Detection of Discriminative Sequence Motifs in Proteins Obtained from Prokaryotes Grown at Various Temperatures

LI-CHENG WU,[1] JORNG-TZONG HORNG,[1,2] SHIR-LY HUANG,[2] HSIEN-DA HUANG,[3] BAW-JHIUNE LIU[4]

[1]*Department of Computer Science and Information Engineering, National Central University, Taiwan, Republic of China*

[2]*Department of Life Science, National Central University, Taiwan, Republic of China*

[3]*Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Taiwan, Republic of China*

[4]*Department of Computer Science and Engineering, Yuan Ze University, Taiwan, Republic of China*

Received 20 May 2005; Accepted 12 December 2005
DOI 10.1002/jcc.20391
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Recent investigations on the stability of proteins have demonstrated various structural factors, but few have considered sequence factors such as protein motifs. These motifs represent highly conserved regions and describe critical regions that may only exist on proteins that remain functional at high temperatures. This investigation presents a method for identifying and comparing corresponding mesophilic and thermophilic sequence motifs between protein families. Discriminative motifs that are conserved only in the mesophilic or thermophilic subfamily are identified. Analysis of the results shows that, although the subfamilies of most protein families share similar motifs, some discriminative motifs are present in particular thermophilic/mesophilic subfamilies. The thermophilic discriminative motifs are conserved only in thermophilic organisms, revealing that physiochemical principles support thermostability.

 J Comput Chem 27: 798–808, 2006

Key words: protein thermostability; protein motif

Introduction

Proteins are employed extensively in industry as biocatalysts.[1] Chemical reactions must be performed at high temperatures to accelerate industrial processes. However, not many enzymes are stable when heated.[2] Research is required to ensure that proteins remain active and stable when heated, to overcome current limits on their industrial applications. Recent developments on the stability of proteins demonstrate that most thermophilic proteins exhibit numerous van der Waals interactions, hydrogen bonds, salt bridges, or dipole–dipole interactions, potentially contributing to their thermostability, according to comparisons among homologous structures. However, known protein folds are fewer than known sequences of proteins, and such comparisons can only be made among protein families with several known structures at various temperatures. Protein motifs refer to highly conserved regions of protein families, and are typically considered to describe regions that are crucial to the stability and functioning of the protein.

The environmental temperature (T_{env}) is directly related to the melting temperature (T_m) of various proteins across various protein families.[3] The correlation coefficient between the T_{env} and the T_m is 0.91, and the corresponding regression equation between the T_{env} and T_m is $T_m = 24.4 + 0.93T_{env}$.[3] The optimal growth temperature must be very close to the T_{env}, so an obvious correlation exists between T_m and optimal growth temperature. Some investigations[3–6] have taken the source organism optimal growth temperature as thermostability data of the protein for comparing various proteins. Proteins from organisms that have a high optimal growth temperature are more adaptive and better able to remain chemically active at such high temperature. Organisms with an optimal temperature less then 45°C are called mesophiles.[7] Organisms with an optimum temperature equal or more then 45°C are

Correspondence to: J.-T. Horng; e-mail: horng@db.csie.ncu.edu.tw

Contract/grant sponsor: National Science Council of the Republic of China; contract/grant number: NSC94-2213-E-008-006

called thermophiles.[7] Proteins that function under mesophilic conditions tend to have similar structural stabilities, despite the differences among their sequences and structural folds.[8] Proteins from thermophilic organisms (which exist at high ambient temperatures) typically exhibit considerably higher intrinsic thermal stabilities than their mesophilic counterparts, but retain the basic fold characteristics of the entire family.[9]

Recent work[6] has demonstrated that motif is critical to the thermostability of some families of proteins, but only point mutations in the motif region have been considered. What if motif blocks on mesophilic proteins differ from those of thermophilic ones? This investigation seeks to provide information that is both complementary and orthogonal to that provided in ref. 6. The aim is to further analyze the differences among motif blocks of mesophilic and thermophilic proteins, and to determine whether a protein family has a distinguishing mesophilic/thermophilic motif.

This investigation proposes a method for identifying and comparing corresponding mesophilic and thermophilic sequence motifs between Pfam protein families and conserved orthologous groups. Previous research[6] attempted to answer the question, "How well are motif pairs conserved between mesophilic and thermophilic subfamilies?" This investigation takes a different approach to answer the question, "To what extent do the motifs conserved in mesophilic subfamilies differ from those conserved in thermophilic subfamilies?" Both questions are biologically important because motifs commonly refer to the regions of the protein that are critical to functionality and structural stability.[6] The proposed approach successfully identifies some discriminative motifs, which are conserved only in the mesophilic or the thermophilic subfamily but are not conserved in their counterpart temperature subfamilies. The results indicate that although the subfamilies of most protein families share similar motifs, certain thermophilic/mesophilic subfamilies exhibit discriminative motifs. These discriminative motifs are conserved only in thermophilic organisms, which, in fact, demonstrates the physiochemical principles that confer thermostability. Nondiscriminative conservation motifs in specific subfamilies highlight the importance of motif to structure and function, whereas discriminative motifs are associated thermostability. Additionally, comparisons between corresponding mesophilic and thermophilic motifs yield crucial biochemical insights into thermostability, and can be used to test the evolutionary robustness of individual structural comparisons.

Methods

This investigation proposes a method for comparing motifs from mesophilic and thermophilic proteins from various genomes and protein families. The temperatures of proteins from prokaryotic organisms in PGTdb[10] are used in this investigation. The information in PGTdb about source organism optimal growth temperatures[10] of 1023 organisms. Previous orthogonal work[6] uses the protein groups from 44 genomes in COG[11] database (the current version of the COG[11] database has more than 44 genomes included in it). Thus, we also consider the proteins of 66 genomes (updated version) in COG[11] in this work. Every protein in the Pfam and the COG[11] database is assigned to a mesophilic or thermophilic subfamily, and short, highly conserved regions (motifs) are identified in each. Figure 1 depicts the system processing flow.

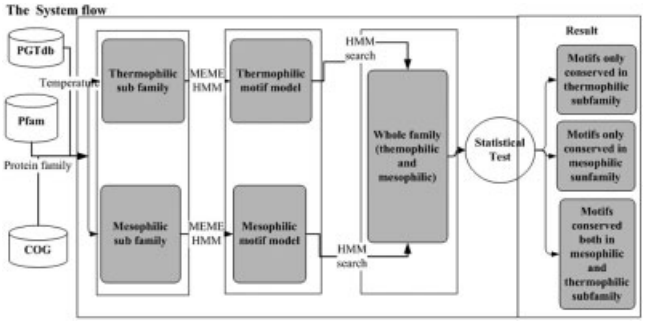

Figure 1. System flow of the proposed analytical process. Families of proteins are divided into subfamilies according to temperature. MEME and profile HMM are used to discover motifs and generate a model of each subfamily. Motif models are used to match self and counterpart subfamilies, and the results are matched by performing a statistical *Z* test to determine the discriminative motifs.

