

Chapter 1 Introduction

Enzymes play crucial role in the life of organisms. Enzymes catalyze biochemical reactions to make biological processes in a cell occur at significant rates. Due to extreme specificity for their substrates, the set of enzymes made in a cell determine the biological pathways in that cell and form the life. The activity of enzymes is greatly related to the optimal growth temperature of their source organisms. Enzymes are used in chemical industry, especially pharmacy, for the specific catalysis. In industrial applications, enzymes need to be active in high temperature environment. However, the activity of enzymes is sensitive to temperature and most of enzymes are from mesophiles that crucially limit practical use of specific enzymes^{1,2}. The need of thermophilic enzymes leads a lot of study on the factors affecting the thermostability of proteins and the design of practical methods to make thermophilic proteins of the desired catalysis function. On the other hand, life can be found in some extreme environment such as hot spring. The thermophile or hyperthermophile is living in 45 to 80°C or over 80°C³. It is conjectured that the thermophile or hyperthermophile could be the first life-form on the earth. Enzymes from these organisms are adapted to the hot environment. Researchers are interested in the evolution of enzyme, protein thermostability mechanism, and what the temperature limitation of life is. Through comparison between thermophilic proteins and their mesophilic homologues, biologist observed many features related to the thermostability of proteins. However, no single outstanding feature account for the thermostability. Most of the commercial enzymes are straightforwardly derived from the native thermophilic gene from thermophilic organisms, or they mutate mesophilic genes according to homologous thermophilic genes³. If there are no thermophilic genes or homologous

genes from thermophiles, researchers will make potential thermophilic ones by indirect random mutation methods (for example: DNA shuffling⁴). It is very difficult to make quickly a useful enzyme by site-directed mutagenesis in practical uses.

1.1 Research Status of Thermostability

The features of thermophilic proteins are obviously interesting in either principium or application. Protein thermal stabilization has been the focus of many experimental and theoretical research works⁵, but the molecular basis of thermal stability appears to be of diverse origin⁶. Although thermophilic proteins and their mesophilic homologues share a high degree of similarity in both sequence and three-dimensional structure, thermophilic proteins are intrinsically more stable and active in high temperature than their mesophilic homologues⁷⁻¹⁰.

There are many theories on protein thermostability. Based on comparison between thermophilic and mesophilic proteins, thermal stability appears to arise from a variety of sequence and structure features like, for example, the presence of stronger electrostatic interactions like more charged residues or salt bridges on the surface^{3,11-13}, more side chain-side chain hydrogen bonds^{14,15}, dipole-dipole interactions or cation- π interactions^{10,16-21}; higher degree of hydrophobic packing in the core regions^{3,11,12,14,22}; increased packing density^{6,23}; more disulfide bridges^{3,24}, shorter loop structures²⁵ or higher conformational rigidity¹²; more secondary structural elements such as α -helix or β -sheet¹⁰; pronounced bias in amino acid composition on the exposed regions²⁶⁻²⁹.

However, despite these many structure and sequence features, no single outstanding feature can adequately account for the thermal stabilization of proteins^{10,13,30-36}. This is

because net thermal stability may result from a multitude of weakly stabilizing interactions, and different protein families may adopt different structural devices for stabilization^{10,37-42}. Another difficult issue facing the structural analysis of thermal stability is the insufficient amount of structural data available for comprehensive comparison of different thermophilic proteins and their mesophilic homologues. Therefore, even various mechanisms of thermostability were discussed in the literature, many authors pointed to only changes in amino acid composition as one of the clearest manifestations of thermal adaptation^{6,23,25,29,34,36,43}.

To overcome this difficulty, researchers constructed some tools or web servers for protein thermostability prediction recently^{44,45}. Most of them use machine learning method, including SVM or neural networks, to predict experimental $\Delta\Delta G$ or ΔT_m upon the stability change of single point mutation. In these works, all of the training information, thermostability experiment data on single point mutation of proteins, is retrieved from ProTherm database⁴⁶, which was launched in 2002 for collecting protein thermostability experimental data. Several thermostability prediction works followed. As Guerois R. *et al.* reported FOLDEF (for FOLD-X energy function)⁴⁷ in 2002, they provided a quantitative estimation of the importance of the intra-protein atom interactions contributing to the stability of proteins and protein complexes. In 2002, Gilis and Rooman also reported a tool PoPMuSiC for rational computer-aided design of single-site mutations⁴⁸. Moreover, Capriotti E. *et al.*⁴⁹ and Bordner, A.J. *et al.*⁵⁰, Hoppe C *et al.*⁵¹ in 2005, Cheng J *et al.* in 2006⁵², Parthiban V. *et al.*^{53,54} all work on predicting $\Delta\Delta G$ change of single point mutations. Magyar C. *et al.* reported SRide in 2005 for identifying stabilizing residues (SRs) in protein structure⁵⁵ without information from ProTherm, but this server did not have

real experimental validation.

1.2 Motivation

All the thermostability prediction tools mentioned above, except the sequence version of I-Mutant 2.0 as to our best knowledge, require 3-D structures as input. For most of proteins without 3-D structure at a time, it is a practical need to have a thermostability prediction method with sequences as the only input. On the other hand, structure study is always in a case-by-case base and the amount of data with structure available from ProTherm is only about 1,000. However, the number of combinations of amino acid single point mutation is already 380. Obviously, using the data with structure from ProTherm only as the training set for a thermostability prediction is inadequate very much. Therefore, besides ProTherm data, one desires to discover some other useful information for protein thermostability prediction. There are far more sequences than structures available; in the meantime, due to the great progress of genome projects, there are sufficient thermophilic genomes sequenced. It enables to elucidate the relationship between protein sequences and their thermostability, and then possibly apply to protein stabilization. In this direction, we derive two thermostability prediction methods with sequence as the only input. First, we develop an index, structural entropy, of secondary structure information content, eight secondary structure types defined by DSSP⁵⁶, in a local sequence. For a protein sequence, its structural entropy can be computed and used to identify residues involved in thermal stabilization of various protein families. Our results show that the positions of the largest structural entropy difference between wild type and mutant usually coincide with the residues relevant to thermostability, and the

lower information content region means the more stable region in local sequence.

Second, we develop a thermostability prediction method based on amino acid coupling patterns, which defined as two types of amino acid separated by one or more amino acids in a protein sequence. Due to the fast accumulation of protein sequences, genome sequence-based analysis is valuable in the study of thermal adaptation of proteins. The often-used sequence-based methods^{26,28,29,57-61} differentiate the amino acid compositions between thermophilic and mesophilic proteins, and show that thermophilic proteomes exhibit significant bias in their amino acid compositions. However, amino acid composition analysis provides a useful but simplified picture of the relative importance of each individual amino acid type in the thermophilic proteins. Such analysis overlooks the coupling effects between amino acid types on thermal stability of proteins. We conduct a statistical analysis of frequencies of coupling patterns appear in mesophiles and thermophiles and identify significant amino acid coupling sequence patterns in thermophilic proteins⁶². Though no single outstanding coupling pattern can adequately account for protein thermostability, we can use a group of amino acid coupling patterns having strong statistical significance (p values $< 10^{-7}$) to distinguish between thermophilic and mesophilic proteins. We found a good correlation between the optimal growth temperatures of the genomes and the occurrences of the coupling patterns (the correlation coefficient is 0.89). Furthermore, we can separate the thermophilic proteins from their mesophilic orthologs using the amino acid coupling patterns.

Chapter 2 Major Features in Thermophilic Enzymes

There are many features in thermophilic enzymes reported. In this chapter we make a brief survey of these features.

2.1 More Salt Bridges (Ion Pairs)

Ionized residues play essential roles in modulating protein stability, folding and function⁶³. A single salt bridge is recommended 3 to 5 kcal/mol stabilized contribution in T4 lysozyme case study⁶⁴. However, salt bridges are also a destabilized factor in viewpoint of desolvation contribution [$\Delta\Delta G(\text{desolvation})$]³. Many researches report that there are more charged residues or salt bridges (also named ion pairs) on the surface of thermophilic enzymes^{3,11-13}. In 2000, Szilágyi *et al.* had a comprehensive comparison on protein structure of 64 mesophilic and 29 thermophilic or hyperthermophilic protein subunits in 25 protein families¹³. Szilágyi *et al.* conclude the salt bridge is more often included in thermophilic or hyperthermophilic proteins than their mesophilic homologues.

The glutamate dehydrogenase from *Pyrococcus furiosus*, a hyperthermophile (PfGDH) and *Clostridium symbiosum*, a mesophile, (CsGDH), of high sequence and structural similarity, is a typical good model for investigating the molecular basis of thermostability⁶⁵. The comparison of these two proteins shows that salt bridges result in huge difference in melting temperatures (about 60° C) of PfGDH and CsGDH. In Figure 1, the diagram of the monomers of PfGDH (the B chain of 1GTM) and CsGDH (the B chain of 1HRD) are illustrated and the salt-bridge-forming residues are showed in ball-and-stick representation. The side chain atoms of the positively charged residues (Arg, Lys, His) residues of blue are showed in blue, and the side chain atoms of the negatively charged

residues (Glu, Asp) are showed in red. C α atoms of the salt-bridge-forming residues are shown in black. There are 29 salt bridges in a hyperthermophilic PfGDH monomer, and 17 salt bridges in the corresponding mesophilic CsGDH monomer^{65,66}. The PfGDH monomer has more salt bridges near the active site than CsGDH monomer.

Across the subunit, salt bridges form a large network in PfGDH (Figure 2). There are 24 residues, belonging to four different subunits, forming 18 salt bridges^{3,67}. These evidences suggest that salt bridges are an important feature for protein thermostability.

In general, the salt bridge is an important feature of protein thermostability. However, it is difficult to find suitable structure position for adding more salt bridges to improve protein thermostability.



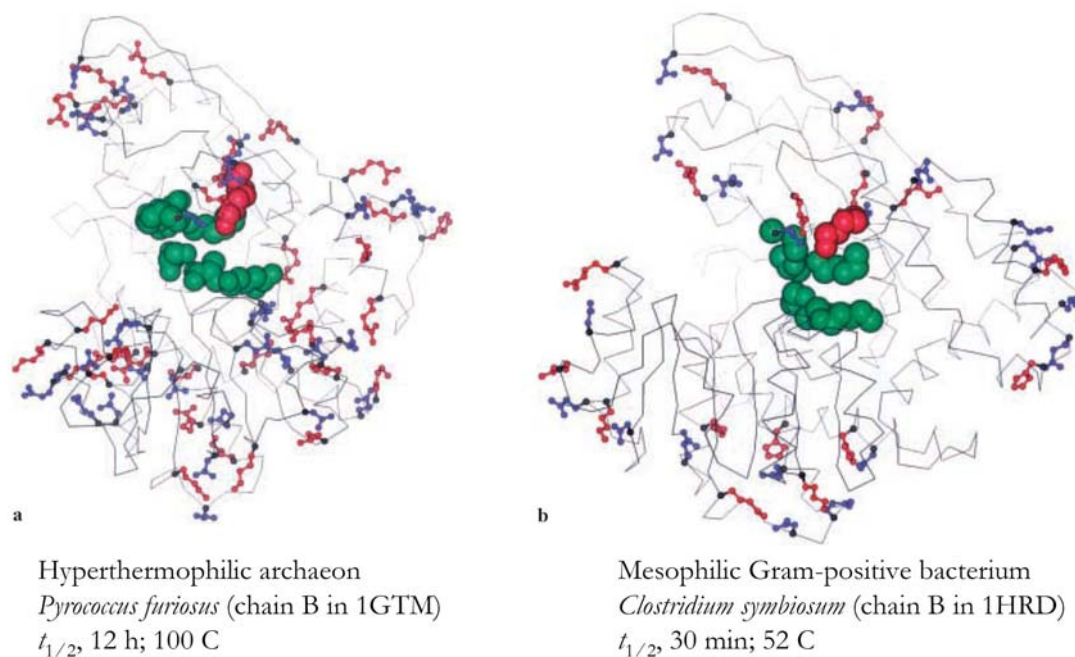


Figure 1. Glutamate dehydrogenase from the hyperthermophilic archaeon *Pyrococcus furiosus* (chain B in 1GTM) (a) and mesophilic Gram-positive bacterium *Clostridium symbiosum* (chain B in 1HRD) (b) are shown with Ca trace representation in a subunit. Salt-bridge-forming residues are shown with ball-and-stick representation. Active-site residues are shown with CPK representation, and the conserved active site Lys is shown in CPK red. Other active-site residues are in green. Thermophilic glutamate dehydrogenase has more salt bridges than the mesophilic glutamate dehydrogenase in the neighborhood of the active site. Reprinted from ref 65.

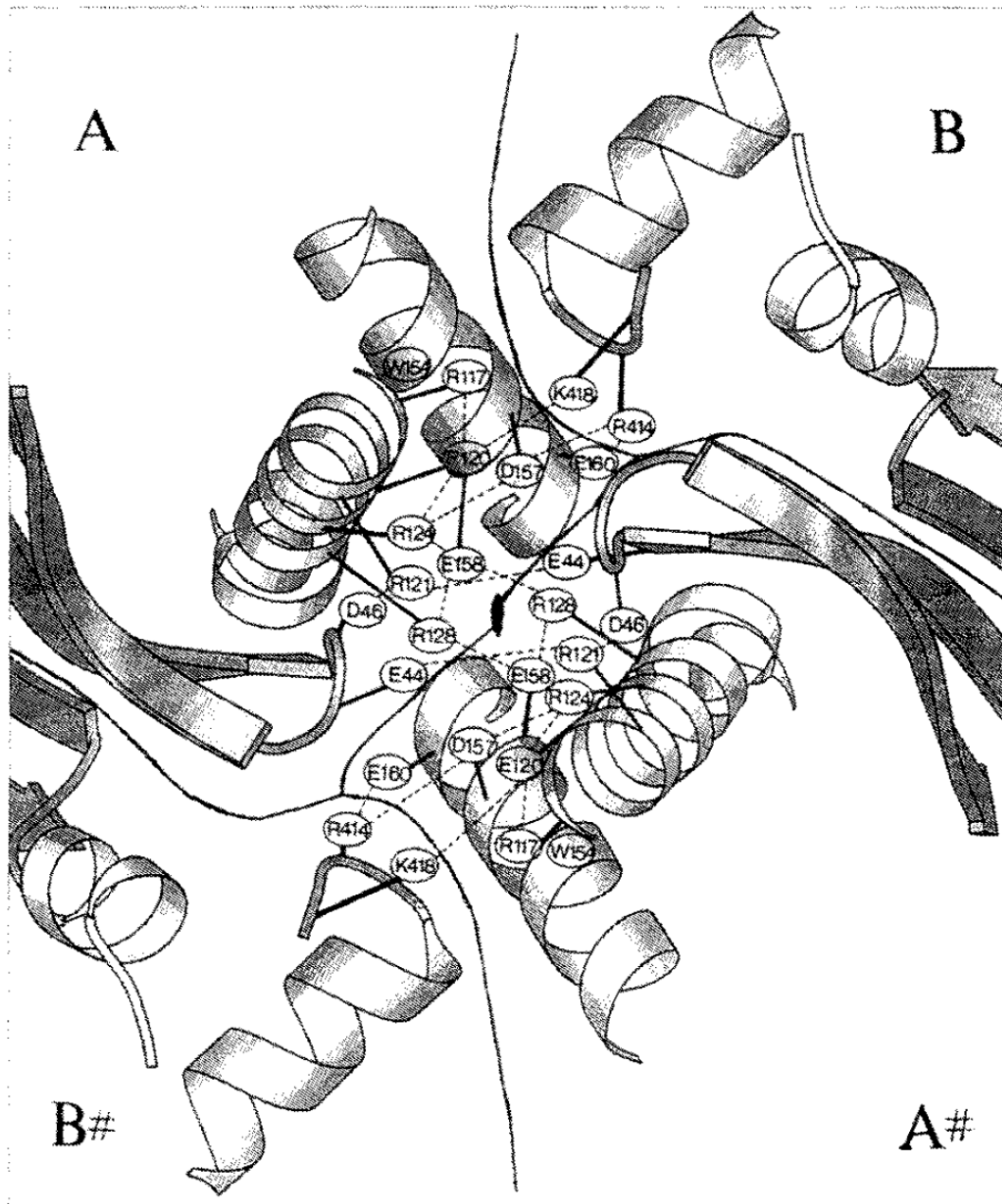


Figure 2. The salt-bridge network in the hexameric *Pyrococcus furiosus* GDH is shown and considered that the salt-bridge network can stabilize the intersubunit interactions. Salt-bridge interactions are represented in dotted lines. The two-fold axis of symmetry between the dimers is indicated by the # symbol. Reprinted from reference 3.

2.2 Side Chain- Side Chain Hydrogen Bonds

In a case of stability study on RNase T1, it is estimated that a single H bond contributes an average of 1.3 kcal/mol to the stabilization⁶⁸. By Tanner *et al.*, an increased number of charged-neutral H-bonds is found in thermophilic GAPDH in comparison with its mesophilic homologue⁶⁹, which suggests that charged residues form H-bonds in protein stabilization.

The H-bond is defined by a distance cutoff in structures that less than 3 Å between the H donor and the H acceptor and by donor-hydrogen-acceptor angle below 90°. Since numbers of hyperthermophilic protein structures have not been refined to sufficiently high resolutions, it is not clear now to answer the role of H bonds in thermostability by structure analysis³.

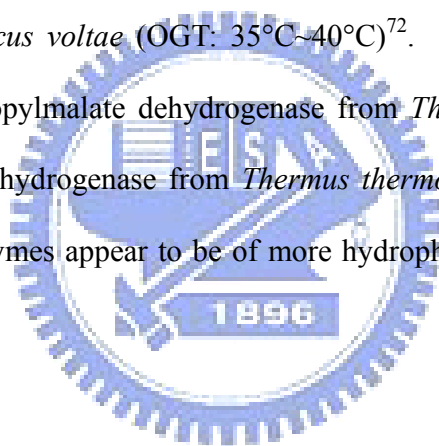
2.3 Dipole-Dipole Interactions or Cation- π Interactions

Cation- π interactions have significant contribution in enhancing thermal stability, because an increased frequency of both exposed aromatic and positively charged residues in thermophiles¹⁰. In cation- π interactions, aromatic rings of Phe, Tyr, and Trp are nonpolar residue types and do not have a net permanent dipole moment. However, they have quadrupole moments that are quite substantial in magnitude¹⁰. The quadrupole is two opposing dipoles originating from either face of the ring. Cations (metal cations or cationic side chains of Arg and Lys) interact with the center of the aromatic ring strongly and there is two fold than salt bridges^{21,70}, because Phe, Tyr, and Trp are low desolvation energies and can easily be stabilized in hydrophobic environment²¹. There are over 70% of all Arg side near aromatic side chains and 26% of all Trps are involved in energetically

significant cation- π interactions on the surface of proteins⁷⁰.

