Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 3578–3583, March 1999
Biochemistry

Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species

PAUL J. HANEY*†, JONATHAN H. BADGER†, GERALD L. BULDAK‡, CLAUDIA I. REICH, CARL R. WOESE, AND GARY J. OLSEN§

Department of Microbiology, University of Illinois, B103 Chemical and Life Sciences Laboratory, 601 South Goodwin Avenue, Urbana, IL 61801

Contributed by Carl R. Woese, January 11, 1999

ABSTRACT **The genome sequence of the extremely thermophilic archaeon *Methanococcus jannaschii* provides a wealth of data on proteins from a thermophile. In this paper, sequences of 115 proteins from *M. jannaschii* are compared with their homologs from mesophilic *Methanococcus* species. Although the growth temperatures of the mesophiles are about 50°C below that of *M. jannaschii*, their genomic G+C contents are nearly identical. The properties most correlated with the proteins of the thermophile include higher residue volume, higher residue hydrophobicity, more charged amino acids (especially Glu, Arg, and Lys), and fewer uncharged polar residues (Ser, Thr, Asn, and Gln). These are recurring themes, with all trends applying to 83–92% of the proteins for which complete sequences were available. Nearly all of the amino acid replacements most significantly correlated with the temperature change are the same relatively conservative changes observed in all proteins, but in the case of the mesophile/thermophile comparison there is a directional bias. We identify 26 specific pairs of amino acids with a statistically significant ($P < 0.01$) preferred direction of replacement.**

Identifying the bases of protein adaptation to higher or lower temperatures is integral to our understandings of protein folding, the relationship of protein structure to function, the design of high temperature biocatalysts, and the history of life on this planet. Most studies of protein thermostability take one of two approaches. A structural/mutational approach uses protein structures to locate differences between high- and low-temperature proteins and thereby propose hypotheses for the bases of thermal adaptation. In turn, these hypotheses guide directed mutagenesis studies, which can yield a detailed portrait of the interactions stabilizing the particular protein. However, the labor and expense of a structural/mutational approach restrict analyses to limited numbers of molecules, resulting in a potentially biased view of thermal stabilizing mechanisms. Current data emphasize a few specific proteins, including bacteriophage T4 lysozyme (1–5), the neutral protease family (6–9), and glyceraldehyde-3-phosphate dehydrogenase (10–13). In the search for universal themes, these highly focused studies and their sometimes conflicting observations offer a restricted view. Also, the lack of consensus among studies has given rise to the recognition (and even resignation) that no set of simple factors distinguish all thermophile and mesophile proteins. If there are general rules to adaptation and thermostability, a broader approach to the problem will be required to elucidate them.

A less costly, yet potentially more comprehensive, approach invokes sequence comparisons of families of homologous high- and low-temperature proteins (14–17). Here, statistical analyses extract recurring amino acid replacement trends—presumably those important for thermal adaptation (signal)—from a background of random genetic drift (noise). This approach has been hampered in the past by the noise accompanying the high levels of sequence divergence separating most available pairs of high- and low-temperature proteins, and, until recently, by a paucity of sequence data from extremely thermophilic organisms. These two approaches also address slightly different questions. Comparative studies generally seek the sequence features that distinguish proteins that work *in vivo* under the different environmental conditions. Structural/mutational studies frequently also seek to resolve the effects on activity, folding, and "irreversible denaturation."

Overcoming the historical limitations of the comparative approach requires (*i*) large quantities of data and (*ii*) closely related organisms with very different growth temperatures. The ability to efficiently sequence whole genomes can provide the necessary quantities of data. In selecting *Methanococcus jannaschii* for complete genome sequencing (18), we also addressed the second need—this extreme thermophile (85°C growth temperature) has a number of mesophilic relatives. In the present work, we complement the *M. jannaschii* genome data by sequencing random clones of genomic DNA from *Methanococcus voltae*, one of these mesophilic relatives. We then combine these new data with the other data available from the mesophiles *M. voltae*, *Methanococcus vannielii* and *Methanococcus maripaludis* (optimal growth temperatures of ≈35°C) to conduct an extensive comparative analysis of protein thermal adaptation. Because of the complete *M. jannaschii* genome sequence, nearly every sequence available from a mesophilic member of the genus can be included in the analysis, providing a sample of over 100 proteins. From these data, we identify 26 pairs of amino acids that display a statistically significant ($P < 0.01$) bias in the direction of change and that we suggest are associated with the differences in growth temperature. In addition, we identify four amino acid properties that distinguish the high- and low-temperature proteins, each of which applies to >80% of the complete proteins.

MATERIALS AND METHODS

DNA Sequencing. Fragments (1.0–1.5 kbp) of *M. voltae* PS (DSM 1537) genomic DNA were generated by using random shearing. Fragments were end-repaired and cloned into pUC18, as described for *M. jannaschii* (18). Cloned DNAs were partially sequenced from plasmid primer sites by using Sequenase version

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF078607–AF078665).

*Present address: Mayo Foundation, 200 First Street, SW, Rochester, MN 55905.

†P.J.H. and J.H.B. contributed equally to this work.

‡Present address: Department of Molecular Genetics (M/C 669), College of Medicine, University of Illinois at Chicago, 900 S. Ashland Avenue, Chicago, IL 60607.

§To whom reprint requests should be addressed. e-mail: gary@phylo.life.uiuc.edu.

2.0 (United States Biochemical). The sequences have been assigned GenBank accession nos. AF078607–AF078665.

