Using Motif-Based Methods in Multiple Genome Analyses: A Case Study Comparing Orthologous Mesophilic and Thermophilic Proteins[†]

David La,[‡] Melanie Silver,[‡] Robert C. Edgar,[§] and Dennis R. Livesay*,[‡]

Department of Chemistry, California State Polytechnic University at Pomona, 3801 West Temple Avenue, Pomona, California 91768, and 195 Roque Moraes Drive, Mill Valley, California 94941

Received December 31, 2002; Revised Manuscript Received May 23, 2003

ABSTRACT: Protein motifs represent highly conserved regions within protein families and are generally accepted to describe critical regions required for protein stability and/or function. In this comprehensive analysis, we present a robust, unique approach to identify and compare corresponding mesophilic and thermophilic sequence motifs between all orthologous proteins within 44 microbial genomes. Motif similarity is determined through global sequence alignment of mesophilic and thermophilic motif pairs, which are identified by a greedy algorithm. Our results reveal only modest correlation between motif and overall sequence similarity, highlighting the rationale of motif-based approaches in comprehensive multigenome comparisons. Conserved mutations reflect previously suggested physiochemical principles for conferring thermostability. Additionally, comparisons between corresponding mesophilic and thermophilic motif pairs provide key biochemical insights related to thermostability and can be used to test the evolutionary robustness of individual structural comparisons. We demonstrate the ability of our unique approach to provide key insights in two examples: the TATA-box binding protein and glutamate dehydrogenase families. In the latter example, conserved mutations hint at novel origins leading to structural stability differences within the hexamer structures. Additionally, we present amino acid composition data and average protein length comparisons for all 44 microbial genomes.

Proteins that function under standard (mesophilic) conditions tend to have similar structural stabilities, despite having different sequences and structural folds (*1*, *2*). Several organisms, mostly archaea, thrive under extreme environmental conditions, e.g., high pressure, high salt concentrations, very high and low temperatures, and extreme pH. Enzymes that function optimally in such adverse conditions mediate the metabolic and biological functions of these organisms. Proteins from thermophilic (extremely high ambient temperatures) organisms generally exhibit substantially higher intrinsic thermal stabilities than their mesophilic counterparts while retaining the basic fold characteristics of the whole family (*3*).

Although the molecular underpinnings of protein thermal stabilization have been the focus of many experimental and theoretical research efforts (for a review see Vielle et al.), the subject is only partially understood (*3*, *4*). In general, it is thought that thermostability is achieved by an increase in the type and numbers of noncovalent interactions (*5*). Analyses of all noncovalent interactions within thermophilic and mesophilic structural pairs reveal that thermophilic proteins generally have increased numbers of van der Waals interactions, hydrogen bonds, salt bridges, dipole−dipole interactions, disulfide bridges, and hydrophobic interactions (*5*−*18*). Other differences include shortening of loop regions, fewer and smaller destabilizing voids within the protein, increased structural water content, and increased incidence of ion binding (*16*, *19*−*21*). Increased conformational rigidity of the protein structure and optimization of the surface electrostatics also appear to parallel thermostability (*22*−*28*). The secondary structure propensity of each amino acid within α-helices and *β*-sheets has also been demonstrated to be linked to added stability (*29*, *30*). Despite key differences between mesophilic and thermophilic structural pairs, the overall fold and the active site of the protein generally remain unchanged (*31*).

To overcome the lack of abundant structural data for orthologous mesophilic and thermophilic protein pairs, Chakaravarty et al. have created high quality homology models taken from 30 complete bacterial genomes (nine of which are thermophilic) (*32*). This study identifies several statistically significant, specific amino acid substitutions, significantly more salt bridges in thermophiles, a slight decrease in loop length, and an increase in previously overlooked cation−*π* interactions. Additionally, statistically significant hydrophobic amino acid substitutions are reported to be consistent with decreased side chain conformational entropy.

Several studies have concentrated on sequence analysis to investigate the origins of thermostability. Much of this work has focused on differences in amino acid composition between mesophilic and thermophilic genomes. It has been observed that arginine and tyrosine are significantly more

[†] This work was supported by an American Chemical Society Petroleum Research Fund type G grant (36848-GB4), an NIH score grant (S06 GM53933), and a supercomputer allocation from the National Center for Supercomputing Applications to D.R.L.

* Corresponding author. Telephone: 909-869-4409. E-mail: drlivesay@csupomona.edu.

[‡] California State Polytechnic University at Pomona.

[§] Roque Moraes Drive, Mill Valley, CA.

10.1021/bi027435e CCC: $25.00 © 2003 American Chemical Society
Published on Web 07/09/2003

common in thermophilic proteins, whereas cysteine and serine occur less frequently (*5*). Jaenicke et al. compared 14 bacterial and archaeal genomes (five of which were thermophilic) in an attempt to uncover some general rules leading to thermostability (*3*). Some general correlations were uncovered, such as an increased abundance of charged residues and a decreased abundance of glutamine in thermophilic genomes, yet a consistent mechanism is still elusive. A recent analysis of the clusters of orthologous groups (COG)[1] of proteins database reveals only a single protein specific to hyperthermophilic (∼100 °C) genomes compared to moderately thermophilic (65−90 °C) and mesophilic genomes (*33*−*35*). The COG database is a compendium of all predicted gene products encoded from 44 microbial genomes. Each of the 3166 COGs consists of at least three orthologous proteins, whereas some of the larger COGs contain several hundred. The lack of novel hyperthermophilic proteins within the COG database further emphasizes the evolutionary connectivity between mesophilic and thermophilic proteins.

In this comparative proteomic study we present a robust, unique approach for comparing orthologous mesophilic and thermophilic proteins from multiple genomes. After assigning every protein within the COG database into a mesophilic or thermophilic subfamily (Table 1), we identify short, highly conserved regions (motifs) within each. Subsequently, we seek to answer the question, “How well are motif pairs conserved between mesophilic and thermophilic subfamilies?” This question is biologically relevant because motifs frequently describe critical regions of the protein necessary to functionality and/or structural stability. Knowing how well motif regions have been conserved is used as a metric to gauge the similarity between critical regions of the protein. Further, assessment of the corresponding mesophilic and thermophilic motif pairs can provide key biochemical insights. Nonvarying positions that are conserved in both subfamilies highlight their importance to structure and/or function, whereas conserved mutations between the two subfamilies might be related to increased thermostability. Our results advocate that motif-based approaches are robust enough for large-scale multigenome analyses, resulting in, that we are aware of, the most comprehensive mesophilic vs thermophilic comparison ever completed.

METHODS

Identifying Conserved Regions (Motifs). All 44 genomes maintained in the COG database (Table 1) are assigned to mesophilic and thermophilic subfamilies for further analysis (*34*, *35*). If either subfamily contains less than two sequences, the COG is eliminated, leaving 1593 COGs (conserved regions cannot exist in a single sequence). Motifs for each subfamily are identified using MEME version 3.0.3 (*36*). Consensus sequences from each motif pair are aligned using a modified version of ALIGN (*37*). Our version (nALIGN) calculates a normalized global and local alignment score that eliminates the sequence-length bias in the total score, essentially calculating a sequence position average score. It could be argued that using a consensus sequence (versus a multiple sequence alignment derived profile) is insufficient to accurately represent the identified motif pairs. However, due to the homology between identified motifs, our results indicate that a consensus representation is sufficient.

Table 1: Genomes Analyzed in This Work

genome name	genome code	no. of proteins in COGs	optimal growth temp (°C)[a]
mesophilic			
Borrelia burgdorferi	Bbu	716	37
Bacillus halodurans	Bha	2998	30
Bacillus subtilis	Bsu	3006	30
Buchnera sp. APS	Buc	585	n/a[b]
Caulobacter crescentus	Ccr	2758	30
Candida albicans	Cda	2839	30
Campylocater jejuni	Cje	1341	37
Chlamydia pneumoniae	Cpn	667	37
Chlamydia trachomatis	Ctr	649	37
Deinococcus radiodurans	Dra	2316	30
Escherichia coli O157	EcZ	3807	37
Escherichia coli K12	Eco	3553	37
Halobacterium sp. NRC-1	Hbs	1768	37
Haemophilus influenzae	Hin	1585	37
Helicobacter pylori 26695I	Hpy	1126	37
Helicobacter pylori J99	jHp	1105	37
Lactococcus lactis	Lla	1668	30
Mycoplasma genitalium	Mge	393	37
Mycobacterium leprae	Mle	1204	n/a[b]
Mesorhizobium loti	Mlo	5175	26
Mycoplasma pneumoniae	Mpn	437	37
Mycobacterium tuberculosis	Mtu	2726	37
Neisseria meningitidis Z2491	NmA	1532	37
Neisseria meningitidis MC58	Nme	1545	37
Pseudomonas aeruginosa	Pae	4626	37
Pasteurella multocida	Pmu	1805	37
Rickettsia prowazekii	Rpr	714	37
Saccharomyces cerevisiae	Sce	2411	20
Streptococcus pyogenes	Spy	1256	37
Synechocystis	Syn	2347	25
Treponema pallidum	Tpa	732	37
Ureaplasma urealyticum	Uur	410	37
Vibrio cholerae	Vch	2979	28
Xylella fastidiosa	Xfa	1645	26
hyperthermophilic/thermophilic			
Aquifex aeolicus	Aae	1374	80
Archaeoglobus fulgidus	Afu	1950	83
Aeropyrm pernix	Ape	1198	90
Methanococcus jannaschii	Mja	1399	83
Methanobacterium thermoautotrophicus	Mth	1453	65
Pyrococcus abyssi	Pab	1511	98
Pyrococcus horikoshii	Pho	1431	98
Thermoplasma acidophilum	Tac	1259	66
Thermotoga maritima	Tma	1579	80
Thermoplasma volcanium	Tvo	1269	60

[a] Optimal growth temperatures taken from German Collection of Microorganisms and Cell Cultures (http://www.dsmz.de/). [b] Optimal growth temperatures for *M. leprae* and *Buchnera* are not defined as neither can live outside of their respective hosts.