First, prokaryotic protein sequences are obtained from Swiss-Prot[12] version 12 and grouped using protein family definition in Pfam[13] version 11. Sequences of proteins in the COG[11] (updated version, 66 genome, 2003-Dec) database are divided into orthogonal groups. The proteins are then assigned a temperature based on source organism's optimal growth temperature in PGTdb[10] version 1.0. Accordingly, each protein family is divided into two subfamilies—thermophilic and mesophilic—by temperature. Protein sequences for which information on the optimal growth temperature their source organisms is absent from the PGTdb are discarded.

"Conserved" cannot be defined on a subfamily that contains only a single protein sequence. Accordingly, motif-discovering tools require more than two sequences to locate the conserved region. Thus, if one of the two subfamilies contains fewer than two sequences, then the analysis of this family cannot proceed and the family is eliminated. The elimination leaves 887 Pfam families and 2191 COG families. Motifs of each subfamily are identified using MEME version 3.0.4.[14,15] MEME is the most commonly employed motif-discovery tool that has been adapted to discovery motifs.[6,16] MEME stands for multiple expectation maximization for motif elicitation, and is applied to identify conserved regions in a set of DNA or protein sequences without gaps. The following parameters (mostly default) are used: minimum and maximum motif widths of 6 and 50, a motif model biased toward zero or one motif occurrence per sequence, FASTA format motif output, and a maximum motif search number of three.

Second, the profile HMM[17,18] is employed to establish a model of each motif region. The simple MAST search tool of the MEME was not used as a model scanning tool herein because the MAST is too sensitive to position, and building the HMM model will relax the model to match various patterns of proteins. The profile HMMs are generated using the HMMER package (http://hmmer.wustl.edu/) version 2.3.2, which is an implementation of profile HMM software for analyzing protein sequences.[17,18] The profile HMM is generated by inputting the motif FASTA file with the default option of domain alignment, MAP (maximum a posterior) and

Gerstein/Sonnhammer/Chothia tree weights (parameter—g—amino).

Then, both the mesophilic subfamily and the thermophilic subfamily are searched using the generated profile HMM model. The HMMER search utility matches each sequence in the family. The motif model generated from the mesophilic subfamily is employed to search the sequences in the whole family with E-values of 0.001. The search result can be further divided into two numbers after reference to which subfamily the matched sequence belong. Thus, the match number of thermophilic sequences represents "the number of thermophilic subfamily sequences that contain the motif that was discovered from mesophilic subfamily sequences." The match result of mesophilic sequences represent "the number of mesophilic subfamily sequences contains whose own motifs have been discovered," ensuring the quality of the motif and supporting further statistic testing. The thermophilic motif model has been processed similarly.

The importance or significance of a discriminative motif must be measured. How many matched sequences in the counterpart subfamily are insignificant? The statistical method is applied herein to yield the related results. The statistical Z test is performed to determine whether a motif is significant[19] (the t-test is used instead of the Z test for subfamilies that have fewer than 23 members[19]). If a random process is associated with only two types of results, then the random process is called a Bernoulli trial.[19] The HMM result is "matched" or "unmatched," so the process is a Bernoulli trial.

Let null hypothesis be the probabilities of motif match sequences in both subfamilies are equal, alternative hypothesis: the probabilities of motif match sequences in both subfamilies are not equal. Let a be the level of significance. The reject region RR will be $|Z| > z_a/2$. If the observable value of Z exceeds $z_a/2$ or the observable value of Z is below $-z_a/2$, then alternative hypothesis is accepted, so the probability that the subfamily includes the motif differs statistically significantly from the probability that it is matched in counterpart subfamily.[19] The a for significance in this investigation is set to 95%. The t-test used for small families (fewer than 23 members) involves the same hypothesis setting,[19] and thus omitted here. A motif that is associated with a statistically significant difference is defined as a discriminative motif.

Results

MEME identified 5309 motifs in 887 Pfam families. Accordingly, most families contain six motifs (three from the mesophilic subfamily, and three from the thermophilic subfamily). Among these 5309 motifs, only 1056 are discriminative. That is, 19.8% of the motifs discovered by MEME are discriminative. Over 80% motifs are not discriminative, and represent the conserved regions of both mesophilic and thermophilic sequences. These nondiscriminative motifs may contain conserved mutations that are associated with thermostability, as displayed in La et al.'s investigation.[6] Mesophilic and thermophilic subfamilies exhibit different motif models, rather than single mutations, because 19.8% of motifs are discriminative. Previous work of La et al.[6] ignored these discriminative motifs, which nevertheless exhibit significant differences across protein subfamilies at various temperatures. Figure 2a presents the distribution of the number of discriminative motifs in Pfam families; 458 of the 887 families contain at least one discriminative motif; 422 of the 1056 discriminative motifs are derived from mesophilic subfamily sequences and 634 are derived from thermophilic subfamily sequences. Figure 2a demonstrates that half of the protein families contain discriminative motifs and most families have fewer than three discriminative motifs; 19.8% of motifs are discriminative, and one-third of protein families contain at least one discriminative motif. If the probability that a motif is discriminative is uniformly 19.8%, then the probability that a protein contains at least one discriminative motif is as high as approximately 73% [1-(1-0.198)^6]. Figure 2a indicates that the distribution of discriminate motifs does not vary uniformly across the protein families. That is, the discriminative motif is a family-specific feature, and some protein families tend to contain more discriminative motifs than others. Figure 2b plots the numbers of mesophilic and thermophilic discriminative motifs. It reveals that the number of families with thermophilic discriminative motifs exceeds the number with mesophilic discriminative motifs.

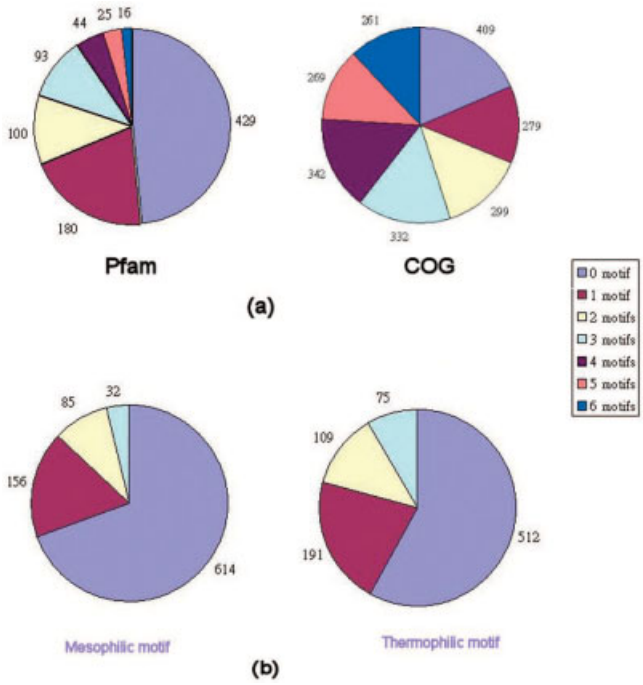

Figure 2. (a) Distribution of discriminative motifs in Pfam and COG families. Only one-half of Pfam families contain discriminative motifs. COG families include a higher proportion of families with discriminative motifs. (b) Distribution of number of discriminative motifs per subfamily. Thermophilic subfamilies exhibit more discriminative motifs than do mesophilic subfamilies. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