2.4 Higher Degree of Hydrophobic Packing in the Core Regions

Hydrophobic interaction is a stabilization mechanism in hyperthermophilic proteins³. Each additional methyl group buried in protein folding contribute an average increase in stability of 1.3 (± 0.5) kcal/mol⁷¹. The estimation is based on cavity-creating mutations in which a large aliphatic residue was replaced with a smaller aliphatic residue. In homologous protein comparison study, adenylate kinase from *Methanococcus jannaschii* (OGT: 85°C) has higher degree of hydrophobic packing in the core region than adenylate kinase from *Methanococcus voltae* (OGT: 35°C~40°C)⁷². It had similar result in the comparison with 3-isopropylmalate dehydrogenase from *Thermus thermophilus* and *E. coli* 3-isopropylmalate dehydrogenase from *Thermus thermophilus*⁷³. According these reports, thermophilic enzymes appear to be of more hydrophobic interaction than mesophilic homologues.



2.5 Increased Packing Density

There are some of hyperthermophilic proteins with better packing than mesophilic proteins. Britton K. L. *et al.* reported that significantly more Ile in *Pyrococcus furiosus* (OGT: 100°C) GDH than in the mesophilic *Clostridium symbiosum* GDH and guess that the effort is due to Ile is more conformations and better fill various voids than Leu⁷⁴. The thermophilic *Methanobacterium fervidus* histone is bulkier hydrophobic side chains in solvent-accessible cavity than mesophilic *Methanobacterium formicicum* histone⁷⁵. Ala31Ile and Lys35Met mutations of the *M. formicicum* histone increased T_m by 11 and 14°C, respectively, while Ile31Ala and Met35Lys mutations of the *M. fervidus* histone

decreased T_m by 4 and 17°C, respectively⁷⁶.

2.6 Disulfide Bridges

Disulfide bridges are an entropic effect to stabilize proteins and decreasing the unfolded state entropy⁷⁷. The serine protease from *Aquifex pyrophilus* contains eight cysteines (none are present in subtilisin BPN⁷⁸). However, some paper report 100°C is the upper limit for the stability of proteins containing disulfide bridges⁷⁹, because disulfide bridges is suspected to be destructed at high temperature.

2.7 Shorter Loop Structures or Higher Conformational Rigidity

Loops and N and C termini unfold first during thermal denaturation. It is reported that in hyperthermophilic proteins, loops are either shortened or better anchored to the rest (like N and C termini) of the protein³. Loop shortening creates more ratio of secondary structure. Loop anchoring is stabilized by ion pairing, H-bonding, or hydrophobic interactions. For one example, the lactate dehydrogenase from *Thermotoga maritima* (OGT: 80°C) has shorter loop structures than its mesophilic homologues⁸⁰. Another example is the C terminus of *Aquifex pyrophilus* superoxide dismutase is 10 or 11 residues longer than mesophilic superoxide dismutases, and C-terminal helix makes contacts with another subunit¹¹.

2.8 More Secondary Structural Elements Such as α -Helix or β -Sheet

In 2002, Chakravarty S. *et al.* reported a statistical result that there are more regular secondary structures in thermophilic than in mesophilic homologues, and the protein in thermophiles especially increase in helical content and decrease in loop content¹⁰. In

thermophilic proteins, the ratio of residues in helical, strand, and loop regions are 38.5, 17.9, and 43.6%; in mesophilic proteins, the ratio of residues in helical, strand, and loop regions are 36.9, 18.2, and 44.6%.

2.9 Bias in Amino Acid Composition

Amino acid composition has been thought to be correlated to thermostability. In the early statistical analysis on amino acid compositions of mesophilic and thermophilic proteins indicated the trend toward substitutions of amino acid such as Gly to Ala and Lys to Arg in mesophilic proteins to their thermophilic homologues³. As more genome sequence data accumulated, the indication can be more reliable.

In 2000, Cambillau C. *et al.* reported *CvP*-bias, a large difference between the proportions of charged (Asp, Glu, Lys, Arg) versus polar (non-charged) (Asn, Gln, Ser, Thr) amino acids in comparison of 58 thermophilic and mesophilic proteins. *CvP*-bias is the most significant signature of the hyperthermophilic organisms proteomes (*OGT* >80 °C)²⁷. In 2003, Karsten S. *et al.* re-analyzed *CvP*-bias on a wider data set of 9 thermophiles, 9 hyperthermophiles, and 53 mesophiles⁸¹.

In 2001, Kreil D. P. *et al.* used hierarchical clustering and principal component analysis (PCA) to show an influence of underlying factors on the global amino acid composition which was based on the proteome of 6 thermophilic archaea, 2 thermophilic bacteria, 17 mesophilic bacteria, and 2 eukaryotic species⁵⁸. They reported that the G + C content can not identify the thermophilic species while their global amino acid compositions can identify thermophilic species.

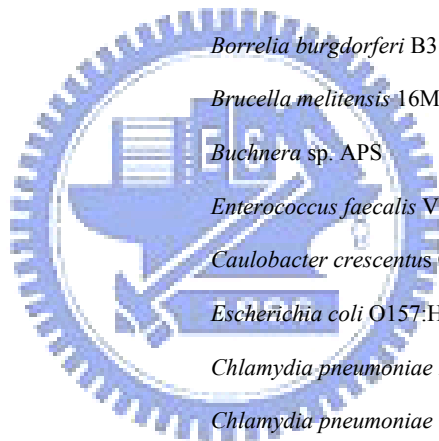
We analyze more genomes from TIGR database for higher statistic confidences. There were 89 prokaryotic genomes in TIGR database on 2002 Oct. 3, and at that time

we found 85 prokaryotic genomes with optimal growth temperature (OGT) record. The genomes are listed in Table 1 and the statistics is showed in Table 2. We collect the open reading frame (ORF) of these genomes. In this set of data, the ORF protein sequence number is 251,836; the average sequence number of each genome is 2,830 (251,836/89). The organisms are divided into two classes, mesophiles of OGT below 45° C and thermophiles of OGT over 45° C.

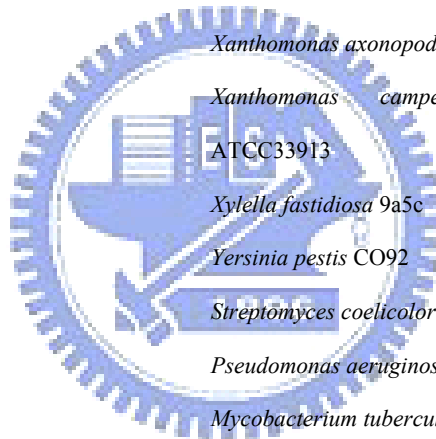
Table 1. The prokaryotic genome in TIGR on 2002 Oct3.

Domain	Optimal growth temperature	Species name
Archaea	80	<i>Sulfolobus tokodaii</i> strain 7
Archaea	30	<i>Methanosarcina mazei</i> Goe1
Archaea	37	<i>Halobacterium</i> sp. NRC-1
Archaea	98	<i>Pyrococcus horikoshii</i> shinkaj OT3
Archaea	100	<i>Pyrococcus furiosus</i> DSM 3638
Archaea	96	<i>Pyrococcus abyssi</i> GE5
Archaea	100	<i>Pyrobaculum aerophilum</i> IM2
Archaea	80-87	<i>Sulfolobus solfataricus</i> P2
Archaea	59	<i>Thermoplasma acidophilum</i> DSM 1728
Archaea	65-70	<i>Methanobacterium thermoautotrophicum</i> delta H
Archaea	35	<i>Methanosarcina acetivorans</i> C2A
Archaea	>100	<i>Methanopyrus kandleri</i> AV19
Archaea	95	<i>Aeropyrum pernix</i> K1
Archaea	85	<i>Methanococcus jannaschii</i> DSM2661
Archaea	60	<i>Thermoplasma volcanium</i> GSS1
Archaea	67.5	<i>Mesorhizobium loti</i> MAFF303099
Archaea	83	<i>Archaeoglobus fulgidus</i> DSM4304
Bacteria	30	<i>Lactococcus lactis</i> subsp. lactis IL1403
Bacteria	37	<i>Helicobacter pylori</i> J99
Bacteria	37	<i>Escherichia coli</i> O157:H7 VT2-Sakai
Bacteria	20-25	<i>Listeria innocua</i> CLIP 11262

Bacteria	37	<i>Corynebacterium glutamicum</i> ATCC 13032
Bacteria	37	<i>Helicobacter pylori</i> 26695
Bacteria	20-25	<i>Listeria monocytogenes</i> EGD-e
Bacteria	35-37	<i>Haemophilus influenzae</i> KW20
Bacteria	37	<i>Fusobacterium nucleatum</i> ATCC 25586
Bacteria	NA	<i>Magnetococcus</i> MC-1
Bacteria	37	<i>Chlamydia muridarum</i> strain Nigg
Bacteria	25-28	<i>Agrobacterium tumefaciens</i> C58 Cereon
Bacteria	25-28	<i>Agrobacterium tumefaciens</i> C58 UWash
Bacteria	95	<i>Aquifex aeolicus</i> VF5
Bacteria	40-55	<i>Bacillus halodurans</i> C-125
Bacteria	37	<i>Bacillus subtilis</i> 168
Bacteria	30	<i>Borrelia burgdorferi</i> B31
Bacteria	37	<i>Brucella melitensis</i> 16M
Bacteria	NA	<i>Buchnera</i> sp. APS
Bacteria	37	<i>Enterococcus faecalis</i> V583
Bacteria	30	<i>Caulobacter crescentus</i> CB15
Bacteria	37	<i>Escherichia coli</i> O157:H7 EDL933
Bacteria	37	<i>Chlamydia pneumoniae</i> AR39
Bacteria	37	<i>Chlamydia pneumoniae</i> CWL029
Bacteria	37	<i>Chlamydia pneumoniae</i> J138
Bacteria	37	<i>Chlamydia trachomatis</i> serovar D
Bacteria	20-35	<i>Chlorobium tepidum</i> TLS
Bacteria	37	<i>Clostridium perfringens</i> 13
Bacteria	37	<i>Mycoplasma pulmonis</i> UAB CTIP
Bacteria	25-35	<i>Deinococcus radiodurans</i> R1
Bacteria	37	<i>Escherichia coli</i> K12-MG1655
Bacteria	42-45	<i>Campylobacter jejuni</i> NCTC 11168
Bacteria	75	<i>Thermoanaerobacter tengcongensis</i> MB4(T)
Bacteria	37	<i>Mycoplasma genitalium</i> G-37
Bacteria	30-37	<i>Staphylococcus aureus</i> N315



Bacteria	37	<i>Streptococcus agalactiae</i> 2603V/R
Bacteria	37	<i>Streptococcus pneumoniae</i> TIGR4
Bacteria	37	<i>Streptococcus pneumoniae</i> R6
Bacteria	37	<i>Streptococcus pyogenes</i> SF370 serotype M1
Bacteria	37	<i>Streptococcus pyogenes</i> MGAS8232
Bacteria	30-37	<i>Staphylococcus aureus</i> MW2
Bacteria	26	<i>Synechocystis</i> sp. PCC6803
Bacteria	30-37	<i>Staphylococcus aureus</i> COL
Bacteria	80	<i>Thermotoga maritima</i> MSB8
Bacteria	37	<i>Treponema pallidum</i> Nichols
Bacteria	37	<i>Ureaplasma urealyticum</i> parvum biovar serovar 3
Bacteria	37	<i>Vibrio cholerae</i> El Tor N16961
Bacteria	35-37	<i>Xanthomonas axonopodis</i> pv. citri 306
Bacteria	35-39	<i>Xanthomonas campestris</i> pv. campestris ATCC33913
Bacteria	28	<i>Xylella fastidiosa</i> 9a5c
Bacteria	28	<i>Yersinia pestis</i> CO92
Bacteria	10-37	<i>Streptomyces coelicolor</i> A3(2)
Bacteria	37	<i>Pseudomonas aeruginosa</i> PAO1
Bacteria	37	<i>Mycobacterium tuberculosis</i> CDC1551
Bacteria	37	<i>Mycobacterium tuberculosis</i> H37Rv (lab strain)
Bacteria	36-38	<i>Mycoplasma pneumoniae</i> M129
Bacteria	NA	<i>Brucella suis</i> 1330
Bacteria	36-37	<i>Neisseria meningitidis</i> MC58
Bacteria	36-37	<i>Neisseria meningitidis</i> serogroup A Z2491
Bacteria	26	<i>Nostoc</i> sp. PCC 7120
Bacteria	30-37	<i>Staphylococcus aureus</i> Mu50
Bacteria	37	<i>Porphyromonas gingivalis</i> W83
Bacteria	37	<i>Mycobacterium leprae</i> TN
Bacteria	32	<i>Ralstonia solanacearum</i> GMI1000
Bacteria	32-34	<i>Rickettsia conorii</i> Malish 7



Bacteria	35	<i>Rickettsia prowazekii</i> Madrid E
Bacteria	37	<i>Salmonella enterica</i> serovar Typhi CT18
Bacteria	37	<i>Salmonella typhimurium</i> LT2 SGSC1412
Bacteria	30	<i>Shewanella oneidensis</i> MR-1
Bacteria	26	<i>Sinorhizobium meliloti</i> 1021
Bacteria	37	<i>Pasteurella multocida</i> PM70
Viruses	NA	SIFV (<i>Sulfolobus islandicus</i> filamentous virus)

Table 2. Prokaryotic genome data statistic

	Recorded	Mesophiles	Thermophiles and hyperthermophiles	Unknown OGT	Total
Archaea	17	3	14	0	17
Bacteria	68	63	5	3	71
Viruse	0	0	0	1	1
Total	85	66	19	4	89

In Figure 3, the box plot indicates that the amino acid frequency distribution of whole genome ORFs. The data for thermophiles and hyperthermophiles are in red, while that for mesophiles is in blue. As shown in this plot, we can conclude that the frequency of Glu (E) and Val (V) are significant more in thermophiles and hyperthermophiles than in mesophiles and that of Gln (Q) and Thr (T) are significantly less. The result is similar to that of Cambillau and Claverie²⁷ and Kreil and Ouzounis⁵⁸.

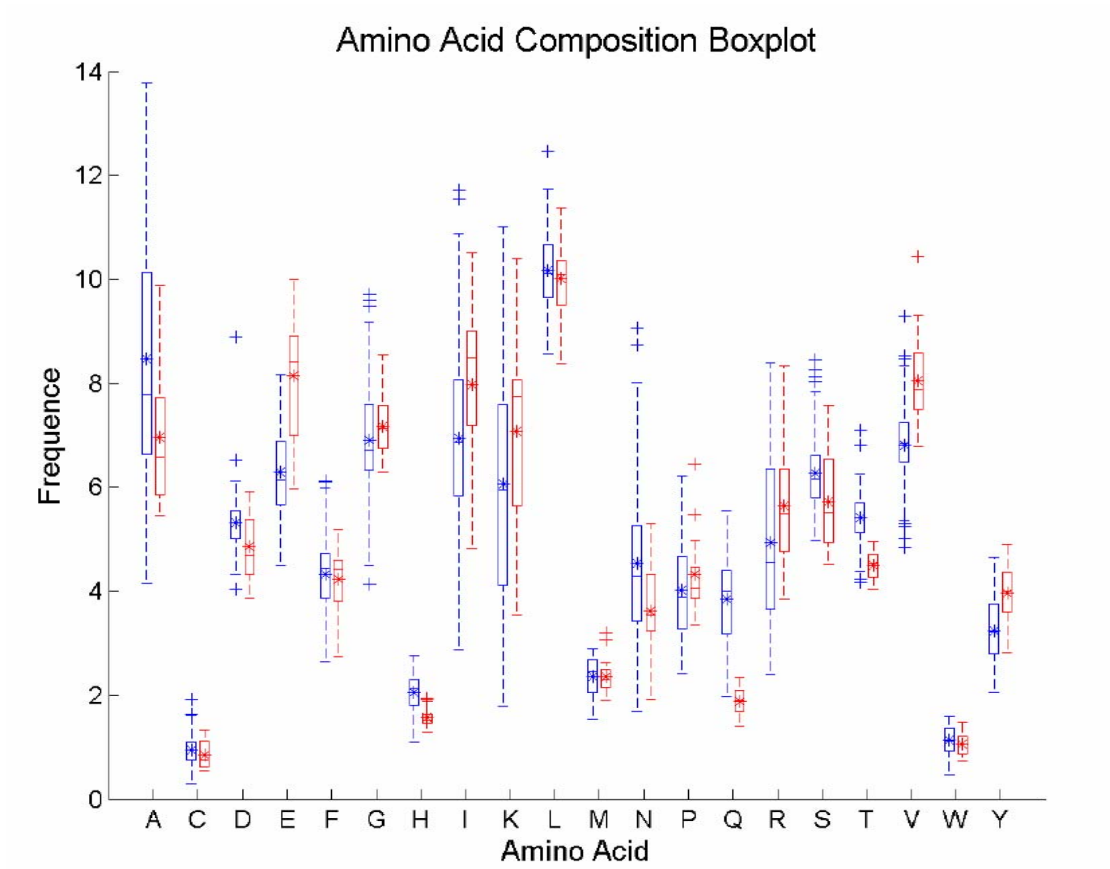


Figure 3. The box-plot represents the amino-acid composition in high growth temperature genomes (red line, thermophiles and hyperthermophiles) and low growth temperature genomes (bleu line, mesophiles).

Chapter 3 Methods

We develop two methods, sequence derived structural entropy (SDSE) and amino acid coupling patterns method, for thermostability prediction in sequence information. In this chapter we introduce the details.

3.1 Sequence Derived Structural Entropy (SDSE)

The structural propensity of an amino acid segment x of length l is described by an n -component structural vector (p_1, p_2, \dots, p_n) , where p_i is the probability of the i^{th} structural descriptor characterized by eight secondary structure types defined by DSSP⁵⁶: β -bridges, extended β -sheet, 3_{10} -helix, α -helix, π -helix, bend, turn, and others. The structural entropy of the segment is calculated by

$$S = -\sum_i p_i \ln p_i \quad (1)$$

where the summation is over the secondary structure types and p_i is the occurrence probability of i^{th} structural descriptor. For an amino acid segment of length l , the expected number of segments is 20^l . If l is large, there may not be enough samples in the structure database to compute the occurrence probability of a particular structural descriptor. In the present work, the computations of entropy were done for tetra-peptide segments for large variability in patterns and reasonable coverage. Both the web-implemented program and the compiled entropy library are available from <http://SDSE.life.nctu.edu.tw/>.