MEME uses expectation maximization to identify conserved regions in a set of ungapped DNA or protein sequences. A stand-alone version of MEME is used with custom settings that include empirical Dirichlet mixture priors weighted according to the megaprior heuristic (*38*), a minimum and maximum motif width of 25 and 50, a motif model biased toward zero or one motif occurrence per sequence, a maximum motif search number of three, and an

[1] Abbreviations: COG, clusters of orthologous groups; MEME, multiple Em for motif elicitation; MOSS, motif overall similarity score; HMM, hidden Markov model; PSS, entire protein sequence score; BLAST, basic local alignment search tool; PAM, point-accepted mutation; BLOSUM, Blocks substitution matrix; TBP, TATA-box binding protein; GDH, glutamate dehydrogenase.

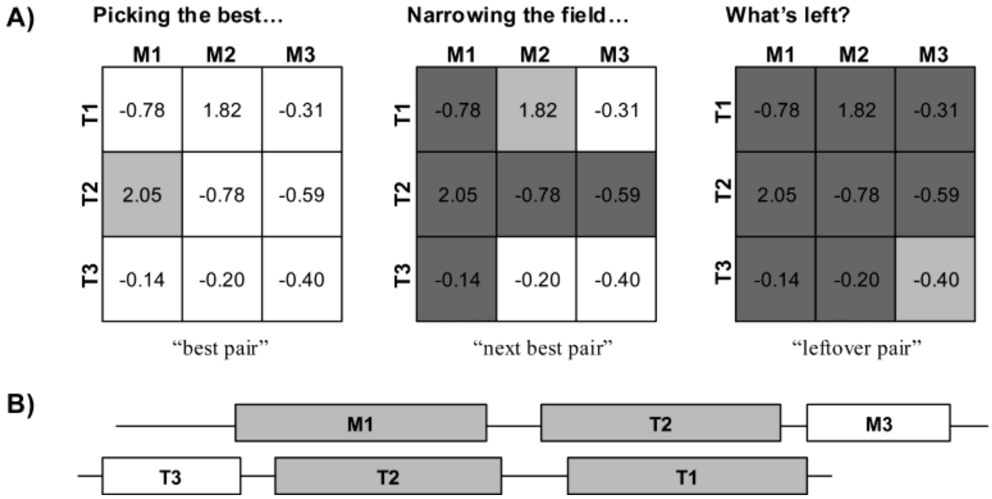


FIGURE 1: (A) Cartoon representing the greedy algorithm used to match the corresponding motifs. (B) The greedy algorithm is necessary due to the heterogeneity in the number of and naming of identified motifs between each subfamily pair. In the above example, motifs M1:T2 and M2:T1 correspond to each other, whereas the leftover pair of M3:T3 does not.

E value threshold of 0.01. If MEME identifies no motifs in either subfamily, the COG is eliminated, leaving 1354 COGs.

To correctly match pairs of orthologous motifs, we must consider several alignment combinations. MEME is parametrized to search for up to three motifs, making it necessary to evaluate up to nine different alignment combinations (three mesophilic versus three thermophilic). Because we are attempting to identify corresponding mesophilic and thermophilic motifs, once a pair has been established, both of those motifs are excluded from being used again. A sample of all nine possible alignment combinations for a COG with three consensus sequences in both subfamilies is illustrated in Figure 1. To identify motif pairs, we use the following greedy procedure. (a) For all possible combinations within a particular COG, select the pairwise alignment with the highest normalized alignment score (*best pair*) and eliminate those motifs from further consideration, (b) if additional alignment combinations exist, repeat step (a) once to find the *next best pair*, and (c) if a last score exists, this is defined as the *leftover pair*. An illustration of our algorithm is shown in Figure 1A. This method is applied to all COGs in our sample to extract the most likely scoring combinations of corresponding mesophilic/thermophilic motif pairs.

We define the motif overall similarity score (MOSS) to indicate the global similarity between all motif pairs within a given family. For cases where the best pair score, next best pair score, and leftover pair score are all positive, MOSS is simply the sum of the three values. However, if any of the three scores are negative, then that particular score is reset to zero. Although seemingly arbitrary, this approach is necessary as we are essentially asking two separate questions: (a) "do the two motifs correspond to each other?" (which we define by positive nALIGN scores) and (b) "if they are a corresponding pair, how similar is this pair?" (which is determined by the magnitude of the positive value). This point is clarified by the example in Figure 1. Here MOSS $= 2.05 + 1.82 + 0.00 = 3.87$, which is indicative of the overall similarity between M1:T2 and M2:T1. However, if the nALIGN score for M3:T3 were included in MOSS, then the resultant value $(2.05 + 1.82 - 0.40 = 3.47)$ would underestimate the similarity between the two correctly matched pairs.

Identifying Conserved Mutations. Conserved mutations between mesophilic and thermophilic subfamilies are identified from the MEME output. Each MEME output includes a motif profile that indicates the occurrence (0−10, with 10 being 100%) of each amino acid at each position within the identified motif. After aligning each subfamily's consensus sequence, we identify the conserved positions between the two corresponding profiles. The threshold used to defined conservation operates on a user input sliding integer scale from the range above. For example, a threshold of 4 will result in more conserved residues than if the threshold were more strict, say 8. All of our results and the tools to analyze them (including the original MEME outputs) are accessible through the web at http://www.csupomona.edu/∼drlivesay/pmap/.

Quantifying the Homology between Whole Sequence Subfamilies. The Coach algorithm is used to align "whole sequence" profiles of each corresponding mesophilic and thermophilic subfamilies. Coach scores and aligns two multiple sequence alignments to each other by constructing a profile hidden Markov model (HMM) (*39*) from one of the alignments. In conventional profile HMM methods, a sequence is aligned to the HMM by finding the most probable way that the HMM can generate the sequence. This is equivalent to finding the optimal assignment of letters in the sequence to emitter states. Coach generalizes this in the natural way to a multiple alignment: the HMM is assumed to have generated all sequences in the alignment and identifies the most probable assignments of columns to emitter states. A profile hidden Markov model is estimated from one alignment (A). The second alignment (B) is aligned to that profile HMM, with the constraint that all letters in a given column of (B) are assigned to the same emitter state in the HMM. If a column in (B) is assigned to a match state, this column is identified as being homologous to the column in (A) which was used as a template for that match state. Otherwise, a column in (B) is assigned to an HMM insert state and is then interpreted as not corresponding to any consensus position in (A). The log-odds score for this alignment is computed by summing the score for each sequence, assuming the path through the HMM induced on that sequence by the alignment. [Sequences are not aligned

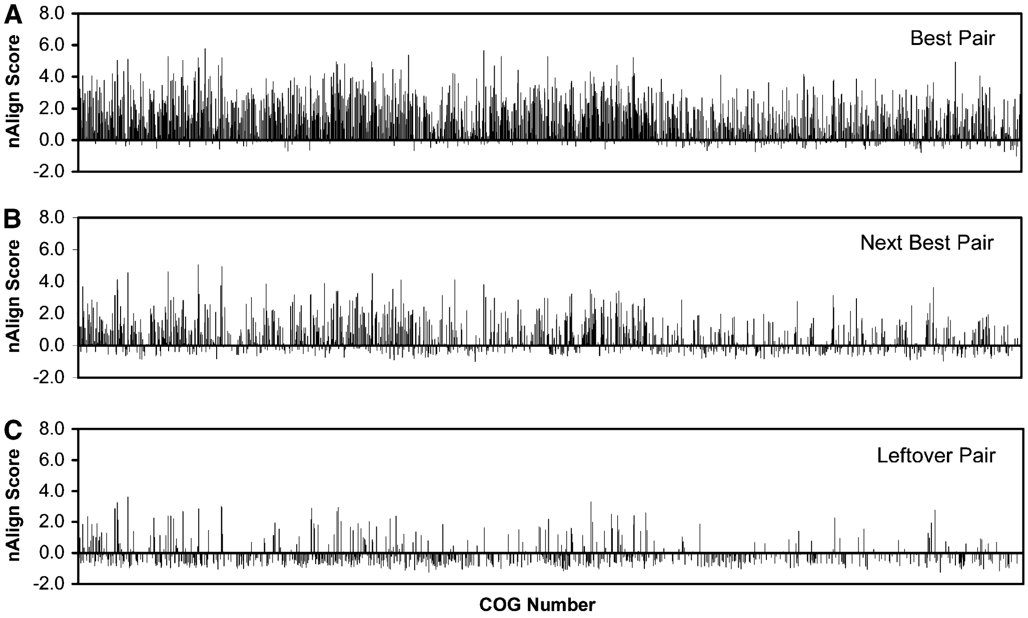


FIGURE 2: Bar diagrams showing the global nALIGN scores for each matched pair for all COGs returning three motifs in both the mesophilic and thermophilic subfamilies. (A) Only 9.2% returned negative nALIGN scores in the best pair, whereas that number significantly increases in the (B) next best pair (30.7%) and (C) leftover pair (75.0%). Summing all of the nonnegative pair scores tabulates the motif overall similarity score (MOSS).