Data for Comparative Analysis. We compiled mesophile sequence data comprised of (*i*) the new sequence data from *M. voltae* (above), (*ii*) unpublished genomic DNA sequence data from *M. maripaludis* (strain MM) kindly provided by W. B. Whitman (University of Georgia), and (*iii*) all unique *M. voltae*, *M. vannielii*, and *M. maripaludis* DNA sequences available (in November 1997) from GenBank (19). The DNA sequences from mesophiles were aligned to *M. jannaschii* proteins by using BLASTX 1.4.8 MP (20, 21). The ungapped alignments provided by BLASTX were examined and edited to remove misaligned regions, to eliminate multiple use of the same residues, and to ensure, to as far as feasible, that the sequences are orthologs, not paralogs.

Amino Acid Exchange Bias. Each type of amino acid replacement was counted in the edited BLASTX alignments. We will discuss amino acid replacements in the direction mesophile → thermophile to maintain consistency with other studies, not to suggest an evolutionary direction. For a given pair of amino acids, the "forward" direction designates the more common of the two replacements in converting mesophile proteins to thermophile proteins. The two-tail binomial distribution was used to calculate the probability that a random sampling of equally probable forward and reverse replacements would give rise to a directional bias (asymmetry) greater than or equal to that observed.

Analysis of Amino Acid Properties. For ease of discussion, each amino acid was assigned to one of three categories: charged (Asp, Glu, Arg, and Lys), uncharged polar (Ser, Thr, Asn, and Gln), and nonpolar (Gly, Ala, Val, Leu, Ile, Phe, Trp, Tyr, Pro, Met, Cys, and His). Our categorizations of Cys and His are based on the similarities of their behaviors to those of the other nonpolar residues.

A computer-readable database of amino acid characteristics, AAINDEX (22), was used to identify amino acid properties associated with thermal adaptation. A "property" in AAINDEX assigns a numerical value to each of the 20 amino acids (e.g., its volume or its partition coefficient in a two-phase solvent system). We added three new charge properties to AAINDEX (differing only in their treatment of His): Asp, Glu, Arg, and Lys were assigned a charge magnitude of 1; His was assigned a charge of 0 ("full charge"), 0.7 ("charge magnitude"), or 1 ("charged"), and all other amino acids were assigned a charge of 0. We also added a property for uncharged polar amino acids, with a value of 1 for Ser, Thr, Gln, and Asn and 0 for all others. To facilitate comparisons, we normalized each property in this expanded AAINDEX so that the 20 values would have a mean of 0 and a standard deviation of 1 (with no weighting for average usage).

To quantitate the association of an amino acid property with growth temperature, we calculated a correlation coefficient of the property value with temperature, treating each observed replacement as two data points: (mesophile temperature, mesophile amino acid property value) and (thermophile temperature, thermophile amino acid property value). Bootstrap resamplings of the replacements (23) were used to estimate the uncertainty of the correlation coefficient. To determine whether the correlation coefficient of one property is significantly greater than that of another property, the fraction of the bootstrap replicates in which the first property showed a greater correlation coefficient than the second (adjusting the sign, if necessary) was determined.

Interpreting the results is complicated by correlations among many of the properties in AAINDEX. Once a significant property was identified, we sought to remove its effects from subsequent evaluations of additional properties. Given any two properties P and Q, with normalized values P_i and Q_i for each amino acid i, we define a new property $Q_{\perp P}$ that is the components of Q that are orthogonal to P (that is, the aspects of Q that are independent of P). Then $Q_{\perp Pi}$, the value of $Q_{\perp P}$ for amino acid i, is

$$Q_{\perp Pi} = Q_i - \left(\sum_{j \epsilon A} P_j Q_j \Big/ \sum_{j \epsilon A} P_j^2 \right) P_i.$$

As with the normalization, this value is not weighted for amino acid usage.

RESULTS

Collection and Editing of the Data. In anticipation of the complete genome sequence of *M. jannaschii*, we produced a library of random genomic DNA fragments from the related mesophile *M. voltae*. Sequences were determined from 68 clone termini, yielding a total of 10,794 nucleotides of new data. For the comparisons with *M. jannaschii* proteins, these data were combined with unpublished *M. maripaludis* DNA sequences from W. B. Whitman (which are now available through GenBank) and with the other mesophilic methanococcal DNA sequences available from GenBank.

For each DNA sequence from a mesophilic *Methanococcus* sp., similar *M. jannaschii* proteins were located by using BLASTX (21). No *M. jannaschii* protein was used more than once. Overlaps between the ends of BLAST high-scoring segment pairs were eliminated. We used ungapped BLAST alignments, rather than gapped alignments, to minimize inclusion of intervening regions of low sequence similarity. Even so, omitted regions were few in number and usually small (1–3 aa). When the protein sequences were full length, the edited alignment generally retained >95% of the residues. Protein pairs with <45% identity (the surface-layer protein and some *nif* gene proteins) were eliminated because alignment was uncertain. Ferredoxins were omitted because of variations in size and the difficulty in defining orthologs. Only one version of the nearly identical flagellin proteins was used.

The final alignments contained 115 complete or partial protein sequences from mesophiles and their high-temperature homologs (a list of the proteins is provided as supplemental data on the PNAS web site, www.pnas.org). They comprise 23,824 aligned pairs of amino acids, of which 7,131 are nonidentical (an average of 70.1% sequence identity). Alignments covering at least 90% of a protein's length provided 78% of the amino acid replacement data.

Amino Acid Composition. Table 1 summarizes the net change in amino acid composition between the mesophile and thermophile proteins. The thermophile proteins are characteristically reduced in Ser, Asn, Gln, Thr, and Met and increased in Ile, Arg, Glu, Lys, and Pro. The magnitudes of the changes range from a 16.5% increase (Arg) to a 32% reduction (Ser). The large sampling of events makes these shifts statistically significant, with random probabilities from 3.7×10^{-4} to 9.5×10^{-39}. Because they provide a reduced sampling, individual proteins show few significant shifts in residue composition.