MEME identified 13,070 motifs in 2191 COGs, using the COG dataset. Of these 13,070 motifs, 6152 are discriminative. That is, 47% of the motifs discovered by MEME are discriminative motifs; 1782 of the 2191 COG families contain at least one discriminative motif. Figure 2 plots the distribution of the number of discriminative motifs in COG families; 2721 of the 6152 discriminative

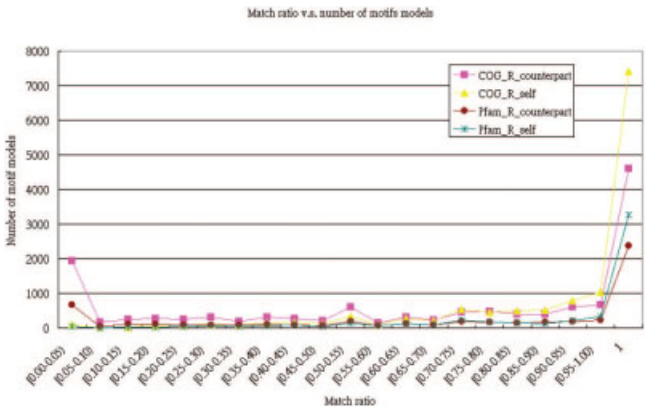

Figure 3. Matching ratio of Pfam and COG families. The right peak refers to the motif shared by both the mesophilic and the thermophilic subfamilies. The left peak indicates that the motif model cannot match any of the counterpart subfamily sequences. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

motifs are derived from mesophilic subfamily sequences, and 3431 are derived from thermophilic subfamily sequences. Figure 2a indicates that over three-quarters of all protein families contain discriminative motifs and most families contain fewer than three discriminative motifs, where results are similar to those obtained for Pfam families. The distribution of discriminative motifs is not uniform. The assumption that 47% of discriminative motifs are uniformly distributed on families yields the result that 97.7% of protein families contain at least one discriminative motif; however, the real distribution is only 81%. This assumption also leads to the result that approximately 1% of protein families contain exactly six discriminative motifs, but in fact, over 12% do. Hence, the distributions of discriminative motifs are also not uniform across COG protein families. By comparing the distribution of Pfam and COG in Figure 2a shows that the orthogonal groups in COG contain more discriminative motifs.

The matching ratio R is defined as (number of sequences in the subfamily that match the motif model)/(total number of sequences in subfamily). Consider the motif generated from subfamily; search itself with a matching ratio defined as R_{self} and search the counterpart subfamily with a matching ratio defined as $R_{\text{counterpart}}$. Figure 3 plots the matching ratio of the Pfam and COG motifs. First, the R_{self} curve in Figure 3 clearly demonstrates that most R_{self} close to 1. The high R_{self} demonstrate the high quality of the motif. Next, the $R_{\text{counterpart}}$ curve is considered.

Figure 3 includes two peaks of $R_{\text{counterpart}}$—the one on the right (close to 1) represents the motif that can completely match the subfamily and the peak on the left (close to 0) represents the motif that can match almost none of the sequences of the counterpart families. Clearly, most motifs either match the counterpart family with a high ratio or have a zero match with the counterpart family. Figure 3 demonstrates that, although numerous motifs have an $R_{\text{counterpart}}$ of approximately 1, some different motifs exist and match no part of the counterpart family. Figure 3 shows that the COG orthogonal group contains motifs that are not conserved in the counterpart family. These motifs have been neglected in the work of La et al.,[6] because they are unique and may not be aligned with motifs from the counterpart subfamily.

The statistical discriminative motif is selected and the matching ratio chart is plotted in Figure 4. The figure clearly indicates that the discriminative motif has a high R_{self} and a low $R_{\text{counterpart}}$. Comparing Figure 3 with Figure 4 reveals that not all of the motifs associated with the right peak of $R_{\text{counterpart}}$ in Figure 3 are taken as discriminative motifs in Figure 4, because some of the motifs may have a low $R_{\text{counterpart}}$, but R_{self} is also quite low and did not pass the statistical Z test or t-test. In Figure 3, 1929 motifs have an $R_{\text{counterpart}}$ of 0−0.05 but only 1787 of these are chosen as discriminative motifs, as shown in Figure 4. Not all of the motifs with an $R_{\text{counterpart}}$ of 0−0.05 failed to pass the statistical Z test, suggesting that this statistical test was required to filter out poor-quality motifs that did not match sequences of both subfamilies.

The mean lengths of the protein and the motifs are also analyzed. The average length of the mesophilic Pfam proteins in PGTdb is 387.41 residues and that of the thermophilic proteins and is 361.82 residues. The standard deviation of the mesophilic protein length is 280.62, whereas that of the thermophilic protein length is 252.95. The z value is 8.56, indicating that the thermophilic proteins are shorter than the mesophilic proteins. The average length of all motifs is 35.49. The average length of motifs built from mesophilic subfamilies is 34.48. The average length of the motifs built from thermophilic subfamilies is 36.63. Although thermophilic sequences are generally shorter then mesophilic sequences, the motifs built from thermophilic subfamilies are slightly longer then those built from mesophilic subfamilies. The average length of discriminative motifs is 30.69, whereas that of nondiscriminative motifs is 36.46. Accordingly, discriminative motifs are much shorter than nondiscriminative motifs. The discriminative mesophilic motifs have a mean length of 31.32, and the discriminative thermophilic motifs have a mean length of 29.94. Thus, the discriminative motifs are short, and the discriminative thermophilic motifs are even shorter. The nondiscriminative motifs exhibit an opposite relationship: the nondiscriminative thermophilic motifs have a mean length of 37.93, which exceeds that, 35.14, of nondiscriminative mesophilic motifs.

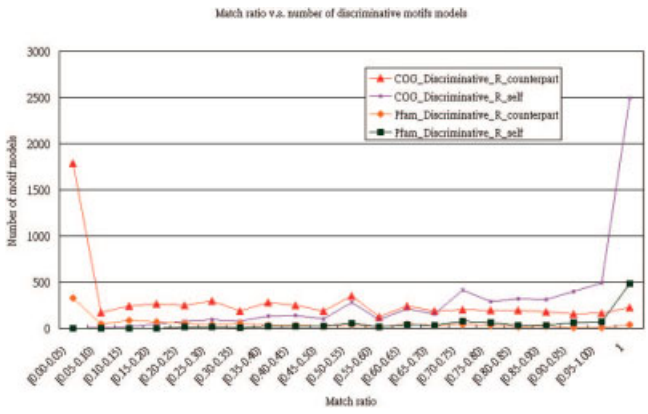

Figure 4. Matching ratio of discriminative motifs of Pfam families. Discriminative motifs clearly have high R_{self} and low $R_{\text{counterpart}}$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