individually to the HMM, as this could cause the alignment defined by (B) to be lost.] Log-odds scores are computed relative to the null model defined by the HMMER package (*40*) and corrected for sequence length and the number of sequences in the alignment so that scores for alignments of different sizes are comparable. The end result is a log-odds score for an alignment against an HMM that can be interpreted in a similar way to the log-odds score of a sequence against an HMM, i.e., as a similarity or relatedness measure which we call the protein similarity score (PSS). In detail, the procedure is as follows.

First, each subfamily is aligned using ClustalW version 1.8 with default parameters (*41*). Sequences are then weighted using the scheme described by Gerstein et al. (*42*). The profile HMM is constructed from either subfamily alignment, assigning match states to columns with fewer than 50% gapped positions (*39*). Dirichlet mixture priors (*43*) are added to estimate probability distributions for transitions and for match state emissions; the observed (weighted) sequence count is scaled to the fixed value of three to prevent the priors from being swamped by observations. This HMM is then used to align the second subfamily by determining the path implied for each sequence by the most probable assignment of columns in the input alignment to HMM emitter states. This gives us the average probability per sequence, P_s. Suppose HMM_t was constructed from the thermophilic subfamily and used to align the ClustalW alignment of the mesophilic subfamily (Aln_m), which results in the combined alignment of the two subfamilies. The following relationship score is calculated:

$$R_{mt} = \log_2 P_s(Aln_m|HMM_t)/P(Aln_m|null)$$

where null is a simple null model that emits sequences according to the observed background probabilities of amino acids. A positive score suggests a relationship between the two subfamilies that is more likely than chance; a negative score suggests that they are distantly related or unrelated. The whole protein similarity score shows a length bias; to correct for this, we divide by the number of match states in the HMM, which we denote by length(HMM):

$$r_{mt} = R_{mt}/\text{length}(HMM_t)$$

This measure requires us to choose one subfamily as the HMM template. We symmetrize by averaging over the two choices, resulting in the PSS:

$$PSS = (r_{mt} + r_{tm})/2$$

RESULTS AND DISCUSSION

Motif Analysis. MEME is parametrized to return up to three motifs but often identifies less, especially in shorter sequences. MEME returns 748 COGs with three identified motifs in each subfamily and 292 with two identified motifs in either (or both) subfamily, and 314 COGs are returned with only one motif in either (or both) subfamily. MEME fails to identify any motifs in either (or both) subfamily 239 times. From Figure 2, it is apparent that the degree of motif similarity (as measured by global alignment) decreases for each subsequent identified motif. Of the 1354 COGs that are analyzed here, 86.3% return a positive global best pair score, 48.2% return a positive global next best pair score, and only 13.8% return a positive global leftover pair score.

The motif overall similarity score (MOSS) is used to represent the overall similarity between all identified motif pairs (according to the greedy algorithm described above). Figure 3A clearly demonstrates that some COGs have very high motif similarity. However, 37.5% of the COGs have MOSSs less than 1.0 and another 18.6% with MOSSs less than 2.0. In fact, only 17.5% of the COGs have MOSS values greater than 4.0. We attempt to identify any bias in our MOSS scores with respect to minimum subfamily size. Figure 3B indicates that the MOSS distribution is inde-

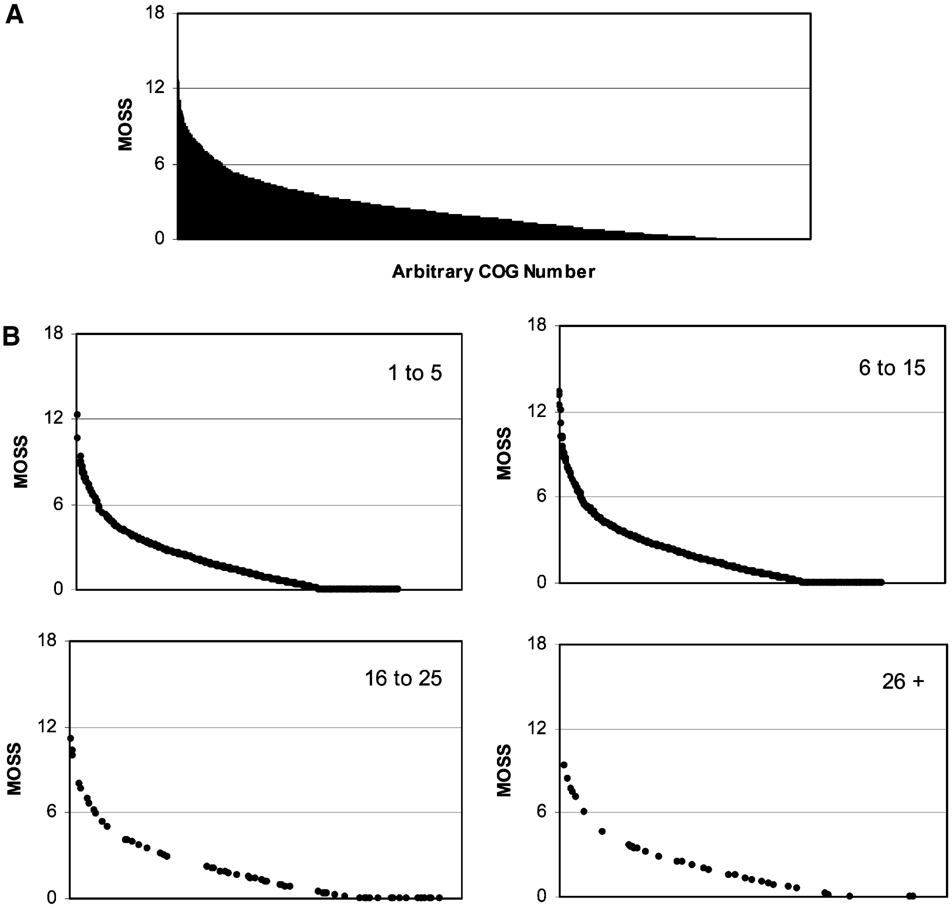


FIGURE 3: (A) Global alignment motif overall similarity scores (MOSSs) plotted in descending order versus an arbitrary COG number. The MOSS values range between 0.00 and 13.34, indicating that some of the overlapping motif regions are very similar, whereas many have no appreciable similarity. (B) Line graphs showing no bias in MOSS distributions with respect to subfamily size. In this instance, each subfamily pair was placed into one of four groups, dependent on the size of the smaller of the two subfamilies. In each of the above figures the *x*-axis is COG number, although not necessarily in the same order. In each case the *x*-axis has been reshuffled such that the COGs are listed in descending order according to their *y*-axis values.

pendent of the minimum number of sequences in either subfamily from each COG.

All proteins within each COG are homologous on the basis of their original classification; however, the sequence similarity within each COG is still variable. It is possible that the extent of similarity between the motif regions might parallel the similarity across the whole protein sequence. In other words, COGs with large MOSSs may correspond to proteins with high overall homology. If this is the case, our method will have failed to quantify anything novel. To test for this possibility, we compare MOSS against the protein similarity score (PSS), which quantifies the similarity between whole sequence profiles representing each subfamily. There is a direct correlation between MOSS and PSS in many instances (Figure 4). However, the overall correlation is modest, $R = 0.477$. This result provides a list of protein families where the similarity between the motif regions is significantly greater than or less than that of the entire sequence (points far away from the correlation line). The modest correlation demonstrated in Figure 4 strongly supports the value of motif methods in comprehensive analyses. Beyond the direct applicability to thermostability, this result confirms the usefulness of motif-based methods in other postgenomic endeavors, i.e., functional annotation, classification of gene products, and active site prediction (*44*).

