, and employed homology modelling to assign structures to the sequences. Their analysis differs from the one presented here in that they did not perform any analysis regarding which biophysical properties can explain the temperature-dependent differences observed. Also, they used a more limited dataset including only 21 psychrophilic proteins.

Materials and methods

Protein sequence and structural data

Protein sequences used in this study were selected from the UniProt protein sequence database (Apweiler et al. 2004) and protein structures were obtained from the Protein Data Bank (PDB) (Berman et al. 2000). The sequences were separated into bacterial and archeal origin. The Pfam protein domain database (Bateman et al. 2004) was used to associate sequences from UniProt and structures in the PDB. For each domain with a known structure, Pfam contains cross-references to the PDB (code and chain identifier for the relevant structure entry). Pfam was also used to limit the computational time required, by restricting the search for close homologues to the protein pairs that share a domain in Pfam.

Temperature data

The prokaryotic growth temperature database (PGTdb; Huang et al. 2004) contains the OGT and common laboratory growth conditions for more than 1,000 prokaryotic species. Not all entries in PGTdb have known OGT; therefore the midpoint of the temperature laboratory growth interval of a species was used in the analysis if OGT was not known.

Temperature and amino acid frequency correlation

We performed a correlation analysis on the relationship between the OGT and amino acid frequency for each of the amino acids. The sequences was the about 330,000 sequences from UniProt as of May 2006 originating from species that are included in PGTdb. The sequences were grouped by the source species, using the taxonomy annotations in UniProt to identify the species. Only those species with known growth temperature and with sequences with a total length of more than 1,000 amino acids were used in the analysis. The relative percentage of each amino acid in each species was calculated. It was investigated whether there exists a correlation between the frequency of each of the amino acids and OGT. The null hypothesis of this setup is that the amino acid frequency does not get systematically higher or lower with increasing OGT. The correlation and *P*-values were found using the *R* statistical package (R-core-team 2006), using Spearman's nonparametric correlation and test for *P*-values, and rejected the null hypothesis when the *P*-value was 0.0001 or lower. We could then be reasonably sure that no single one of the 80 correlation-based tests giving *P*-values below this threshold was due to chance (false positive). Correlations were calculated for the full dataset and for the subsets of species in the interval 0–40 and 35–102°C. In addition, archeal species were analysed separately since these dominate the thermophiles. A correlation is a number between –1.0 and 1.0, and is how well the variation of one variable can be explained as a linear expression of the other. A value of 1.0 means that there is a perfect linear relationship, –1.0 means that there is a perfect negative linear relationship, while a 0 means that there is no relationship at all between the two variables. Plots of amino acid frequency against OGT are included in the supplementary material.

Amino acid classification

The 20 naturally occurring amino acids were divided into groups sharing similar biophysical properties (Table 1). The division into the groups is inspired by the VENN diagram of Taylor (1986), but unlike this grouping, we tried to make groups that predominantly have a particular biophysical property, rather than all amino acids with the property, in order to easier identify which biophysical properties that is most important for adaptation to different temperatures. The ''nonpolar'' group is the amino acids without any polar groups in the side chain. These can favourably be buried in the core of the proteins. The ''small'' amino acids are those with 3 or less nonhydrogen atoms the side chain, while the ''long chained'' are those with 5 or more. The ''positive'' group consists of arginine and lysine, where both are good formers of salt-bridges and cation-π bonds (Gromiha et al. 2002). Histidine is excluded because it lacks these properties. The ''polar'' group includes the polar amino acids, but not those that are charged, and thus the amino acids in this group cannot form salt bridges. Proline and glycine are in separate groups, since these amino acids are associated with main chain rigidity and flexibility, respectively.

Table 1 Groups of amino acids sharing the same biochemical properties used to analyse the occurrence of amino acids in the core and on the surface of proteins from species adapted to different temperatures

Property	Amino acids
Nonpolar	Val, Ile, Leu, Met, Phe, Ala
Aromatic	Phe, His, Trp, Tyr
Positively charged	Lys, Arg
Negatively charged	Asp, Glu
Polar	Ser, Thr, Asn, Gln, Cys
Small	Gly, Cys, Ser, Ala, Thr, Val
Long chain	Lys, Arg, Met, Glu, Gln
Beta-branched	Ile, Val, Thr
Proline	Pro
Glycine	Gly

The division is based on the Venn diagram, but it is modified to be able to visualise differences in the data set

Amino acid substitution in close homologues

Pairs of sequences in UniProt from species with at least 20°C difference in growth temperature were aligned, and retained only if the sequences were least 60% identical in the aligned region and this region covered at least 80% of each sequence. We used a pairwise alignment algorithm similar to Needleman and Wunsch (1970), using the PAM120 matrix (Dayhoff 1978) for scoring. To avoid an all against all pair-wise alignment of the sequences in UniProt, we compared only those, which have a common domain (cd) in the Pfam protein family database. To remove redundant sequences, the redundancy identification program cd-hit (Li and Godzik 2006) was used to cluster both the groups of ''hot'' and ''cold'' sequences using an 80% cut-off, and no homologous pair from the same pair of clusters was allowed. Only pairs where one of the proteins had a known structure were used. This resulted in 324 homologous pairs where both sequences were in the interval 0–40°C (psychrophiles-mesophiles) and 648 pairs where both sequences had a growth temperature of 35–102°C (mesophiles-thermophiles). The two groups were analysed separately.

Surface and core regions of the proteins were identified by calculating the solvent accessible surface area (SASA) using the DSSP program (Kabsch and Sander 1983). The maximal surface area differs significantly between the amino acids so the percentage of the area of the free amino acid was used. Amino acids with at least 20% exposure were considered to be on the surface, and those with 7% or less were considered to be in the core of the protein. The surface exposure was computed from the whole protein complex if there was more than one chain in the structure.