Specific Amino Acid Replacements. The preceding abstraction overlooks the importance of residues that are gained in some contexts and lost in others. To better understand the individual contributions, Table 2 reports all 380 replacement types. The residues favored in thermophile proteins are usually listed further to the right (and lower) than amino acids favored in the mesophile sequences. This ordering conforms to the preferred direction of replacement for 138 of the 167 (83%) amino acid pairs that show any directional bias (red entries above the diagonal and blue entries below). Although alternative orderings would be consistent with more replacements, they would scatter the charged residues. The bold entries indicate the 26 pairs of amino acids (14% of all pairs) that have a directional replacement bias with a random probability <0.01.

The ratio of forward replacements to reverse replacements (as defined in *Materials and Methods*) for a given pair of amino acids reveals the most biased replacements (Table 3). These highly biased replacements suggest very distinct selective pressures on the amino acids in the mesophilic and thermophilic methanococci. However, most of these replacements are so rare, even in

Table 1. Change in amino acid composition going from mesophile to thermophile proteins

Amino acid	Gains	Losses	Ratio	P^*	Net change	Change, %
Ile	842	658	1.28	2.2×10^{-6}	184	9.5
Glu	739	562	1.31	1.0×10^{-6}	177	9.1
Arg	383	214	1.79	4.5×10^{-12}	169	16.5
Lys	789	620	1.27	7.4×10^{-6}	169	8.3
Pro	167	96	1.74	0.000014	71	7.0
Tyr	224	177	1.27	0.021	47	5.8
Ala	504	458	1.10	0.15	46	2.8
Trp	23	11	2.09	0.058	12	8.3
Leu	560	548	1.02	0.74	12	0.6
Cys	72	69	1.04	0.87	3	0.9
Phe	200	202	0.99	0.96	−2	−0.3
Asp	429	432	0.99	0.95	−3	−0.2
Val	666	670	0.99	0.93	−4	−0.2
His	80	92	0.87	0.40	−12	−2.8
Gly	201	264	0.76	0.0040	−63	−3.4
Met	174	248	0.70	0.00037	−74	−11.3
Gln	158	234	0.68	0.00015	−76	−13.1
Thr	336	431	0.78	0.00068	−95	−8.4
Asn	313	481	0.65	2.7×10^{-9}	−168	−15.9
Ser	271	664	0.41	9.5×10^{-39}	−393	−31.7

*The random probability of a directional bias greater than or equal to that observed (calculated using the two-tailed binomial distribution).

the favored direction, that they cannot be major contributors to protein thermal adaptation. Only Ser → Lys and Ser → Ala contribute an average of about one net replacement per typical 300-aa protein, and only three other highly biased replacements contribute a net compositional shift of 20 or more residues in the entire data set (1 per 4 typical proteins).

The 20 most frequent replacements (Table 3) are conservative replacements, and they contribute 3,991 of the 7,131 observed replacements (that is, 11% of the pairs contribute 56% of the replacements). Although common, most of these replacements are significantly biased in direction (12 of 20 have random probabilities <0.05). Although the ratio of forward-to-reverse changes for these replacements is less than those in the left third of Table 3 (with the exception of Ser ↔ Ala, which appears in both lists), their high frequency makes even a small bias in direction statistically significant, and suggests the biological importance of small cumulative effects.

To simultaneously emphasize the magnitude of replacement bias and the frequency of replacement, we examined the amino acid pairs with the largest numerical difference between forward and reverse replacements (Table 3, right one-third). We suggest that these replacements have the broadest roles in thermal adaptation. Again, the list is dominated by conservative replacements, although some less conservative changes are interspersed. Despite often low ratios between forward and reverse changes, 17 of these 20 replacements have significant ($P < 0.005$) directional bias. Their frequent occurrence means that these replacements can be accepted in many contexts, while their significant bias suggests that they are useful to thermal adaptation. Yet, even these numerically most biased replacements are far from universal. Individually, only the replacements Ser → Ala and Lys → Arg are sufficiently common to contribute a net shift of one residue per typical 300-aa protein (seven other replacements are within a factor of two of this level). Overall, these 20 replacements contribute a net shift in the forward, thermostabilizing direction of 868 residues (of the 7,131 replacements analyzed). Under the (overly simple) model that thermostabilization is caused by this excess of forward changes, these amino acid pairs would contribute 10.9 stabilizing changes per 300-aa protein. Because these 20 amino acid pairs contribute 47.4% of all replacements, it might be expected that directional bias in the remaining replacements could provide a similar number of additional stabilizing changes. Although this is only a fraction of the 90 replacements observed per full-length protein, it still suggests that *in vivo*, temperature has had an observable effect at roughly 20 positions (≈5–10% of the sites) in these proteins, more than just a few key sites (9).

Changes in Overall Amino Acid Properties. Whereas Tables 1–3 address specific amino acid replacements, we sought underlying themes associated with the events. To minimize the influence of preconceived notions, we surveyed the extensive list of amino acid properties compiled in AAINDEX (22) for the properties most significantly correlated with temperature, given the observed replacements (see *Materials and Methods*). Although ≈50 properties were highly correlated with temperature (with correlation coefficients >10 times the estimated uncertainty), most of the strongest correlations could be placed into four classes: decrease in uncharged polar residues, increase in charged residues (24), increased residue hydrophobicity (24–26), and increased residue volume (27–29). Table 4 gives the correlation coefficient of the property values with temperature for one example of each class. When the differences in correlation coefficients were tested by directly comparing the correlation coefficients for each replicate in a bootstrap resampling (23), the decrease in uncharged polar residues was significantly more correlated with temperature than were charge, hydrophobicity, and volume, with each of these latter three properties showing comparable correlations (data not shown).