It is fairly straightforward to understand how COGs with increased motif similarity with respect to the whole sequence (high MOSS/PSS ratio) have come about. These results parallel the consensus view of protein evolution as they highlight the importance of conserved motifs and the plasticity of the intervening sequence space. On the other hand, it is not obvious why a sequence would have less motif similarity with respect to the whole protein (low MOSS/PSS ratio). One factor behind this observation arises from how strictly we define global motif similarity. For example, in COG1156 motifs M1 and T1 correspond to each other. However, only approximately half of each motif correctly aligns (the motifs only partially overlap with each other), resulting in a low global nALIGN score (0.934). In such instances, using a local alignment scheme makes much more sense. Low MOSS/PSS ratios also occur when motif pairs do not correspond to each other are present, yet the remainder of the alignment aligns well (such as in Figure 1B).

Testing the Validity of the Identified Motifs and Motif Pairs. We use cross-validation to test the validity of the identified motifs and the accuracy of using a consensus sequence representation. A data set consisting of a random portion (~80%) of each subfamily is created and used by MEME to identify the motif regions (same parameters as above). The resultant motifs are then BLASTed against the remaining sequences in order to determine the predictive

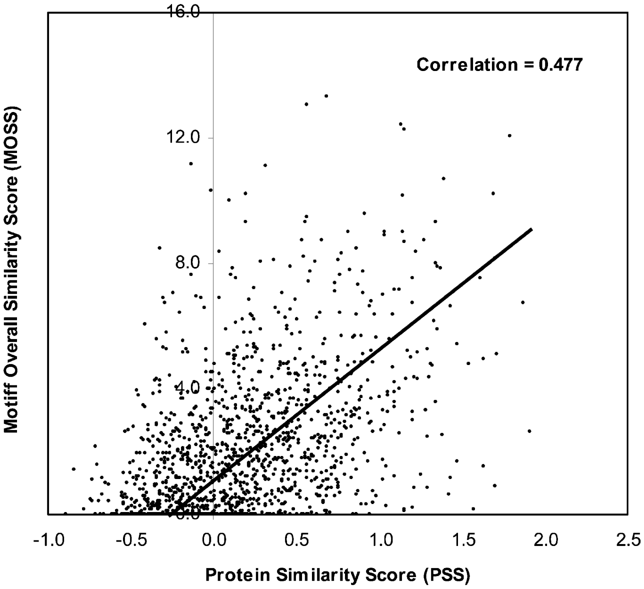


FIGURE 4: Global motif overall similarity score (MOSS) plotted against the overall sequence homology (protein similarity score, PSS) for each analyzed COG. There is a direct correlation between MOSS and PSS in many instances, but the overall correlation is modest ($R = 0.477$). This result demonstrates the applicability of motif-based methods (versus more common multiple sequence alignments) in multigenome analyses.

Table 2: Testing the Validity of the Identified Motifs and Motif Pairs

(A) Testing the Predictive Ability of the Identified Motifs[a]

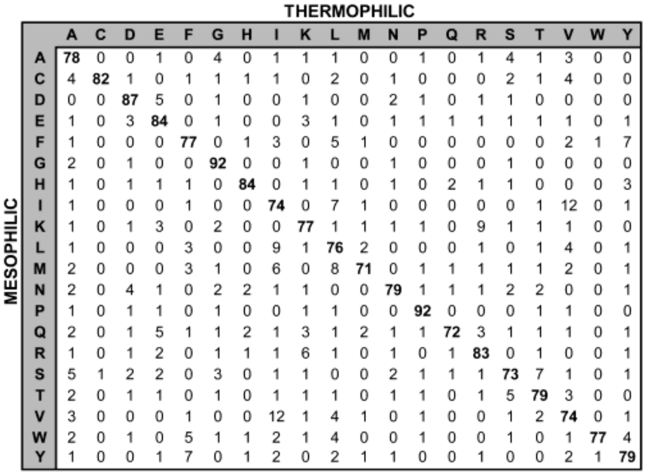
subfamily	motif 1	motif 2	motif 3
mesophilic	93.3 ± 0.9	86.0 ± 1.4	79.0 ± 1.7
thermophilic	94.9 ± 1.0	90.8 ± 1.6	86.7 ± 2.0

(B) Correlation of nALIGN Results Using Different Scoring Matrices[b]

	PAM 250	BLOSUM 50	Gonnett
PAM 250			
BLOSUM 50	0.988		
Gonnett	0.983	0.992	

[a] Average percentages with 95% confidence intervals for motif cross-validation analysis. MEME is used to identify conserved regions from 80% (randomly chosen) of each subfamily. The consensus sequence from the returned motifs are BLASTed against the remaining 20%. The scores presented here indicate how often each BLASTed consensus sequence was observed within the remaining sequences. [b] Correlation coefficients comparing the nALIGN results for each of the three scoring matrices.

ability of our motif representation. As expected, the predictive ability of the highest scoring motif is larger than each subsequent motif but is substantial in all instances (Table 2A). Additionally, we test for bias in how similarity is defined in the nALIGN results using various scoring matrices (PAM 250, BLOSSUM 50, and Gonnett) (*45*−*47*). Changing the scoring matrix results in no overall differences; the pairwise correlation between the three scoring matrix results is always greater than 0.98 (Table 2B).



THERMOPHILIC

MESOPHILIC	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**78**	0	0	1	0	4	0	1	1	1	0	0	1	0	1	4	1	3	0	0
C	4	**82**	1	0	1	1	1	1	0	2	0	1	0	0	0	2	1	4	0	0
D	0	0	**87**	5	0	1	0	0	1	0	0	2	1	0	1	1	0	0	0	0
E	1	0	3	**84**	0	1	0	0	3	1	0	1	1	1	1	1	1	1	0	1
F	1	0	0	0	**77**	0	1	3	0	5	1	0	0	0	0	0	0	2	1	7
G	2	0	1	0	0	**92**	0	0	1	0	0	1	0	0	0	1	0	0	0	0
H	1	0	1	1	1	0	**84**	0	1	1	0	1	0	2	1	1	0	0	0	3
I	1	0	0	0	1	0	0	**74**	0	7	1	0	0	0	0	0	1	12	0	1
K	1	0	1	3	0	2	0	0	**77**	1	1	1	1	0	9	1	1	1	0	0
L	1	0	0	0	3	0	0	9	1	**76**	2	0	0	0	1	0	1	4	0	1
M	2	0	0	0	3	1	0	6	0	8	**71**	0	1	1	1	1	1	2	0	1
N	2	0	4	1	0	2	2	1	1	0	0	**79**	1	1	1	2	2	0	0	1
P	1	0	1	1	0	1	0	0	1	1	0	0	**92**	0	0	0	1	1	0	0
Q	2	0	1	5	1	1	2	1	3	1	2	1	1	**72**	3	1	1	1	0	1
R	1	0	1	2	0	1	1	1	6	1	0	1	0	1	**83**	0	1	0	0	1
S	5	1	2	2	0	3	0	1	1	0	0	2	1	1	1	**73**	7	1	0	1
T	2	0	1	1	0	1	0	1	1	1	0	1	1	0	1	5	**79**	3	0	0
V	3	0	0	0	1	0	0	12	1	4	1	0	1	0	0	1	2	**74**	0	1
W	2	0	1	0	5	1	1	2	1	4	0	0	1	0	0	1	0	1	**77**	4
Y	1	0	0	1	7	0	1	2	0	2	1	1	0	0	1	0	0	2	1	**79**

FIGURE 5: Conserved mutations within overlapping motif regions (past the 60% conservation threshold) indicating the conserved differences between the mesophilic and thermophilic pairs. Each value represents the percentage of conserved mesophilic residue mutations (or conservation). This substitution matrix is different from most traditional scoring matrices (i.e., PAM or BLOSUM) because it is not symmetrical. This is because, here, all tabulated mutations have a directionality; all results represent mutations from the mesophilic consensus sequence to the thermophilic. For example, the (mesophilic to thermophilic) mutation of Asn to Asp is twice as common as the opposite.

Analysis of Conserved Mutations between the Mesophilic and Thermophilic Subfamilies. Several experimental studies have used systematic mutation and stability analysis to pinpoint the exact amino acid mutations leading to thermostability (*23*, *25*, *26*, *48*−*54*). Due to the time requirements of such work, these methods are only accessible to a single (or very small set of) protein(s). Inasmuch as each of these methods is impractical for complete genome analyses, we identify conserved mutations to highlight evolutionarily conserved differences between the two subfamilies. These results provide key residue differences that are potentially related to increased thermostability. In addition, our results provide strictly conserved (nonvarying) positions, which are expected to be critically linked to structure and/or function of the whole family.

Overall conserved mutation results (past the 60% conservation threshold) are presented in Figure 5. The most conserved (least varying) residues are structurally unique glycine and proline residues, whereas glutamine and methionine are the least conserved. The most common substitutions are similar to those in standard scoring matrices, with some subtle differences. The most common conserved mutation is between valine and isoleucine. Conserved mutation of either valine or isoleucine to leucine occurs less frequently (relative to the valine/isoleucine substitution rate) than predicted by standard scoring matrices, likely due to leucine's lowered β-sheet propensity and increased side chain conformational entropy. Conserved mutations between chemically similar arginine/lysine and phenylalanine/tyrosine also occur frequently. Mutations resulting in thermophilic cysteine, asparagine, or glutamine residues rarely occur.