We illustrate the computational procedure with an example: Given a sequence "...CRLPGTPEAICATYTGCI...", imagine we are interested in computing the struc-

tural entropy at the “I” position for this sequence. If $l = 4$, there are four sequence windows covering this particular residue **I**, whose structural profile vectors are given by $\mathbf{p}_4^{\text{PEAI}}$, $\mathbf{p}_3^{\text{EAIC}}$, $\mathbf{p}_2^{\text{AICA}}$ and $\mathbf{p}_1^{\text{ICAT}}$, respectively (see Figure 4). We compute the average structural profile vector at **I** by

$$\bar{\mathbf{p}} = \frac{1}{4} \left(\mathbf{p}_4^{\text{PEAI}} + \mathbf{p}_3^{\text{EAIC}} + \mathbf{p}_2^{\text{AICA}} + \mathbf{p}_1^{\text{ICAT}} \right) \quad (2)$$

This is equivalent to a weighted average over a seven residue window where the nearer neighboring residues are given more weight. The structural entropy S at **I** for the query sequence is then computed by

$$S = - \sum_{j=1}^8 \bar{\pi}_j \ln \bar{\pi}_j, \quad (3)$$

where $\bar{\pi}_j$ is the j th component of $\bar{\mathbf{p}}$. We built the library of structural profile using the SCOP-35 dataset⁸², which is the non-redundant subset comprised of sequences with pairwise sequence identities <35%. Using the SCOP-35 dataset can help avoid sampling bias due to homologue redundancy. For sequence fragments with lengths 3, 4 and 5, the numbers of distinct patterns are 8×10^3 , 1.6×10^5 and 3.2×10^6 , respectively, and their coverage by SCOP-35 is 99%, 86% and 19%, respectively. In this work, the structural profile library is built for tetra-peptides ($l = 4$) for the consideration of sufficient sequence coverage and sequence patterns. For sequences of lower occurrence, we used the pseudocount method⁸³ described before to estimate the occurrence probability. The complete flowchart for computing the structural entropy of a query sequence is shown schematically in Figure 5.

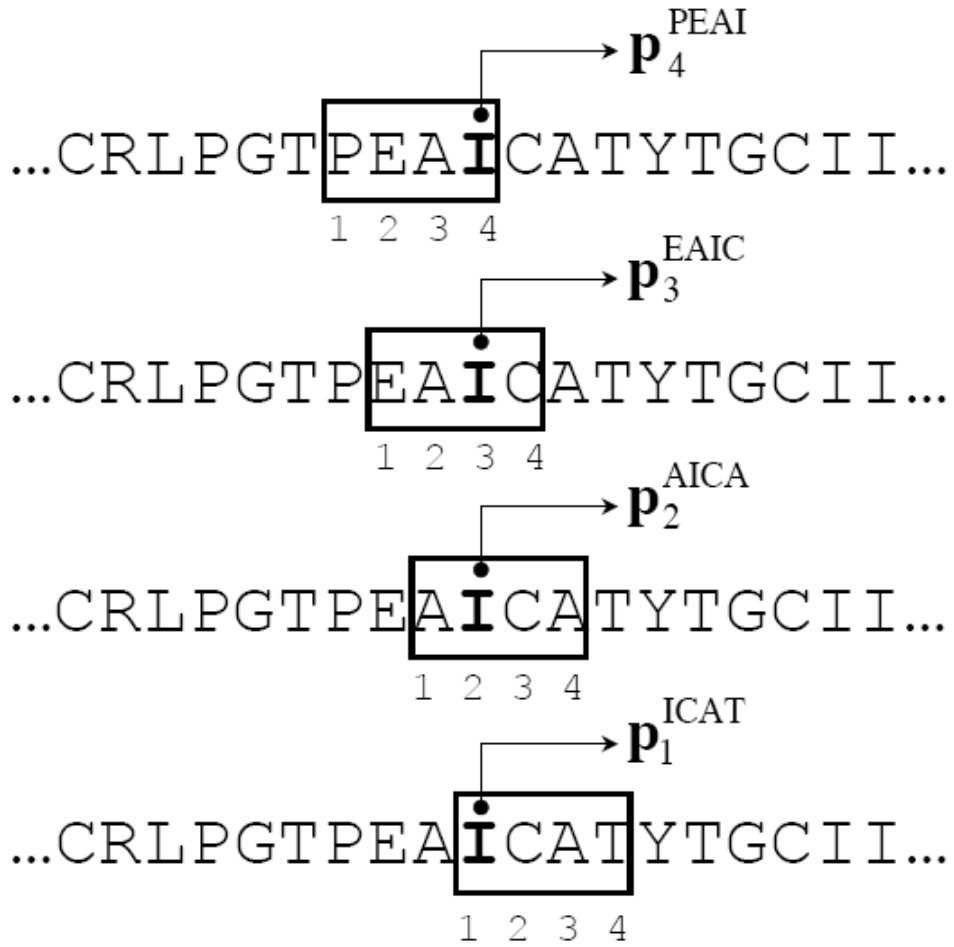


Figure 4. An example to compute the structural entropy of a particular residue ("I" in bold face) of a protein sequence. If the length of the sequence window is $l = 4$, there are possible four sequence windows covering this particular residue "I": PEAI, EAIC, AICA and ICAT. The structural profile vectors of "I" for these sequence fragments are $\mathbf{p}_4^{\text{PEAI}}$, $\mathbf{p}_3^{\text{EAIC}}$, $\mathbf{p}_2^{\text{AICA}}$ and $\mathbf{p}_1^{\text{ICAT}}$, respectively. The structural entropy of "I" is computed using Eq. 2 and 3.

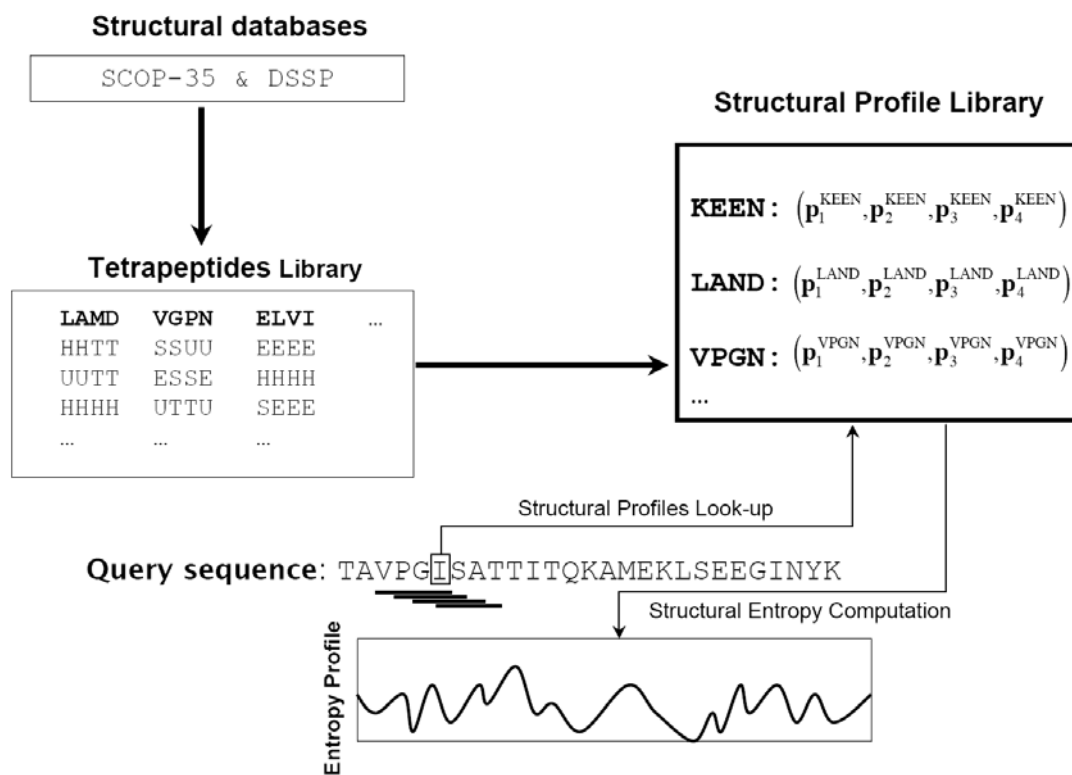


Figure 5. The schematics of calculating the structural entropy profile of a query sequence. We built the tetra-peptide library together with their secondary structural elements from the SCOP-35 and DSSP databases. We then built the library of structural profiles for all tetra-peptides. For a query sequence, we can compute the structure entropy of each position from the structural profile library by averaging four successive sequence windows, indicated by four stacking thick lines.

3.2 Amino Acid Coupling Patterns

The sequence coupling pattern is defined as any two types of amino acids separated by one or more amino acids. We conduct a statistical analysis on the thermophilic and mesophilic microbial genomes to identify significant sequence-coupling patterns for thermostability.

3.2.1 Coupling Patterns [XdZ]

Let [XdZ] denote the amino acid-coupling pattern of amino acids type X and Z that are separated by d amino acids. Since the protein sequence is directional, the sign of d is determined by the relative positions of X and Z . If X is closer to the N terminal side, d is defined to be positive, and if X is closer to the C terminal side, it is defined to be negative. Let $N(XdZ)$ be the number of occurrences of the pattern [XdZ]. We define the conditional probability R_{XdZ} as

$$R_{XdZ} = \frac{N(XdZ)}{N(Xd\cdot)}, \quad (4)$$

where $N(Xd\cdot) = \sum_Y N(XdY)$ and $Y \in \{20 \text{ types of amino acid}\}$. The coupling

strength C_{XdZ} between X and Z of the pattern [XdZ] is given by

$$C_{XdZ} = \frac{R_{XdZ}}{P(Z)}, \quad (5)$$

where $P(Z)$ is the probability of the occurrence of amino acid Y . If $C_{XdZ} \geq 1$, then X and Z are positively correlated with respect to the distance d , and if $C_{XdZ} < 1$ they are negatively correlated. We use R_{XdZ}^T and R_{XdZ}^M to denote the means of R_{XdZ} over thermophilic and mesophilic, respectively. To compute the relative occurrence of $[XdZ]$ in thermophilic proteins, we define

$$\rho_{XdZ} = \frac{R_{XdZ}^T}{R_{XdZ}^M}, \quad (6)$$

The ρ value of pattern $[XdZ]$ gives a measure of its relative occurrence in thermophiles compared with mesophiles. If $\rho_{XdZ} > 1$, $[XdZ]$ is increased in thermophilic proteins, and if $\rho_{XdZ} < 1$, it is decreased in mesophilic proteins. We will refer to ρ_{XdZ} as the thermophilic coefficient, or simply the ρ value of $[XdZ]$.

To check the statistical significance of $[XdZ]$, we carry out the Wilcoxon rank-sum test on R_{XdZ} and C_{XdZ} between thermophilic and mesophilic genomes. The Wilcoxon rank-sum test is a non-parametrical statistical test for two independent samples. The resultant p -value is used to determine whether the null hypothesis is true. For example, in the Wilcoxon rank-sum test, if p -value is less than 10^{-2} , we have 99% confidence that the coupling patterns present in the thermophilic and mesophilic samples are significantly different. We have studied $20 \times 20 \times 40 = 16,000$ amino acid-coupling patterns $[XdZ]$ for X, Z over all 20 types of amino acid and $-20 \leq d \leq 20$. When the separation greater than 20 amino acids, we find that $C_{XdZ} \sim 1$, indicating the correlation between amino ac-

ids becomes insignificant when $|d| \geq 20$. The p -values of the Wilcoxon rank-sum test for R_{XdZ} and C_{XdZ} are called their $RS(R)$ and $RS(C)$ values, respectively.

3.2.2 Thermophilic Dominant (Thermo) and Mesophilic Dominant (Meso) Patterns score T and M

The Thermo and Meso pattern scores T and M of protein sequence are thermostability score calculated based on significant Thermo and Meso coupling patterns observed in the sequence.

Let P_T and P_M denote the set of significant Thermo and Meso coupling patterns, respectively. Let S denote a protein sequence under investigation, $P(S)$ denote the set of coupling patterns observed in sequence S and $P_T(S)$ denote the intersection of P_T and $P(S)$. $P_M(S)$ denote the intersection of P_M and $P(S)$. Let $k = 8,000$ denote the number of amino acid coupling patterns $[XdZ]$ for X, Z over all 20 types of amino acid and $-10 \leq d \leq 10$.

The Thermo and Meso pattern scores T and M are defined as following

$$T = \frac{100 \times k \times |P_T(S)|}{|P(S)| \times |P_T|}, \quad (7)$$

$$M = \frac{100 \times k \times |P_M(S)|}{|P(S)| \times |P_M|}, \quad (8)$$

3.2.3 Conditional Probability Score R

The conditional probability score R of protein sequence is a thermostability score calculated based on significant coupling patterns under conditional probability observed in the sequence.

Let P_R denote the set of significant coupling patterns under conditional probability. Let

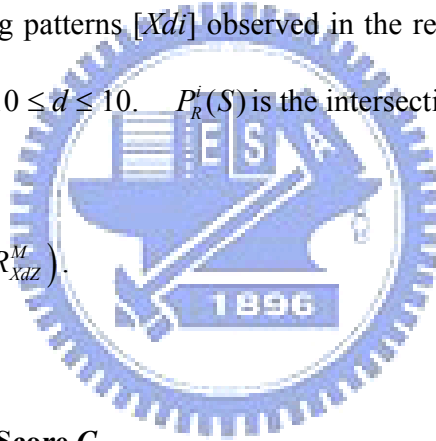
S denote a protein sequence under investigation, $P(S)$ denote the set of coupling patterns observed in sequence S and $P_R(S)$ denote the intersection of P_R and $P(S)$. Then R is defined as

$$R = \frac{k}{|P(S)| \times |P_R|} \times \sum_{XdZ \in P_R(S)} \log(R_{XdZ}^T / R_{XdZ}^M). \quad (9)$$

The conditional probability profile of protein sequence is a vector of conditional probability score for each residue. For residue i , the conditional probability score R_i is a thermostability score that calculated based on significant coupling patterns under conditional probability observed in the residue.

$P^i(S)$ is the set of coupling patterns $[Xdi]$ observed in the residue i , where X over all 20 types of amino acid and $-10 \leq d \leq 10$. $P_R^i(S)$ is the intersection of P_R and $P^i(S)$. Then R_i is defined as

$$R_i = \sum_{XdZ \in P_R^i(S)} \log(R_{XdZ}^T / R_{XdZ}^M). \quad (10)$$



3.2.4 Coupling Strength Score C

The coupling strength score C of protein sequence is a thermostability score that calculated based on significant coupling patterns under coupling strength observed in the sequence.

Let P_C denote the set of significant coupling patterns under coupling strength. Let S denote a protein sequence under investigation, $P(S)$ denote the set of coupling patterns observed in sequence S and $P_C(S)$ denote the intersection of P_C and $P(S)$. Then C is defined as

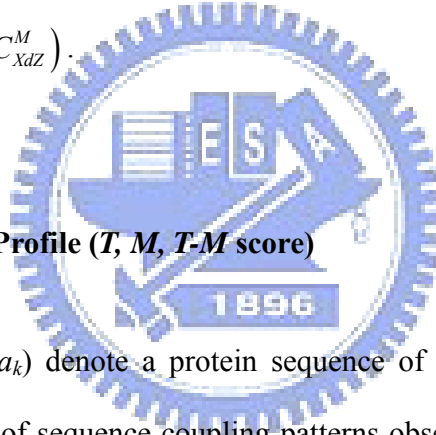
$$C = \frac{k}{|P(S)| \times |P_C|} \times \sum_{XdZ \in P_C(S)} \log(C_{XdZ}^T / C_{XdZ}^M), \quad (11)$$

The coupling strength profile of protein sequence is a vector of coupling strength score for each residue. For residue i , the coupling strength score C_i is a thermostability score that calculated based on significant coupling patterns under coupling strength observed in the residue.

Let $P^i(S)$ be the set of coupling patterns $[Xdi]$ observed in the residue i , where X over all 20 types of amino acid and $-10 \leq d \leq 10$, and $P_C^i(S)$ be the intersection of P_C and $P^i(S)$.

Then C_i is defined as

$$C_i = \sum_{XdZ \in P_C^i(S)} \log(C_{XdZ}^T / C_{XdZ}^M). \quad (12)$$



3.2.5 Dominant Pattern Profile (T , M , T - M score)

Let $S = (a_1, \dots, a_i, \dots, a_k)$ denote a protein sequence of size n under investigation.

Let $P_i(S)$ denote the set of sequence coupling patterns observed at residue a_i , namely,

$\{[a_j(i-j)a_i] | 1 \leq j \neq i \leq n\}$. Let $P_{iT}(S)$ denote the intersection of P_T and $P_i(S)$, and

$P_{iM}(S)$ denote the intersection of P_M and $P_i(S)$. The thermo-score T_i and

meso-score M_i at residue a_i are defined as below.

$$T_i = |P_{iT}(S)|, \quad (13)$$

$$M_i = |P_{iM}(S)|. \quad (14)$$

The $T - M$ score at residue a_i is defined as $T_i - M_i$. The dominant pattern profile of this protein sequence is given as $T_i, M_i, T_i - M_i, i = 1, \dots, n$.

