Temperature Adaptation at Homologous Sites in Proteins from Nine Thermophile–Mesophile Species Pairs

John H. McDonald*

Department of Biological Sciences, University of Delaware

*Corresponding author: E-mail: mcdonald@udel.edu.

Accepted: 2 May 2010

Abstract

Whether particular amino acids are favored by selection at high temperatures over others has long been an open question in protein evolution. One way to approach this question is to compare homologous sites in proteins from one thermophile and a closely related mesophile; asymmetrical substitution patterns have been taken as evidence for selection favoring certain amino acids over others. However, most pairs of prokaryotic species that differ in optimum temperature also differ in genome-wide GC content, and amino acid content is known to be associated with GC content. Here, I compare homologous sites in nine thermophilic prokaryotes and their mesophilic relatives, all with complete published genome sequences. After adjusting for the effects of differing GC content with logistic regression, 139 of the 190 pairs of amino acids show significant substitutional asymmetry, evidence of widespread adaptive amino acid substitution. The patterns are fairly consistent across the nine pairs of species (after taking the effects of differing GC content into account), suggesting that much of the asymmetry results from adaptation to temperature. Some amino acids in some species pairs deviate from the overall pattern in ways indicating that adaptation to other environmental or physiological differences between the species may also play a role. The property that is best correlated with the patterns of substitutional asymmetry is transfer free energy, a measure of hydrophobicity, with more hydrophobic amino acids favored at higher temperatures. The correlation of asymmetry and hydrophobicity is fairly weak, suggesting that other properties may also be important.

Key words: thermophiles, amino acids, substitutional asymmetry.

Introduction

Thermophilic organisms live at 50 °C to over 100 °C, temperatures that would quickly denature most proteins from mesophiles. There is considerable interest in determining what enables proteins from thermophiles to function at high temperatures, both for the practical benefit of engineering proteins for high-temperature industrial processes and as an evolutionary and biochemical puzzle.

One way to investigate whether some amino acids are more favorable than others at higher temperature is to compare the overall proportions of amino acids in protein sequences from prokaryotes living at different temperatures (Cambillau and Claverie 2000; Fukuchi and Nishikawa 2001; Chakravarty and Varadarajan 2002; Singer and Hickey 2003; Berezovsky et al. 2007). An amino acid that is more abundant in species living at higher temperatures is then interpreted to be adaptive to the higher temperatures. However, a major problem with this approach is that prokaryotes vary widely in genome-wide GC content, and amino acids with GC-rich codons are generally more abundant in organisms with GC-rich genomes (Lobry 1997; Singer and Hickey 2000). There is conflicting evidence about whether genome-wide GC content shows any relationship with habitat temperature (Musto et al. 2006; Wang et al. 2006), but the strong association of GC content and amino acid abundance will obscure any relationship between temperature and amino acid abundance if the variation in GC content is ignored.

The effects of temperature and GC content can be separated using multivariate statistical techniques, such as principal component analysis (Kreil and Ouzounis 2001; Saunders et al. 2003), correspondence analysis (Tekaia et al. 2002; Lobry and Chessel 2003; Tekaia and Yeramian 2006; Boussau et al. 2008; Puigbò et al. 2008), and other techniques (Naya et al. 2006; Zeldovich et al. 2007). However, these approaches suffer from "phylogenetic pseudoreplication"; they treat multiple species from the same clade and same habitat as if they were independent samples, and it has long been known that this can cause serious statistical problems (Felsenstein 1985; Harvey and Pagel 1991). To illustrate why this is a problem, imagine biologists who were interested in temperature adaptation of terrestrial vertebrates. If those biologists surveyed vertebrates from a variety of habitats and looked for associations with temperature, they would see a higher proportion of species that shed their skin living in warmer areas. However, it would be erroneous to conclude from this that shedding skin is an adaptation to high temperature; the association would merely result from sampling large numbers of Squamata (lizards and snakes) in warm areas and few squamates in cold areas. Similarly, in studies of temperature and amino acid composition, some clades are found predominantly among thermophiles, and some are predominant among mesophiles; for example, of the 204 species studied by Zeldovich et al. (2007), 63% of the thermophiles and 5% of the mesophiles are archaea, whereas 0% of the thermophiles and 54% of the mesophiles are proteobacteria. A multivariate statistical technique that treated each species as an independent data point could produce an apparent association of particular amino acids with higher temperatures, when in reality that association might result from a difference between clades that may have nothing to do with temperature.

A second form of evidence used to compare amino acid composition in mesophiles and thermophiles is substitutional asymmetry (Argos et al. 1979; Haney et al. 1999; McDonald et al. 1999). Protein sequences from one mesophile and one thermophile are aligned, and the observation of more aligned sites with amino acid A in the mesophile and B in the thermophile than the opposite pattern provides evidence that B is favored over A in the higher temperature organism. Because only aligned sites in homologous proteins are considered, the effect of gain or loss of proteins of different amino acid composition does not obscure the results. In addition, each mesophile–thermophile pair of species can be phylogenetically independent of others that have been compared, an important consideration when using comparative methods to infer adaptation. (To say that mesophile–thermophile pair A and B are "phylogenetically independent" of other pairs means that A and B are more closely related to each other than either is to any of the other species in the data set.) This approach has found extensive evidence for substitutional asymmetry (Haney et al. 1999; McDonald et al. 1999; McDonald 2001; Nishio et al. 2003; Mizuguchi et al. 2007), but the problem remains that for those pairs of amino acids whose codons have different GC content, overall differences in GC content between the mesophile and thermophile could still be the cause of substitutional asymmetry. Here, I use logistic regression of the proportion of substitutions in one direction versus the overall difference in GC content to predict the substitutional asymmetry in a pair of species with identical genomic GC content. This method should help determine whether amino acids that are favored at higher temperatures share biochemical properties.