Using this type of analysis on specific examples, we are able to discern some general conclusions regarding the molecular basis of added stability within particular protein families. Conserved mutation results are most telling when used in concert with structural studies. Together, one can exactly identify particular factors contributing to thermostability within each set of solved structures and test the evolutionary robustness of the identified differences. Also, key insights hidden from limited structural analyses are generally exposed. A critical examination into the molecular basis of thermostability for each family is well beyond the

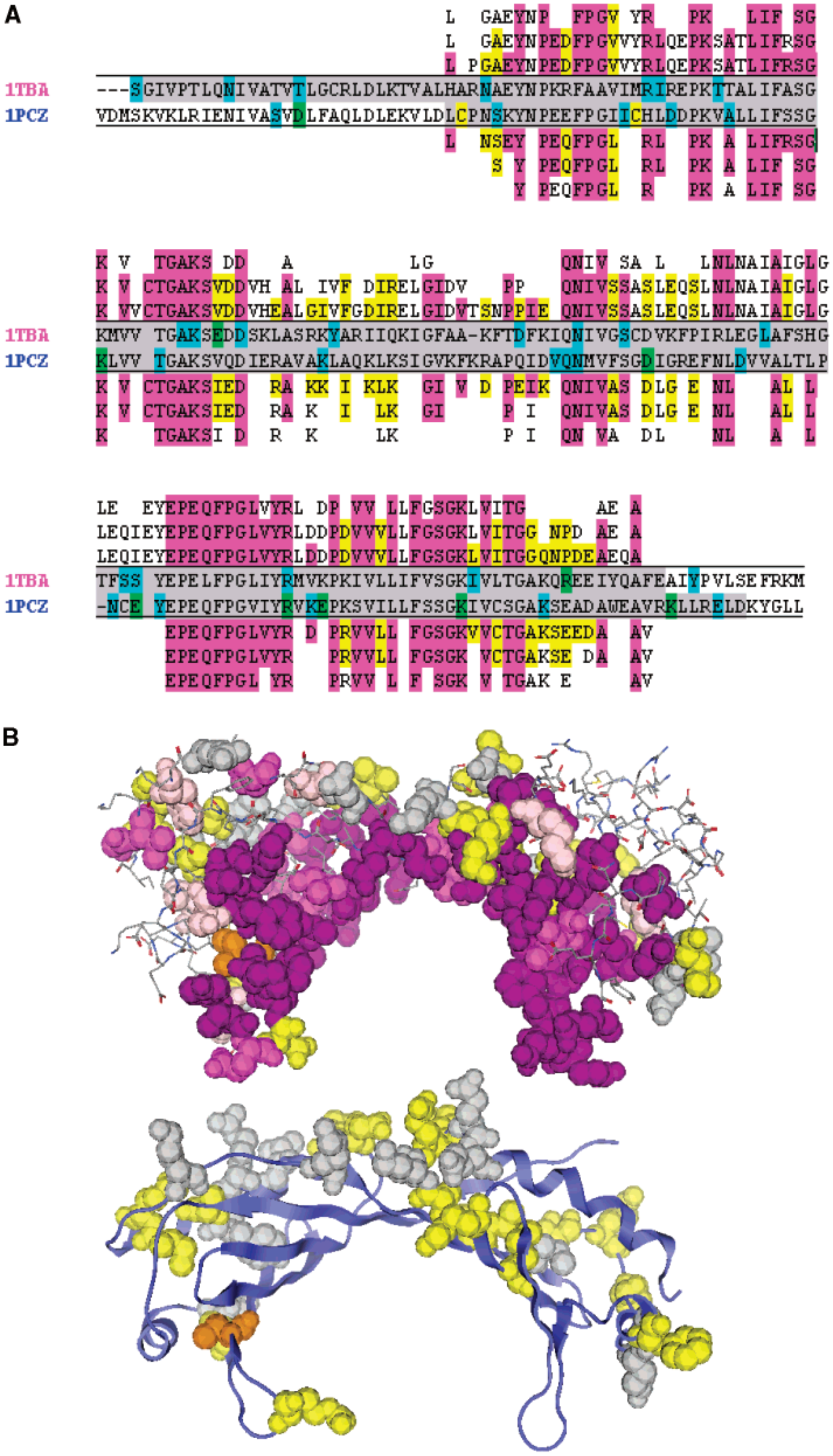

A
L GAEYNP FPGV YR PK LIF SG
L GAEYNPEDFPGVVYRLQEPKSATLIFRSG
L PGAEYNPEDFPGVVYRLQEPKSATLIFRSG
1TBA ---SGIVPTLQNIVATVTLGCRLDLKTVALHARNAEYNPKRFAAVIMRIREPKTTALIFASG
1PCZ VDMSKVKLRIENIVASVDLFAQLDLEKVLDLCPNSKYNPEEFPGIICHLDDPKVALLIFSSG
L NSEY PEQFPGL RL PK A LIFRSG
S Y PEQFPGL RL PK A LIF SG
Y PEQFPGL R PK A LIF SG
K V TGAKS DD A LG QNIV SA L LNLNAIAIGLG
K V CTGAKSVDDVH AL IVF DIRELGIDV PP QNIVSSASLEQSLNLNAIAIGLG
K VVCTGAKSVDDVHEALGIVFGDIRELGIDVTSNPPIE QNIVSSASLEQSLNLNAIAIGLG
1TBA KMVV TGAKSEDDSKLASRKYARIIQKIGFAA-KFTDFKIQNIVGSCDVKFPIRLEGLAFSHG
1PCZ KLVV TGAKSVQDIERAVAKLAQKLKSIGVKFKRAPQIDVQNMVFSGDIGREFNLDVVALTLP
K V CTGAKSIED RA KK I KLK GI V D PEIK QNIVAS DLG E NL AL L
K V CTGAKSIED RA K I LK GI P I QNIVAS DLG E NL AL L
K TGAKSI D R K LK P I QN VA DL NL A L
LE EYEPEQFPGLVYRL DP VV LLFGSGKLVITG AE A
LEQIEYEPEQFPGLVYRLDDPDVVVLLFGSGKLVITGG NPD AE A
LEQIEYEPEQFPGLVYRLDDPDVVVLLFGSGKLVITGGQNPDEAEQA
1TBA TFSS YEPELFPGLIYRMVKPKIVLLIFVSGKIVLTGAKQREEIYQAFEAIYPVLSEFRKM
1PCZ -NCE YEPEQFPGVIYRVKEPKSVILLFSSGKIVCSGAKSEADAWEAVRKLLRELDKYGLL
EPEQFPGLVYR D PRVVLL FGSGKVVCTGAKSEEDA AV
EPEQFPGLVYR PRVVLL FGSGK VCTGAKSE DA AV
EPEQFPGL YR PRVV L F SGK V TGAK E AV
B

FIGURE 6: (A) Structural alignment from MOE of the mesophilic TATA-box binding proteins from *S. cerevisiae* (1TBA) and the thermophilic Archaeon *P. woesei* (1PCZ). Thermostability arises from an added disulfide bond (yellow) and three added salt bridges (green). There are nearly identical numbers of hydrogen bonds (blue). Identified motifs are shaded gray. Amino acids above (mesophilic) and below (thermophilic) the structural alignment represent conserved consensus sequence identity at 40%, 60%, and 80% conservation thresholds. Nonvarying positions are colored pink, whereas the conserved mutations are colored yellow. (B) Nonvarying positions and conserved mutations at 40%, 60%, and 80% thresholds (nonvarying, light pink, pink, and magenta; conserved mutations, gray, yellow, and orange) are highlighted on the thermophilic structure. The bottom structure is the same as the previous, but the conserved positions have been removed for clarity.

scope of this work. Here, our aim is to demonstrate the utility of our robust, unique approach toward uncovering general rules leading to increased thermostability in dissimilar protein families. As such, we present two examples of how these results can be used to supplement our current understanding within the TATA-box binding protein and glutamate dehy-