Many of the properties correlated with temperature also are correlated with each other and therefore cannot be considered independently. In an attempt to partially disentangle these effects, we used a simple formula (see *Materials and Methods*) to define the components of a property orthogonal to any chosen second property. Table 4 also shows the effects on the temperature correlations when each of the four properties is removed from the other three. Removing the effect of uncharged polar residues substantially lowers the observed temperature correlations of charge, volume, and especially hydrophobicity. Removing the effects of volume or hydrophobicity noticeably lowered the correlation of the other, but did little to the correlations of polar and charged residues with temperature. The charge property had little effect on the others.

DISCUSSION

Overview of Amino Acid Replacements. Although most comparative analyses of proteins require the accumulation of amino acid replacements at noncritical sites to recognize the conserved essential residues, this random drift can confound investigations of protein thermal adaptation (5, 30–32). To discern recurring themes associated with temperature, three factors are essential:

Table 2. Amino acid replacements distinguishing mesophile and thermophile proteins

	Thermophile amino acid																			
	Uncharged polar				Nonpolar												Charged			
Mesophile aa	Ser	Gln	Asn	Thr	Cys	Gly	Ala	His	Met	Tyr	Phe	Val	Leu	Pro	Ile	Trp	Asp	Glu	Lys	Arg
Uncharged polar																				
Ser	575	12	44	**90**	21	29	**176**	7	11	13	6	19	18	**32**	**15**	1	**39**	**46**	**72**	13
	—	*2.40*	*1.10*	***2.05***	*1.40*	*1.00*	***3.38***	*7.00*	*1.10*	*3.25*	*3.00*	*2.71*	*3.60*	***4.57***	***5.00***	*>1.0*	***2.17***	***3.07***	***7.20***	*3.25*
Gln	5	348	11	7	1	3	3	13	9	6	3	2	10	4	6	1	11	67	**58**	**14**
	0.42	—	*1.83*	*0.54*	*1.00*	*1.00*	*0.25*	*2.60*	*1.29*	*>6.0*	*3.00*	*0.33*	*3.33*	*0.80*	*3.00*	*>1.0*	*1.22*	*1.56*	***2.07***	***7.00***
Asn	40	6	575	31	2	32	13	19	4	18	3	6	**15**	11	11	0	84	**70**	**90**	**26**
	0.91	*0.55*	—	*1.19*	*2.00*	*1.03*	*0.93*	*1.73*	*2.00*	*2.25*	*>3.0*	*1.50*	***5.00***	*2.75*	*5.50*	—	*1.29*	***2.00***	***2.00***	***3.71***
Thr	**44**	13	26	706	10	9	47	2	11	3	3	**60**	21	12	**50**	0	16	**42**	**46**	16
	0.49	*1.86*	*0.84*	—	*1.67*	*1.12*	*0.89*	*1.00*	*2.75*	*3.00*	*0.33*	***1.88***	*1.75*	*1.71*	***2.63***	—	*2.29*	***2.33***	***2.00***	*2.29*
Nonpolar																				
Cys	15	1	1	6	256	5	11	1	1	4	2	9	1	0	5	0	2	1	3	1
	0.71	*1.00*	*0.50*	*0.60*	—	*1.67*	*1.00*	*>1.0*	*>1.0*	*1.33*	*1.00*	*1.80*	*0.17*	—	*2.50*	—	*1.00*	*1.00*	*3.00*	*0.50*
Gly	29	3	31	8	3	1,597	**60**	1	4	5	1	4	5	6	2	1	25	28	35	13
	1.00	*1.00*	*0.97*	*0.89*	*0.60*	—	***2.14***	*0.33*	*>4.0*	*2.50*	*>1.0*	*1.33*	*1.25*	*1.20*	*1.00*	*0.50*	*1.14*	*1.08*	*1.59*	*3.25*
Ala	**52**	12	14	53	11	**28**	1,215	2	7	7	5	71	21	**36**	22	1	18	**50**	39	9
	0.30	*4.00*	*1.08*	*1.13*	*1.00*	***0.47***	—	*>2.0*	*0.88*	*1.75*	*1.67*	*1.13*	*1.24*	***2.77***	*1.05*	*>1.0*	*1.80*	***2.08***	*1.70*	*1.12*
His	1	5	11	2	0	3	0	342	0	**28**	5	3	6	3	3	0	4	5	8	5
	0.14	*0.38*	*0.58*	*1.00*	*<1.0*	*3.00*	*<0.5*	—	—	***2.80***	*1.25*	*1.50*	*1.00*	*1.00*	*3.00*	—	*2.00*	*2.50*	*2.67*	*2.50*
Met	10	7	2	4	0	0	8	0	406	4	8	25	**97**	0	51	1	2	11	12	6
	0.91	*0.78*	*0.50*	*0.36*	*<1.0*	*<0.2*	*1.14*	—	—	*2.00*	*2.67*	*2.08*	***1.76***	*<1.0*	*1.55*	*>1.0*	*>2.0*	*1.83*	*0.86*	*6.00*
Tyr	4	0	8	1	3	2	4	**10**	2	637	72	7	22	1	18	6	5	4	8	0
	0.31	*<0.2*	*0.44*	*0.33*	*0.75*	*0.40*	*0.57*	***0.36***	*0.50*	—	*1.06*	*1.17*	*1.29*	*0.50*	*1.80*	*2.00*	*1.67*	*0.44*	*0.67*	*<0.2*
Phe	2	1	0	9	2	0	3	4	3	68	581	15	53	1	24	4	0	4	9	0
	0.33	*0.33*	*<0.3*	*3.00*	*1.00*	*<1.0*	*0.60*	*0.80*	*0.38*	*0.94*	—	*1.50*	*1.29*	*1.00*	*1.00*	*1.00*	*<0.3*	*1.33*	*2.25*	*<0.5*
Val	7	6	4	**32**	5	3	63	2	12	6	10	1,172	86	12	356	2	6	17	**33**	8
	0.37	*3.00*	*0.67*	***0.53***	*0.56*	*0.75*	*0.89*	*0.67*	*0.48*	*0.86*	*0.67*	—	*1.01*	*1.50*	*1.13*	*2.00*	*2.00*	*0.81*	***3.30***	*8.00*
Leu	5	3	**3**	12	6	4	17	6	**55**	17	41	85	1,461	8	**227**	2	3	11	**28**	15
	0.28	*0.30*	***0.20***	*0.57*	*6.00*	*0.80*	*0.81*	*1.00*	***0.57***	*0.77*	*0.77*	*0.99*	—	*2.00*	***1.34***	*2.00*	*0.75*	*1.38*	***2.80***	*1.88*
Pro	**7**	5	4	7	0	5	**13**	3	1	2	1	8	4	918	8	0	5	9	12	2
	0.22	*1.25*	*0.36*	*0.58*	—	*0.83*	***0.36***	*1.00*	*>1.0*	*2.00*	*1.00*	*0.67*	*0.50*	—	*1.33*	—	*0.71*	*0.60*	*1.00*	*2.00*
Ile	**3**	2	2	**19**	2	2	21	1	33	10	24	316	**170**	6	1,283	1	4	14	23	5
	0.20	*0.33*	*0.18*	***0.38***	*0.40*	*1.00*	*0.95*	*0.33*	*0.65*	*0.56*	*1.00*	*0.89*	***0.75***	*0.75*	—	*>1.0*	*0.80*	*0.93*	*1.15*	*1.25*
Trp	0	0	0	0	2	0	0	0	3	4	1	1	0	0	135	0	0	0	0	
	<1.0	*<1.0*	—	—	—	*2.00*	*<1.0*	—	*<1.0*	*0.50*	*1.00*	*0.50*	*0.50*	—	*<1.0*	—	*<1.0*	—	*<1.0*	*<1.0*
Charged																				
Asp	**18**	9	65	7	2	22	10	2	0	3	3	3	4	7	5	1	895	219	47	5
	0.46	*0.82*	*0.77*	*0.44*	*1.00*	*0.88*	*0.56*	*0.50*	*<0.5*	*0.60*	*>3.0*	*0.50*	*1.33*	*1.40*	*1.25*	*>1.0*	—	*1.30*	*1.47*	*1.25*
Glu	**15**	43	**35**	**18**	1	26	**24**	2	6	9	3	21	8	15	15	0	169	1,374	133	19
	0.33	*0.64*	***0.50***	***0.43***	*1.00*	*0.93*	***0.48***	*0.40*	*0.55*	*2.25*	*0.75*	*1.24*	*0.73*	*1.67*	*1.07*	—	*0.77*	—	*1.07*	*1.12*
Lys	**10**	**28**	**45**	**23**	1	22	23	3	14	12	4	**10**	**10**	12	20	1	32	124	1,408	**226**
	0.14	***0.48***	***0.50***	***0.50***	*0.33*	*0.63*	*0.59*	*0.38*	*1.17*	*1.50*	*0.44*	***0.30***	***0.36***	*1.00*	*0.87*	*>1.0*	*0.68*	*0.93*	—	***1.70***
Arg	4	**2**	**7**	7	2	4	8	2	1	6	2	1	8	1	4	1	4	17	**133**	809
	0.31	***0.14***	***0.27***	*0.44*	*2.00*	*0.31*	*0.89*	*0.40*	*0.17*	*>6.0*	*>2.0*	*0.12*	*0.53*	*0.50*	*0.80*	*>1.0*	*0.80*	*0.89*	***0.59***	—