If substitutional asymmetry between mesophilic and thermophilic proteins results from temperature adaptation based on the fundamental biochemical properties of the amino acids, the same patterns should be found in all mesophile–thermophile comparisons after controlling for differences in GC content. Differences in other aspects of the environment, such as salinity, hydrostatic pressure, pH, oxygen, and nutrient source, could cause patterns of asymmetry that are unrelated to temperature and therefore different in different mesophile–thermophile pairs. In addition, biosynthetic costs of amino acids are high enough to cause selection on amino acid usage (Akashi and Gojobori 2002; Seligmann 2003; Heizer et al. 2006; Swire 2007), so organisms which differ in biosynthetic pathways, or which differ in whether they are autotrophic or heterotrophic for a particular amino acid, may have different patterns of substitutional asymmetry. A second goal of this paper is to see how consistent the patterns of substitutional asymmetry are among different species, which may help determine how much of the asymmetry is due to temperature adaptation and how much is due to other factors.

Materials and Methods

Choice of Mesophile–Thermophile Pairs The NCBI Entrez Genome Project database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj) was searched for thermophilic archaea and bacteria (optimum growth temperature, T_{opt}, greater than or equal to 50 °C) with complete, published genome sequences. Species from higher taxa in which all species with published genomes are thermophiles, such as Aquificae and Crenarchaeota, were excluded. The closest mesophile ($T_{opt} \leq 40$ °C) with a complete published genome sequence was identified for each thermophile using published phylogenies. Where a thermophile had more than one mesophile that was equally closely related or vice versa, the species pair was chosen with the most similar habitat, physiology, and genomic GC content. Where more than one strain of a species had been sequenced, the strain with the earliest published sequence was used. Nine phylogenetically independent pairs of mesophiles with thermophiles were identified (table 1); at the time the database was searched, there were no other mesophile–thermophile species pairs with published genomes that were phylogenetically independent of the nine used here.

Identification and Alignment of Homologous Proteins For seven of the mesophile–thermophile pair of species, the Entrez Gene Plot function (http://www.ncbi.nlm.nih.gov/sutils/geneplot.cgi) was used to obtain a list of reciprocal best matches of protein sequences. Each list was sorted, and where a sequence from one species had multiple best matches from the other species (which can

Table 1
Species Pairs Used in This Study

Species	T_{opt}	GC	Genome Reference
Sulfurovum sp. NBC37-1	33	43.8	Nakagawa et al. (2007)
Nitratiruptor sp. SB155-2	55	39.7	Nakagawa et al. (2007)
Streptomyces avermitilis	26	70.7	Omura et al. (2001)
Thermobifida fusca	50–55	67.5	Lykidis et al. (2007)
Methanococcus maripaludis	35–40	33.1	Hendrickson et al. (2004)
Methanocaldococcus jannaschii	85	31.4	Bult et al. (1996)
Deinococcus radiodurans	30–37	67.0	White et al. (1999)
Thermus thermophilus	68	69.4	Henne et al. (2004)
Desulfitobacterium hafniense Y51	37	47.4	Nonaka et al. (2006)
Pelotomaculum thermopropionicum	55	53.0	Kosaka et al. (2008)
Synechocystis sp. PCC6803	26	47.7	Kaneko et al. (1996)
Thermosynechococcus elongatus	55	53.9	Nakamura et al. (2002)
Bacillus subtilis	25–35	43.5	Kunst et al. (1997)
Geobacillus kaustophilus	60	52.1	Takami et al. (2004)
Clostridium tetani	37	28.7	Bruggeman et al. (2003)
Thermoanaerobacter tengcongensis	75	37.6	Bao et al. (2002)
Methanosphaera stadtmanae	36–40	27.6	Fricke et al. (2006)
Methanothermobacter thermautotrophicus	65–70	49.5	Smith et al. (1997)

NOTE.—GC, GC content of the major chromosome (excluding plasmids and extrachromosomal elements). T_{opt} and GC from the NCBI Genome Project database, except T_{opt} for *Sulfurovum* and *Nitratiruptor* (Nakagawa et al. 2007); *Desulfitobacterium* (Suyama et al. 2001), *Geobacillus* (Takami et al. 2004), and *Synechocystis* (growth temperature recommended by the American Type Culture Collection). T_{opt}, optimum growth temperature.

happen when there are multiple identical protein sequences), all but one of the matching pairs were deleted. Proteins encoded by small extrachromosomal elements in *Methanocaldococcus jannaschii* or plasmids in the other species were deleted.

For the *Pelotomaculum thermoproprionicum* versus *Desulfitobacterium hafniense* and *Nitratiruptor* versus *Sulfurovum* comparisons, Geneplot was not available. I therefore used Blast to obtain a list of the best match for each protein sequence in the other species and then sorted the two lists in a spreadsheet to identify the reciprocal best matches.

No attempt was made to eliminate proteins whose genes may have been acquired recently by horizontal gene transfer (HGT). Whether a gene could be identified as acquired through HGT would depend on how divergent the source species was and whether its sequences were available; therefore, painstaking investigation of each gene would only result in eliminating some, but not all, such genes. Leaving genes acquired through HGT in the data set would tend to obscure patterns of consistent substitutional asymmetry by introducing noise into the data rather than creating patterns by statistical artifacts that would not be there otherwise.

The complete set of protein sequences was downloaded from Entrez Genome for each species, and a Pascal program was written to use the list of reciprocal best matches, create a file for each pair of protein sequences, extract the protein sequences, and put them in the appropriate files.