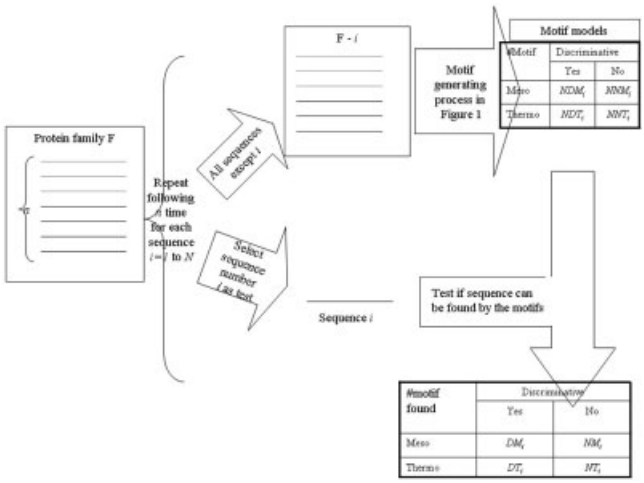

Figure 5. The process flow of leave-one-out crossvalidation. The process of choosing sequence *i* repeat *n* times and each sequence in family F will be the test sequence *i* once.

A leave-one-out crossvalidation is conducted to further assess the predictive ability and the modeling consistency of the discriminative motif model. The validation process is as follows:

1 For each sequence *i* in each the family F with *n* sequences: (a) assume the sequence *i* is not part of family F, construct motif models and test the model with rest of the sequences of F except *i*. The number of motif models are NDM_i discriminative mesophilic, NDT_i discriminative thermophilic, NNM_i nondiscriminative mesophilic, and NNT_i nondiscriminative thermophilic. (b) Determined whether *i* can be found the motif models. The number of motif models that can be found on *i* are DM_i discriminative mesophilic, DT_i discriminative thermophilic, NM_i nondiscriminative mesophilic, and NT_i nondiscriminative thermophilic.
2. Sum the results in the previous step. Let $DM = \Sigma DMi$, $DT = \Sigma DTi$, $NM = \Sigma NMi$, $NT = \Sigma NTi$, $NDM = \Sigma NDMi$, $NDT = \Sigma NDTi$, NNM $= \Sigma NNMi$, and $NNT = \Sigma NNTi$.
3. Calculate the predict ability (*PA*) of specific type of motif as the number of specific type motif divided by the total number of specify motif generated. Let predict ability of discriminative mesophilic motif, discriminative thermophilic motif, nondiscriminative mesophilic motif, and nondiscriminative thermophilic motif denoted *PA_DM*, *PA_DT*, *PA_NM*, and *PA_NT*, respectively. Thus, $PA\_DM = DM/NDM$, $PA\_DT = DT/NDT$, $PA\_NM = NM/NNM$, and $PA\_NT = NT/NNT$.

The process flow of crossvalidation is shown in Figure 5. The result of predictive ability of each family will be difference. We plot the family size (number of sequences) vs. predictive ability of each family in Figure 6. In Figure 6, families are being grouped to summarize the results. There are less family of size larger then 60; accordingly, the group region of is larger for family size larger then 60. Figure 6 shows that the predict ability of discriminative thermophilic motif increase when the family size arise. Figure 6 also shows that nondiscriminative motif have predictive ability more then 60%. Nondiscriminative motifs and discriminative mesophilic motif have the largest predictive ability on family size, about 31 to 45. The predictive ability of discriminative thermophilic motif is lower then other types of motifs, but is catching up with others when family size is larger. There are more mesophilic sequences in a protein family. Accordingly, we can expect a predictive ability increase when more temperature information is available. The average of all *PA_DM* and *PA_DT* is 0.57, and average of all *PA_NM* and *PA_NT* is 0.78. The average predictive ability of all types of motif is 0.68.

We also present the intrasubfamily crossvalidation process. The process is similar to the above leave-one-out process except it is only performed in the subfamily. The order of the motif in the crossvalidation model may not equal the order of the original motif model, and aligning two protein motifs is currently not feasible. Therefore, the average number of discriminative motifs of each family is plotted vs. the original number of discriminative motifs of the family in Figure 7. Figure 7 demonstrate that a family of three discriminative motifs is more stable than one of one or two discriminative motifs. Only 102 subfamilies contain three discriminative motifs in Pfam families, as shown in Figure 2. Families of only one discriminative motif may disappear during the crossvalidation process. The predict ability of the subfamily containing three discriminative motifs is 70%. Thus, the predictive abilities of families with two or one discriminative motifs decreases, and the "discriminative" characteristics may be lost in the leave-out-out crossvalidation.

The above analysis indicates that the discriminative motif is not a general phenomenon for all protein families. Discriminative motifs are present only in some protein families. The discriminative motif results are most informative when applied in concert

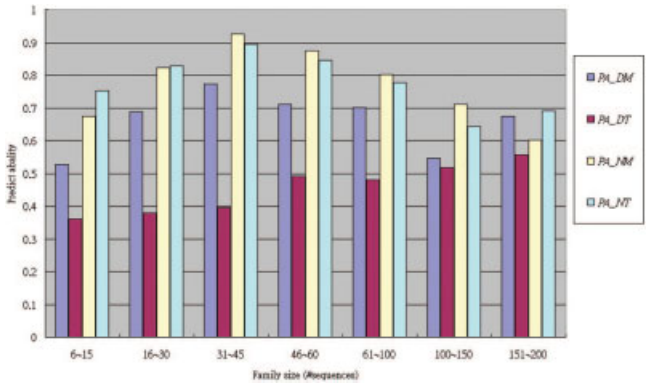

Figure 6. Result of crossvalidation. The predict ability (*PA*) of specific type of motif is the number of specific type motif divided by the total number of specify motif generated. The predict ability of discriminative mesophilic motif, discriminative thermophilic motif, nondiscriminative mesophilic motif, and nondiscriminative thermophilic motif denoted *PA_DM*, *PA_DT*, *PA_NM*, and *PA_NT*, respectively. The predict ability are grouped and averaged by family size. The predict ability of thermophilic discriminative motif rises while family size increases. Mesophilic discriminative motifs have higher predict ability then the thermophilic discriminative motifs. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

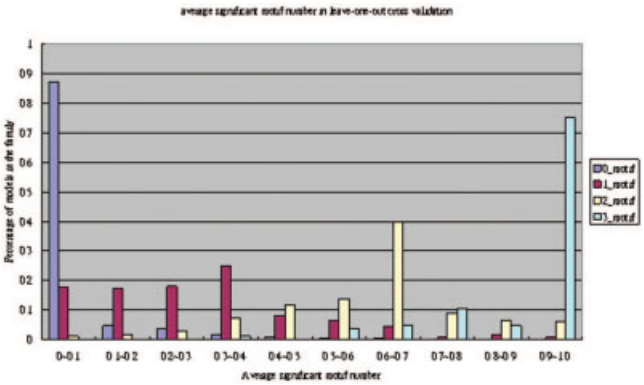

Figure 7. Mean number of discriminative motifs in subfamily. Seventy-five percent of subfamilies of three discriminative motifs still have three discriminative motifs after one sequence was removed in crossvalidation; 17.6% of families that contain a single discriminative motif lose its discriminative characteristic during crossvalidation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

with structural investigations. Specific factors that contribute to the thermostability of a protein in a particular family can be identified from structural studies,[3,4,20,21] from conserved mutations in the motif region,[6] and from the consideration of discriminative motifs. The molecular basis of thermostability is examined for each family, and the hidden relationship between ion pair and the presence of a discriminative motif is elucidated. The protein sequence and structural mapping were obtained from iProClass (http://pir.georgetown.edu/iproclass/)[22] and protein structures were obtained from the Protein Data Bank (PDB) (http://www.rcsv.org/pdb/).[23] The protein structure with the maximum sequence similarity was selected and duplicate structures removed for each protein sequence. A total of 612 subunit structures are associated with the Pfam protein sequences in PGTdb. The generated motif models are employed to search for the sequences of the 612 subunit structures; only 308 subunit structures include discriminative motifs.