3.3 Optimal Growth Temperature Dependent Amino Acid Composition (OGT_{Comp})

For a protein sequence $S = (a_1, \dots, a_n)$, OGT_{Comp} , the optimal growth temperature prediction formula based on amino acid composition, is a formula stated by Nakashima H. *et al.*⁵⁷ We find that it is also useful in protein thermostability prediction.

$$\begin{aligned}
 w(Ala) &= -0.96, w(Cys) = -0.85, w(Asp) = -2.57, w(Glu) = 1.77, \\
 w(Phe) &= 0.64, w(Gly) = 0.63, w(His) = -1.79, w(Ile) = 2.60, \\
 w(Lys) &= 1.22, w(Leu) = 1.26, w(Met) = 0.62, w(Asn) = -1.27, \\
 w(Pro) &= 1.49, w(Gln) = -3.51, w(Arg) = 1.37, w(Ser) = -0.83, \\
 w(Thr) &= -0.48, w(Val) = 2.10, w(Trp) = 1.95 \text{ and } w(Tyr) = 2.53
 \end{aligned}$$

$$OGT_{comp}(S) = \frac{1}{n} \sum_{i=1}^n w(a_i) \times 100 + 0.45 \quad (15)$$

For a protein sequence $S = (a_1, a_2, \dots, a_n)$, the amino acid composition profile $C = (c_1, c_2, \dots, c_n)$ are the vector of the calculated OGT_{Comp} of amino acid residue i .

$$c_i = OGT_{comp}(a_i) = w(a_i) \quad (16)$$

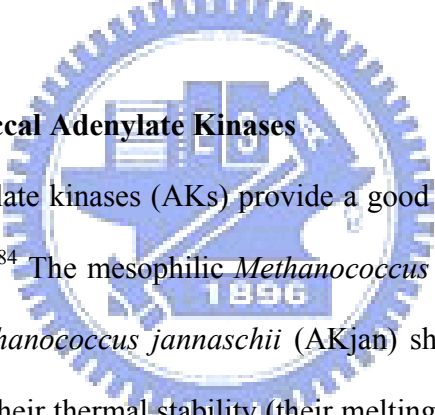
Chapter 4 Results and Discussion

In this chapter we discuss the results in sequence derived structural entropy (SDSE) and amino acid coupling patterns method.

4.1 Sequence Derived Structural Entropy (SDSE)

To explore the relationship between the structural entropy profile and the protein thermal stability, we present three examples: methanococcal adenylate kinases and their chimeric constructs,^{72,84} Ribonuclease HIs and their chimeric constructs,⁸⁵ and holocytochrome c551 and its single/multiple amino acid mutants.¹⁴

4.1.1 Case 1: Methanococcal Adenylate Kinases



The methanococcal adenylate kinases (AKs) provide a good model system to study protein thermostabilization.^{72,84} The mesophilic *Methanococcus voltae* (AKvol) and the extremely thermophilic *Methanococcus jannaschii* (AKjan) share 61% sequence identity, but differ significantly in their thermal stability (their melting temperatures are 69 °C and 103 °C, respectively). The structure of AK is characterized by the CORE domains (residues 1-38, 86-134 and 145-192), the nucleoside monophosphate (NMP)-binding domain (residue 39-85) and the LID domain (residue 135-144). Figure 6 shows the computed structural entropy profiles of AKjan (*SAKjan*) and AKvol (*SAKvol*) as well as their entropy difference $\Delta S = SAKjan - SAKvol$. Most residues of the AKjan sequence are seen to have lower structural entropy than those of the AKvol sequence, especially in the CORE domains. We observed that most of the residues (filled circles) involved in the thermal stabilization of AKs^{72,84} occur at or close to the ΔS minima. Figure 7a-b

shows the colorimetric mapping of ΔS on the tertiary structure of AK (1KI9).⁷² The colour of the sphere in the figure represents the sign of ΔS (blue for negative ΔS and red for positive ΔS). The size of the spheres indicates the magnitude of ΔS . As seen in the figures, the large blue spheres (or the residues with large negative ΔS) are usually in close proximity to each other, especially in the N and C terminal regions. These results are encouraging, since they indicate that our approach may provide a simple, straightforward means to identify the residues involved in thermal stabilization.

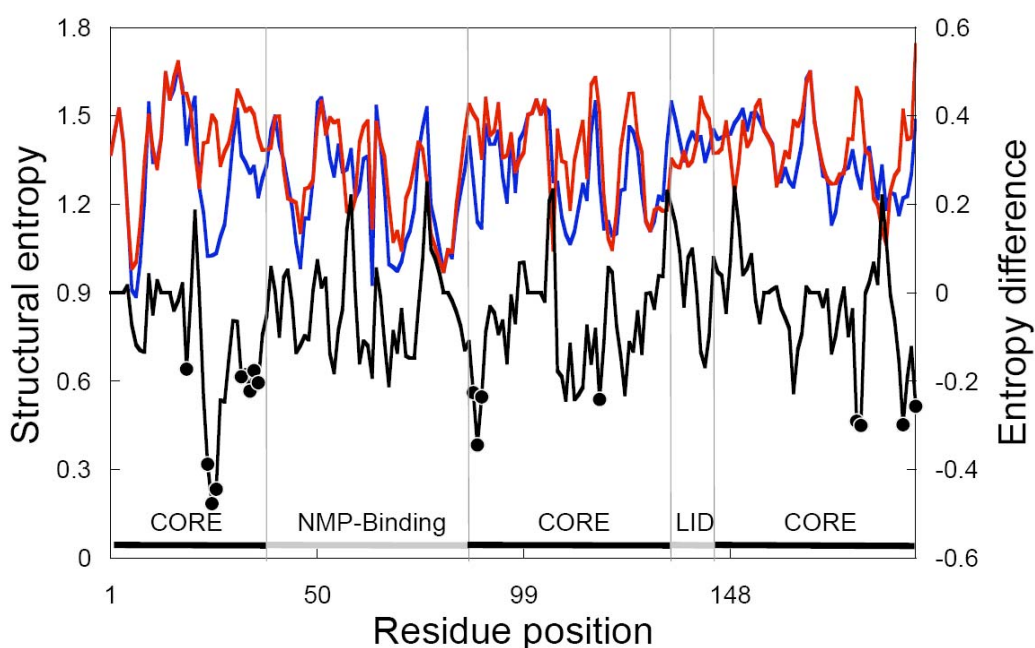


Figure 6. The Structural entropy profiles of AKjan (*SAKjan*, blue line), AKvol (*SAKvol*, red line) and their entropy difference ΔS (black line). Filled circles are the residues related to thermostabilization.^{72,84} The domains of AK are indicated by the lines above the x-axis.

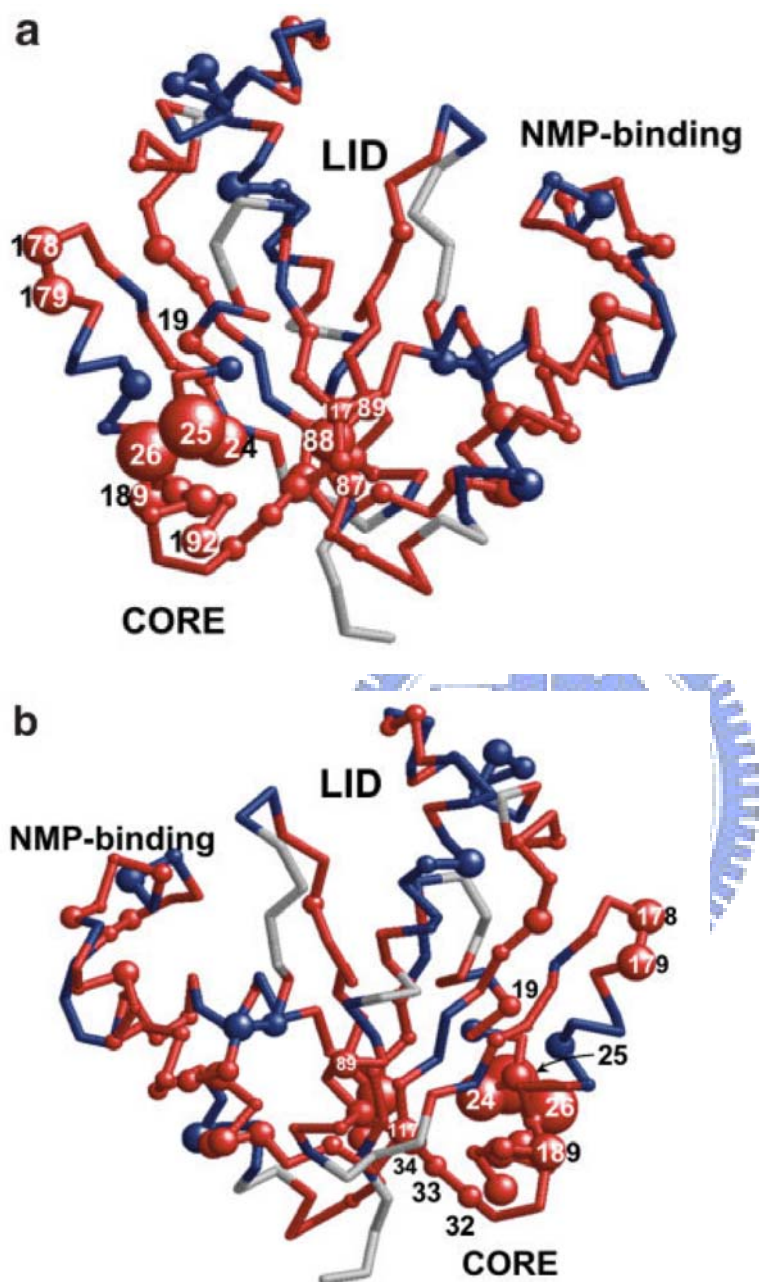


Figure 7. The colorimetric mapping of ΔS between AKjan and AKvol on the tertiary structure of the methanococcal AK (1KI9)⁸⁴. The color and size of the sphere represent the sign (red for negative and blue for positive) and the magnitude of ΔS , respectively. Two views are shown (A and B), the latter is rotated by 180° from the first. The figures were produced by RASMOL⁸⁶.

4.1.2 Case 2: Ribonuclease HI

Kimura *et al.*⁸⁵ have constructed a variety of chimeric proteins of *Escherichia coli* Ribonuclease HI (EI RNase HI) by substituting the corresponding R1-R9 regions from *Thermus thermophilus* Rnase HI (TH RNase HI), an exceptionally thermal stable protein. Both enzymes share a 52% sequence identity. Experimental results⁸⁷ show that the replacement of four regions (R4-R7) increases the melting temperatures from 52.0 °C to 68.7 °C. Figure 8 compares the structural entropy profiles of EI RNase HI and the chimeric R4-R7 protein. The structural entropy profile of the chimeric protein also shows very large entropy reduction in the R5, R6 and R7 regions of the four substitution regions.

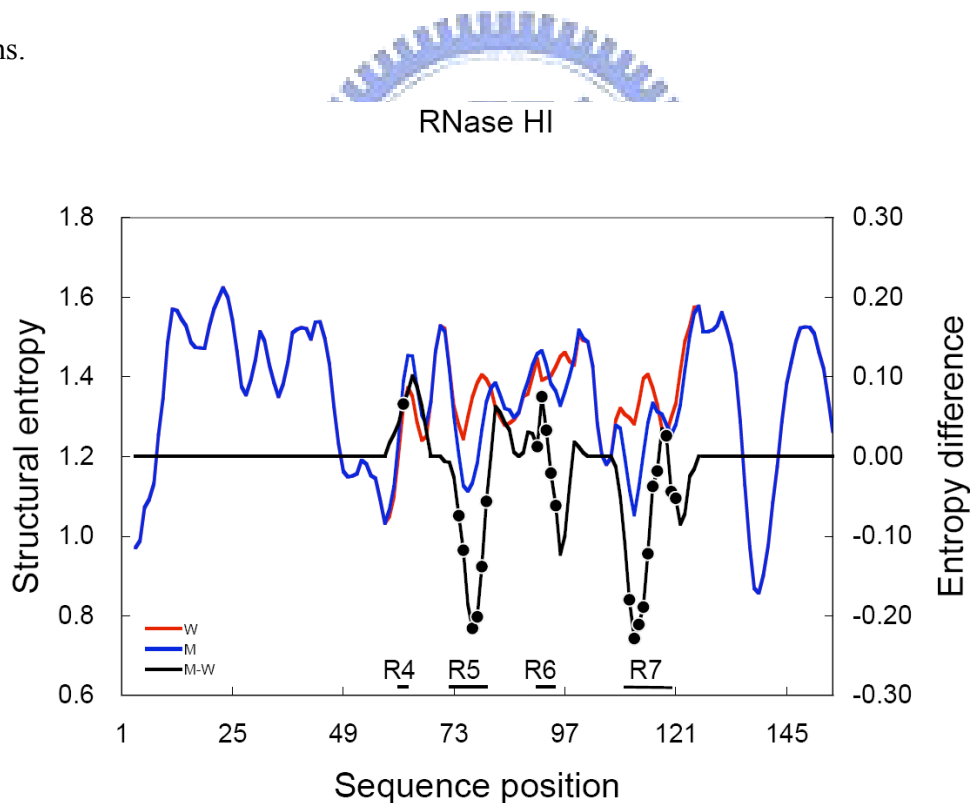


Figure 8. The Structural entropy profiles of RNase HI (red line), the R₄-R₇ mutant proteins (blue line) and their entropy difference ΔS (black line). The residues of R₄, R₅, R₆ and R₇ regions are shown in filled circles⁸⁵.

4.1.3 Case 3: Holocytochrome c551

Hasegawa *et al.*¹⁴ have systematically substituted the amino acids of *Pseudomonas aeruginosa* cytochrome c551 (PA c551) based on the structure of thermophilic *Hydrogenobacter thermophilus* cytochrome c552 (HT c552). Using this approach, they succeeded in constructing several single and multiple amino acid mutants of increased thermostability compared with that of PA c551. Figure 9 compares the structural entropy profiles of PA c551 and the mutant proteins (F7A, V13M, F34Y, E43Y and V78I). As shown in the figure, two mutations F7A and F34Y show the largest entropy reduction, which are consistent with the experiment that these two mutations result in the largest ΔT_m of all single amino acid mutants. Structural analysis shows that the FA7 mutation results in tighter hydrophobic packing and the F34Y mutation forms a new hydrogen bond between the hydroxyl group of the tyrosine residue and the guanidyl base of R47.

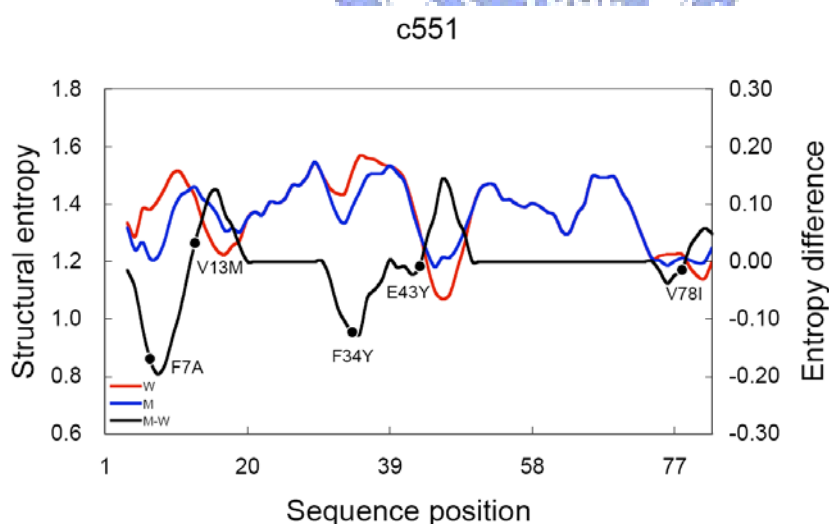


Figure 9. The Structural entropy profiles of PA c551 (red line) and its mutant proteins (blue line), and their entropy difference ΔS (black line). The mutated residues are indicated by the filled circles.¹⁴

4.1.4 The Relationship Between Structural Entropy and Thermal Stability

Haney *et al.*⁷² constructed a number of chimeric proteins of varying melting temperatures from AKjan and AKvol. These sequences share 68 % to 81% sequence identity and their melting points range from 69 °C to 103 °C (Table 3). Figure 10a shows the plot of ΔT_m versus α for these sequences, where ΔT_m is the difference of the melting temperatures between the particular sequence and the reference sequence (AKvol), and α is their difference of the average structural entropies per amino acid. We observed an excellent linear relationship between them. Note that in the figure the slope of the line is negative, indicating that lower structural entropy is related to higher thermostability. Figure 10b compares the observed melting temperatures with those computed from the linear model.

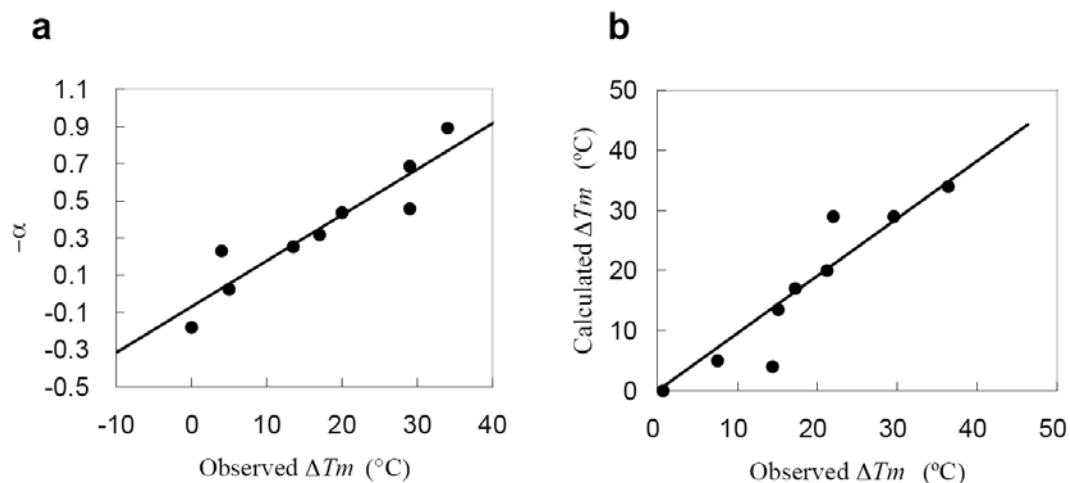


Figure 10. (a) The $\Delta T_m - \alpha$ plots for AKjan, AKvol and their chimeric proteins. ΔT_m is in °C and α in arbitrary unit. The correlation between ΔT_m and α is $r = 0.909$. (b) Comparison of the calculated melting temperature computed from the linear model with the observed melting temperature.

Table 3. The melting temperatures of AKs and their chimeric constructs.

Proteins ^a	T_m (°C)
AKvol	69.0
J36V	73.0
V160V	74.0
JVJ	89.0
V36J	98.0
J160V	96.0
VJV	82.5
AKjan	103.0

^a The melting temperatures of AKvol, AKjan and their chimeric constructs.⁸³ A 36 residue N-terminal residue region (1-36) or a 32-residue C-terminal region (161-192) was swapped to produce chimeric proteins. The notation J36V represents AKjan sequence through residue 36 followed by the remaining AKvol sequence. For the double chimera like JVJ, it represents AKjan through residue 36, AKvol through 160, and AKjan residues 161-192. The similar logic applies for the nomenclature of the other chimeras.