Each pair of protein sequences was aligned using ClustalW (Chenna et al. 2003), with the default parameters. Protein pairs with less than 35% identical sites and proteins less than 20 amino acids long were deleted. Ambiguously aligned sites adjacent to gaps were then omitted, with the omitted sites extending from the gap to the nearest pair of adjacent sites that were both identical in the two sequences, using the program AmbiguityRemover. The number of unambiguously aligned sites exhibiting each of the 190 possible pairwise patterns of difference was then counted using the program AsymmetryCounter. Both programs are available for download from http://udel.edu/~mcdonald/asymmetry.html.

Statistical Analysis For each pair of amino acids in each pair of species, the exact binomial test (for $N < 1,000$; McDonald 2009, p. 24–32) or G-test of goodness-of-fit (for $N > 1,000$; McDonald 2009, p. 46–51) was used to test the significance of the deviation from the expected 1:1 ratio.

To distinguish between asymmetry resulting from genomic GC differences and asymmetry due to other causes, the LOGISTIC procedure of SAS (SAS Institute 2009) was used to perform logistic regression for each pair of amino acids, with the difference in genomic GC content between the thermophile and the mesophile as the independent variable and the proportion of substitutions in one direction as the dependent variable. Logistic regression (McDonald 2009,

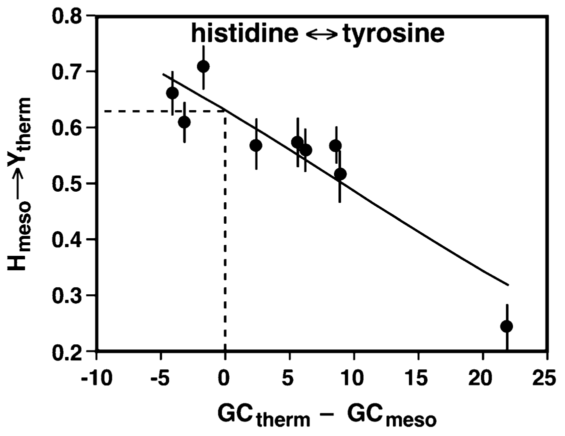

Fig. 1.—Example of logistic regression of substitutional asymmetry and difference in GC content. $GC_{therm} - GC_{meso}$, the percent difference in GC content between the thermophile and the mesophile in each species pair. $H_{meso} \rightarrow Y_{thermo}$, the proportion of sites in each species pair that have histidine in the mesophile and tyrosine in the thermophile, as a proportion of all aligned sites that have histidine in one species and tyrosine in the other. Error bars are 95% confidence intervals of the binomial proportion. The solid line is the logistic regression line, given by solving $\ln[Y/(1 - Y)] = a + bX$ for Y, where Y is $H_{meso} \rightarrow Y_{thermo}$, X is $GC_{therm} - GC_{meso}$, a is the intercept, and b is the slope. The dashed line shows the estimation of the expected asymmetry in a species pair with zero difference in GC content.

p. 247–255) finds the best-fitting equation of the form $\ln[Y/(1 - Y)] = a + bX$, where Y is the probability of obtaining a particular value of a nominal variable for a given value of the measurement variable, a is the intercept, b is the slope, and X is the value of the measurement variable. For example, the logistic regression equation for the amino acids histidine and tyrosine (fig. 1) predicts the probability (Y) that a histidine/tyrosine site has histidine in the mesophile and tyrosine in the thermophile for any value of X, the difference in GC content between two species. The significance of the slope was used to test whether there was a significant relationship between the difference in GC content and the pattern of asymmetry. The significance of the intercept was used to test whether the predicted asymmetry for a mesophile–thermophile pair with equal GC contents was significantly different from the 1:1 ratio expected under the neutral model of molecular evolution.

To identify amino acids that deviated from the overall pattern in particular species pairs, the residual (difference between the observed proportion of substitutions in one direction and the proportion predicted by the logistic regression model) was calculated for each amino acid pair in each species pair and then averaged across the 19 pairs involving each amino acid. For this analysis, the proportion of sites with the target amino acid in the thermophile and the other amino acid in the mesophile was used.

Amino Acid Properties The logistic regression equation for each pair of amino acids was used to predict the expected proportion of substitutions in each direction in a hypothetical species pair that did not differ in GC content. These predicted proportions were multiplied by the total number of substitutions across the nine species pairs for that amino acid pair to yield a synthetic data set. The AAindex list of amino acid indices (Kawashima et al. 2008) was downloaded from http://www.genome.ad.jp/dbget/aaindex.html. Indexes that measure the propensity of amino acids to occur in particular proteins or parts of proteins were deleted, as were those with missing or estimated values. For each index, the difference between the values of the index for each pair of amino acids was used as the independent variable in a simple logistic regression. The dependent variable was taken from the synthetic data set, the expected number of substitutions in each direction in a species pair that does not differ in GC content.

Results

Extensive Substitutional Asymmetry Related to Difference in GC Content There is extensive substitutional asymmetry; of the 1,710 total comparisons (190 pairs of amino acids in nine species pairs), 1,038 are significantly ($P < 0.05$) different from the expected 1:1 ratio (supplementary table 1, Supplementary Material online). Each of the 190 pairs of amino acids is significantly asymmetrical in at least one of the nine species pairs, and 125 of the pairs of amino acids are asymmetrical in at least five species pairs.

Some of the asymmetry is associated with differences in GC content. Of the 190 pairs of amino acids, 153 differ in average GC content of their codons (e.g., histidine [H] has an average of 1.5 GC in its codons [CAC, CAT] vs. tyrosine [Y], which has an average of 0.5 GC in its codons [TAC, TAT]). The logistic regression of substitutional asymmetry versus difference in genome-wide GC content has a significant slope for 122 out of these 153 pairs of amino acids (supplementary table 2, Supplementary Material online), indicating that the proportion of substitutions in each direction depends on the difference in genome-wide GC content. Figure 1 shows an example of this; the proportion of H ↔ Y sites with H in the mesophile and Y in the thermophile decreases for species pairs in which the thermophile has greater GC than the mesophile. Of the 37 amino acid pairs with no difference in average GC content of their codons, 15 have a significant slope.