Next, the number of ion pairs in each protein structure is calculated. The ion pair is defined by the distance between charged atoms in refs. 24 and 25. The ion pair distribution throughout the subunit, the motif region, and the discriminative motif region are calculated. Table 1 presents the distribution of the ion pairs of 308 protein structures. Table 1 reveals that the distributions of ion pairs do not differ significantly markedly across the three regions. Accordingly, the single case study in La et al.'s [6] work shown that ion pairs rich in motif regions do not indicate a global distribution difference across all protein families. Table 1 also indicates that the thermophilic protein statistically contains more ion pairs per residue, where the result is consistent with results presented elsewhere.[5,25] The distribution of the number of ion pair numbers on the motif and in the discriminative motif region are similar to those of the whole sequence, but the following question remains: does the ion pair in the motif region contain more information that could be used to increase the accuracy of the prediction of thermostability? The Bayesian thermostability prediction model[24] was applied to the different ion pair distributions to its accuracy. The normalized number of different types (His-Asp, Arg-Asp, Lys-Asp, His-Glu, Arg-Glu, Lys-Glu) and strengths (2–4, 4–6, and 6–8 Å) of ion pairs are calculated as ion pair properties for each protein structures at different temperatures.[24] A probabilistic Bayesian statistical method for efficiently predicting the thermostability of proteins according to the properties of their ion pairs has been presented,[24] and high accuracy has been achieved in the crossvalidation of three protein families (α-amylase, glyceraldehyde 3-phosphate dehydrogenase, and Xylanase), indicating that ion pairs critically influence the thermostability in particular protein families. Figure 8 presents the results predicted using the ion pair Bayesian models established by the three protein family in ref. 24. Figure 8 shows that, although the distribution of ion pairs in the motif region does not diff markedly from that over the entire sequence, incorporating the number of ion pairs on the motif increased the overall accuracy of the Bayesian prediction model. This result confirms La et al.'s[6] findings that the ion pair in the motif region is associated with thermostability. The ion pair on the discriminative motif further increases the overall accuracy of the Bayesian model, suggesting that the discovery of the discrimina-

Table 1. Ion Pair Densities of Various Protein Regions.

		Ion pair number	Residue number of the region	Ion pairs per residue	Residues distance between ion pairs
Entire subunit		8361	88307	0.09	10.56
	Mesophilic entire subunit	3536	46710	0.08	13.21
	Thermophilic entire subunit	4825	41597	0.12	8.62
Motif region		3877	43286	0.09	11.16
	Mesophilic motif	2781	30505	0.09	10.97
	Thermophilic motif	2818.5	28473	0.10	10.10
Discriminative Motif region		1435	16791	0.09	11.70
	Discriminative Mesophilic motif	853.5	9672	0.09	11.33
	Discriminative Thermophilic motif	822	8342	0.10	10.15

The density of ion pairs does not vary among subunits, the motif region, and the discriminative motif region. Thermophilic protein contains more ion pairs per residue than does mesophilic protein.

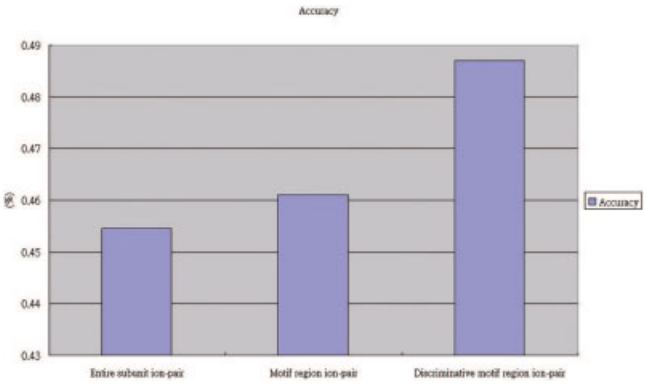

Figure 8. Calculated normalized numbers of ion pairs in different regions. In ref. 24, a Bayesian model was established in which the normalized number of ion pairs in the protein structures was input the thermostability was predicted. Overall accuracies obtained using the normalized number of ion pairs of different regions are compared using the Bayesian prediction model in ref. 24. Considering ion pairs in the discriminative motif region slightly increases the accuracy of the Bayesian predictive model. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

tive motif may improve the accuracy of the prediction of thermostability. The accuracy is not as high as 75% as in ref. 24, because the model whose accuracy is predicted in ref. 24 uses only three families, whereas the experiment in Figure 8 involves proteins from 176 protein families. The density of disulfide bonds is also determined. A total of 193 disulfide bonds are observed on 308 structures; 85 of them are located on the motif region and 12 are located on the discriminative motif region. The distribution of disulfide bonds varies across the motif region, but this assertion cannot be confidently drawn based on such limited statistics.

Two examples of how these results can be applied to improve our understanding of thioredoxin and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) are presented, to demonstrate further the effectiveness of the proposed approach.

The structures of thioredoxin obtained from PDB[23] has PDBID 1F6M, 1JPE, 1L6P, 1QUW, 1T7P, 2TRX, 1TXX. and 1XOB. The stability of thioredoxin from *Bacillus acidocaldarius* (pdbid: 1QUW) with a sequence length of 105 residues, has been discussed.[26] The thermostability of three mutant points, R82, K18, and D102, has been investigated,[27] and the wild-type structure has been demonstrated to be more stable than structure of the mutants. Only one thermophilic motif of length 11 residues in the family is discriminative. Among the three mutant residues, R82, located on the only discriminative thermophilic motif (position: 73–83), was discovered using the proposed presented approach, revealing that residues located on discriminative motifs affect the stability of the protein.