The result of this analysis were six substitution matrices $M_{a,b}$, where a is the amino acid in the coldest sequences, and b is the amino acid in the hottest. For each of the two temperature sets, we constructed one substitution matrix for the amino acids found to be on the surface and one for those found to be in the core. The matrices are included in the supplementary material.

Barplots

The barplot in Fig. 1 was made by summing all substitutions to and from each group of amino acids when going from the sequence adapted to the colder environment to the hotter. If there is no systematic temperature dependent shift, the two values will follow a binomial distribution with the same probability (P = 0.5) for a mutation to or away from the group. Using this as our null hypothesis, we used a binomial test to obtain a P-value for the observed substitutions. Considering the number of tests performed we set the P-value threshold to 0.001. The barplot itself show the percent-wise difference between mutations in each direction. If the group is most common in the species adapted to the hotter environment, the bar goes to the right, while if the opposite is true, the bar goes to the left. If no bar is present, then no statistical significant relationship was found. The underlying data can be found in the supplementary material.

Circle plots

The circle plots were computed in a similar way to the bar plots, but instead of looking at one group versus all other groups, the difference between all pairs of groups were calculated. This gave a binomial distribution in the same manner as for the bar plot described above, and we could do a binomial test in the same way. If the test rejected the null hypothesis, we concluded that there is a ‘‘flow’’ of amino acid from one group to the other when going from sequences adapted to a colder to a hotter environment, and we drew an arrow between the groups. On all the arrows, we printed the net flow between the groups. The first number in each circle representing an amino acid group, is the number of amino acids in that group in the species

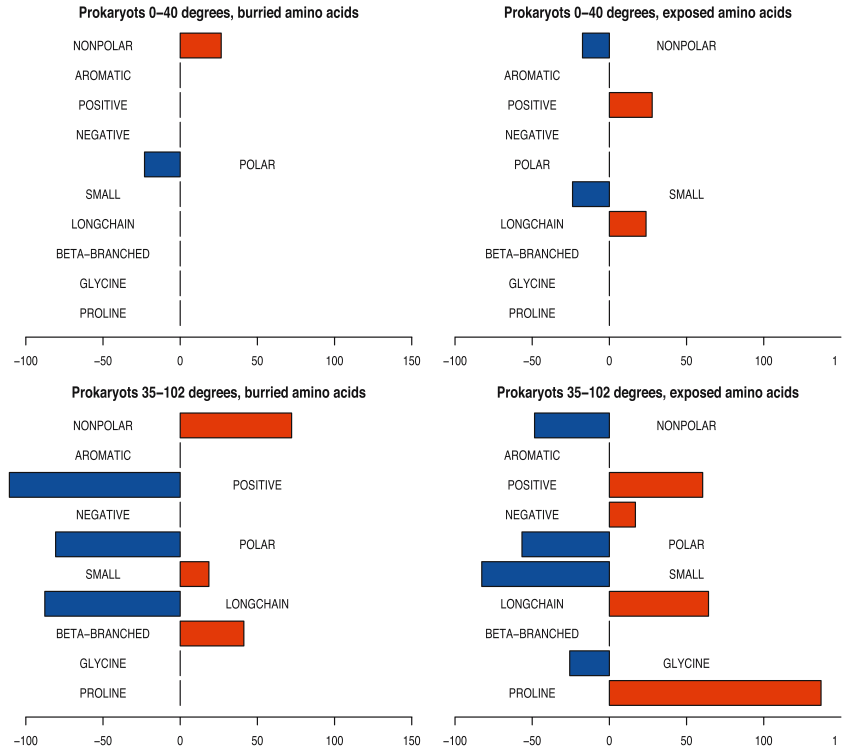


Fig. 1 Barplot of the relative of the different amino acid groups in sequences from species separated by 20°C or more in OGT. The analysis is divided into psychrophilic versus mesophilic comparison (0–40°C) and mesophilic versus thermophilic comparison (35–102°C). *Red right pointing bars* illustrate the preference in the species adapted to higher temperatures, and *blue, left-pointing* illustrate the preference in the species adapted to the colder environment. The plots on the left side is for amino acids in the core of the proteins, while those on the right side is for amino acids on the surface of the proteins

adapted to the colder environment, while the number of number in parentheses is the change when going to higher temperature. The threshold for the *P*-values was set to 0.0005, so that we could be reasonably sure that none of the arrows are false positives. The sizes of the arrows and the circles reflect the relative size of the flows and frequency of the groups, respectively.

Results

Correlation analysis

Table 2 shows a correlation analysis of growth temperature versus amino acid frequencies and their *P*-values. All the species were analysed together, but they were also separated into groups with OGT in the range 0–40°C, OGT in the range 35–102°C and archeal species. A positive correlation means that the frequency of the amino acid increases with higher temperatures, while a negative correlation means that it decrease with higher temperatures. If no significant correlation was found, the field was left blank. We considered *P*-values below 0.01 as significant.

There is a clear trend through the whole temperature range for some of the amino acids. The smaller amino acids like Ala and Gly seem to be decreasing in frequency as OGT increases while Ile, Glu and Lys are increasing. The frequencies of His decreases and Tyr increases as OGT increases in the bacterial sequences, but there is no significant trend for these amino acids in the archeal sequences. Ser, Thr, Asn and Gln seem to be less frequent at higher temperatures, but the trend does not extend to the lower end of the temperature scale. The differences are generally clearer between thermophiles and mesophiles than between mesophiles and psychrophiles, as can be seen from the higher correlation coefficients. Some of the amino acids have very different frequencies between species adapted to the same temperatures, potentially masking differences caused by temperature adaptation. This is especially the case for Arg, but also Met and to some extent Lys.