The top of each table cell is the number of times the amino acid for the row was found in a mesophile protein and was replaced by the amino acid for the column in the corresponding thermophile protein. The bottom of each table cell (in *italics*) is the ratio of that replacement to the opposite replacement. Red values indicate replacements favored in the mesophilic to thermophilic direction, while blue values are replacements that are favored in the opposite direction. Replacements with significant directional bias ($P < 0.01$) are in **bold**.

having multiple, diverse proteins (to reproduce important events); having proteins with high sequence identity (to minimize random drift); and having proteins adapted to very disparate temperatures (to maximize signal). In addition, the G+C content of the respective DNAs should be similar (ref. 33; P.J.H. and J.H.B., unpublished results). The archaeal genus *Methanococcus* meets all of these criteria.

Alignments of protein sequences from mesophilic *Methanococcus* spp. with their homologs in the extreme thermophile *M. jannaschii* sample over 7,000 amino acid replacements. Among these, we identify specific changes that were repeatedly utilized in adaptation of the proteins to environments differing by ≈50°C. The 26 pairs of amino acids with significant replacement direction bias are generally consistent with, but are much more comprehensive than, previous lists (e.g., ref. 15). Taken as a whole, the observed replacements decrease the content of uncharged polar residues, increase the content of charged residues, increase residue hydrophobicity, and increase residue volume for the proteins in *M. jannaschii* relative to their mesophilic counterparts. These shifts are general—they are seen in 88%, 83%, 89%, and 92%, respectively, of those alignments that cover at least 90% of a protein. Although each trend has been observed in previous studies (e.g., refs. 10, 16, 17, 32, 34–42), they were not reported to apply to such a large fraction of the proteins. The results that we observe might even understate the underlying trends because we have not verified that the mesophile proteins analyzed are fully adapted to their moderate-temperature environment. We would expect less adapted proteins to dilute the data contributed by more fully adapted proteins.