Thioredoxin from *Escherichia coli* (pdbid: 2TRX) has been the subject of numerous studies, as presented in ProTherm (http://gibk26.bse.kyutech.ac.jp/jouhou/jouhoubank.html).[28] Two site-directed mutants (L78K and L78R) have been designed to study the effect of placing a charged residue in the hydrophobic core of the protein; thermal denaturation of both of these mutant proteins at pH 7.0 reduces the stability by approximately 4 kcal $\cdot$ mol^{-1} below that of the oxidized wild type.[29] The investigation[29] demonstrated that the replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding.[30] A single-point mutation of glutamate 85 to arginine increases the stability of the thioredoxin.[31] Polar residues of thioredoxin are important to the specificity of folding; five polar core residues (D26I, C32A, C35A, T66L, and T77V) have been studied.[32] D26I is more stable than wild-type, whereas the other point mutants are less stable. Other mutation points associated with thermostability include proline 34,[33] methionine 37,[34] and proline 40.[34] Among the 11 mutant residues (D26, C32, P34, C35, M37, P40, T66, P76, T77, L78, and E85), four are in the discriminative motif region (position: 76–85), which was identified using presented approach. Again, the sequence length of 2TRX is 108 residues, and the random hit probability of the residues in the motif area is only 1/10. A single

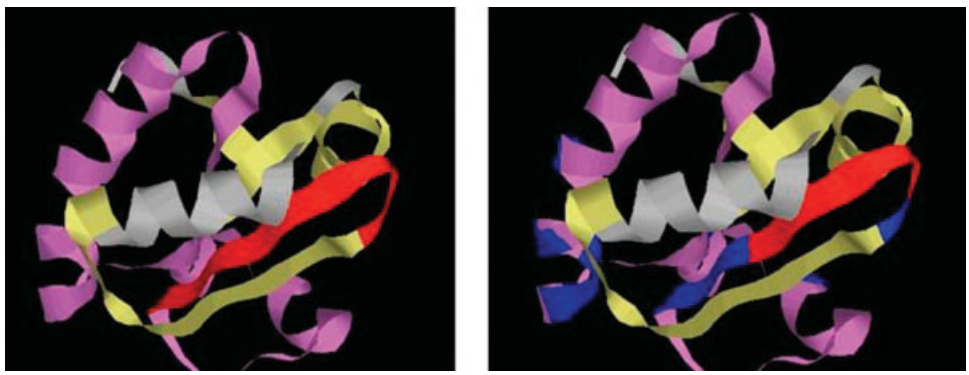

Figure 9. Schematic representation of thioredoxin structure 2THX. The mesophilic nondiscriminative motif region, thermophilic nondiscriminative motif region, and discriminative motif region are shown in yellow, violet, and red, respectively. The picture on the right shows the mutation residues of (D26, C32, P34, C35, M37, P40, T66, P76, T77, L78, and E85) in blue for comparison. Four tenths of mutant targets are located on a discriminative motif that occupies only one-tenth of the sequences. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

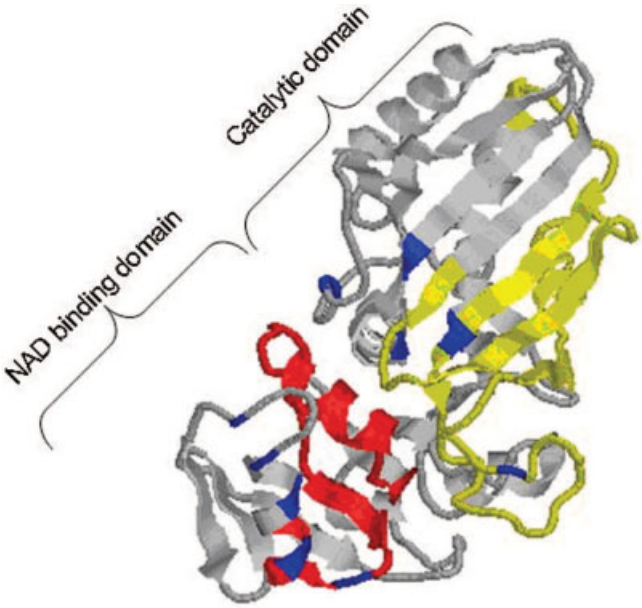

Figure 10. Schematic representation of the subunit O of the holo form of GAPDH. The crystallographic structure was obtained from the Brookhaven Protein Data Bank, access code 1GD1.[23] The yellow region represents the mesophilic discriminative motif and the red region represents the thermophilic discriminative motif. The blue residues are key mutant targets related to thermostability in ref. 35.

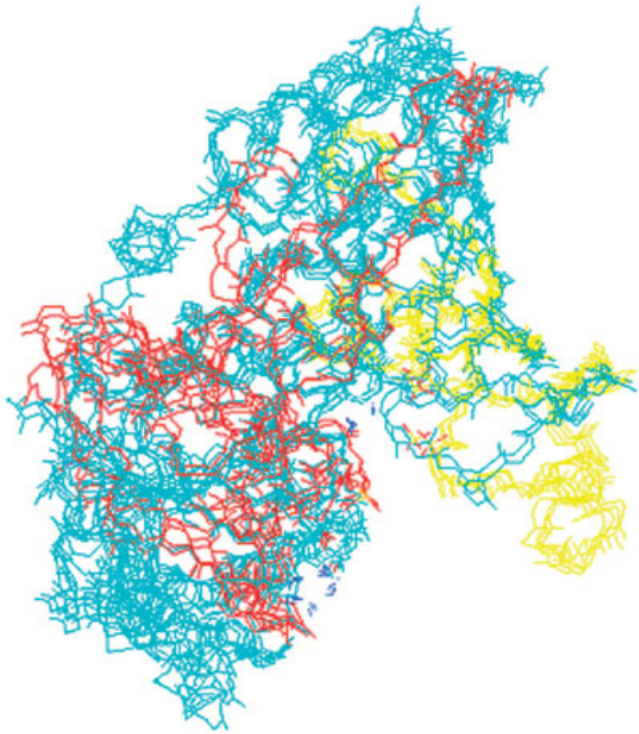

Figure 11. Multiple structure alignment of six GAPDH structures: 1B7G, 1CF2, 1CER, 1HDG, 1GD1, and 1GAD. The red residues represent sequences identified by the thermophilic discriminative model. The region identified by the mesophilic discriminative motif is clearly present on the surface. Part of the thermophilic discriminative motif region is also on the surface, and most of it is outside the unidentified region. Motif model and the yellow residues represent those identified by the mesophilic discriminative motif.[36]

mutation of E85 increases the stability by forming additional H-bonds,[31] revealing that the discriminative motif is an ideal mutant target for increasing/reducing thermostability. Figure 9 presents the positions of the mutation targets in earlier investigations and the discriminative motifs identified herein.