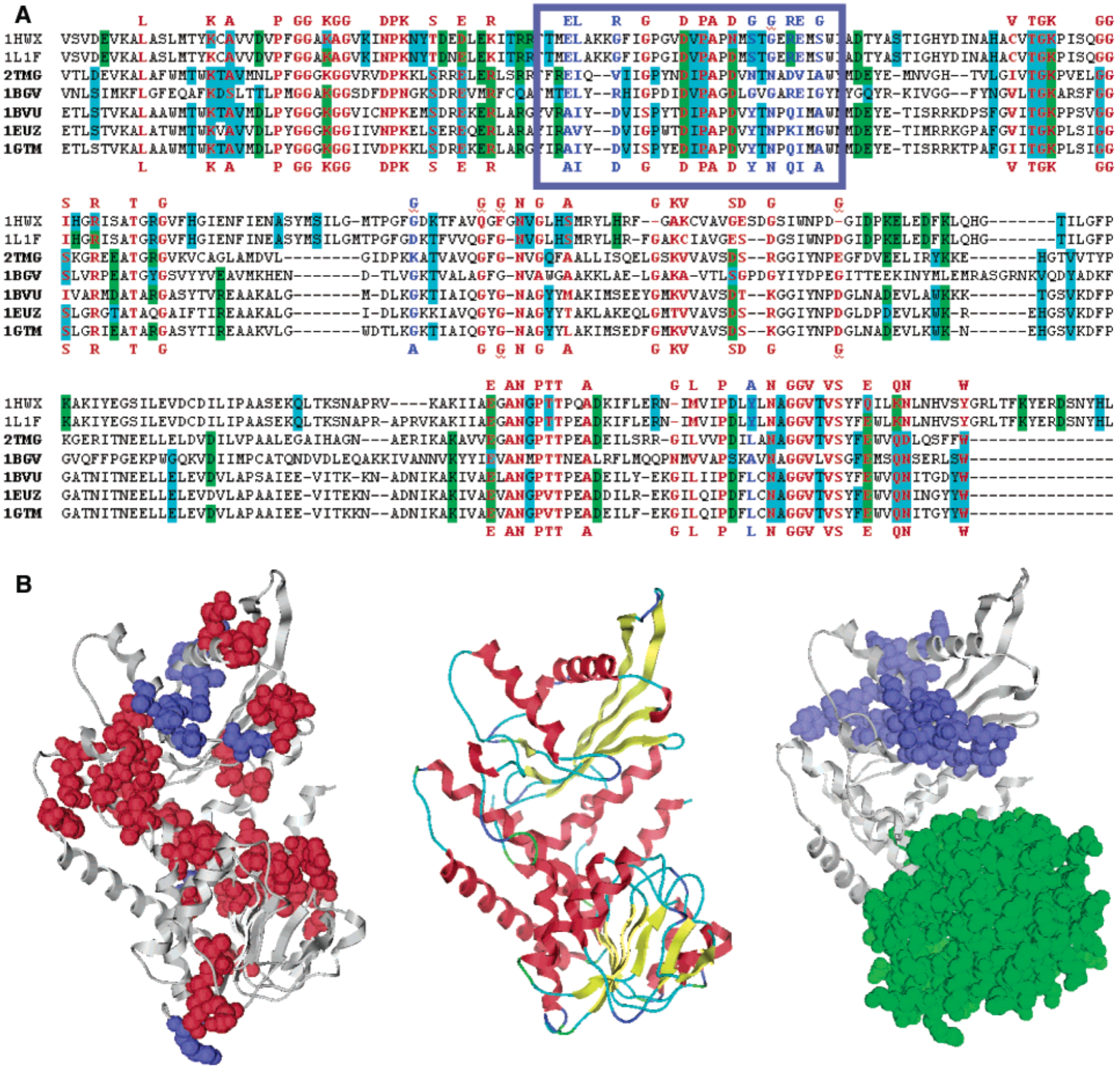

A
L KA P GG KGG DPK S E R EL R G D PA D G G RE G V TGK GG
1HWX VSVDEVKALASLMTYKCAVVDVPFGGAKAGVKINPKNYTDEDLEKITRRTMELAKKGFIGPGVDVPAPNMSTGEREMSWIADTYASTIGHYDINAHACVTGKPISQGG
1L1F VSVDEVKALASLMTYKCAVVDVPFGGAKAGVKINPKNYTDNELEKITRRTMELAKKGFIGPGIDVPAPDMSTGEREMSWIADTYASTIGHYDINAHACVTGKPISQGG
2TMG VTLDEVKALAFWMTWKTAVMNLPFGGGKGGVRVDPKKLSRRELERLSRRFREIQ--VIIGPYNDIPAPDVNTNADVIAWYMDEYE-MNVGH--TVLGIVTGKPVELGG
1BGV VNLSIMKFLGFEQAFKDSLTTLPMGGAKGGSDFDPNGKSDREVMRFCQAMTELY--RHIGPDIDVPAGDLGVGAREIGYMYGQYR-KIVGG--FYNGVLTGKARSFGG
1BVU ETLSTVKALAAWMTWKTAVMDLPYGGGKGGVICNPKEMSDREKERLARGYVRAIY--DVISPYTDIPAPDVYTNPQIMAWMMDEYE-TISRRKDPSFGVITGKPPSVGG
1EUZ ETLSTVKALATWMTWKVAVVDLPYGGGKGGIIVNPKELSEREQERLARAYIRAVY--DVIGPWTDIPAPDVYTNPKIMGWMMDEYE-TIMRRKGPAFGVITGKPLSIGG
1GTM ETLSTVKALAAWMTWKTAVMDLPYGGGKGGIIVDPKKLSDREKERLARGYIRAIY--DVISPYEDIPAPDVYTNPQIMAWMMDEYE-TISRRKTPAFGIITGKPLSIGG
L KA P GG KGG DPK S E R AI D G D PA D Y N QI A V TGK GG
S R T G G G G N G A G KV SD G G
1HWX IHGRISATGRGVFHGIENFIENASYMSILG-MTPGFGDKTFAVQGFGNVGLHSMRYLHRF--GAKCVAVGESDGSIWNPD-GIDPKELEDFKLQHG-------TILGFP
1L1F IHGRISATGRGVFHGIENFINEASYMSILGMTPGFGDKTFVVQGFG-NVGLHSMRYLHR-FGAKCIAVGES--DGSIWNPDGIDPKELEDFKLQHG-------TILGFP
2TMG SKGREEATGRGVKVCAGLAMDVL--------GIDPKKATVAVQGFG-NVGQFAALLISQELGSKVVAVSDS--RGGIYNPEGFDVEELIRYKKE-------HGTVVTYP
1BGV SLVRPEATGYGSVYYVEAVMKHEN-------D-TLVGKTVALAGFG-NVAWGAAKKLAE-LGAKA-VTLSGPDGYIYDPEGITTEEKINYMLEMRASGRNKVQDYADKF
1BVU IVARMDATARGASYTVREAAKALG-------M-DLKGKTIAIQGYG-NAGYYMAKIMSEEYGMKVVAVSDT--KGGIYNPDGLNADEVLAWKKK-------TGSVKDFP
1EUZ SLGRGTATAQGAIFTIREAAKALG-------I-DLKGKKIAVQGYG-NAGYYTAKLAKEQLGMTVVAVSDS--RGGIYNPDGLDPDEVLKWK-R------EHGSVKDFP
1GTM SLGRIEATARGASYTIREAAKVLG-------WDTLKGKTIAIQGYG-NAGYYLAKIMSEDFGMKVVAVSDS--KGGIYNPDGLNADEVLKWK-N------EHGSVKDFP
S R T G A G G N G A G KV SD G G
E AN PTT A G L P A N GGV VS E QN W
1HWX KAKIYEGSILEVDCDILIPAASEKQLTKSNAPRV----KAKIIAEGANGPTTPQADKIFLERN-IMVIPDLYLNAGGVTVSYFQILKNLNHVSYGRLTFKYERDSNYHL
1L1F KAKIYEGSILEVDCDILIPAASEKQLTKSNAPR-APRVKAKIIAEGANGPTTPEADKIFLERN-IMVIPDLYLNAGGVTVSYFEWLKNLNHVSYGRLTFKYERDSNYHL
2TMG KGERITNEELLELDVDILVPAALEGAIHAGN---AERIKAKAVVEGANGPTTPEADEILSRR-GILVVPDILANAGGVTVSYFEWVQDLQSFFW--------------
1BGV GVQFFPGEKPWGQKVDIIMPCATQNDVDLEQAKKIVANNVKYYIEVANMPTTNEALRFLMQQPNMVVAPSKAVNAGGVLVSGFEMSQNSERLSW--------------
1BVU GATNITNEELLELEVDVLAPSAIEE-VITK-KN-ADNIKAKIVAELANGPTTPEADEILY-EKGILIIPDFLCNAGGVTVSYFEWVQNITGDYW--------------
1EUZ GATNITNEELLELEVDVLAPAAIEE-VITEKN--ADNIKAKIVAEVANGPVTPEADDILR-EKGILQIPDFLCNAGGVTVSYFEWVQNINGYYW--------------
1GTM GATNITNEELLELEVDVLAPAAIEE-VITKKN--ADNIKAKIVAEVANGPVTPEADEILF-EKGILQIPDFLCNAGGVTVSYFEWVQNITGYYW--------------
E AN PTT A G L P L N GGV VS E QN W
B

FIGURE 7: (A) Segment of the structural alignment of the seven glutamate dehydrogenase sequences (the names of the five thermophilic are highlighted in bold). Amino acids involved in conserved hydrogen bonds and salt bridges are shown in blue and green, respectively. The conservation threshold within the thermophilic structures is four of five, whereas both mesophilic positions must be involved for the interaction to be tabulated. Amino acids above (mesophilic) and below (thermophilic) the structural alignment represent the conserved consensus sequence identity for each subfamily at the 60% conservation threshold. Nonvarying motif regions are colored red, whereas conserved mutations are colored blue. There are only 10 positions with conserved mutations above the 60% threshold. Eight of those form a structural cluster at the C-terminal end of helix 5 and the N-terminal end of helix 6 (indicated by a blue box). (B) On the left, the nonvarying positions (red) and conserved mutations (blue) are highlighted on the glutamate dehydrogenase structure from *T. maritima*. On the right is the sequence segment in which the cluster of 8 conserved mutations is highlighted (blue) with respect to domain II (green). All three monomer structures are displayed in the same orientation.

drogenase families. The examples presented here demonstrate the types of insights that can be gleaned from our comprehensive analysis.