Substitution analysis of core and surface amino acids of close homologues

The substitution analysis of aligned pairs of close homologues from proteins adapted to temperatures at least 20°C apart resulted in a 20 × 20 asymmetric matrix where each element (*i*, *j*) represented the number of amino acids of type *i* in the species adapted to the lower temperature being substituted by *j* in the species adapted to the higher temperature. Surface and core regions of the protein sequences were identified on the basis of 3D structure data, and the regions were analysed separately. The analysis was also divided into two intervals, one where both homologues are from species with growth temperature in the range 0–40°C (psychrophile versus mesophile analysis), and one where both are from species with growth temperature in the range 35–102°C (mesophile versus termophile analysis). The amino acids were further divided into groups after biochemical properties as described in Table 1. The increase or decrease of the number of amino acids (in %) within a group when the temperature increases, is illustrated in Fig. 1. A right-pointing bar illustrate that the group is preferred in the species adapted to higher temperatures, and the left-pointing bars illustrate that the group is preferred in the species adapted to the colder environment. The data gives a slight indication that there is an inverse relationship between groups of amino acids preferred in the core and on the surface of the proteins. This relationship is not, however, proportional. This is not so obvious in the comparison between the psychrophilic and mesophilic proteins, where the trend only is observed for nonpolar amino acids. Nonpolar residues are preferred in the core of cold adapted proteins. The trend is much clearer in the comparison between mesophilic and thermophilic proteins. As for the psychrophilic proteins, nonpolar residues are more preferred in mesophilic proteins than in the thermophilic proteins, and long chains are less preferred. In addition, thermophilic proteins seem to find small amino acids less favourable, but acidic amino acids more favourable than mesophilic proteins. In thermophilic proteins, polar amino acids seem to be unfavourable both in the core and on the surface of the proteins compared to their mesophilic counterparts. We cannot see this effect when comparing psychrophiles and mesophiles.

Figure 2 shows the net substitution in the core going from psychrophiles to mesophiles. The core of cold adapted proteins has less nonpolar, less small and more polar amino acids than the mesophilic proteins, thus, making the core less hydrophobic. Most of the polar residues seem to be replaced by small and nonpolar amino acids when going towards higher temperatures.

Figure 3 illustrates the shift in amino acids frequency for the proteins on surface of psychrophilic and mesophilic species. The surface of the psychrophilic proteins has more of both polar and nonpolar residues than the mesophilic proteins. The cold adapted proteins also have more small amino acids on the surface than the more temperate proteins. The majority of the amino acids in the cold adapted proteins are substituted by longer amino acids in the mesophilic proteins. The surface of the mesophilic proteins also contains a significantly higher number of positively charged amino acids than the psychrophilic proteins do.

Table 2 Correlation between optimal growth temperature and amino acid frequency

Amino acids	Correlation All species	*P*-value All species	Correlation 0–40°C	*P*-value 0–40°C	Correlation 40–102°C	*P*-value 40–102°C	Correlation Archea	*P*-value Archea
Alanine	–0.31	2.2e-16	–0.29	6.32e-14	–0.47	1.16e-8		
Cysteine	–0.17	2.63e-6						
Aspartate					–0.55	2.2e-16	–0.53	3.81e-8
Glutamate	0.35	2.2e-16	0.28	9.3e-13	0.45	6.1e-8		
Phenylalanine								
Glycine	–0.19	1.4e-7	–0.21	3.2e-8				
Histidine	0.30	2.2e-16	–0.24	5.0e-10	–0.46	3.9e-8		
Isoleucine	0.28	4.8e-15	0.25	1.1e-10	0.50	2.7e-9		
Lysine	0.31	2.2e-16	0.36	2.2e-16	0.49	1.8e-14		
Leucine			0.16	7.2e-5			0.43	1.6e-5
Methionine								
Aspagine			0.30	1.5e-14				
Proline	–0.15	4.23e-5	–0.31	1.8e-15				
Glutamine	–0.21	8.9e-9			–0.71	2.2e-16	–0.56	3.1e-9
Arginine	–0.19	1.16e-7	–0.35	2.23-16				
Serine					–0.37	1.4e-5	–0.38	1.0e-4
Threonine	–0.14	4.5e-5			–0.61	2.2e-16	–0.38	1.0e-4
Valine								
Thryptophan			–0.28	7.6e-13			0.39	9.9e-5
Tyrosine	0.33	2.2e-16	0.25	1.5e-10	0.48	9.5e-9		

A positive correlation value means that the frequency of the amino acid increase with higher temperatures while a negative value means that the frequency decrease. Only *P*-values below 0.0001 were conceded significant, and thus, were listed in the table