The structural thermostability of GAPDH (PDBid: 1GD1, 334AA) has been extensively investigated.[35,36] The polypeptide chain associated with each subunit is folded into two domains—the NAD-binding and the catalytic domains. Olivier Roitel et al.[36] investigated the contribution of intra- and intersubunit interactions to GAPDH thermostability. Among the 10 mutant residues that contributed to thermostability,[36] two were in the discriminative thermophilic motif region (32AA). One mutant residue was in the mesophilic discriminative motif region. These mutant targets are associated with thermostability and weaken cooperative interactions between the catalytic and the cofactor domains, reducing the efficiency of the binding of NAD.[36] Levashov et al.[35] also considered the denaturing characteristics of GAPDH. Of the two key residues, C149 and H176, H176 is located in the mesophilic discriminative motif region. The sequence results agree with the results of the structural studies. Figure 10 depicts the thermophilic structure 1GD1, which has been extensively analyzed.[36] 1GD1 is a thermophilic protein of 334 residues that contains one discriminative thermophilic motif (pos: 7–38, 32AA) and occasionally two mesophilic discriminative motifs (pos: 163–188, 25AA and 190–239, 50AA). Accordingly, 1GD1 is not a very good example of the discriminative motif as a thermophilic protein should contain more thermophilic discriminative motifs and fewer mesophilic discriminative motifs. However, 1GD1 is an ideal case for showing the motif in a single structure with a known structural mutation. Figure 10 demonstrates that discriminative motifs between

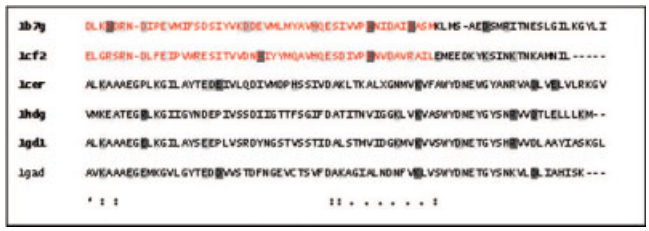

Figure 12. Multiple sequence alignment of last 80 residues from the C-terminal of the six GAPDH structures. Structures in bold type represent thermophilic sequences. Residues described using red characters represent the thermophilic discriminative motif region. Dark gray, medium gray, and light gray represent residues with strong, medium, and weak ion pair strengths. More ion pairs are located on this thermophilic discriminative motifs region than other regions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

mesophilic and thermophilic sequences are generally on the surface of the protein. Most of the intrasubunit positions considered in ref. 36 are generally close to the their positions on the discriminative motif region. Previous works[37–39] point out the importance of surface ion pairs on thermostability. Most of these positions are exposed to solvent, indicating the importance of the charge composition on the surface of thermophilic proteins. Additionally, the results herein provide evidence that cannot be obtained from small protein families, which are associated with a limited structural data set and the observed structural differences may not be conserved in global evolution. Figure 11 presents the multiple structure alignment of the six structures of the GAPDH family, 1B7G, 1CF2, 1CER, 1HDG, 1GD1, and 1GAD. The multiple structure alignment in Figure 11 was elucidated using MASS (http://bioinfo3d.cs.tau.ac.il/MASS)[40] and is presented using a Swiss-Pdb Viewer (http://www.expasy.org/spdby/).[41] Figure 11 presents depicts the discriminative mesophilic motif located on the surface of the subunit in the catalytic domain and the discriminative thermophilic motif, which is generally located in the NAD domain and on the surface. Figure 12 presents the partial multiple sequence alignment of the GAPDH family (NAD domain at the C-terminal of the protein). The red characters refer to the discriminative motif region. Dark gray, medium gray, and light gray represent strong, medium, and weak ion pair strengths, respectively. The positions of the ion pairs on the discriminative motif clearly differ from those of the ion pairs on the corresponding aligned region. Accordingly, the ion pairs on the discriminative motif may contribute differently to the thermostability of homologs.

The amino acid composition and average protein length of the Pfam proteins in PGTdb are also presented. Figure 13 presents the analysis of the amino acid composition of the mesophilic sequences. Similar analyses have been reported elsewhere.[6,9,20,42–44] All of the results herein agree with those reported elsewhere, except that the variance is smaller here and the motif vs. nonmotif distinctions are also provided here. Figure 13 shows the plot with a 95% confidence interval. A statistical Z test is conducted to compare the amino acid compositions; unsurprisingly, the contents of all amino acids, except glycine, differ statistically significantly across subfamilies. Figure 13 also presents the amino acid composition associated with each different motif regions. Thermophilic sequences contain more Glutamic acid (E) and valine (V) than mesophilic sequences. Motif regions contain more cysteine (C), proline (P) and threonine (T) than whole sequences. Discriminative motif regions contain more cysteine (C) and proline (P) but less alanine (A) and glycine (G) than nondiscriminative motif regions. The distribution of glycine between different regions is of interest: nondiscriminative motif regions contain most glycine, followed by discriminative motif regions in the middle, and whole sequences have the least. However, the glycine content of mesophilic sequences did not differ markedly from that of and thermophilic, regardless of the region. Alanine (A) and threonine (T) exhibit similar patterns to glycine, peaking on the nondiscriminative region. The distribution of proline (P) reveals that more proline is present on the thermophilic motif region, and a very large amount is present on the discriminative thermophilic region. Cystine (C) and phenylalanine (F) exhibit similar patterns. Accordingly, cystine (C) phenylalanine (F) and proline (P) may be the critical residues in the discriminative motifs.

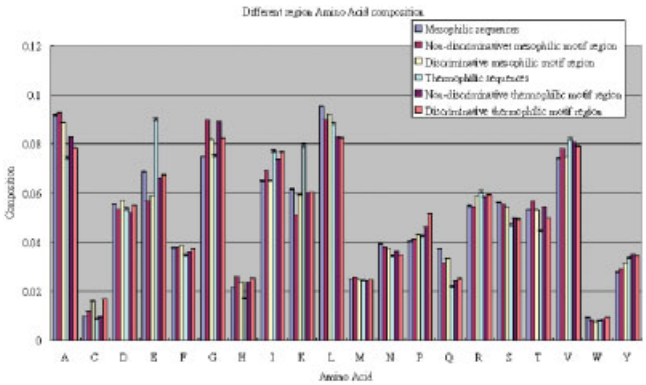

Figure 13. Amino acid composition of Pfam proteins in PGTdb. Results for the whole sequence are almost identical to those of La et al.[6] except the variance is smaller here. Thermophilic sequences contain more Glutamic acid (E) and valine (V) than mesophilic sequences. Motif regions contain more cysteine (C), proline (P), threonine (T), and tyrosine(Y) but less lysine and leucine (L) than complete sequences. The region of discriminative motifs contains more cysteine(C) and proline (P) but less alanine (A) and glycine (G) than regions of nondiscriminative motifs motif regions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 2 presents differences betwen the results of this investigation and those of La et al.'s.[6] La et al. focused on conserverd mutations in motifs, whereas this study focused on discriminative motifs across subfamilies at different temperatures. Hence, this work complements that of La et al.,[6] and provides orthogonal information concerning the thermostability of motifs. Data on amino acid contents are presented and average protin lengths compared.