Of the 122 pairs of amino acids with differing average GC content and significant slopes, 114 are in the expected direction: sites with the GC-rich amino acid in the mesophile and the GC-poor amino acid in the thermophile become less common in the species pairs where the thermophile has higher genome-wide GC content than the mesophile (supplementary table 2, Supplementary Material online). Seven of the eight pairs of amino acids that show the opposite pattern involve methionine. Sites with aspartic acid, cysteine, glutamic

Table 2

The Substitutional Asymmetry Predicted for a Mesophile–Thermophile Pair with No Difference in GC Content, Based on the Intercept of the Logistic Regression of Asymmetry Versus Difference in GC Content

SN	0.508	DG	0.509	GC	0.529	HK	0.517	KP	0.558*
SD	0.521*	DQ	0.558*	GV	0.553*	HC	0.537	KY	0.547*
ST	0.542*	DM	0.565*	GI	0.546*	HV	0.554*	CV	0.579*
SG	0.482*	DH	0.578*	GF	0.564*	HI	0.578*	CI	0.504
SQ	0.561*	DE	0.574*	GL	0.589*	HF	0.581*	CF	0.526
SM	0.569*	DA	0.547*	GR	0.608*	HL	0.554*	CL	0.539*
SH	0.562*	DK	0.561*	GW	0.555	HR	0.559*	CR	0.446*
SE	0.582*	DC	0.576*	GP	0.601*	HW	0.598*	CW	0.487
SA	0.593*	DV	0.565*	GY	0.603*	HP	0.591*	CP	0.492
SK	0.603*	DI	0.538	QM	0.536*	HY	0.632*	CY	0.520
SC	0.590*	DF	0.619*	QH	0.556*	EA	0.479*	VI	0.507*
SV	0.610*	DL	0.571*	QE	0.518*	EK	0.513*	VF	0.502
SI	0.609*	DR	0.622*	QA	0.511	EC	0.554	VL	0.511*
SF	0.607*	DW	0.693*	QK	0.516*	EV	0.505	VR	0.555*
SL	0.610*	DP	0.618*	QC	0.598*	EI	0.515	VW	0.512
SR	0.624*	DY	0.653*	QV	0.538*	EF	0.542*	VP	0.540*
SW	0.641*	TG	0.460*	QI	0.584*	EL	0.518*	VY	0.538*
SP	0.604*	TQ	0.521*	QF	0.588*	ER	0.550*	IF	0.490
SY	0.676*	TM	0.511	QL	0.581*	EW	0.571*	IL	0.522*
ND	0.500	TH	0.554*	QR	0.579*	EP	0.566*	IR	0.536*
NT	0.546*	TE	0.553*	QW	0.631*	EY	0.574*	IW	0.506
NG	0.502	TA	0.545*	QP	0.610*	AK	0.520*	IP	0.521
NQ	0.549*	TK	0.563*	QY	0.644*	AC	0.453*	IY	0.529*
NM	0.576*	TC	0.523	MH	0.500	AV	0.522*	FL	0.512*
NH	0.611*	TV	0.595*	ME	0.522	AI	0.500	FR	0.526
NE	0.545*	TI	0.607*	MA	0.517	AF	0.526*	FW	0.498
NA	0.552*	TF	0.580*	MK	0.540*	AL	0.536*	FP	0.517
NK	0.587*	TL	0.591*	MC	0.537	AR	0.571*	FY	0.500
NC	0.594*	TR	0.607*	MV	0.574*	AW	0.525	LR	0.513
NV	0.629*	TW	0.604*	MI	0.583*	AP	0.605*	LW	0.515
NI	0.612*	TP	0.611*	MF	0.596*	AY	0.546*	LP	0.505
NF	0.593*	TY	0.619*	ML	0.607*	KC	0.532	LY	0.508
NL	0.626*	GQ	0.529*	MR	0.556*	KV	0.487	RW	0.576*
NR	0.650*	GM	0.514	MW	0.569*	KI	0.503	RP	0.503
NW	0.624*	GH	0.548*	MP	0.628*	KF	0.541*	RY	0.549*
NP	0.604*	GE	0.544*	MY	0.617*	KL	0.501	WP	0.551
NY	0.685*	GA	0.561*	HE	0.493	KR	0.599*	WY	0.495
DT	0.508	GK	0.543*	HA	0.494	KW	0.623*	PY	0.482

NOTE.—The number is the predicted proportion of sites with the first amino acid in the mesophile and the second amino acid in the thermophile; an asterisk indicates that the proportion is significantly different from 0.50 ($P < 0.05$). Amino acids are ordered from least preferred (serine, S) to most preferred (tyrosine, Y) in thermophiles.