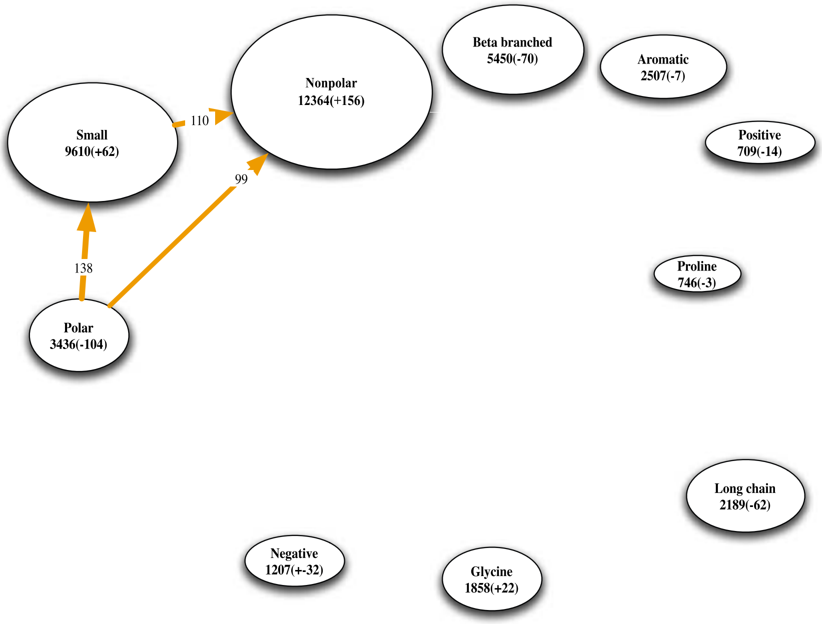


Fig. 2 Systematic shift in mutation frequency between different groups of amino acids when comparing the amino acids in the core of proteins from psycrophilic and mesophilic species. If there are significantly more amino acids with long-side chains than short-side chains in the homologues adapted to the warmer environment, there will be an *arrow* from ''small'' to ''longchain'' with the net difference in mutations going in each direction printed on the *arrow*. The first number, below the group name in the circle, shows the number of occurrences in the proteins adapted to the colder environment. The number in parenthesis is the difference to the proteins adapted to the warmer environment

The core of the thermophilic proteins is characterised by a slightly (relative values) higher number of small, polar and β-branched amino acids than the mesophilic proteins (Fig. 4), and a significantly lower number of long and polar amino acids. Most of the long amino acids in the mesophilic proteins are substituted by nonpolar amino acids in the thermophilic proteins. A significant number of polar amino acids in the mesophilic proteins are replaced by nonpolar amino acids, but also β-branched amino acids, in the thermophilic proteins. These substitutions will probably make the core of the thermophilic proteins more hydrophobic and more tightly packed than the mesophilic proteins.

The surface of the thermophilic proteins (Fig. 5) is characterised by higher numbers of large and charged amino acids than the mesophilic proteins. At the same time the number of both polar and nonpolar amino acids is lower. The majority of polar amino acids in the mesophilic proteins are replaced by long and positively charged amino acids in the thermophiles.

Discussion

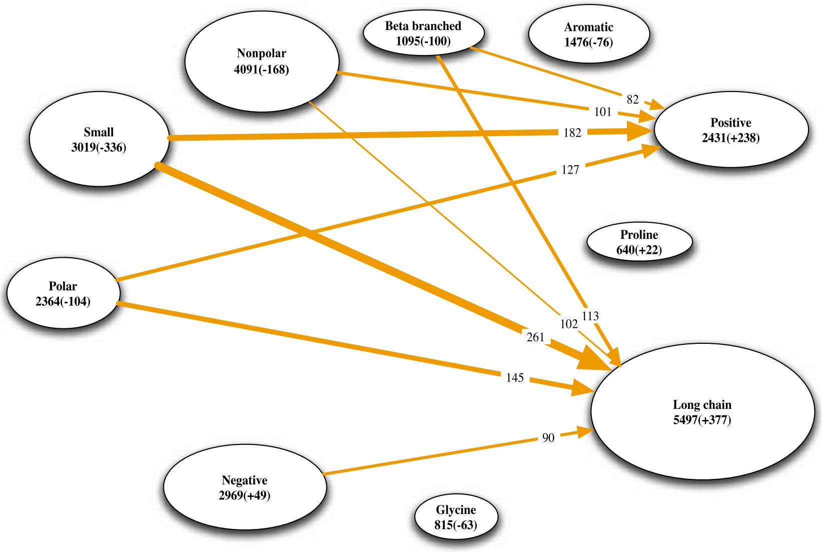
Nonpolar amino acids

It has been suggested that there are more nonpolar amino acids on the surface of the majority of cold-adapted proteins (see for example the review by Siddiqui and Cavicchioli 2006), and our results are in accordance with this. However, in the correlation analysis, only Ile and Ala have clear temperature dependent trends. Ile is more common at higher temperatures, while Ala is more common in cold-adapted proteins. Pack and Yoo (2004) found that long and branched side chains are favourable at higher temperatures, and this is also in accordance with our observations. We also observe that the core becomes more hydrophobic the higher the growth temperature of the organism is. Since the core in general is more hydrophobic than the surface in proteins, this means that the difference between the core and the surface, in terms of hydrophobicity, gets more pronounced the higher the temperature the species is adapted to.

Charged amino acids

The thermal motion of the protein may be high at elevated temperatures, and it is therefore essential that the proteins have a means of stabilising the 3D structure. Salt bridges have been suggested to play an important part in this (Kumar 2001). Thus, it would be expected that charged residues would be more frequently found on the surface of the proteins from the thermophilic species. The correlation analysis shows that this is the case for Glu and Lys residues throughout the whole data set, but not for Asp and Arg. However, there is a trend towards more charged amino acids as the temperature increases when considering only the amino acids on the surface, and this is particularly clear for the positively charged amino acids, both when comparing psychrophilic versus mesophilic proteins and mesophilic versus thermophilic proteins. It is surprising that the increase in negatively charged residues is not equal to the increase in positively charged residues, even though in the mesophilic group the positive and negative residues are practically balanced (Fig. 5). One would assume that, to form the maximum number of new salt bridges, the in-

Fig. 3 Systematic shift in mutation frequency between different groups of amino acids when comparing the amino acids on the surface of proteins from psycrophilic and mesophilic species. See legend for Fig. 2 or the method section for details