Conclusions

This proteomic analysis identifies protein motifs on Pfam families and COG orthogonal groups. Comparisons of motifs of corresponding mesophilic and thermophilic subfamilies yield key features that are associated with the thermostability of proteins. The discovered two-peak match ratio of the motifs clearly demonstrates the value of the discriminative motif methods in the analyses performed herein. The leave-one-out cross-validation shows the predictive ability of the method. The discriminative of larger family has higher predictive ability. Subfamilies containing three discriminative motifs are also shown to have higher predictive ability. Additionally, the utility of this approach is demonstrated by the ion pair thermostability prediction model, and works of structural examples that yield insights into thermostability. The results obtained using the thermostability prediction model highlight the importance of the discriminative motif, and the discovery of discriminative motifs improves the accuracy of the prediction of thermostability. The results also show the importance of discriminative

Table 2. Comparison with La et al.'s Work.[6]

	La et al.'s work[6]	Our work
Dataset	COG	COG and Pfam
Organisms	44 organisms	221 organisms
Protein family	1354 COGs	2191 COGs and 948 Pfam family
Motif search method	MEME	MEME
Motif model construct	Profile HMM	Profile HMM
Motif comparison	nalign, HMM	HMM search model
Scoring/filtering	MOSS score	*Z* test
Motif not found on counter part subfamily	Discard	Statistical significant test and shows discriminative motif
Result	Conserved mutations	Discriminative motifs
Result on Web	Yes	Yes
Additional results	Amino acid composition and average protein length	Ion-pair prediction model on discriminative motifs, disulfide bond distribution, amino acid composition and average protein length

Table 2 presents differences between the results of this investigation and those of La et al.[6]. La et al.[6] focused on conserved mutations in motifs, whereas this study focused on discriminative motifs across subfamilies at different temperatures. Hence, this work complements that of La et al.[6] and provides orthogonal information concerning the thermostability of motifs. Data on amino acid contents are presented and average protein lengths compared.

motifs in structure stabilization. Moreover, an example of the GAPDH family is presented, suggesting that the discriminative motif enriches the structural analysis and identifies more factors that are associated with thermostability. Each motif and the related alignments in this investigation are also presented in Web pages in PGTdb.[10]

References

1. Burton, S. G. Trends Biotechnol 2003, 21, 543.
2. Voronov, S.; Zueva, N.; Orlov, V.; Arutyunyan, A.; Kost, O. FEBS Lett 2002, 522, 77.
3. Gromiha, M. M.; Oobatake, M.; Sarai, A. Biophys Chem 1999, 82, 51.
4. Szilagyi, A.; Zavodszky, P. Struct Fold Des 2000, 8, 493.
5. Vieille, C.; Zeikus, G. J. Microbiol Mol Biol Rev 2001, 65, 1.
6. La, D.; Silver, M.; Edgar, R. C.; Livesay, D. R. Biochemistry 2003, 42, 8988.
7. Madigan, T.; Martinko, M.; Parker, J. M.; Brock J. Biology of Microorganisms; Prentice-Hall Inc.: Englewood Cliffs, NJ, 2000.
8. Onuchic, J. N.; Luthey–Schulten, Z.; Wolynes, P. G. Annu Rev Phys Chem 1997, 48, 545.
9. Jaenicke, R.; Bohm, G. Curr Opin Struct Biol 1998, 8, 738.
10. Huang, S. L.; Wu, L. C.; Liang, H. K.; Pan, K. T.; Horng, J. T.; Ko, M. T. Bioinformatics 2004, 20, 276.
11. Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J. J.; Natale, D. A. BMC Bioinformatics 2003, 4, 41.
12. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M. C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. Nucleic Acids Res 2003, 31, 365.
13. Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths–Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. Nucleic Acids Res 2004, 32, D138.
14. Bailey, T. L.; Elkan, C. Proc Int Conf Intell Syst Mol Biol 1994, 2, 28.
15. Bailey, T. L.; Gribskov, M. J Comput Biol 1998, 5, 211.
16. Kataeva, I. A.; Blum, D. L.; Li, X. L.; Ljungdahl, L. G. Protein Eng 2001, 14, 167.
17. Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids; Cambridge University Press: New York, 1998.
18. Krogh, A.; Brown, M.; Mian, I. S.; Sjolander, K.; Haussler, D. J Mol Biol 1994, 235, 1501.
19. Wackerly, D. D.; Mendenhall, W.; Scheaffer, R. L. Mathematical Statistics with Applications; Duxbury Press: New York, 1996.
20. Kumar, S.; Tsai, C. J.; Nussinov, R. Protein Eng 2000, 13, 179.
21. Kumar, S.; Nussinov, R. Cell Mol Life Sci 2001, 58, 1216.
22. Wu, C. H.; Huang, H.; Nikolskaya, A.; Hu, Z.; Barker, W. C. Comput Biol Chem 2004, 28, 87.
23. Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. Acta Crystallogr D Biol Crystallogr 2002, 58(Pt 6), 899.
24. Huang, S. L.; Wu, L. C.; Huang, H. D.; Liang, H. K.; Ko, M. T.; Horng, J. T. Appl Bioinfomatics 2004, 3, 21.
25. Vogt, G.; Woell, S.; Argos, P. J Mol Biol 1997, 269, 631.
26. Nicastro, G.; De Chiara, C.; Pedone, E.; Tato, M.; Rossi, M.; Bartolucci, S. Eur J Biochem 2000, 267, 403.
27. Pedone, E.; Cannio, R.; Saviano, M.; Rossi, M.; Bartolucci, S. Biochem J 1999, 339, 309.
28. Gromiha, M. M.; Uedaira, H.; An, J.; Selvaraj, S.; Prabakaran, P.; Sarai, A. Nucleic Acids Res 2002, 30, 301.
29. Ladbury, J. E.; Wynn, R.; Thomson, J. A.; Sturtevant, J. M. Biochemistry 1995, 34, 2148.
30. Kelley, R. F.; Richards, F. M. Biochemistry 1987, 26, 6765.
31. Pedone, E.; Saviano, M.; Rossi, M.; Bartolucci, S. Protein Eng 2001, 14, 255.

32. Bolon, D. N.; Mayo, S. L. Biochemistry 2001, 40, 10047.
33. Lin, T. Y.; Kim, P. S. Proc Natl Acad Sci USA 1991, 88, 10573.
34. Chakrabarti, A.; Srivastava, S.; Swaminathan, C. P.; Surolia, A.; Varadarajan, R. Protein Sci 1999, 8, 2455.
35. Levashov, P.; Orlov, V.; Boschi–Muller, S.; Talfournier, F.; Asryants, R.; Bulatnikov, I.; Muronetz, V.; Branlant, G.; Nagradova, N. Biochim Biophys Acta 1999, 1433, 294.
36. Roitel, O.; Ivinova, O.; Muronetz, V.; Nagradova, N.; Branlant, G. Biochemistry 2002, 41, 7556.
37. Alsop, E.; Silver, M.; Livesay, D. R. Protein Eng 2003, 16, 871.
38. Pace, C. N.; Alston, R. W.; Shaw, K. L. Protein Sci 2000, 9, 1395.
39. Torrez, M.; Schultehenrich, M.; Livesay, D. R. Biophys J 2003, 85, 2845.
40. Dror, O.; Benyamini, H.; Nussinov, R.; Wolfson, H. Bioinformatics 2003, 19, 95.
41. Guex, N.; Peitsch, M. C. Electrophoresis 1997, 18, 2714.
42. Fukuchi, S.; Nishikawa, K. J Mol Biol 2001, 309, 835.
43. Jaenicke, R.; Bohm, G. Methods Enzymol 2001, 334, 438.
44. Chakravarty, S.; Varadarajan, R. Biochemistry 2002, 41, 8152.