Uncharged Polar Residues. Of the properties examined, the uncharged-polar residue content (Ser, Thr, Asn, and Gln) had the

Table 3. Amino acid replacements that are most biased in ratio, most frequent, or most biased in number between mesophile and thermophile proteins

Replacements most biased in ratio					Most frequent replacements					Replacements most biased in number				
Replacement	Forward	Reverse	Ratio	P^*	Replacement	Forward	Reverse	Ratio	P	Replacement	Forward	Reverse	Gain	P
Val–Arg	8	1	8.0	0.039	Val–Ile	356	316	1.1	0.13	Ser–Ala	176	52	124	6.0×10^{-17}
Ser–Lys	72	10	7.2	1.0×10^{-12}	Leu–Ile	227	170	1.3	0.0049	Lys–Arg	226	133	93	1.1×10^{-6}
Gln–Arg	14	2	7.0	0.0042	Asp–Glu	219	169	1.3	0.013	Ser–Lys	72	10	62	1.0×10^{-12}
Gln–Tyr	6	0	>6.0	0.031	Lys–Arg	226	133	1.7	1.1×10^{-6}	Leu–Ile	227	170	57	0.0049
Arg–Tyr	6	0	>6.0	0.031	Glu–Lys	133	124	1.1	0.62	Asp–Glu	219	169	50	0.013
Asn–Ile	11	2	5.5	0.022	Ser–Ala	176	52	3.4	6.0×10^{-17}	Ser–Thr	90	44	46	0.000087
Asn–Leu	15	3	5.0	0.0075	Val–Leu	86	85	1.0	1.0	Asn–Lys	90	45	45	0.00013
Ser–Ile	15	3	5.0	0.0075	Met–Leu	97	55	1.8	0.00082	Met–Leu	97	55	42	0.00082
Ser–Pro	32	7	4.6	0.000070	Asn–Asp	84	65	1.3	0.14	Val–Ile	356	316	40	0.13
Ala–Gln	12	3	4.0	0.035	Tyr–Phe	72	68	1.1	0.80	Asn–Glu	70	35	35	0.00082
Asn–Arg	26	7	3.7	0.0013	Asn–Lys	90	45	2.0	0.00013	Gly–Ala	60	28	32	0.00085
Ser–Leu	18	5	3.6	0.011	Ala–Val	71	63	1.1	0.55	Thr–Ile	50	19	31	0.00024
Ser–Ala	176	52	3.4	6.0×10^{-17}	Ser–Thr	90	44	2.0	0.000087	Ser–Glu	46	15	31	0.000088
Val–Lys	33	10	3.3	0.00061	Gln–Glu	67	43	1.6	0.028	Gln–Lys	58	28	30	0.0016
Ser–Arg	13	4	3.2	0.049	Asn–Glu	70	35	2.0	0.00082	Thr–Val	60	32	28	0.0046
Ser–Tyr	13	4	3.2	0.049	Ala–Thr	53	47	1.1	0.62	Ala–Glu	50	24	26	0.0034
Gly–Arg	13	4	3.2	0.049	Phe–Leu	53	41	1.3	0.26	Ser–Pro	32	7	25	0.000070
Ser–Glu	46	15	3.1	0.000088	Thr–Val	60	32	1.9	0.0046	Thr–Glu	42	18	24	0.0027
His–Tyr	28	10	2.8	0.0051	Gly–Ala	60	28	2.1	0.00085	Gln–Glu	67	43	24	0.028
Leu–Lys	28	10	2.8	0.0051	Gln–Lys	58	28	2.1	0.0016	Ala–Pro	36	13	23	0.0014

*The random probability of directional bias greater than or equal to that observed. In the left one-third of the table, only replacements with $P < 0.05$ are included.

strongest correlation with the observed amino acid replacements. Nearly every other amino acid is preferred over these in the thermophile sequences (52 of the 64 possible replacements). There is a net loss of nine uncharged polar residues in a typical (300-aa) thermophile protein.

Although the importance to thermostability of intramolecular hydrogen bonds and polar surface area (solvent hydrogen bonding) has been emphasized (6, 17), the marked reduction in uncharged-polar residues seen here argues against increased hydrogen bonding at 85°C. Although some of the hydrogen bonds could be retained by changes to charged amino acids, most uncharged-polar residue losses involve replacement by nonpolar residues. These latter replacements with bulkier, more hydrophobic residues are apt to decrease solvent access to the interior of the protein and to increase the extent of the hydrophobic core. In addition to stabilizing folding, decreasing polar residues helps avoid the deamidations and backbone cleavages involving Asn and Gln, which can be catalyzed by Ser and Thr (34, 35).

Charged Residues. The thermophile proteins have an 8% increase in their content of fully charged residues (His does not possess a "full" charge) compared with the mesophile proteins. One of every 14 amino acid replacements increased the charge in the thermophilic homolog (6.5 residues per 300-aa protein). Replacements of the uncharged polar residues in the mesophilic sequences by charged ones in the thermophile are responsible for 75% of this increase. All 16 types of polar ↔ charged replacements favor the charged residue in the thermophile (Table 2), 10 of them significantly so ($P < 0.01$). Less conservative nonpolar → charged replacements contribute the remaining 138 new charges to the thermophile sequences, mostly through Gly → charged and Ala → charged replacements.

Although previous studies of thermophile proteins have observed increased numbers of charged groups and ionic bonds, their importance to thermal stability has been debated (2, 3, 36, 43). The recognition of networks of interconnected salt bridges in thermostable proteins has renewed interest in the role of ionic interactions (43). Some of the charge gain could also be an indirect consequence of the decrease in uncharged polar residues: polar → charged replacements would provide less labile residues while retaining hydrogen-bonding capacity. Finally, Lys residues can also contribute to local hydrophobic interactions.