Table 3

Average Asymmetry and Transfer Free Energy for Each Amino Acid

Amino Acid	Average Asymmetry	Transfer Free Energy
Serine (S)	0.416	0.04
Asparagine (N)	0.417	−0.01
Aspartic acid (D)	0.430	0.54
Threonine (T)	0.450	0.44
Glycine (G)	0.451	0.00
Glutamine (Q)	0.459	−0.10
Methionine (M)	0.470	1.30
Histidine (H)	0.485	1.10
Glutamic acid (E)	0.497	0.55
Alanine (A)	0.500	0.73
Lysine (K)	0.504	1.50
Cysteine (C)	0.523	0.70
Valine (V)	0.529	1.69
Isoleucine (I)	0.531	2.97
Phenylalanine (F)	0.542	2.65
Leucine (L)	0.544	2.49
Arginine (R)	0.551	0.73
Tryptophan (W)	0.562	3.00
Proline (P)	0.565	2.60
Tyrosine (Y)	0.575	2.97

NOTE.—Average asymmetry is the predicted proportion, in a pair of species with equal GC contents, of substitutions from other amino acids in the mesophile to the given amino acid in the thermophile. Transfer free energy is from Simon (1976). Amino acids are ordered from least preferred (serine) to most preferred (tyrosine) in thermophiles.

acid, glutamine, leucine, serine, or threonine in the mesophile and methionine in the thermophile become more common as the thermophile–mesophile GC difference increases, even though the methionine codon has a slightly smaller GC content than the codons for the other amino acids.

The logistic regression for 139 out of 190 pairs of amino acids had a significant intercept (supplementary table 2, Supplementary Material online), meaning that a mesophile–thermophile species pair with no difference in genomic GC content would be expected to have significant asymmetry. The intercept of each logistic regression was used to estimate the substitutional asymmetry predicted for a mesophile–thermophile pair with no difference in GC content (table 2). The average of the 19 intercepts for each amino acid gives a measure of how strongly that amino acid is preferred in mesophiles or thermophiles; for example, only 41.6% of the substitutions involving serine would have serine in the thermophile and some other amino acid in the mesophile (table 3).

Consistency among Pairs of Species The residual (the difference between the observed asymmetry and that predicted by the logistic regression) was calculated for each pair of amino acids in each species pair, and the average residual was calculated for each amino acid in each species pair. In some species pairs, the average residual for some amino acids is quite a bit larger or smaller than expected (fig. 2). For example, in the *Streptomyces–Thermobifida* species pair, there are fewer sites with lysine (K) in the thermophile and other amino acids in the mesophile than predicted by the logistic regression, whereas there are more such sites than predicted in the *Deinococcus–Thermus* species pair. Out of 180 average residuals (20 amino acids in nine species pairs), 98 have a 95% confidence interval that does not include 0.

Amino Acid Properties After removing indices with missing or estimated values, and indices that represent frequencies in different parts of proteins, the AAindex database (Kawashima et al. 2008) contains 238 measures of

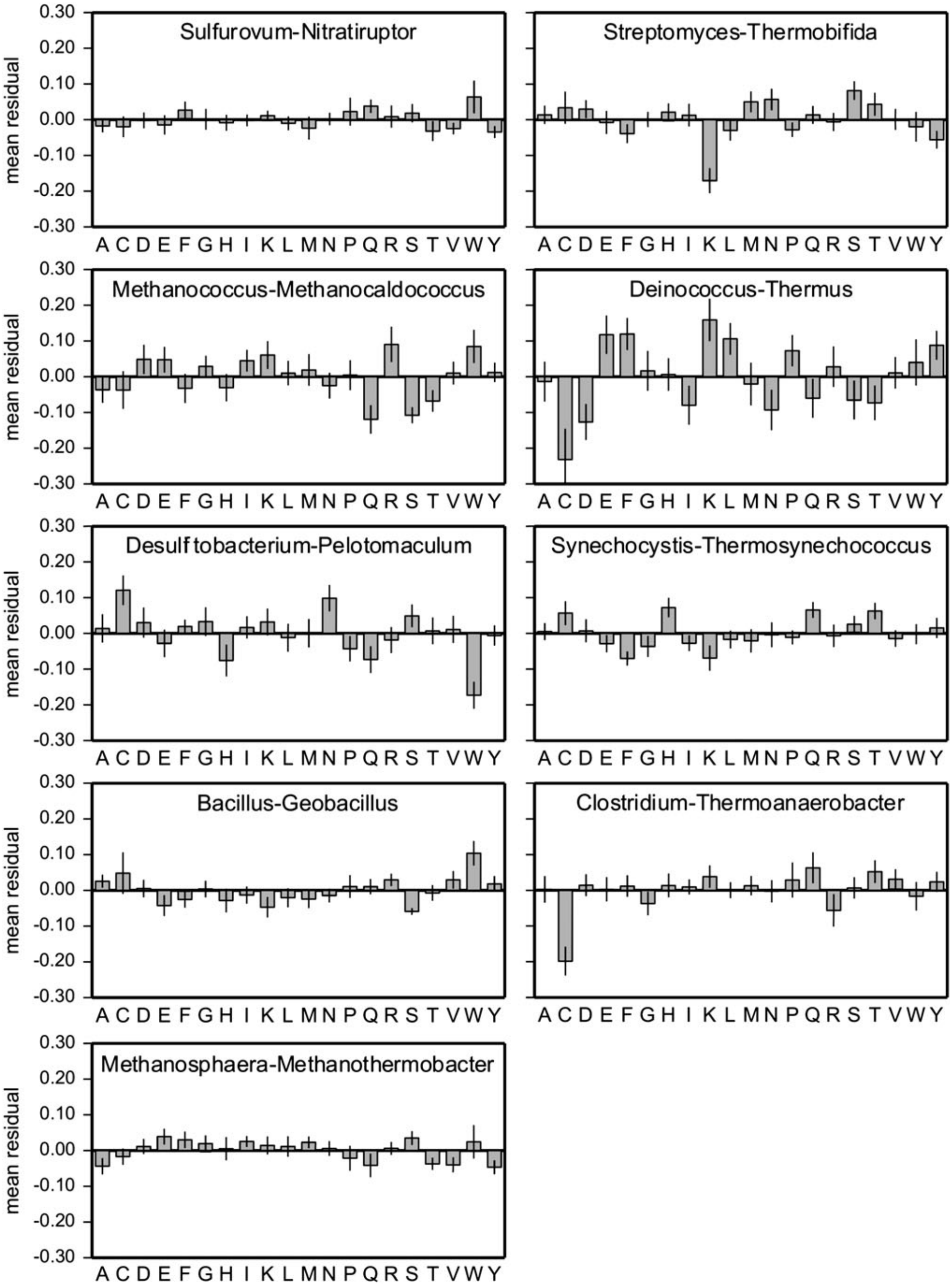

Fig. 2.—Mean of the 19 residuals (differences between the observed number of substitutions and that expected from the logistic regression) for each amino acid in each species pair. Values above 0 indicate that sites with that amino acid in the thermophile and other amino acids in the mesophile are more common than expected from the logistic regression of all species. Error bars are 95% confidence intervals.