DeDecker et al. have solved the thermophilic structure of the conserved TATA-box binding protein (TBP) from *Pyrococcus woesei* (*9*). Comparison with the mesophilic TBP from *Saccharomyces cerevisiae* indicates that thermostability arises from an added cysteine bridge, three added salt bridges, and more compact protein packing (*55*). Our sequence results complement these structural studies, while also providing key insights not available to the limited structural data set. Our results indicate that the observed structural differences do not represent a global evolutionary theme. Neither the cysteine bridge nor two of the four salt bridges are conserved within the thermophilic subfamily (Figure 6). Conversely, our results confirm that the most strictly conserved residues form the DNA binding site. Conserved differences between the mesophilic and thermophilic sequences generally line the protein surface opposite the DNA binding site. Many of the conserved mutations within the TBP family favor charged residues in the thermophilic subfamily. Of the 31 conserved differences (40% threshold), there are 16 charged residues in the thermophilic subfamily compared to only 10 in the mesophilic. Most of these positions are solvent exposed, emphasizing the importance of charge composition on the surface of thermophilic proteins.

Seven unique structures of glutamate dehydrogenase (GDH) have been solved (*56*−*62*), five of which are

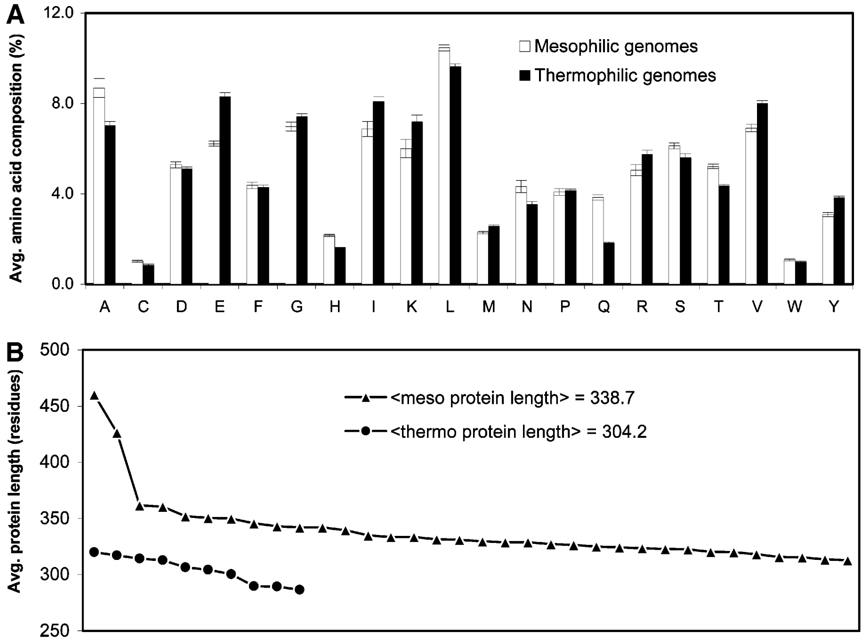


FIGURE 8: Complete genome statistics comparing (A) amino acid abundance (with 95% confidence intervals) and (B) average protein length between mesophilic and thermophilic genomes. Excluding the two eukaryotic yeast genomes (*S. cerevisiae* and *C. albicans*) from this analysis, which have significantly larger average protein lengths, still results in a length bias between mesophilic and thermophilic average protein length, 332 vs 304 residues, respectively. All of our data compare favorably with previously published results. These data are taken from all 44 genomes in the COG database, representing 78847 distinct protein sequences.

thermophilic. Each GDH functions as a homohexamer. The mesophilic structures are from eukaryotic organisms (human and cow), yet their structures are nearly identical to the thermophilic proteins. Several studies have been reported describing the stability of glutamate dehydrogenases from thermophilic organisms. Structural comparisons reveal increased numbers of salt bridges and extended ionic networks in the thermophilic protein (*62*). Consalvi et al. have shown using differential scanning calorimetry that the stability of domain II is lower than that of the hexameric protein (*63*), indicating that at least a portion of the added stability of GDH from *Thermotoga maritima* arises from interdomain interactions. We find that the nonvarying positions within each motif pair make up a highly conserved structural core (Figure 7). Of the 62 conserved positions identified, 9 are charged (and known to be involved in salt bridges), 5 are proline, and 17 are glycine. Conversely, there are only 10 positions with conserved differences (60% threshold). Eight of these positions form a structural cluster at the C-terminal end of helix 5 and the N-terminal end of helix 6, which are separated by a short β-strand. In the monomer and trimer structures, these positions are predominantly solvent exposed; however, on hexamer formation, these positions become buried. The work of Consalvi et al. described above investigates the stabilities of domain II only (*63*). The conserved mutation cluster is outside domain II and is spatially isolated from it, which suggests being related to the large stability differences between domain II and the hexameric GDH structure.

Genome Statistics. We also present amino acid composition and mean protein length of the 44 genomes. Similar descriptions have been reported (*3*, *5*, *22*, *32*, *64*, *65*), although none, that we are aware of, equal the size of our data set. In general, all of our results are consistent with those previously reported. The most obvious difference between mesophilic and thermophilic genomes is the drastic increase of glutamate, lysine, and arginine residues in thermophilic genomes, generally assumed to correspond to an increase in salt bridges (Figure 8A). Our data show an obvious bias against polar−neutral amino acids (i.e., serine, threonine, asparagine, glutamine, and cysteine), which has been explained by their tendency to undergo oxidation or deamidation at high temperatures (*5*). Additionally, we observe a significant increase in thermophilic content of valine and isoleucine, which, compared to leucine, have higher β-sheet propensities and decreased side chain conformational entropies. It has also been argued that shorter protein length increases the compactness of protein structures (*19*, *64*, *66*). Analyzing an earlier version of the COG database, Das et al. observed this bias within their data set of 12 genomes (*64*), although they proceed to argue that protein length might be more related to species kingdom than environment. We hesitate to make claims regarding the length bias and merely report that what we observe (Figure 8B). However, we do wish to point out the striking increase in average protein length, presumably due to multidomain proteins (*35*), of the two eukaryotic yeast genomes, *S. cerevisiae* and *Candida albicans*. Even if *S. cerevisiae* and *C. albicans* are excluded from this analysis, the average mesophilic protein length (332 residues) is still greater than the average thermophilic length (314 residues). We are very cautious about extrapolating too much from the above data because 44 genomes may not be representative of the microbial universe. However, our data set, along with the smaller ones that preceded it, does indicate overall differences between mesophilic and thermophilic genomes. The significance of these conclusions will only be validated by more thorough analyses.

Conclusions. In this comprehensive comparative proteomic analysis we identify protein motifs encoded within 34 mesophilic and 10 thermophilic microbial genomes. Motif comparisons between corresponding mesophilic and thermophilic pairs provide key insights with respect to protein thermostability. The modest correlation between motif and entire protein similarity clearly demonstrates the value of motif methods in such comprehensive multigenome analyses. Further, the utility of our unique approach is highlighted by two example studies providing key evolutionary insights concerning the onset of thermostability. Our results with the TATA-box binding protein highlight the importance of conserved active site residues and the key role of optimized surface electrostatics in structural stabilization, while also indicating that several differences identified from a single structural pair do not represent a family-wide mechanism. In addition, we identify a cluster of conserved mutations between the glutamate dehydrogenase subfamily pairs that seems likely to be related to differences between the experimentally determined stabilities of domain II and the physiological hexamer structure. We also present amino acid composition data and average protein length comparisons for all 44 microbial genomes. The latter results confirm, with improved statistics due to the larger size of our data set, previously reported results.

ACKNOWLEDGMENT

We thank Dr. Kimmen Sjölander (University of California at Berkley) for helpful discussions relating to this work and Dr. Shankar Subramaniam (University of California at San Diego) for careful reading of the manuscript and for a number of valuable suggestions.