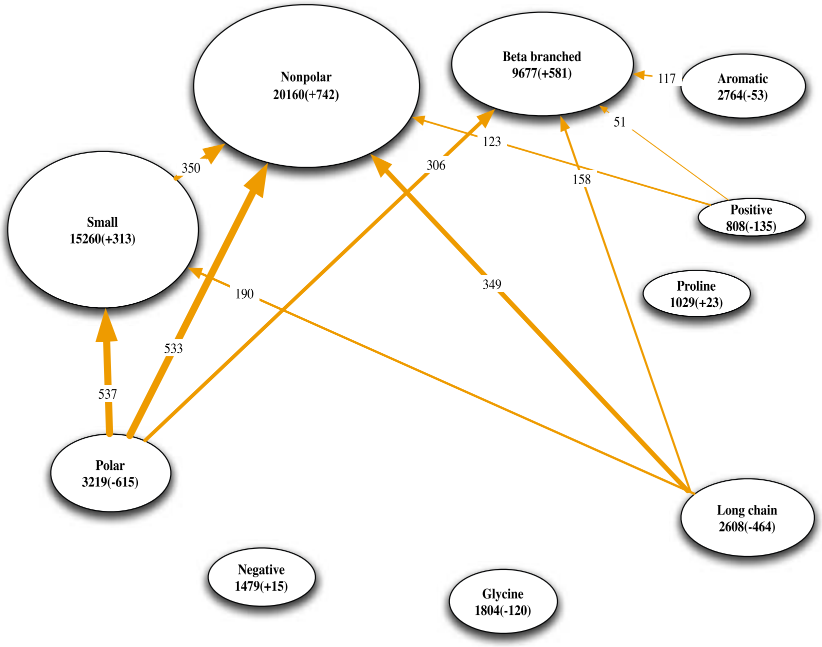


Fig. 4 Systematic shift in mutation frequency between different groups of amino acids when comparing the amino acids in the core of proteins from mesophilic and thermophilic species. See legend for Fig. 2 or the method section for details

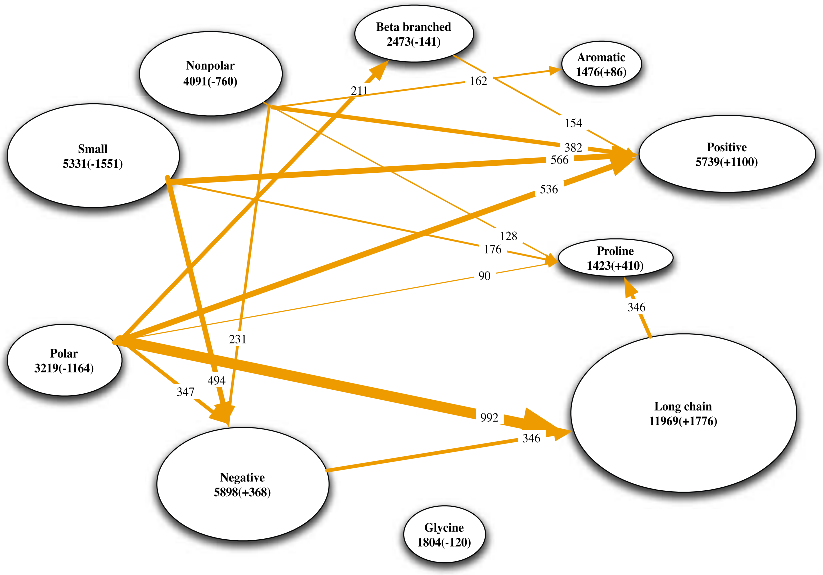


Fig. 5 Systematic shift in mutation frequency between different groups of amino acids when comparing the amino acids on the surface of proteins from mesophilic and thermophilic species. See legend for Fig. 2 or the method section for details

crease in positively and negatively charged residues would be equal. Positively charged residues can, however, favourable cation-π interactions with aromatic amino acids (Gromiha et al. 2002), and especially Arg is suggested to form such interactions. No significant increase in aromatic amino acids can be seen, but the correlation analysis show a significant increase in Tyr and a significant decrease of His residues, and opposite trends for these amino acids may be hidden by the choice of aromatic residue as one group. His cannot form cation-π bonds, and at pH above its pKa (normally 6–7) and up, it is uncharged and thus unable to form salt bridges. Berezovsky et al. (2007) suggested that charged residues could contribute to stabilisation by what he calls negative design. Negative design is when the unfolded and partially unfolded states are made less favourable. Charged residues of the same sign are repelling each other in partially unfolded states and thus making a bigger barrier for unfolding. In that case the positive and negative residues do not need to balance each other, and other favourable properties of positive residues, like their long chains will count. Also the entropy change for going from a folded surface position to an unfolded state is exceptionally low for Arg, and also quite low for Lys, but higher for Glu and Asp.

According to Lee et al. (2004), a direct substitution from Asp to Glu made the protein thioredoxin from *Escherichia coli* more stable, while the mutation in the opposite direction, from Glu to Asp, in the hyperthermophilic homologue from *Methanococcus jannaschii* made that protein less stable. In the core of the proteins negatively charged amino acids are less common in the homologue adapted to the highest temperature (Lee et al. 2004). Pack and Yoo (2004) observe a higher proportion of well-buried Glu residues. The data presented here do not support this hypothesis, as there are fewer charged residues in the protein core of thermophilic proteins.

Aromatic amino acids

Clusters of aromatic amino acids have been suggested to be important for stabilising protein structures, and especially for adaptation to hot environments (Kannan and Vishveshwara 2000; Brinda and Vishveshwara 2005). The correlation analysis and the substitution analysis we performed neither confirm nor reject this theory, but Tyr displays an increase in frequency as the temperature increase, while Phe show no correlation with OGT.

Glycine and proline

Glycine lacks a side chain and is therefore considered to be more flexible than the other amino acids due to more rotational freedom. The side chain of Pro, on the other hand, forms a ring structure that folds back on the main chain and restrain its movements and prevents the formation of hydrogen bonds to the main chain nitrogen atom. It has been suggested that cold-adapted enzymes are more flexible than their more temperate homologues to keep the activity on a high enough level (Feller et al. 1999; Georlette et al. 2004; Olufsen et al. 2005), and one way to make proteins more flexible is to let them contain more glycines and less prolines. There is a slight decrease in Gly with increasing temperatures, but not in the substitution analysis below 40°C. For Pro, the correlation analysis does not show a clear trend, but in the substitution analysis, Pro replaces other amino acids at a moderate level in thermophiles on the surface. Earlier studies of single protein families, have found increased Pro content in thermophiles (Watanabe et al. 1997; Haney et al. 1997; Bogin et al. 1997), but Kumar and Nussinov (2001) could not find any general trend of more Pro. Our substitution data suggest that there is a slight but significant replacement of other amino acids in favour of Pro from most of the amino acid groups in thermophiles, which is consistent with increased presence of Pro to stabilise surface loops.