biochemical and physical properties of amino acids. Treating the difference in each index for each of the pairs of amino acids as 190 values causes all kinds of statistical problems with nonindependence, so the results of the logistic regression of substitutional asymmetry versus index differences should be viewed as an exercise in data exploration not hypothesis testing. The strongest relationship between the difference in amino acid index and the predicted substitutional asymmetry is with transfer free energy (Simon 1976), a measure of hydrophobicity. In general, amino acids with higher transfer free energy tend to be substituted at high temperatures for amino acids with lower transfer free energy (fig. 3). However, differences in transfer free energy do not explain all the substitutional asymmetry. Of 139 pairs of amino acids with a significant intercept in the logistic regression (meaning that the substitutional asymmetry is

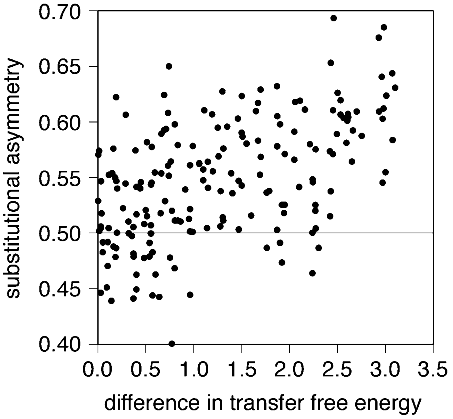

Fig. 3.—Substitutional asymmetry (proportion of all A ↔ B sites that have A in the mesophile and B in the thermophile) versus the difference in transfer free energy of the amino acids (B-A), where B is the amino acid with greater transfer free energy.

predicted to be significant for a mesophile–thermophile pair with no difference in genome-wide GC content), 14 have the opposite pattern: the amino acid with lower transfer free energy is found more often at higher temperatures. The next strongest associations are with several other measures of hydrophobicity (Zimmerman et al. 1968; Jones 1975; Argos et al. 1982; Takano and Yutani 2001), all of which are highly correlated with transfer free energy.

Discussion

Each of the nine mesophile–thermophile species pairs exhibits a large amount of substitutional asymmetry; for most pairs of amino acids, there are more homologous sites with one amino acid in the mesophile and the other amino acid in the thermophile than the opposite. Substitutional asymmetry has been previously observed in small numbers of proteins from *Methanococcus* versus *Methanocaldococcus* (Haney et al. 1999; McDonald et al. 1999), *Bacillus* versus *Geobacillus* (McDonald et al. 1999), and *Deinococcus* versus *Thermus* (McDonald 2001). Here, I use translated protein sequences from the entire genomes of these species pairs and add six additional mesophile–thermophile pairs from a broad variety of habitats.

Differences in genome-wide GC contents are one cause of substitutional asymmetry; all the species pairs used here differ to some degree in GC content, and it has long been known that amino acids with GC-rich codons are more common in species with GC-rich genomes (Lobry 1997; Singer and Hickey 2000). It is not clear whether differences in genome-wide GC content are caused by selection or mutational bias (Rocha and Danchin 2002; Lind and Andersson 2008), and it is not clear to what extent increased habitat temperatures cause increased GC contents (Musto et al. 2006; Wang et al. 2006). What is clear is that any attempt to identify selection on amino acids as a cause of substitutional asymmetry must remove the effects of GC content.

Here, logistic regression modeling is used to control statistically for the effects of differing GC content, with the difference in GC content as the independent variable and the direction of substitution as the dependent variable. For the majority of amino acid pairs, the logistic regression predicts that a mesophile–thermophile pair of species that did not differ in GC content would exhibit extensive substitutional asymmetry. The significant intercepts in the logistic regression models mean that the preferences for one amino acid over another are fairly consistent across the nine pairs of species.

Substitutional asymmetry in one mesophile–thermophile pair could be caused by any number of habitat differences; for example, the mesophile *Methanococcus maripaludis* was isolated from a salt marsh (Jones, Paynter, and Gupta 1983), whereas the thermophile *M. jannaschii* was originally isolated from a deep-sea vent 2,600 m below the ocean surface (Jones, Leigh, et al. 1983). A difference in hydrostatic pressure may favor some amino acids over others (Di Giulio 2005); if hydrostatic pressure were an important selective factor, *M. maripaludis* and *M. jannaschii* would have patterns of asymmetry different from the other mesophile–thermophile pairs, which do not differ in the hydrostatic pressure of their habitats. The consistency of the patterns of asymmetry across species pairs suggests that much of the asymmetry results from selection caused by the different habitat temperatures.