REFERENCES

1. Dill, K. A., and Chan, H. S. (1997) *Nat. Struct. Biol. 4*, 10−19.
2. Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem. 48*, 545−600.
3. Jaenicke, R., and Bohm, G. (1998) *Curr. Opin. Struct. Biol. 8*, 738−748.
4. Vieille, C., and Zeikus, G. J. (2001) *Microbiol. Mol. Biol. Rev. 65*, 1−43.
5. Kumar, S., Tsai, C. J., and Nussinov, R. (2000) *Protein Eng. 13*, 179−191.
6. Britton, K. L., Baker, P. J., Borges, K. M., Engel, P. C., Pasquo, A., Rice, D. W., Robb, F. T., Scandurra, R., Stillman, T. J., and Yip, K. S. (1995) *Eur. J. Biochem. 229*, 688−695.
7. Chang, C., Park, B. C., Lee, D. S., and Suh, S. W. (1999) *J. Mol. Biol. 288*, 623−634.
8. Declerck, N., Machius, M., Wiegand, G., Huber, R., and Gaillardin, C. (2000) *J. Mol. Biol. 301*, 1041−1057.
9. DeDecker, B. S., O'Brien, R., Fleming, P. J., Geiger, J. H., Jackson, S. P., and Sigler, P. B. (1996) *J. Mol. Biol. 264*, 1072−1084.
10. Delboni, L. F., Mande, S. C., Rentier-Delrue, F., Mainfroid, V., Turley, S., Vellieux, F. M., Martial, J. A., and Hol, W. G. (1995) *Protein Sci. 4*, 2594−2604.
11. Knegtel, R. M., Wind, R. D., Rozeboom, H. J., Kalk, K. H., Buitelaar, R. M., Dijkhuizen, L., and Dijkstra, B. W. (1996) *J. Mol. Biol. 256*, 611−622.
12. Lim, J. H., Yu, Y. G., Han, Y. S., Cho, S., Ahn, B. Y., Kim, S. H., and Cho, Y. (1997) *J. Mol. Biol. 270*, 259−274.
13. Numata, K., Hayashi-Iwasaki, Y., Kawaguchi, J., Sakurai, M., Moriyama, H., Tanaka, N., and Oshima, T. (2001) *Biochim. Biophys. Acta 1545*, 174−183.
14. Schafer, T., Bonisch, H., Kardinahl, S., Schmidt, C., and Schafer, G. (1996) *Biol. Chem. 377*, 505−512.
15. Szilagyi, A., and Zavodszky, P. (2000) *Struct. Folding Des. 8*, 493−504.
16. Tanner, J. J., Hecht, R. M., and Krause, K. L. (1996) *Biochemistry 35*, 2597−2609.
17. Tomschy, A., Bohm, G., and Jaenicke, R. (1994) *Protein Eng. 7*, 1471−1478.
18. Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H., and Robb, F. T. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 12300−12305.
19. Thompson, M. J., and Eisenberg, D. (1999) *J. Mol. Biol. 290*, 595−604.
20. Usher, K. C., de la Cruz, A. F., Dahlquist, F. W., Swanson, R. V., Simon, M. I., and Remington, S. J. (1998) *Protein Sci. 7*, 403−412.
21. Vieille, C., Epting, K. L., Kelly, R. M., and Zeikus, J. G. (2001) *Eur. J. Biochem. 268*, 6291−6301.
22. Fukuchi, S., and Nishikawa, K. (2001) *J. Mol. Biol. 309*, 835−843.
23. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M., Thurlkill, R. L., Scholtz, J. M., and Pace, C. N. (1999) *Protein Sci. 8*, 1843−1849.
24. Hardy, F., Vriend, G., van der Vinne, B., Frigerio, F., Grandi, G., Venema, G., and Eijsink, V. G. (1994) *Protein Eng. 7*, 425−430.
25. Perl, D., Mueller, U., Heinemann, U., and Schmid, F. X. (2000) *Nat. Struct. Biol. 7*, 380−383.
26. Perl, D., and Schmid, F. X. (2001) *J. Mol. Biol. 313*, 343−357.
27. Sheinerman, F. B., and Honig, B. (2002) *J. Mol. Biol. 318*, 161−177.
28. Zavodszky, P., Kardos, J., Svingor, and Petsko, G. A. (1998) *Proc. Natl. Acad. Sci. U.S.A. 95*, 7406−7411.
29. Strop, P., Marinescu, A. M., and Mayo, S. L. (2000) *Protein Sci. 9*, 1391−1394.
30. Warren, G. L., and Petsko, G. A. (1995) *Protein Eng. 8*, 905−913.
31. Russell, R. J., and Taylor, G. L. (1995) *Curr. Opin. Biotechnol. 6*, 370−374.
32. Chakravarty, S., and Varadarajan, R. (2002) *Biochemistry 41*, 8152−8161.
33. Forterre, P. (2002) *Trends Genet. 18*, 236−237.
34. Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000) *Nucleic Acids Res. 28*, 33−36.
35. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., and Koonin, E. V. (2001) *Nucleic Acids Res. 29*, 22−28.
36. Bailey, T. L., and Gribskov, M. (1998) *J. Comput. Biol. 5*, 211−221.
37. Myers, E. W., and Miller, W. (1988) *Comput. Appl. Biosci. 4*, 48−54.
38. Bailey, T. L., and Gribskov, M. (1996) *Proc. Int. Conf. Intell. Syst. Mol. Biol. 4*, 15−24.
39. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994) *J. Mol. Biol. 235*, 1501−1531.
40. Eddy, S. R. (1996) *Curr. Opin. Struct. Biol. 6*, 361−365.
41. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res. 22*, 4673−4680.
42. Gerstein, M., Sonnhammer, E. L., and Chothia, C. (1994) *J. Mol. Biol. 236*, 1067−1078.
43. Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D. (1996) *Comput. Appl. Biosci. 12*, 327−345.
44. Mulder, N. J., and Apweiler, R. (2002) *Genome Biol. 3*, 1−8.
45. Dayhoff, M. O., Schwartz, R., and Orcutt, B. C. (1978) *Atlas of protein sequence and structure* (Dayhoff, M. O., Ed.) Vol. 5, no. 3, NBRF, Washington, DC.
46. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992) *Science 256*, 1443−1445.
47. Henikoff, S., and Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 10915−10919.
48. Chen, J., Lu, Z., Sakon, J., and Stites, W. E. (2000) *J. Mol. Biol. 303*, 125−130.
49. Martin, A., Sieber, V., and Schmid, F. X. (2001) *J. Mol. Biol. 309*, 717−726.
50. Ozawa, T., Hakamada, Y., Hatada, Y., Kobayashi, T., Shirai, T., and Ito, S. (2001) *Protein Eng. 14*, 501−504.
51. Sanchez-Ruiz, J. M., and Makhatadze, G. I. (2001) *Trends Biotechnol. 19*, 132−135.
52. Shibuya, H., Kaneko, S., and Hayashi, K. (2000) *Biochem. J. 349*, 651−656.
53. Spector, S., Wang, M., Carp, S. A., Robblee, J., Hendsch, Z. S., Fairman, R., Tidor, B., and Raleigh, D. P. (2000) *Biochemistry 39*, 872−879.
54. Stewart, R. J., Varghese, J. N., Garrett, T. P., Hoj, P. B., and Fincher, G. B. (2001) *Protein Eng. 14*, 245−253.
55. Liu, D., Ishima, R., Tong, K. I., Bagby, S., Kokubo, T., Muhandiram, D. R., Kay, L. E., Nakatani, Y., and Ikura, M. (1998) *Cell 94*, 573−583.

56. Britton, K. L., Yip, K. S., Sedelnikova, S. E., Stillman, T. J., Adams, M. W., Ma, K., Maeder, D. L., Robb, F. T., Tolliday, N., Vetriani, C., Rice, D. W., and Baker, P. J. (1999) *J. Mol. Biol. 293*, 1121−1132.
57. Lebbink, J. H., Knapp, S., van der Oost, J., Rice, D., Ladenstein, R., and de Vos, W. M. (1999) *J. Mol. Biol. 289*, 357−369.
58. Nakasako, M., Fujisawa, T., Adachi, S., Kudo, T., and Higuchi, S. (2001) *Biochemistry 40*, 3069−3079.
59. Peterson, P. E., and Smith, T. J. (1999) *Struct. Folding Des. 7*, 769−782.
60. Smith, T. J., Schmidt, T., Fang, J., Wu, J., Siuzdak, G., and Stanley, C. A. (2002) *J. Mol. Biol. 318*, 765−777.
61. Stillman, T. J., Baker, P. J., Britton, K. L., and Rice, D. W. (1993) *J. Mol. Biol. 234*, 1131−1139.
62. Yip, K. S., Stillman, T. J., Britton, K. L., Artymiuk, P. J., Baker, P. J., Sedelnikova, S. E., Engel, P. C., Pasquo, A., Chiaraluce, R., and Consalvi, V. (1995) *Structure 3*, 1147−1158.
63. Consalvi, V., Chiaraluce, R., Giangiacomo, L., Scandurra, R., Christova, P., Karshikoff, A., Knapp, S., and Ladenstein, R. (2000) *Protein Eng. 13*, 501−507.
64. Das, R., and Gerstein, M. (2000) *Funct. Integr. Genomics 1*, 76−88.
65. Jaenicke, R., and Bohm, G. (2001) *Methods Enzymol. 334*, 438−469.
66. Nagi, A. D., and Regan, L. (1997) *Folding Des. 2*, 67−75.

BI027435E