Side chain size and branching

Amino acids with long side chains are generally believed be to better at forming van der Waal interactions with other amino acids due to the longer (and nonpolar and aliphatic) chains, and hence, stabilise the protein. Amino acids with longer side chains may therefore be favourable for thermostabilisation. Our data suggest that this is the case, and we observe this mainly on the surface, where the long-chained amino acids are clearly more common in the proteins adapted to the hotter environment. The opposite effect is, on the other hand, seen for amino acids in the core, where the long chains seem to be more frequent in the proteins adapted to the colder environment. The increase in Arg and Lys can be explained by the potential favourable interactions their long aliphatic chains can be involved in, as well as their positive charge. Small amino acids are also less favoured on the surface of proteins from the species adapted to the warmest environment. The length of the side chains seems to be of less importance in the core of the proteins, but small amino acids seems to be favoured over long chained amino acids. β-branched amino acids also appear to be favourable in the core of thermophilic proteins. The entropy difference between an unfolded, unburied residue configuration and a buried, folded situation, is low for β-branched amino acids compared to all other residues (D'Aquino et al. 1996). Thus, the -TΔS term in the Gibb's free energy equation favours these residues at temperatures where thermophiles must be stable. Additionally, β-branched amino acids can improve the packing interactions, and hence, providing a more dense and stable core.

Polar amino acids

One clear trend in our data set is that polar amino acids are less favoured in thermophiles than in mesophiles. This is apparent both on the surface and in the core. The same trend is to some extent seen when comparing the core amino acids of psychrophilic and mesophilic proteins, where polar residues are less favoured in the mesophiles. The polar amino acids in the core of the psychrophilic (psychrophilic versus mesophilic) and mesophilic (mesophilic versus thermophilic) proteins are generally replaced by nonpolar amino acids in both studies, and to some extent small amino acids in the psychrophilic versus mesophilic study and β-branched in the mesophilic versus thermophilic study. The polar surface proteins in the psychrophilic proteins are primarily replaced by long and positively charged amino acids. There is not an equally clear trend for the polar surface amino acids of mesophiles since these are replaced by amino acids from almost all groups in the thermophilic proteins, but also there is a large portion of the amino acids replaced by large and positively charged amino acids.

Concluding remarks

In this work we report a statistically significant relationship between growth temperature and amino acid frequency. Only a few amino acids have a direct correlation between frequency and growth temperature over the whole range in this study; Gly and Ala decrease while Lys, Ile and Glu increase when the temperature increases. However, dividing the amino acids into groups with similar biochemical properties and considering buried and exposed amino acids separately, comparison of psychrophilic versus mesophilic and mesophilic versus thermophilic reveals clearer results. It should be noted though that the assignment of residues belonging to the core or the surface may be somewhat inaccurate because some of the monomeric proteins deposited in the PDB are subunits of larger protein complexes, other proteins appear to be oligomeric, maybe due to crystal packing interactions, and in some of the proteins the biological entity is still unresolved, so we assume that these ''errors'' will cancel each other in a larger data set, and hence, will not significantly alter the results observed in this work. Since almost all of the proteins in this study are soluble proteins rather than membrane proteins, the result if this study will reflect the trends in soluble proteins.

Perhaps the most striking observation is that the number of nonpolar amino acids always is higher in the core of the proteins adapted to the higher temperature, and that the opposite effect is seen for the surface amino acids where the proteins adapted to lower temperatures have more nonpolar amino acids. A similar, but opposite, effect is seen for long amino acids, where the long chains are more frequent on the surface of the proteins adapted to the warmer environment and the long chains are more frequent in the core of the cold adapted proteins. There seems to be a clear tendency of more charged amino acids on the surface of thermophilic proteins than mesophilic proteins, but a similar trend with less charged amino acids in the psychrophilic proteins is not significant.

Our data suggests that an important adaptation to temperature is to make the core of the protein more hydrophobic with higher temperatures. The hydrophobic effect is thought to be the main driving force in protein folding, as well as keeping the protein correctly folded. The hydrophobic effect is caused by the water molecules tendency to form hydrogen bonds with themselves, rather than having interactions with aliphatic or aromatic groups. The enthalpic and entropic forces driving this process is near balancing each other at all temperatures, but experimental evidence (Southall and Dill 2002; Xu and Dill 2005), suggest that the hydrophobic effect is strongest around room temperature. In that case a more hydrophobic core at higher temperatures is necessary just to compensate the weakened hydrophobic forces. The surface also features more long-chained and less small residues in proteins from species with higher OGT, and this may also stabilise the protein by stronger van der Waal forces. One could think that this would also be favourable in the core, but we see no such effect, but rather a weak trend in the opposite direction. Charged amino acids replace other amino acids on the surface of the thermophilic proteins. This can either result in more stabilising salt bridges, cation-π bonds with aromatic amino acids or favourable interactions with the solvent. Some of the charged amino acids have, in addition, long aliphatic side chains that can form favourable van der Waal interactions with other amino acids. Especially Arg is often replacing other amino acids in themophiles, while the opposite trend can be seen in cold adapted proteins. In an experiment on alpha amylase from *Pseudoalteromonas haloplanktis*, (Siddiqui et al. 2006) found that replacing lysine with homoarginene, increased the thermostability of the protein. Homoarginine is chemically similar to arginine, except that the aliphatic chain is one carbon longer. Moreover, placing Arg in a surface position is especially favourable from an entropic point of view. Compared to all other residues, the entropic cost of folding a residue into this position is the lowest among all amino acids (D'Aquino et al. 1996).