If the entropy linear model is a general one, the structural entropy will provide a useful measure for the thermal stability. To check this, we compiled a comprehensive dataset comprising 1,153 protein sequences with varying melting temperatures. These sequences contain the following families include adenylate kinases,⁷² cytochrome c551,¹⁴ RNase HI,⁸⁸ staphylococcal nuclease⁸⁹, alpha-amylase,⁸⁷ arc repressor,⁹⁰ rubredoxin variant (PFRD-XC4),⁹¹ and human fibroblast growth factor 1⁹², ligase⁹³, glutamate dehydrogenase⁴³, alcohol dehydrogenase⁹⁴, histone-like bacterial DNA-binding protein⁹⁵, Fyn SH3 domain⁹⁶, cold-shock protein *Bs-CspB*^{97,98}, malate dehydrogenase⁹⁹, cytochrome P450¹⁰⁰, WW domain¹⁰¹, bovine pancreatic trypsin inhibitor^{102,103} and phytase¹⁰⁴ and other families from the ProTherm database⁴⁶. Each family contains highly homologous sequences: the wild-type proteins and its mutants (either single/multiple point mutations or chimeric constructs). These sequences are listed in the supplementary material. For each family, we computed the linear regression of ΔT_m on α . From this linear model, we computed their melting temperatures. Figure 11 compares the calculated and observed melting temperatures of the sequences of the dataset. The linear regression correlation coefficients between the calculated and observed melting temperatures are: $r = 0.7209$ and the p -value = 0.143×10^{-3} .

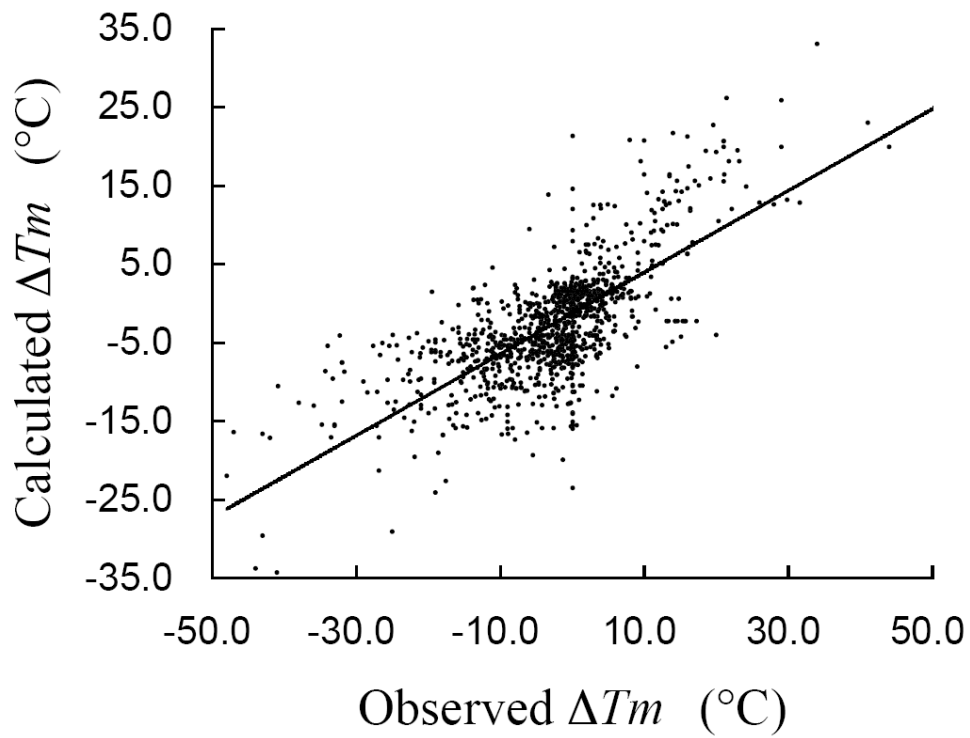


Figure 11. The calculated ΔT_m versus the observed ΔT_m for 1153 protein sequences. The calculated ΔT_m is computed from the linear regression of ΔT_m on α for each family. The linear regression correlation coefficients are: $r = 0.7209$ and the p -value = 0.143×10^{-3} .

On close examination of the results, we found that, for the sequences displaying the best linear relationship between α and ΔT_m , the mutated residues are usually resulted in more hydrophobic packing^{72,89} or conformational rigidity.^{90,91,93,94} On the other hand, if the mutated residues are involving in electrostatic interactions, some case like rubredoxin⁹¹ still shows relatively good linear relationship. Experiment⁹¹ shows that the thermostabilization of the mutant rubredoxin comes from a surface salt bridge involving the protein's backbone, which reduces the entropic cost. But other case like the cold shock protein (*Bs-CspB*)^{97,98} shows little correlation between α and ΔT_m . The increased thermal stability of the mutant *Bs-CspB* is due to electrostatics network arising from the mutated surface residues. The linear entropy model computed from sequences obviously cannot account for the long-range stabilization from such intricate structural features. We noticed that the linear entropy model may also not be applicable to some polymeric proteins like malate dehydrogenase⁹⁹, whose stabilization comes from ionic interactions across the dimer-dimer interface.

Though various interactions enhancing protein thermostability exhibit themselves as different structural features, our results show that the local structural entropy may be used as a generalized measure of the thermal stability. Since structure conservation reflects the effects of both the intrinsically stable (context-independent) sequence patterns and the long-range generic contributions (context-dependent) from surrounding residues,¹⁰⁵ the structural entropy provides a convenient structural measure of the thermal stability. Though the structure entropy profile by itself could be related to functional factors as well as structural factors, the structural entropy *differences* between mesophilic and thermophilic homologues will augment structural features involved in structural stabiliza-

tion. Our approach offers a straightforward way to compute the structural entropy directly from the query sequence and may be used as a useful tool to screen mutant candidates for thermophilic sequences in a high throughput way.

4.2 Amino Acid Coupling Patterns in Thermophilic Proteins

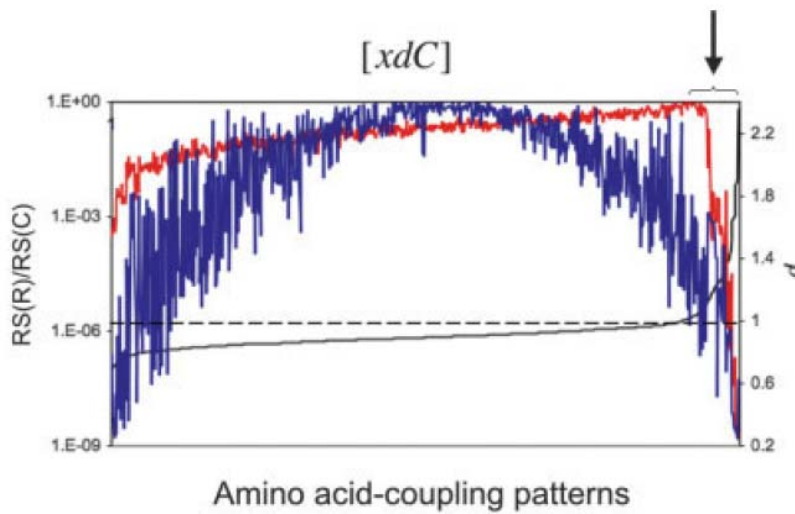
We present some of temperature significant coupling patterns in detail, and the performance in differentiating between thermophilic proteins and their mesophilic orthologs.

4.2.1 The ρ Profiles of Amino Acid Coupling Patterns

Using Eq. 6, we are able to construct the ρ profile of the amino acid-coupling patterns. The ρ profile is useful in providing a global picture of the relative occurrences of the coupling pattern in thermophiles compared with mesophiles. An example of the ρ profile for $[xdC]$ is shown in Figure 12A, which shows the ρ values together with $RS(R)$ and $RS(C)$. Most $[xdC]$ s have $\rho < 1$ and, hence, appear to be decreased in thermophiles. These results are consistent with previous reports^{3,10,58} that the Cys composition is in generally decreased in thermophiles. However, we note that there exist some statistical significant patterns with $\rho > 1.4$ (indicated by the arrow in Figure 12A). We zoom in this region in Figure 12B. These patterns are mostly of the form $[CdC]$, some instances of which are $[C3C]$, $[C4C]$ and $[C7C]$. Rosato *et al.*²⁴ previously reported that that cysteine clustering is closely related to the growth temperature of the organism. Structural analysis²⁴ showed that the increased stability of the cysteine clusters is probably due to their involvement in coordination of metal ions such as zinc, iron or FeS groups, or in disulfide bonds. This example shows that our approach is able to

identify and provide a detailed description of sequence features in thermophilic proteins than the conventional composition analysis. In the following sections, we will discuss the ρ profiles of coupling patterns of the general coupling pattern $[xdZ]$, where Z denotes the particular amino acid type and x is any amino acid type.

A



B

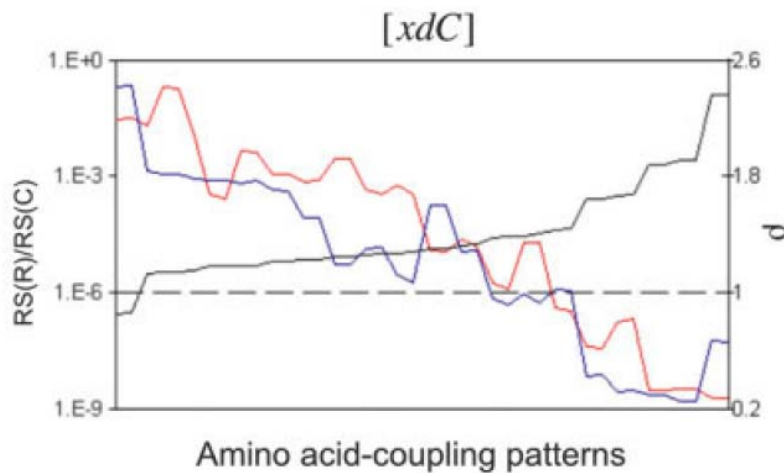


Figure 12. (A) The ρ , $RS(R)$ and $RS(C)$ profiles of the amino acid-coupling pattern $[xdC]$. The ρ values are plotted in black (scale on the right), and the $RS(R)$ and $RS(C)$ values in red and blue, respectively (logarithmic scale on the left). The abscis-

sas are the amino acid coupling patterns $[xdC]$ sorted according to ascending ρ values. The dotted line indicates the threshold $\rho = 1$. The arrow indicates the region of the statistical significant patterns with $\rho > 1.4$. (B) The zoom-in view of this region.

4.2.2 The ρ Profiles of $[xdZ]$

Figure 13A-S show the ρ profiles of the coupling patterns $[xdZ]$. For clarity, we plot the $RS(R)$, but not $RS(C)$ values of each pattern.

4.2.3 $[xdE]$ and $[xdV]$

Figure 13A shows the ρ profile of $[xdE]$. The larger than unit ρ values clearly indicate that $[xdE]$ occurs more in thermophiles. There is little surprise, since previous reports^{10,58} indicated that Glu content is usually higher in the thermophilic proteins. Specifically, the most statistically significant ($RS(C)$ and $RS(R) < 10^{-5}$) patterns are $[KdE]$, $[RdE]$, $[EdE]$ and $[DdE]$. The structural implications of these coupling patterns are clear: these patterns frequently occur in the helical conformations, and, especially, the first two patterns can easily form salt bridges within themselves. It is reported¹⁰ that both local salt bridges and helical conformations are significantly increased in thermophilic proteins, and that the proportions of charged pairs of RK, DE and EE tend to be higher in thermophiles. Note that we derived similar results directly from sequences without the use of structural information.

The ρ profile of $[xdV]$ (Figure 13B) is similar to that of $[xdE]$, though the nonpolar valine and the charged glutamate are completely different types of amino acids. These are the coupling patterns $[DdV]$, $[KdV]$, $[NdV]$ and $[YdV]$ that are significantly increased in thermophiles. The structural implications of these patterns are not clear, though these patterns frequently occur in α -helices or β -sheets, and higher proportion of secondary structures is known to be an important contributor to increased thermal stability.¹⁰

4.2.4 $[xdP]$ and $[xdC]$

The ρ profile of $[xdP]$ is shown in Figure 13C, which is similar to that of $[xdC]$ (Figure 12). Most instances of $[xdP]$ are increased in thermophiles ($\rho > 1$), though with relatively high p -values. It is reported¹⁰ that the Pro composition is increased in thermophilic proteins. There exist a few statistical significant patterns with $\rho > 1.4$ (indicated by the arrow in the figure), which are $[CdP]$ (see previous section) and $[PdP]$. We found from structural analysis that $[PdP]$ s (or proline clusters) are often involved in the formation of polyproline II helix. The helical conformation together with the reduced conformational entropy may contribute to protein stability.

4.2.5 $[xdQ]$, $[xdT]$ and $[xdH]$

The coupling patterns involving polar amino acids are usually decreased in thermophiles. It is reported^{3,26,58} that the Gln composition as well as other polar amino acids like Ser,

Gln, Asn, Thr and Cys are decreased in thermophiles. A typical case $[xdQ]$ is shown in Figure 13D. Specifically, the coupling patterns with p -values $< 10^{-6}$ are $[EdQ]$, $[GdQ]$, $[RdQ]$ and $[QdQ]$. The homo-amino acid coupling pair $[QdQ]$ presents a special case in sequence-coupling patterns. Figure 14. shows the homo-amino acid coupling patterns for twenty amino acid types. Only $[EdE]$, $[CdC]$ and $[PdP]$ show statistically significant instances that are increased in thermophilic proteins (see also the previous sections).

Most instances of $[xdT]$ have $\rho < 1$ (Figure 13E). We notice that the particular pattern [(charged residue) dT] is significantly decreased in thermophiles. For example, $[E1T]$ has $\rho = 0.68$, $RC(R) = 3 \times 10^{-8}$ and $RC(C) = 1 \times 10^{-5}$. Though Glu is usually increased in thermophile, the coupling pattern $[E1T]$ is in fact decreased in thermophiles. The ρ profile of $[xdH]$ (Figure 13F) shows similar shape to that of $[xdT]$. Interestingly, we also observe that [(charged residue) dH] is also significantly decreased in thermophilic proteins.

4.2.6 Other Coupling Patterns

$[xdL]$ (Figure 13G) does not show any significant bias toward thermophiles. However, a particular instance $[CdL]$ is decreased in thermophiles with statistical significance. Other patterns like $[xdM]$, $[xdF]$, $[xdW]$ and $[xdG]$ also show similar neutral ρ profiles (Figure 13H-K). Figure 10M shows that $[xdI]$, unlike $[xdL]$, is increased in thermophiles. For the patterns involving aromatic amino acids, $[xdF]$ and $[xdW]$ are

increased in thermophilic proteins (Figure 13I-J), but [*xdY*] (Figure 13M) is increased. For patterns involving charged amino acids, [*xdE*], [*xdK*] and [*xdR*] (Figure 13A, 2N-O) are increased in thermophilic proteins, but, interestingly, [*xdD*] (Figure 13P) is decreased. For patterns involving polar amino acids, [*xdS*] and [*xdN*] (Figure 13Q-R) are in general decreased in thermophilic proteins. [*xdA*] (Figure 13S) is similar to that of [*xdN*] and is decreased in thermophiles, despite that alanine and asparagine are two different types of amino acids.



Figure 13A.

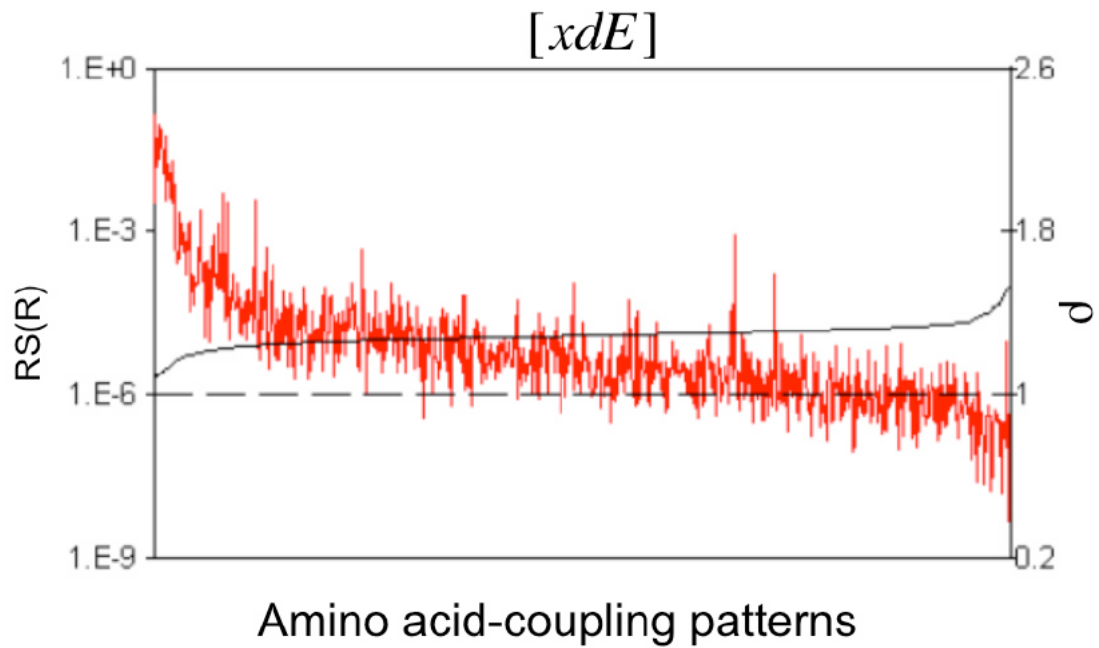


Figure 13B.