Although the patterns of asymmetry are consistent enough across species pairs to produce logistic regression models with significant intercepts, the amounts of asymmetry in each species pair are not exactly as predicted by the logistic regression; many amino acids are favored more or less strongly in some species pairs than would be expected. The optimal temperatures of the species pairs differ by different amounts, from 15 to 55 °C, so it would have been startling if they all exhibited the exact same amount of asymmetry. The species pairs differ in how recently they diverged from a common ancestor, and the species pairs also vary in other aspects that may affect selection on amino acid use: aerobic versus anaerobic; autotrophic versus heterotrophic; marine, freshwater, and terrestrial; and deep sea versus shallow water. Species pairs in which the ancestral species was thermophilic and one lineage then adapted to lower temperatures may show different patterns of temperature adaptation than species pairs in which the ancestor was mesophilic and one lineage adapted to higher temperatures (Berezovsky and Shakhnovich 2005). There is also increasing evidence that biosynthetic costs may affect amino acid use (Akashi and Gojobori 2002; Seligmann 2003; Heizer et al. 2006; Swire 2007), and the costs of particular amino acids will depend on factors that may be unrelated to temperature, such as the biosynthetic pathways used (for autotrophs) and environmental availability and uptake costs (for heterotrophs). Including all the possibly relevant variables when there are only nine species pairs would result in a logistic model that was completely overdetermined, with many spurious correlations; separating the substitutional asymmetry caused by temperature adaptation from the asymmetry resulting from other causes will require examining the genomes of a much larger number of mesophile–thermophile species pairs than currently available.

These results show that amino acids with greater hydrophobicity (higher transfer free energy) tend to be preferred in thermophiles, which is consistent with several earlier studies (Argos et al. 1979; Gromiha, Oobatake, Kono, et al. 1999; Haney et al. 1999; Tekaia et al. 2002; Nakashima et al. 2003; Sadeghi et al. 2006; Berezovsky et al. 2007). There are, however, numerous exceptions to this rule. This is consistent with previous research that has failed to identify a single physicochemical property of the amino acids that would explain all the differences in amino acid abundance between mesophiles and thermophiles (Böhm and Jaenicke 1994; Zhou et al. 2008). One possible explanation is that thermal adaptation of amino acids is based on complicated tradeoffs between different properties (Gromiha, Oobatake, and Sarai 1999). Another possibility is that the cost of synthesizing amino acids plays a major role; the relative synthesis costs of amino acids change as temperatures increase (Amend and Shock 1998), and amino acids with lower synthesis costs tend to be more abundant, even in heterotrophs (Swire 2007). Values for the cost of synthesis of each amino acid in each species at a variety of temperatures are not available; as this information accumulates, it may become possible to understand the role that relative biosynthetic costs of amino acids play in temperature adaptation of proteins.

There are numerous reports of charged amino acids being more common in thermophiles than in mesophiles (Cambillau and Claverie 2000; Das and Gerstein 2000; Szilágyi and Závodszky 2000; Fukuchi and Nishikawa 2001; Vielle and Zeikus 2001; Chakravarty and Varadarajan 2002; Tekaia et al. 2002; Nakashima et al. 2003; Suhre and Claverie 2003; Sadeghi et al. 2006; Berezovsky et al. 2007). That pattern is not apparent here; of 47 significant intercepts in the logistic regression involving one charged amino acid (arginine, aspartic acid, glutamic acid, and lysine) and one noncharged amino acid, 24 have the charged amino acid becoming more common in the thermophiles, but 23 have the charged amino acid becoming less common in the thermophiles (table 2). If histidine, which is weakly charged at physiological pH, is included in the charged amino acids, the result is the same: Of 57 significant intercepts, 28 have the charged amino acid becoming more common in the thermophiles, but 29 have the charged amino acid becoming less common in the thermophiles. Most of the studies reporting increased proportions of charged amino acids in thermophiles have relied heavily on hyperthermophiles, which have optimum growth temperatures of 85 °C to >100 °C, whereas the nine species pairs used here include only one hyperthermophile, *M. jannaschii*, with an optimum growth temperature of 85 °C. It may be that increasing the overall proportion of charged amino acids is only an important adaptation at very high temperatures.

Supplementary Material

Supplementary tables 1 and 2 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Literature Cited

Akashi H, Gojobori T. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci U S A. 99:3695–3700.

Amend JP, Shock EL. 1998. Energetics of amino acid synthesis in hydrothermal ecosystems. Science. 281:1659–1662.

Argos P, Rao JKM, Hargrave PA. 1982. Structural prediction of membrane-bound proteins. Eur J Biochem. 128:565–575.

Argos P, et al. 1979. Thermal stability and protein structure. Biochemistry. 25:5698–5703.

Bao Q, et al. 2002. A complete sequence of the *T. tengcongensis* genome. Genome Res. 12:689–700.

Berezovsky IN, Shakhnovich EI. 2005. Physics and evolution of thermophilic adaptation. Proc Natl Acad Sci U S A. 102:12742–12747.

Berezovsky IN, Zeldovich KB, Shakhnovich EI. 2007. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput Biol. 3:498–507.

Böhm G, Jaenicke R. 1994. Relevance of sequence statistics for the properties of extremophilic proteins. Int J Pept Protein Res. 43:97–106.

Boussau B, Blanquart S, Necsulea A, Lartillot N, Guoy M. 2008. Parallel adaptations to high temperatures in the Archaean eon. Nature. 456:942–945.

Bruggeman H, et al. 2003. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Natl Acad Sci U S A. 100:1316–1321.

Bult CJ, et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science. 23:1058–1073.

Cambillau C, Claverie JM. 2000. Structural and genomic correlates of hyperthermostability. J Biol Chem. 275:32383–32386.

Chakravarty S, Varadarajan R. 2002. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry. 41:8152–8156.

Chenna R, et al. 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497–3500.

Das R, Gerstein M. 2000. The stability of thermophilic proteins: a study based on comprehensive genome comparison. Funct Integr Genomics. 1:76–88.