In recent years there has been a debate whether there is a correlation between GC content and OGT. It would be natural to think it is, since GC-pairs have three hydrogen bonds, while AT-pairs have only two, and the former would be more stable at elevated temperatures. Studies performed so far display opposite conclusions. Galtier and Lobry (1997) and Hurst and Merchant (2001) conclude that GC-content is not an adaptation to high temperatures, while others (Musto et al. 2004) more recently have concluded that it is. Wang et al. (2006) and later Musto et al. (2006) have recently made reviews with opposite conclusions. Regardless of whether high GC-content is an adaptation to high temperatures or not, the variation due to other factors is significant, and can vary from below 30 to above 70%, even between species with similar OGT. Many of the other potential factors are described in the reviews mentioned above and their references.

What is agreed upon is that GC-content will influence amino composition. Kreil and Ouzounis (2001) found that GC-content was the most influential factor on amino acid composition, with OGT as the second most important. If GC content is important for adaptation to high temperatures, this means that some of the selective pressure is on the nucleotide level, not only on protein level, and must be taken into account when studying the amino acid composition. Hence, the amino acids with GC-rich codons (Phe, Leu, Ile, Asn, Lys and Tyr) should be more common in thermophiles, while those with low GC-content (Gly, Ala, Pro and Arg) should be less common. This is to some extent what is seen in our data since Ile and Lys are more frequent at higher temperature while Gly and Ala are less frequent. The case of Arg is particularly interesting, since this amino acid has biophysical properties that would make it favourable in thermophiles, but low GC-content will make it less favourable. This may be the reason that we see a clear trend of Arg substitute other amino acids, while the plot of Arg content against OGT (supplementary material) shows that Arg frequency is quite varied between species regardless of OGT. Still, our data display low correspondence between GC-rich codons and the amino acid frequency at higher temperatures, and it is therefore not clear how strong the selective pressure is on the nucleic acid level. It should be noted though that we have not analysed the GC-content of the genes of the proteins in this study. In general the correlation analysis is more prone to the effect of external factors like GC-content and sequencing bias than the substitution analysis. Whole genome analyses could possibly improve this, as done by Zeldowich et al. (2007).

The prevalence of long-chained, charged amino acids in thermophilic proteins can be understood in terms of folding thermodynamics. The -TΔS of folding for a protein will be more destabilising at elevated temperatures, since the ΔS-term is negative for an unfolded-folded transition. In other words, restricting a residue into a protein fold incurs a higher entropic penalty per residue at higher temperatures compared to lower temperatures. This cost must be compensated by a greater per-residue enthalpic stabilisation. Thus, from the discussion of the previous paragraph, there are multiple ways in which the long-chained, positively charged residues are able to lower the enthapic term in the ΔG of folding for a protein. We propose that favouring the long-chained, charged residues and also removing residues not strictly necessary for the protein fold represent two strategies for thermophilic proteins to maintain a negative ΔG of folding at elevated temperatures. Removal of residues is reported in a study by Thompson and Eisenberg (1999), which reported that many thermophilic proteins have shorter loops than the mesophilic counterpart.

In our data, the differences between mesophiles and psychrophiles are less dramatic than the differences between mesophiles and thermophiles. This may be partly due to the smaller difference in temperature, but other factors are likely to play a role. Our analysis focuses on the temperature-dependent difference in amino acid composition and substitution patterns, and the effect of other forces affecting the selection of amino acids are largely averaged out. The patterns that we do see are consistent with the suggestion that the main force in selection of amino acids in proteins adapted to high temperatures is driven by the need for structure stabilisation, while proteins adapted to colder environments need to modify binding or interaction sites and the flexibility at the active site or other localised areas. Stabilisation affects the whole protein, and is more likely to be detected by the methods used in this study. To address the degree and nature of local changes in the vicinity of active sites, other methods are needed. The study is also unable to detect whether, e.g. an Arg residue really is a part of a salt bridge, and a study that take amino acid contacts into account may reveal trends that are undetected in this study.

Another important factor in temperature adaptation may be the packing of the protein core. Several studies (Schumann et al. 1993; Korkegian et al. 2005; Pack and Yoo 2005) have demonstrated that replacing one aliphatic amino acid with another can improve the packing and thus the thermostability. However, such improvements are dependent on the existent packing in the wildtype protein, and may not cause any systematic shift in amino acid frequency or substitution patterns, except possibly a slight preference for branched amino acids.

Acknowledgements Thanks to William R. Taylor and Trond H. Bø for valuable input. This research was funded by the Norwegian functional genomics program (FUGE) of the Research Council of Norway, through the Envision project and through the technology platform for bioinformatics.