Residue Hydrophobicity. Of the many hydrophobicity scales, some are strongly correlated to the differences between mesophile and thermophile proteins (e.g., refs. 24–26), whereas others are not (e.g., refs. 44 and 45). By using the scale of Zimmerman (24), replacements of polar residues in the mesophile proteins with nonpolar residues in the thermophile contribute 50% of the hydrophobicity increase, although only Ser → Ala, Thr → Ile, Ser → Pro, and Thr → Val occur frequently. An additional 27% of the hydrophobicity increase results from polar ↔ charged replacements, primarily because of polar → Lys replacements. A comparable increase in hydrophobicity is caused by the tendency of nonpolar ↔ nonpolar amino acids replacements to favor a more hydrophobic residue in the thermophile sequence (e.g., Leu → Ile, Gly → Ala and Met → Leu). Other replacements have only a minor influence on hydrophobicity.

In aqueous environments, two factors make hydrophobicity a critical issue to thermostability: hydrophobic effects (*i*) destabilize

Table 4. Correlation of amino acid residue properties with organism growth temperature

Amino acid property	Correlation coefficient	Correlation coefficient after removing effect of:			
		Polar residues	Full charge	Hydrophobicity	Volume
Uncharged polar residues	−0.128 ± 0.008	—	−0.110 ± 0.008	−0.098 ± 0.008	−0.110 ± 0.008
Full charge	0.092 ± 0.008	0.061 ± 0.007	—	0.109 ± 0.008	0.084 ± 0.008
Hydrophobicity	0.086 ± 0.006	0.020 ± 0.007	0.103 ± 0.007	—	0.058 ± 0.006
Residue volume	0.079 ± 0.006	0.044 ± 0.006	0.070 ± 0.006	0.047 ± 0.005	—

The standard deviation of each value is based on 1,000 bootstrap resamplings of the replacement data (23). Details are provided in *Materials and Methods*. The properties are uncharged polar residues (this paper), full charge (this paper), hydrophobicity (24), and residue volume (27).

unfolded forms and (*ii*) increase with temperature (32, 37–41). Most of the gain in hydrophobicity is caused by conservative amino acid replacements: minor changes that could increase van der Waals contacts and packing density without requiring major structural rearrangements. However, significant hydrophobicity is also contributed by less conservative polar → nonpolar replacements, suggesting that relatively polar regions in mesophile proteins can be integrated into the hydrophobic core in thermophile proteins. Interestingly, special importance seems to be attached to conservation of structure around the β carbon of the amino acid (compare the frequencies of replacements Thr ↔ Val and Thr ↔ Ile with Thr ↔ Leu and Ser ↔ Val); perhaps this helps to conserve the geometry of the protein backbone.

Residue Volume. Based on the data of Bigelow (27), the total residue volume increase in a typical thermophile protein corresponds to ≈20 additional methylene groups. Replacements of uncharged polar residues in the mesophile sequences (25% of all replacements) contribute 62% of the volume increase, with an additional 39% coming from replacements of one nonpolar residue by another. In the thermophile proteins, charged residues (33% of the replacements) contribute 56% of the volume increase.

Residue volume increase was observed in 92% of the full-length proteins analyzed, making it the most recurrent trend. Although correlations of residue volume with residue hydrophobicity and charge (especially Arg and Lys) make it difficult to assess how much of the volume increase is a primary effect and how much is secondary, volume is certainly important in its own right because of the ability of larger residues to exclude water from the protein interior, to fill cavities, and to reduce the entropic freedom of the unfolded protein backbone (42). This is consistent with the observation that our attempts to remove the influence of other properties lowered, but did not eliminate, the correlation of residue volume increase with temperature (Table 4).

Concluding Remarks. Although this study is based on more data, more closely related proteins, and a survey of more properties than previous comparative studies of protein thermal adaptation, limitations remain. First, even with this amount of data many of the amino acid replacements are sampled only a few times; additional data will be required to confirm or to refute some of the observed directional trends. Although this is unlikely to affect the overall trends or the replacements that we discussed, it does limit discovery of more idiosyncratic changes that might prove to be critical in specific contexts. Increasing the size of the data set will better address these issues and will allow us to ask whether some trends are dependent on specific sequence or structural contexts (16). To improve the sampling of replacements, we are generating additional sequence data from *M. maripaludis*.

A second concern is that the current sequence data are heavily weighted by proteins from multimeric protein structures (e.g., ribosomal proteins and subunits of the methyl reductase complex). Required interactions of these proteins might bias the observed trends; however, eliminating these proteins—individually or in groups—from the data set had little effect on the trends. This and other issues could be better addressed by mapping the sites of changes onto the three-dimensional structures of related proteins.

Finally, while restricting this study to the genus *Methanococcus* permitted the analysis of large numbers of very similar sequences, other environmental factors or unusual characteristics of the organisms might have affected the amino acid replacements observed. One obvious factor is that *M. jannaschii* is a barophile, although experimental data suggest that its proteins are not particularly adapted to high pressure (46). Similarly, the 30% genomic G+C content of these organisms might have introduced biases to specific amino acid usage, although it should be the same for the organisms compared. Although these complicating factors can be addressed, it would be advantageous to verify the trends by studying other pairs of related mesophilic and extremely thermophilic organisms. However, this is not simple; for the other known extreme thermophiles, either mesophilic relatives are not known, or they have very different genomic G+C contents. This latter effect can systematically bias amino acid usage (33) and thereby confound the analysis of thermal adaptation.