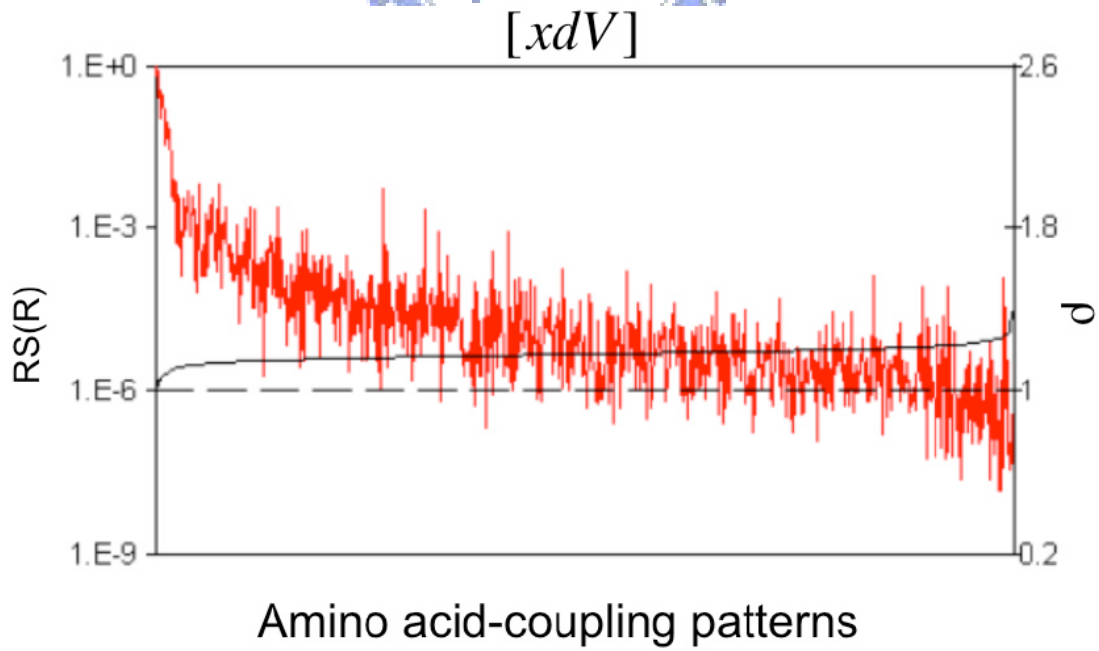


Figure 13C.

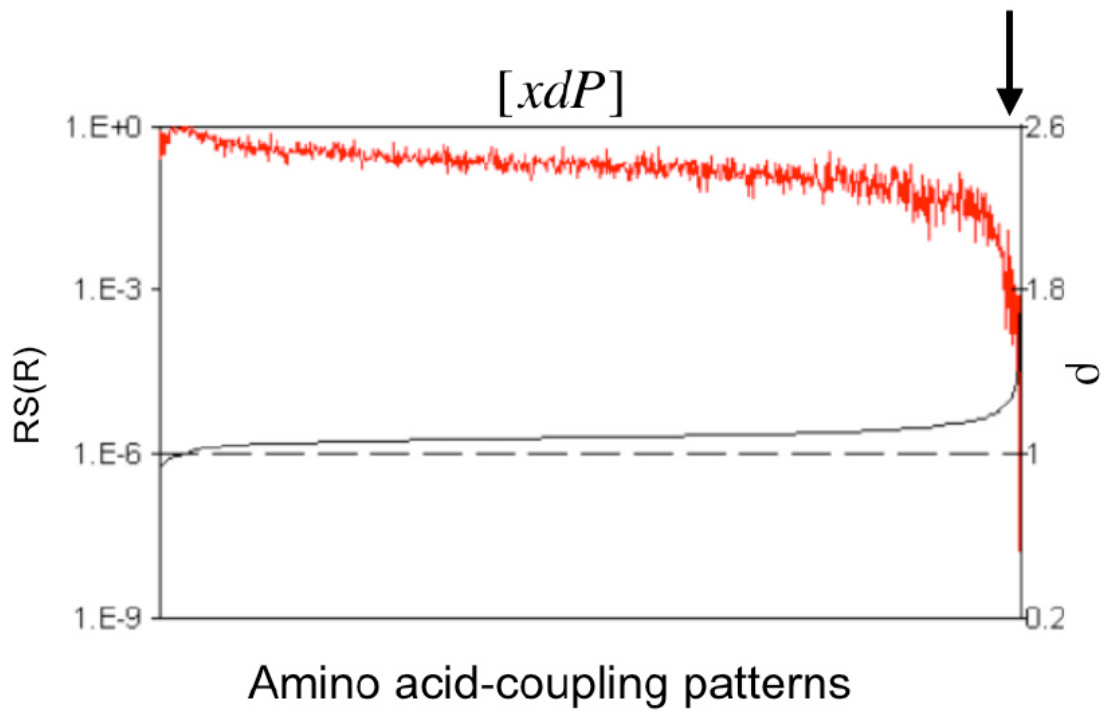


Figure 13D.

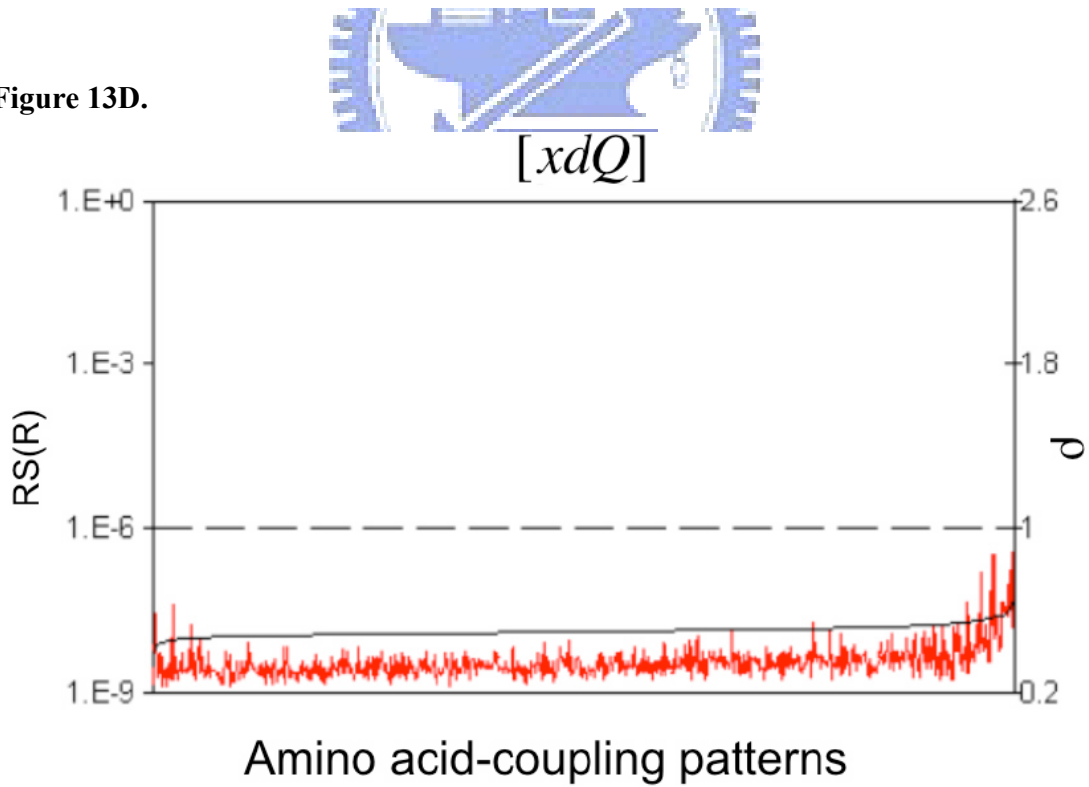


Figure 13E.

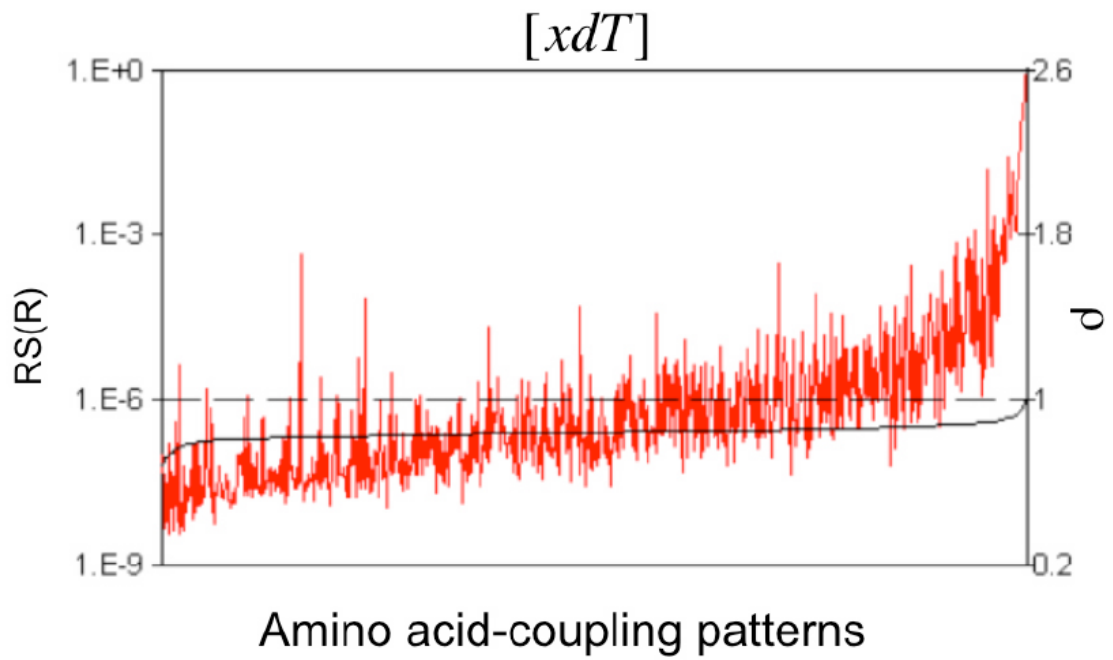


Figure 13F.

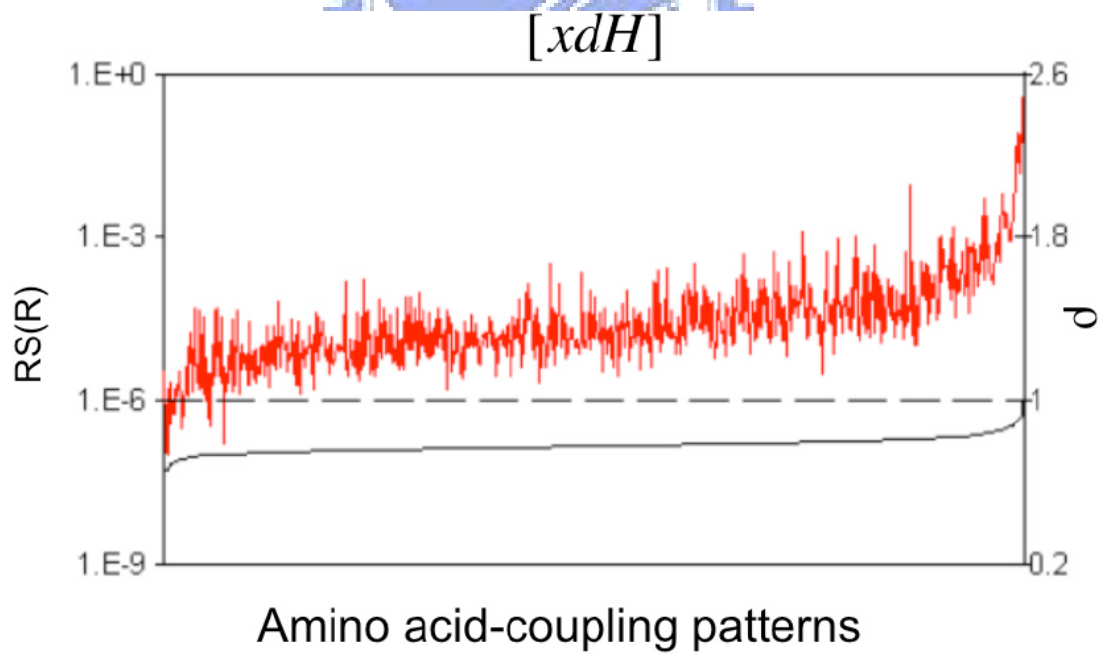


Figure 13G.

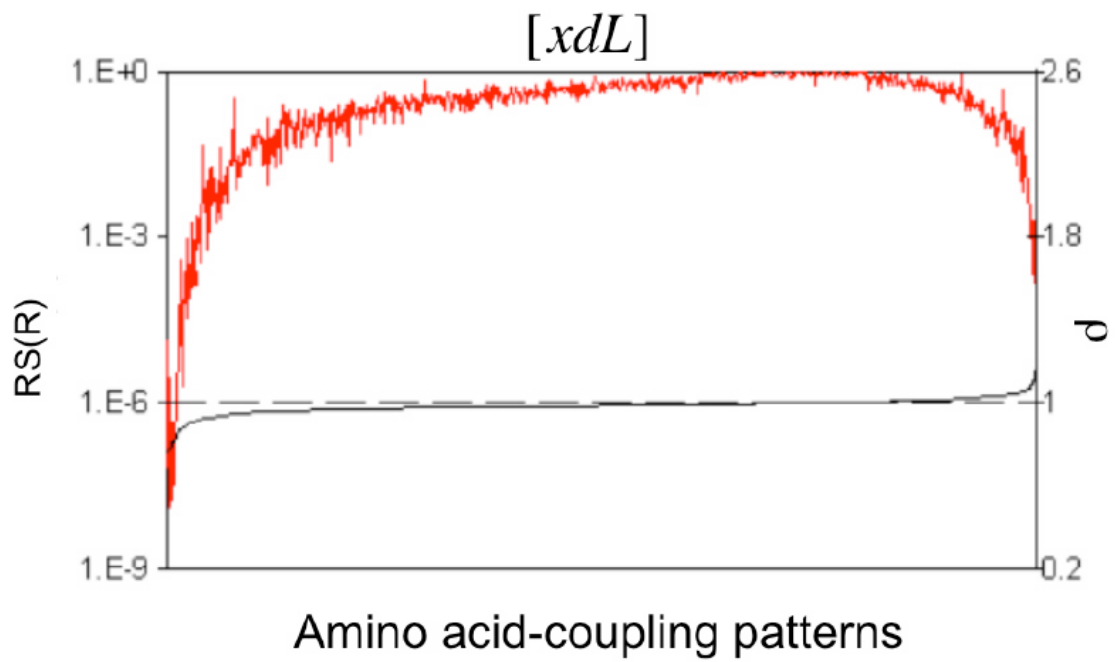


Figure 13H.

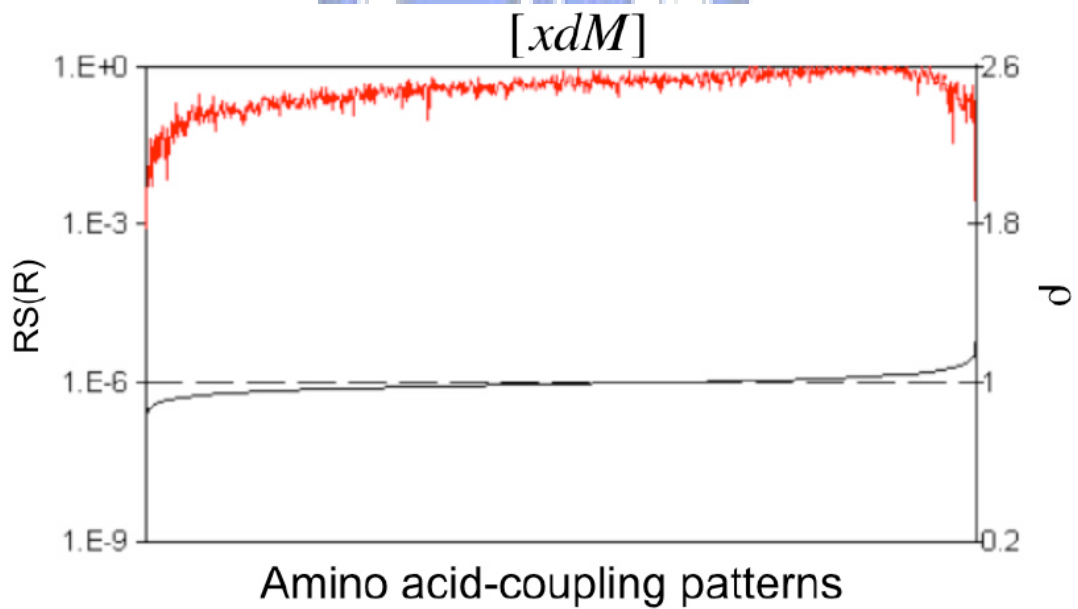


Figure 13I.

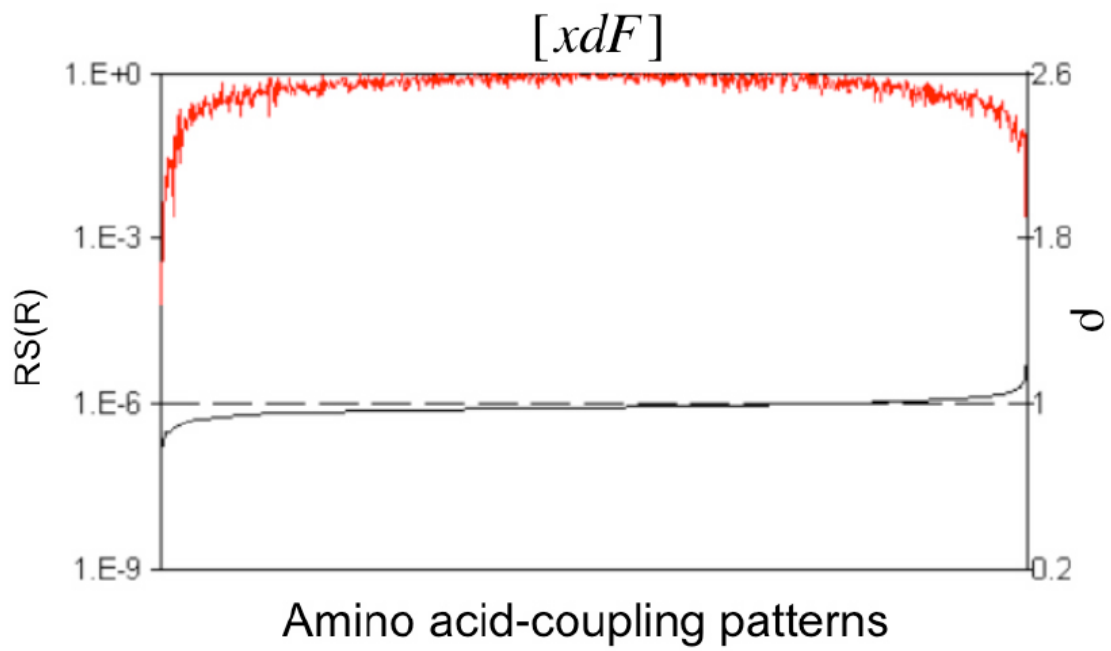


Figure 13J.