References

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS (2004) UniProt: the universal protein knowledgebase. Nucleic Acids Res 32(Database issue):D115–D119

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR (2004) The Pfam protein families database. Nucleic Acids Res 32(Database issue):D138–D141

Berezovsky IN, Shakhnovich EI (2005) Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci USA 102(36):12742–12747

Berezovsky IN, Zeldovich KB, Shakhnovich EI (2007) Positive and negative design in stability and thermal adaptation of natural proteins. PLOS Comput Biol doi:10.1371/journal.pcbi.0030052.eor

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242

Bogin O, Peretz M, Burstein Y (1997) Thermoanaerobacter brockii alcohol dehydrogenase: characterization of the active site metal and its ligand amino acids. Protein Sci 6(2):450–458

Brinda KV, Vishveshwara S (2005) Oligomeric protein structure networks: insights into protein-protein interactions. BMC Bioinformatics (6):296

D'Aquino JA, Gomez J, Hilser VJ, Lee KH, Amzel LM, Freire E (1996) The magnitude of the backbone conformational entropy change in protein folding. Proteins 25(2):143–156

Dayhoff SA (1978) Atlas of protein sequence and structure. National Biomedicine Research Foundation, Washington, DC, USA

Deming JW (2002) Psychrophiles and polar regions. Curr Opin Microbiol 5(3):301–309

Farias ST, Bonato MC (2003) Preferred amino acids and thermostability. Genet Mol Res 2(4):383–393

Feller G, d'Amico D, Gerday C (1999) Thermodynamic stability of a cold-active alpha-amylase from the Antarctic bacterium Alteromonas haloplanctis. Biochemistry 38(14):4613–4619

Feller G, Gerday C (2003) Psychrophilic enzymes: hot topics in cold adaptation. Nat Rev Microbiol 1(3):200–208

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44(6):632–636

Georlette D, Blaise V, Collins T, D'Amico S, Gratia E, Hoyoux A, Marx JC, Sonan G, Feller G, Gerday C (2004) Some like it cold: biocatalysis at low temperatures. Fems Microbiol Rev 28(1):25–42

Gianese G, Argos P, Pascarella S (2001) Structural adaptation of enzymes to low temperatures. Protein Eng 14(3):141–148

Gromiha MM, Thomas S, Santhosh C (2002) Role of cation-pi interactions to the stability of thermophilic proteins. Prep Biochem Biotechnol 32(4):355–362

Haney P, Konisky J, Koretke KK, Luthey-Schulten Z, Wolynes PG (1997) Structural basis for thermostability and identification of potential active site residues for adenylate kinases from the archaeal genus Methanococcus. Proteins 28(1):117–130

Huang SL, Wu LC, Liang HK, Pan KT, Horng JT, Ko MT (2004) PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics 20(2):276–278

Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 268(1466):493–497

Jaenicke R, Bohm G (1998) The stability of proteins in extreme environments. Curr Opin Struct Biol 8(6):738–748

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Kannan N, Vishveshwara S (2000) Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng 13(11):753–761

Kashefi K, Lovley DR (2003) Extending the upper temperature limit for life. Science 301(5635):934

Korkegian A, Black ME, Baker D, Stoddard BL (2005) Computational thermostabilization of an enzyme. Science 308(5723):857–860

Kreil DP, Ouzounis CA (2001) Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 29(7):1608–1615

Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat. Cell Mol Life Sci 58(9):1216–1233

Kumar S, Tsai CJ, Nussinov R (2001) Thermodynamic differences among homologous thermophilic and mesophilic proteins. Biochemistry 40(47):14152–14165

Lee DY, Kim KA, Yu YG, Kim KS (2004) Substitution of aspartic acid with glutamic acid increases the unfolding transition temperature of a protein. Biochem Biophys Res Commun 320(3):900–906

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659

Li WF, Zhou XX, Lu P (2005) Structural features of thermozymes. Biotechnol Adv 23(4):271–281

Marx JC, Blaise V, Collins T, D'Amico S, Delille D, Gratia E, Hoyoux A, Huston AL, Sonan G, Feller G, Gerday C (2004) A perspective on cold enzymes: current knowledge and frequently asked questions. Cell Mol Biol 50(5):643–655

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573(1–3):73–77

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun 347(1):1–3

Nakashima H, Fukuchi S, Nishikawa K (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J Biochem (Tokyo) 133(4):507–513

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Olufsen M, Smalas AO, Moe E, Brandsdal BO (2005) Increased flexibility as a strategy for cold adaptation: a comparative molecular dynamics study of cold- and warm-active uracil DNA glycosylase. J Biol Chem 280(18):18042–18048

Pack SP, Yoo YJ (2004) Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. J Biotechnol 111(3):269–277

Pack SP, Yoo YJ (2005) Packing-based difference of structural features between thermophilic and mesophilic proteins. Int J Biol Macromol 35(3–4):169–174

R-core-team (2006) R: a language and environment for statistical computing. http://www.r-project.org. Wien, R foundation for statistical computing

Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 119(3):256–270

Schumann J, Bohm G, Schumacher G, Rudolph R, Jaenicke R (1993) Stabilization of creatinase from Pseudomonas putida by random mutagenesis. Protein Sci 2(10):1612–1620

Siddiqui KS, Cavicchioli R (2006) Cold-adapted enzymes. Annu Rev Biochem 75:403–433

Siddiqui KS, Poljak A, Guilhaus M, De Francisci D, Curmi PM, Feller G, D'Amico S, Gerday C, Uversky VN, Cavicchioli R (2006) Role of lysine versus arginine in enzyme cold-adaptation: modifying lysine to homo-arginine stabilizes the cold-adapted alpha-amylase from Pseudoalteramonas haloplanktis. Proteins 64(2):486–501

Singer GA, Hickey DA (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17(11):1581–1588

Singer GAC, Hickey DA (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 317(1–2):39–47

Southall NT, Dill KA (2002) Potential of mean force between two hydrophobic solutes in water. Biophys Chem 101–102:295–307

Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119(2):205–218

Thompson MJ, Eisenberg D (1999) Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J Mol Biol 290(2):595–604

Wang HC, Susko E, Roger AJ (2006) On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem Biophys Res Commun 342(3):681–684

Watanabe K, Hata Y, Kizaki H, Katsube Y, Suzuki Y (1997) The refined crystal structure of bacillus cereus oligo-1,6-glucosidase at 2.0 A resolution: structural characterization of proline-substitution sites for protein thermostabilization. J Mol Biol 269(1):142–153

Xu H, Dill KA (2005) Water's hydrogen bonds in the hydrophobic effect: a simple model. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 109(49):23611–23617

Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3(1):e5