Di Giulio M. 2005. A comparison of proteins from *Pyrococcus furiosus* and *Pyrococcus abyssi*: barophily in the physicochemical properties of amino acids and in the genetic code. Gene. 346:1–6.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125:1–15.

Fricke WF, et al. 2006. The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. J Bacteriol. 188:642–658.

Fukuchi S, Nishikawa K. 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol. 309:835–843.

Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. 1999. Relationship between amino acid properties and protein stability: buried mutations. J Protein Chem. 18:565–578.

Gromiha MM, Oobatake M, Sarai A. 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem. 82:51–67.

Haney PJ, et al. 1999. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. Proc Natl Acad Sci U S A. 96:3578–3583.

Harvey PH, Pagel MD. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.

Heizer EM, et al. 2006. Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. Mol Biol Evol. 23:1670–1680.

Hendrickson EL, et al. 2004. Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J Bacteriol. 186:6956–6969.

Henne A, et al. 2004. The genome sequence of the extreme thermophile *Thermus thermophilus*. Nat Biotechnol. 22:547–553.

Jones DD. 1975. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J Theor Biol. 50:167–183.

Jones WJ, Leigh JA, Mayer F, Woese CR, Wolfe RS. 1983. *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch Microbiol. 136:254–261.

Jones WJ, Paynter MJB, Gupta R. 1983. Characterization of *Methanococcus maripaludis* sp. nov., a new methanogen isolated from salt marsh sediment. Arch Microbiol. 135:91–97.

Kaneko T, et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3:109–136.

Kawashima S, et al. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 36:D202–D205.

Kosaka T, et al. 2008. The genome of *Pelotomaculum thermopropionicum* reveals niche-associated evolution in anaerobic microbiota. Genome Res. 18:442–448.

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

Kunst F, et al. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature. 390:249–256.

Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci U S A. 105:17878–17883.

Lobry JR. 1997. Influence of genomic G+C content on average amino acid composition of proteins from 59 bacterial species. Gene. 205:309–316.

Lobry JR, Chessel D. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J Appl Genet. 44:235–261.

Lykidis A, et al. 2007. Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. J Bacteriol. 189:2477–2486.

McDonald JH. 2001. Patterns of temperature adaptation in proteins from the bacteria *Deinococcus radiodurans* and *Thermus thermophilus*. Mol Biol Evol. 18:741–749.

McDonald JH. 2009. Handbook of biological statistics, 2nd ed. Baltimore (MD): Sparky House Publishing.

McDonald JH, Grasso AM, Rejto LK. 1999. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol. 16:1785–1790.

Mizuguchi K, Sele M, Cubellis MV. 2007. Environment specific substitution tables for thermophilic proteins. BMC Bioinformatics. 8:S15.

Musto H, et al. 2006. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 347:1–3.

Nakagawa S, et al. 2007. Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. Proc Natl Acad Sci U S A. 104:12146–12150.

Nakamura Y, et al. 2002. Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. DNA Res. 9:123–130.

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

Naya H, Gianola D, Romero H, Urioste JI, Musto H. 2006. Inferring parameters shaping amino acid usage in prokaryotic genomes via Bayesian MCMC methods. Mol Biol Evol. 23:203–211.

Nishio Y, et al. 2003. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. Genome Res. 13:1572–1579.

Nonaka H, et al. 2006. Complete genome sequence of the dehalorespiring bacterium *Desulfitobacterium hafniense* Y51 and comparison with *Dehalococcoides ethenogenes* 195. J Bacteriol. 188:2262–2274.

Omura S, et al. 2001. Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. Proc Natl Acad Sci U S A. 98:12215–12220.

Puigbò P, Pasamontes A, Garcia-Vallve S. 2008. Gaining and losing the thermophilic adaptation in prokaryotes. Trends Genet. 24:10–14.

Rocha EP, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18:291–294.

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

SAS Institute. 2009. SAS/STAT 9.2 user's guide, 2nd ed. Cary (NC): SAS Institute.

Saunders NFW, et al. 2003. Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Res. 13:1580–1588.

Seligmann H. 2003. Cost-minimization of amino acid usage. J Mol Evol. 56:151–161.

Simon Z. 1976. Quantum biochemistry and specific interactions. Tunbridge Wells (UK): Abacus Press.

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

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

Smith DR, et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol. 179:7135–7155.

Suhre K, Claverie JM. 2003. Genomic correlates of hyperthermostability, an update. J Biol Chem. 278:17198–17202.

Suyama A, et al. 2001. Isolation and characterization of *Desulfitobacterium* sp. strain Y51 capable of efficient dehalogenation of tetrachloroethene and polychloroethanes. Biosci Biotechnol Biochem. 65:1474–1481.

Swire J. 2007. Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. J Mol Evol. 64:558–571.

Szilágyi A, Závodszky P. 2000. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure. 9:493–504.

Takami H, et al. 2004. Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. Nucleic Acids Res. 32:6292–6303.

Takano K, Yutani K. 2001. A new scale for side-chain contribution to protein stability based on the empirical stability analysis of mutant proteins. Protein Eng. 14:525–528.

Tekaia F, Yeramian E. 2006. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics. 7:307.

Tekaia F, Yeramian E, Dujon B. 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene. 297:51–60.

Vielle C, Zeikus GJ. 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev. 65:1–43.

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. 342:681–684.

White O, et al. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science. 286:1571–1577.

Zeldovich KB, Berezovsky IN, Shakhnovich EI. 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 3:62–72.

Zhou X-X, Wang Y-B, Pan Y-J, Li W-F. 2008. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. Amino Acids. 34:25–33.

Zimmerman JM, Eliezer N, Simha R. 1968. The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol. 21:170–201.

Associate Editor: Bill Martin