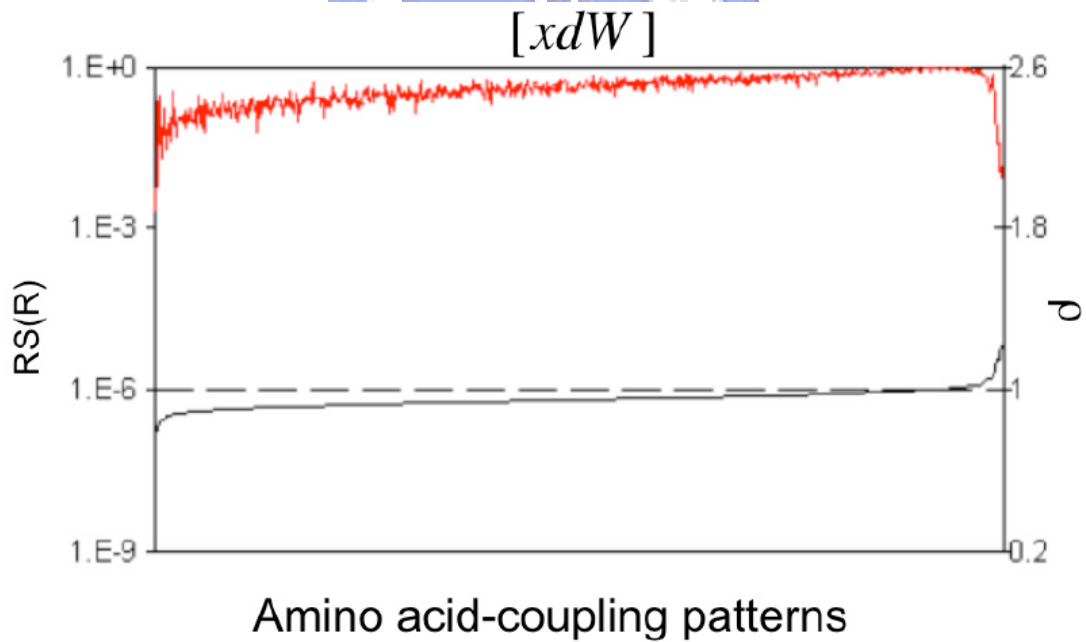


Figure 13K.

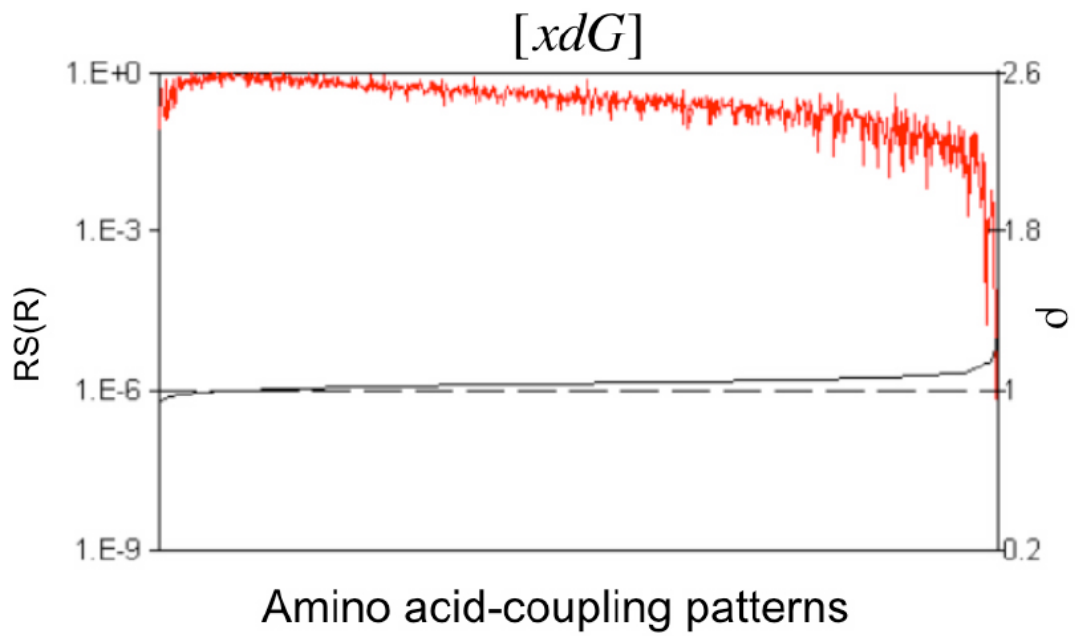


Figure 13L.

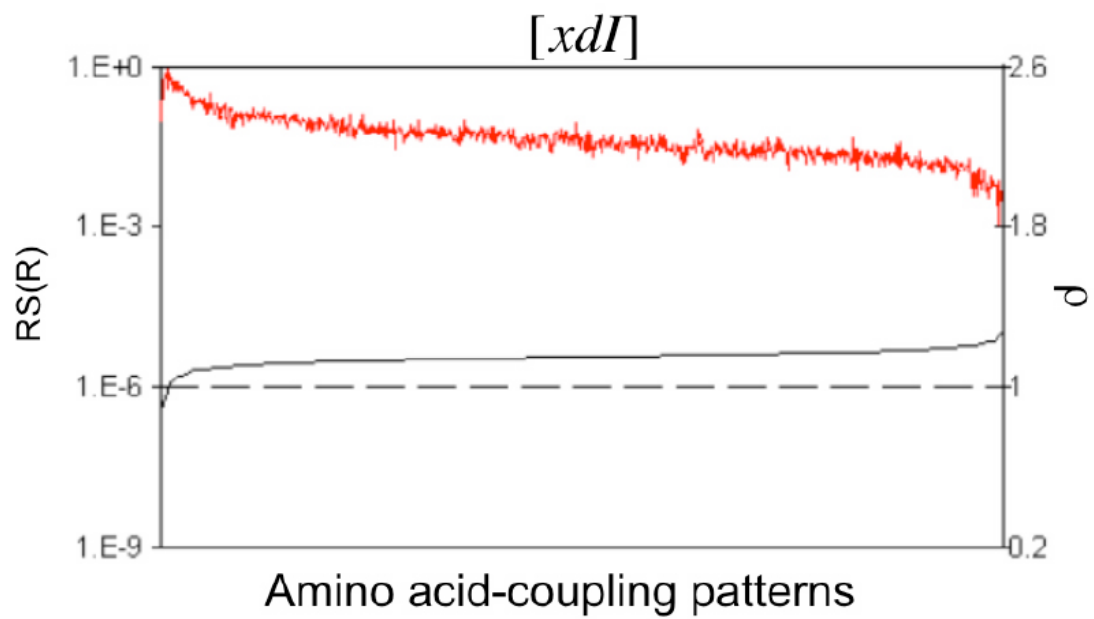


Figure 13M.

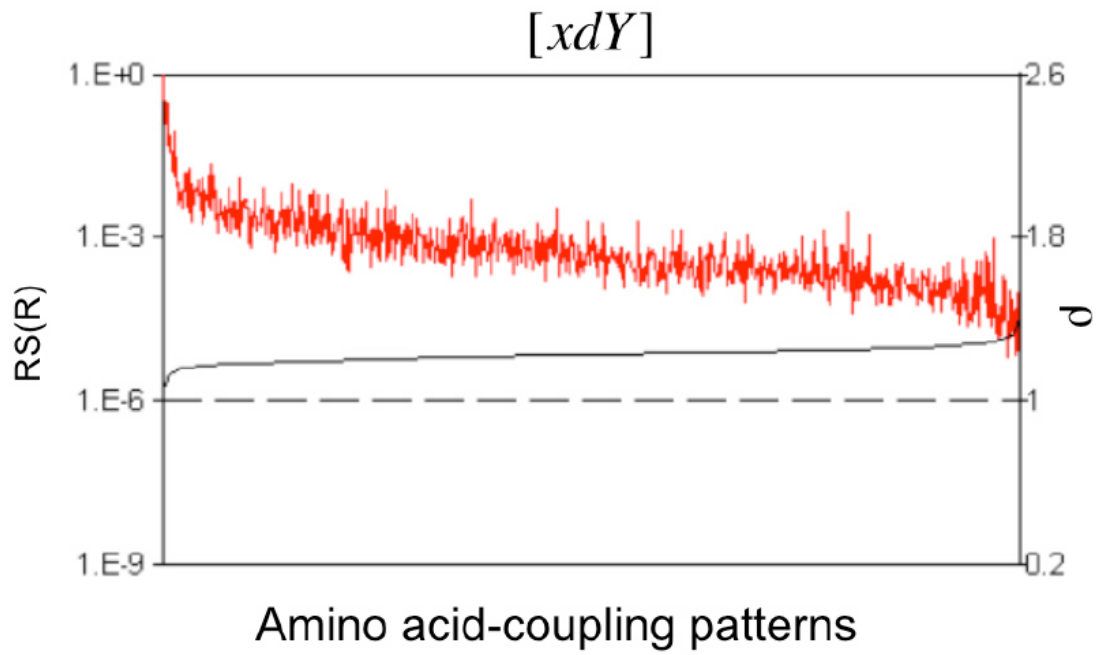


Figure 13N.

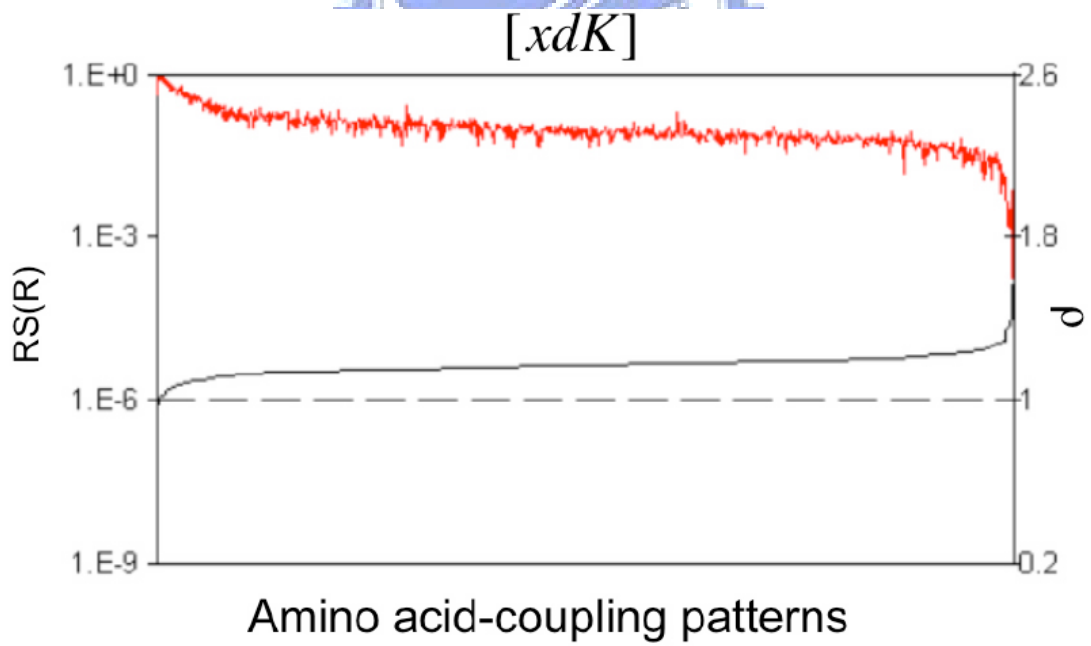


Figure 13O.

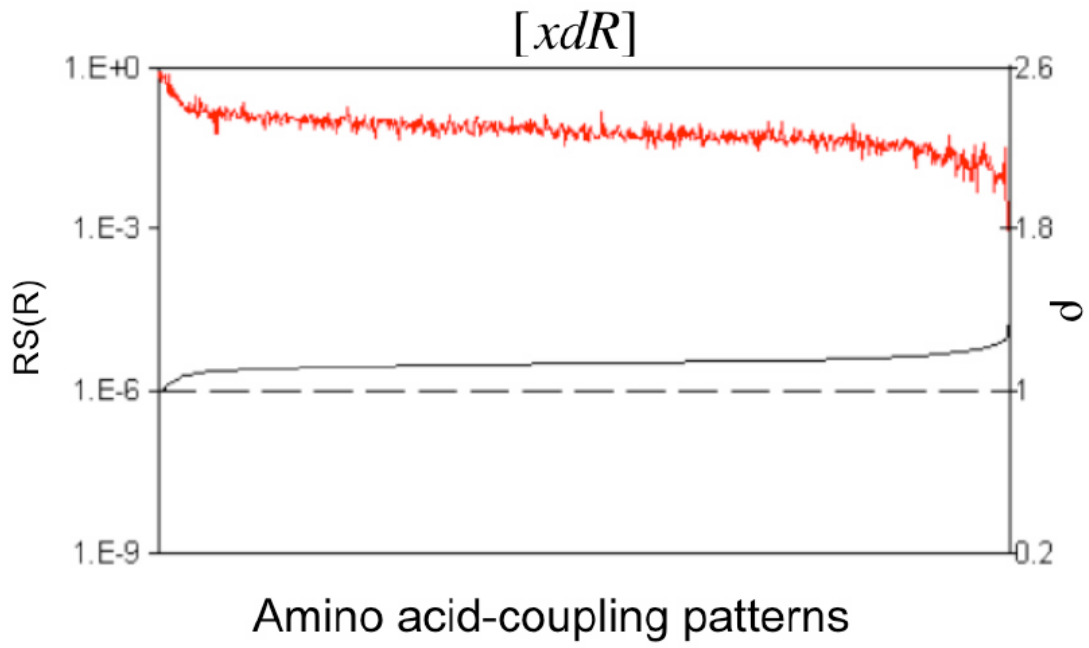


Figure 13P.

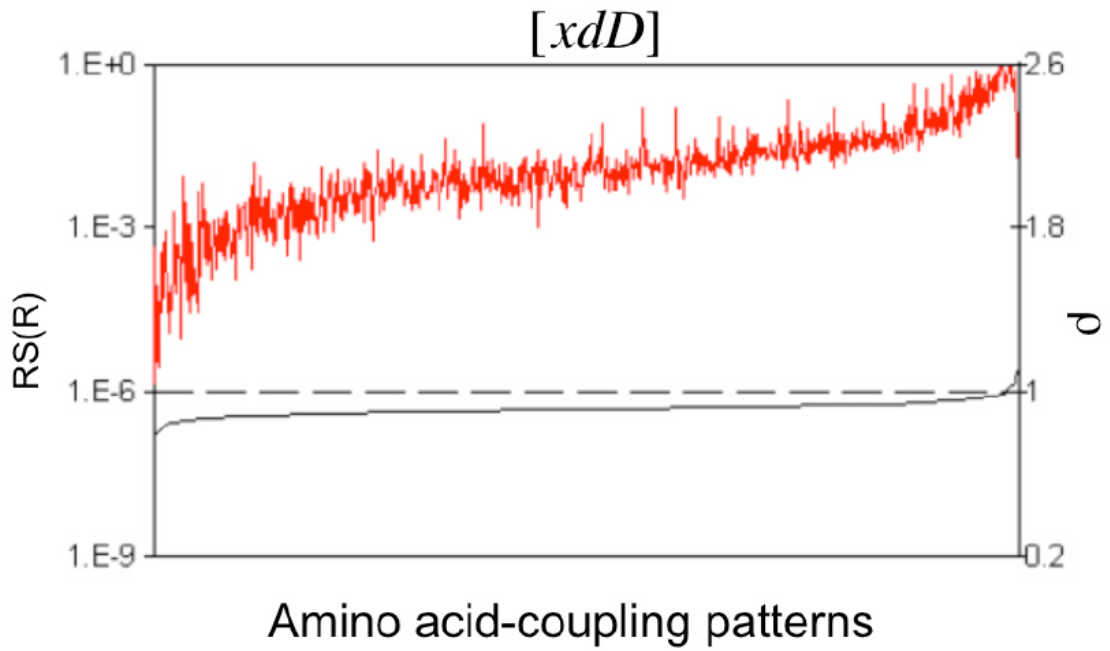


Figure 13Q.

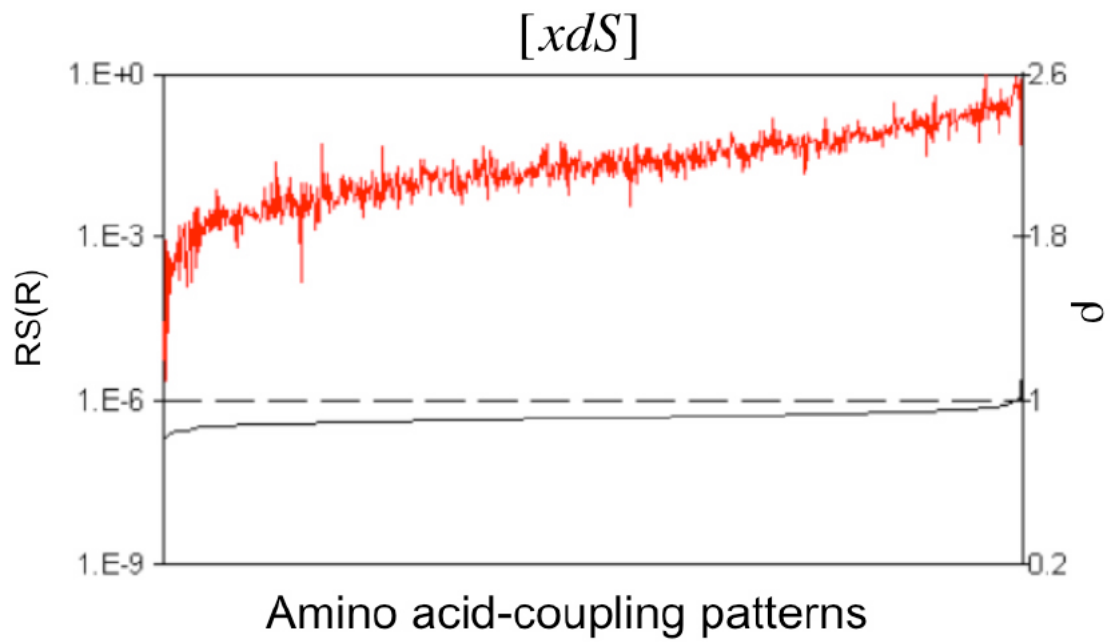


Figure 13R.