We are grateful to Dr. W. B. Whitman for supplying DNA sequence data from *M. maripaludis* before publication. This work was supported by funding from National Aeronautics and Space Administration (NAGW-2554) and the Department of Energy (DEFC02–95ER61963). J.H.B. received training grant support from the National Institutes of Health (GM07283) and the Department of Education (DE-P200A-40706).

1. Nicholson, H., Becktel, W. J. & Matthews, B. W. (1988) *Nature (London)* **336,** 651–655.
2. Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. (1990) *Biochemistry* **29,** 2403–2308.
3. Dao-pin, S., Sauer, U., Nicholson, H. & Mattews, B. W. (1991) *Biochemistry* **30,** 7142–7153.
4. Klemm, J. D., Wozniak, J. A., Alber, T. & Goldenberg, D. P. (1991) *Biochemistry* **30,** 589–594.
5. Heinz, D. W., Baase, W. A. & Matthews, B. W. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 3751–3755.
6. Paupit, R. A., Karlsson, R., Picot, D., Jenkins, J. A., Niklaus-Reimer, A.-S. & Jansonius, J. N. (1988) *J. Mol. Biol.* **199,** 525–537.
7. Frommel, C. & Sander, C. (1989) *Proteins* **5,** 22–37.
8. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence, M. L. & Bryan, P. N. (1989) *Biochemistry* **28,** 7205–7213.
9. Van den Burg, B., Vriend, G., Veltman, O. R., Venema, G. & Eijsink, V. G. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 2056–2060.
10. Walker, J. E., Wonacott, A. J. & Harris, J. I. (1980) *Eur. J. Biochem.* **108,** 581–586.
11. Korndorfer, I., Steipe, B., Huber, R., Tomschy, A. & Jaenicke, R. (1995) *J. Mol. Biol.* **246,** 511–521.
12. Szilagyi, A. & Zavodszky, P. (1995) *Protein Eng.* **8,** 779–789.
13. Tanner, J. J., Hecht, R. M. & Krause, K. L. (1996) *Biochemistry* **35,** 2597–2609.
14. Zuber, H. (1978) in *Biochemistry of Thermophily* (Academic, New York), pp. 267–285.
15. Argos, P., Rossman, M. G., Grau, U. M., Zuber, H., Frank, G. & Tratschin, J. D. (1979) *Biochemistry* **18,** 5698–5703.
16. Menendez-Arias, L. & Argos, P. (1989) *J. Mol. Biol.* **206,** 397–406.
17. Vogt, G., Woell, S. & Argos, P. (1997) *J. Mol. Biol.* **269,** 631–643.
18. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al.* (1996) *Science* **273,** 1058–1073.
19. Benson, D. A., Boguski, M., Lipman, D. J. & Ostell, J. (1996) *Nucleic Acids Res.* **24,** 1–5.
20. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
21. Gish, W. & States, D. J. (1993) *Nat. Genet.* **3,** 266–272.
22. Tomii, K. & Kanehisa, M. (1996) *Protein Eng.* **9,** 27–36.
23. Efron, B. & Gong, G. (1983) *Am. Stat.* **37,** 36–48.
24. Zimmerman, J. M., Eliezer, N. & Simha, R. (1968) *J. Theor. Biol.* **21,** 170–201.
25. Jones, D. D. (1975) *J. Theor. Biol.* **50,** 167–183.
26. Argos, P., Rao, J. K. & Hargrave, P. A. (1982) *Eur. J. Biochem.* **128,** 565–575.
27. Bigelow, C. C. (1967) *J. Theor. Biol.* **16,** 187–211.
28. Goldsack, D. E. & Chalifoux, R. C. (1973) *J. Theor. Biol.* **39,** 645–651.
29. Grantham, R. (1974) *Science* **185,** 862–864.
30. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. (1990) *Science* **247,** 1306–1310.
31. Russell, R. & Barton, G. (1994) *J. Mol. Biol.* **244,** 332–350.
32. Haney, P., Konisky, J., Koretke, K. K., Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Proteins* **28,** 117–130.
33. Lobry, J. R. (1997) *Gene* **205,** 309–316.
34. Tomazic, S. J. & Klibanov, A. M. (1988) *J. Biol. Chem.* **263,** 3086–3091.
35. Wright, H. T. (1991) *Crit. Rev. Biochem. Mol. Biol.* **26,** 1–52.
36. Perutz, M. F. & Raidt, H. (1975) *Nature (London)* **255,** 256–259.
37. Ikai, A. (1980) *J. Biochem.* **88,** 1895–1898.
38. Baldwin, R. L. (1986) *Proc. Natl. Acad. Sci USA* **83,** 8069–6072.
39. Privalov, P. L. & Gill, S. J. (1988) *Adv. Protein Chem.* **39,** 191–232.
40. Britton, K. L., Baker, P. J., Borges, K. M. M., Engel, P. C., Pasquo, A., Rice, D. M., Robb, F. T., Scandurra, R., Stillman, T. J. & Yip, K. S. P. (1995) *Eur. J. Biochem.* **229,** 688–695.
41. Kotsuka, T., Akanuma, S., Tomuro, M., Yamagishi, A. & Oshima, T. (1996) *J. Bacteriol.* **178,** 723–727.
42. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.
43. Yip, K. S. P., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquo, A., Chiaraluce, R., Consalvi, V., *et al.* (1995) *Structure* **3,** 1147–1158.
44. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
45. Radzicka, A. & Wolfenden, R. (1988) *Biochemistry* **27,** 1664–1670.
46. Konisky, J., Michels, P. C. & Clark, D. S. (1995) *Appl. Environ. Microbiol.* **61,** 2762–2764.