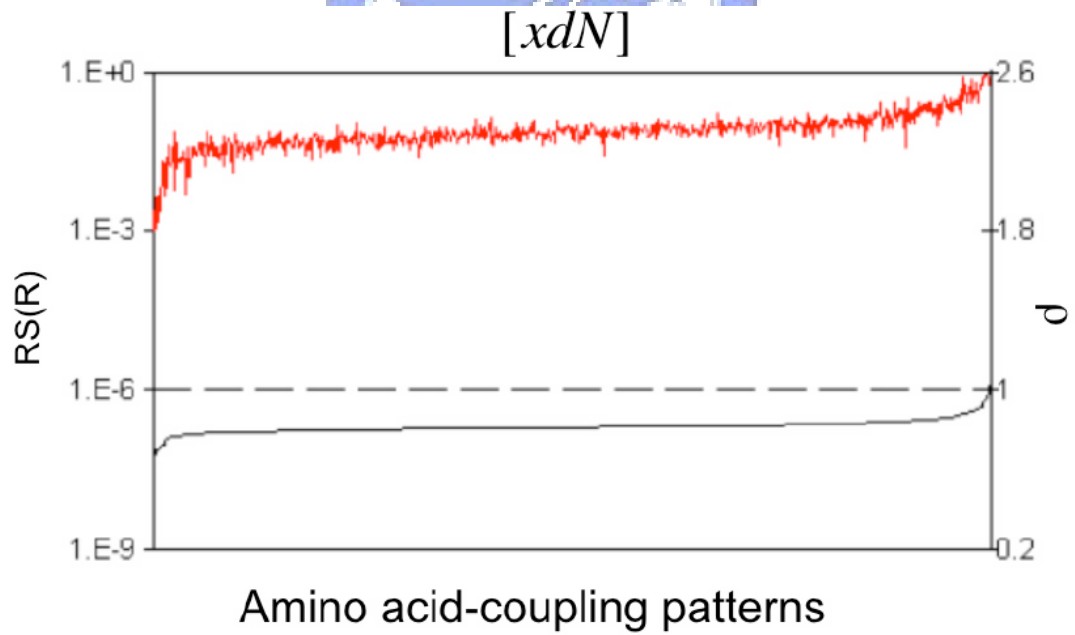


Figure 13S.

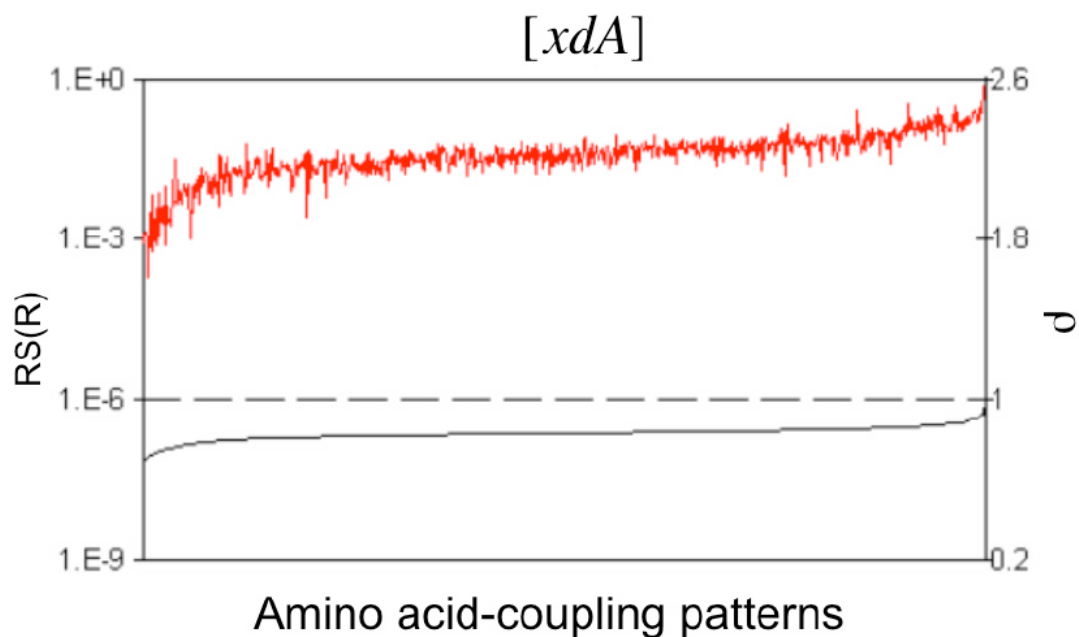


Figure 13. The ρ profiles of the amino acid-coupling patterns $[xdZ]$. For clarity, we plot the $RS(R)$, but not $RS(C)$ values of each pattern. The ρ values are plotted in black (scale on the right) and the $RS(R)$ values in red, respectively (logarithmic scale on the left). The abscissas are the amino acid coupling patterns $[xdZ]$ sorted according to their ascending ρ values. The following amino acid-coupling patterns are shown: (A)[xdE], (B)[xdV], (C)[xdP], (D) [xdQ], (E) [xdT], (F) [xdH], (G) [xdL], (H) [xdM], (I) [xdF], (J) [xdW], (K) [xdG], (L) [xdI], (M) [xdY], (N) [xdK], (O) [xdR], (P) [xdD], (Q) [xdS], (R) [xdN] and (S) [xdA].

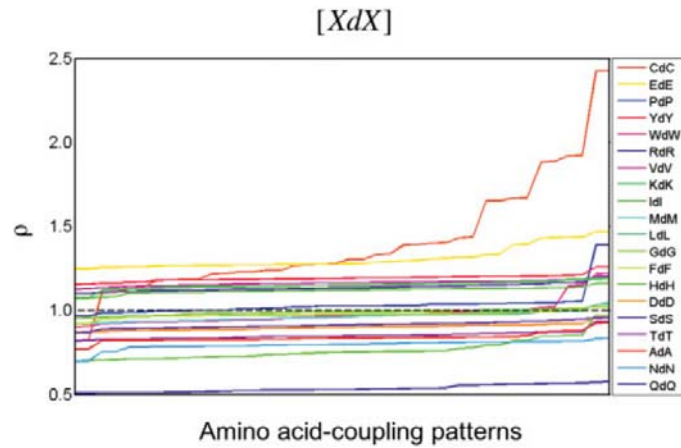


Figure 14. The ρ profiles of 20 homo-amino acid coupling patterns $[XdX]$.

4.2.7 The Significant Amino Acid Coupling Patterns

The net thermal stability of proteins usually results from a multitude of different coupling patterns, and no single outstanding sequence or structural feature can adequately account for thermophilic proteins. We identify from the amino acid coupling pattern the most significant ones with P -values $< 10^{-7}$ for both $RS(R)$ and $RS(C)$. We denote this set by Ω , which contains the following thermophilic amino acid coupling patterns: $[C(-2)P]$, $[C1P]$, $[C3C]$, $[C4C]$, $[C6C]$, $[C7C]$, $[K(-7)E]$, $[K(-4)E]$, $[K3E]$, $[K4E]$ and $[H(-4)V]$, and the following mesophilic amino acid coupling patterns: $[C(-4)L]$, $[C(-3)L]$, $[C(-2)L]$, $[C2L]$, $[C3L]$, $[D(-5)T]$, $[D(-4)T]$, $[E(-8)T]$, $[E(-4)T]$, $[E1Q]$, $[E3T]$, $[E4T]$, $[G(-3)Q]$, $[K(-4)T]$, $[K2T]$ and $[K3T]$.

4.2.8 Identification of Thermophilic and Mesophilic Proteins Using the Amino Acid Coupling Patterns

Most sequenced thermophilic genomes are archaea (as also reflected in our thermophile

data set – 12 archaea and 3 bacteria), and it is possible that some of amino acid coupling patterns between thermophilic and mesophilic proteins may be due to phylogenetic differences instead of temperature adaptation. We compute C_{Ω} for the set Ω for both bacteria and archaea. Figure 15 shows the C_{Ω} -OGT plot for both archaea and bacteria genomes. The amino acid coupling patterns can clearly distinguish between thermophiles and mesophiles of both bacteria and archaea. The results show that we can identify the amino acid coupling patterns that are indeed due to temperature adaptation. Furthermore, we observe a good linear correlation between C_{Ω} and OGT (the correlation coefficient is 0.89). This is encouraging, since the linear relationship is obtained without adjustable parameters⁵⁷.



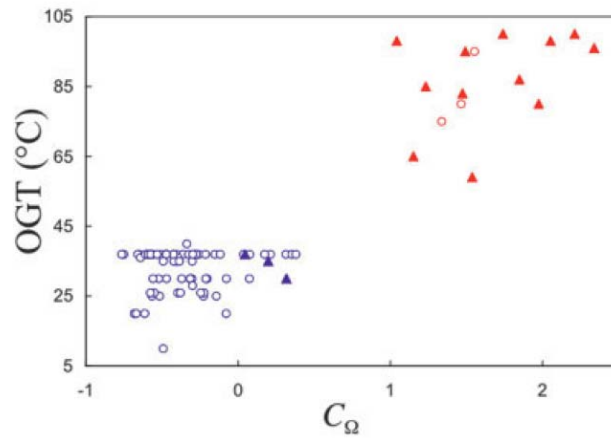
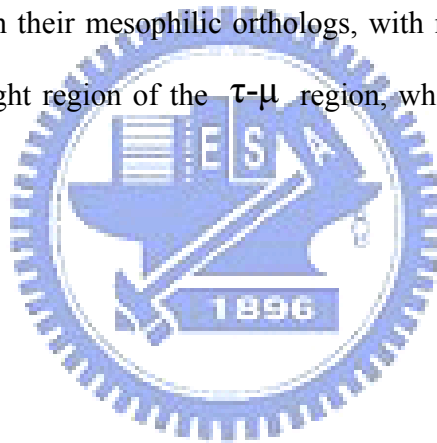


Figure 15. The C_{Ω} – OGT plot of the thermophiles and mesophiles. The circles represent bacterial genomes and the triangle the archaea genomes. Thermophiles are colored in red, and mesophiles in blue.



To distinguish thermophilic proteins and their mesophilic orthologs presents a much harder challenge, because these orthologs usually share higher degrees of sequence similarity. Define τ and μ as the occurrences of thermophilic and mesophilic amino acid patterns of the set Ω , respectively. We compute τ and μ for both thermophilic and mesophilic orthologs of the following COG families¹⁰⁶ – COG0003, COG0068, COG0121, COG0121, COG0156, COG0430 and COG1042. The eukaryotic sequences are excluded in calculation. The τ - μ plot of these COG families is shown in Figure 16, with each point (τ, μ) representing one ortholog. Thermophilic proteins are generally well separated from their mesophilic orthologs, with most thermophilic orthologs clustering in the lower right region of the τ - μ region, while the mesophilic orthologs the upper left regions.



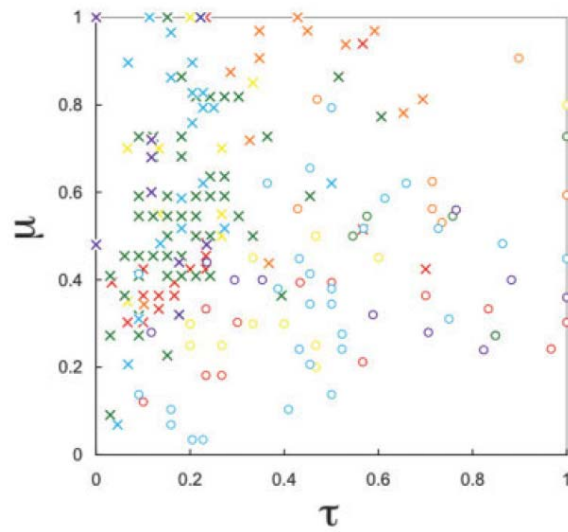


Figure 16. The τ - μ plot of the COG families: COG0003 (red), COG0068 (orange), COG0121 (yellow), COG0156 (green), COG0430 (violet) and COG1042 (turquoise). The thermophilic proteins are shown in circles and the mesophilic orthologs in cross. The occurrences of the thermophilic and mesophilic amino acid coupling patterns are normalized by dividing the maximal occurrences of the corresponding patterns of each COG family.

4.2.9 GC Contents and the Amino Acid Coupling Patterns

Though GC content is the dominant influence on the amino acid composition, it has been shown that GC pressure and thermophily are essentially independent of each other.⁵⁸ We compute C_{Ω} for both bacteria and archaea. Figure 17 compares C_{Ω} s and the corresponding GC contents of the genomes. While C_{Ω} clearly distinguishes between thermophiles and mesophiles, both thermophiles and mesophiles scatter over a range of similar the GC contents.

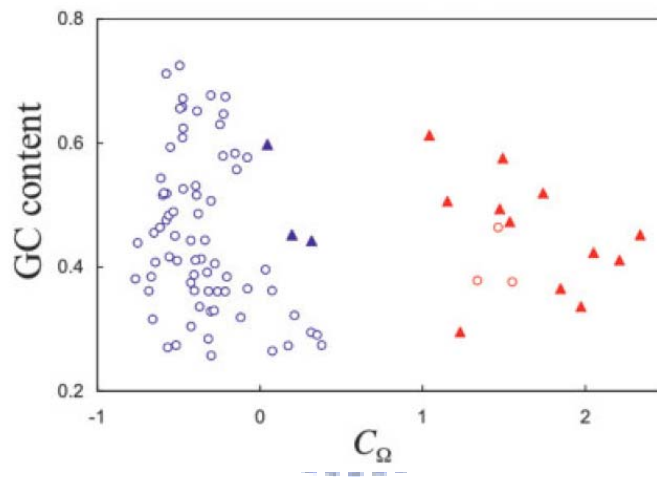
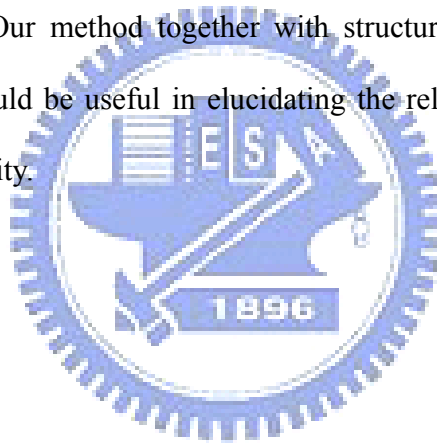


Figure 17. The C_{Ω} – GC content plot of the thermophiles and mesophiles.

4.3 Conclusion

We have developed a sequence-based approach to identify the amino acid-coupling patterns in thermophilic proteins. This approach is especially useful when no three-dimensional structures are available. The often-used composition analysis, ignoring the coupling effects of the nearby amino acids, presents only a simplified picture of amino acid features related to the thermal stability of proteins. Our approach provides a more detailed description of the relationship between coupling patterns and protein stability. Using this approach, we are able to identify statistically significant coupling patterns in thermophiles. Our method together with structural analysis and amino acid composition analysis should be useful in elucidating the relationship between sequence features and protein stability.



Chapter 5 Protein Mutation Stability Change Prediction Based on Sequence Information

In this chapter, we propose a prediction method using SVMs to predict the direction of the protein stability changes $\Delta\Delta G$ based on the sequence coupling patterns⁶². As previously described, we find that thermophiles and mesophiles have their own preferred sequence coupling patterns and that these sequence-coupling patterns are useful in distinguishing thermophilic and mesophilic sequences⁶². For example, the sequence-coupling pattern EXXK, where X represents any amino acid type, is a preferred sequence-coupling patterns (p -value $< 10^{-8}$) in the thermophilic microbial genomes. We will refer to those coupling pattern preferred by thermophiles as the thermo-dominant patterns, and those preferred by mesophiles as the meso-dominant patterns. Then features are merged into the coupling composition for SVMs prediction. Our result shows that the accuracy of prediction for stable mutation is significantly improved, although the data set is unbalanced.

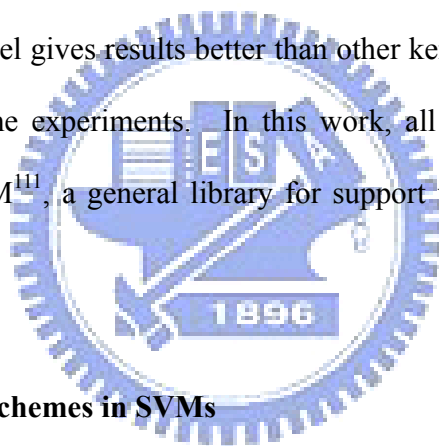
5.1 Methods and Implementation

5.1.1 Data Sets

The data set S2048 in Capriotti's works^{107,108} is used. S2048 includes 2048 single mutations obtained from 64 different proteins with PDB ID. The $\Delta\Delta G$ values of mutants have been experimentally detected and reported in ProTherm database⁴⁶. S2048 is available at <http://gpcr2.biocomp.unibo.it/~emidio/Imutant2.0/dbMutSeq.html>.

5.1.2 Support Vector Machines Predictor

The SVMs¹⁰⁹ try to find the separating hyper-plane with the largest distance between two classes, measured along a line perpendicular to this hyper-plane. However, data to be classified may not be linearly separable. To overcome this difficulty, SVM nonlinearly transforms the original input space into a higher dimensional feature space by the so-called kernel functions. When the training data are mapped into vectors in a higher dimensional space, it is possible that data can be linearly separated. In the training process, only part of the data are used to construct the hyper-plane, hence avoiding the over-fitting problem usually plaguing other machine learning methods. These data constructing the classifier are called support vectors. Preliminary tests show that the radial basis function (RBF) kernel gives results better than other kernels¹¹⁰. Therefore, we use the RBF kernel for all the experiments. In this work, all SVM calculations are performed by using LIBSVM¹¹¹, a general library for support vector classification and regression.



5.1.3 Sequence Coding Schemes in SVMs

The features considered in the SVMs for the mutants' sign of $\Delta\Delta G$ include amino acid composition, dipeptide composition and coupling composition of the wide-type and mutated protein sequences, as well as temperature and pH values, the experimental environment features. A vector for a mutant in the SVMs consists of the differences of compositions of the mutant and its wild-type sequence and the experimental environment features.

5.1.3.1 The Amino Acid Composition

The amino acid composition of a protein consists of 20 components representing the occurrence frequencies of the 20 native amino acids in it.

5.1.3.2 The Dipeptide Composition

The dipeptide composition of a protein consists of 400 components representing the occurrence frequencies of the 400 dipeptides in it.

5.1.3.3 The Coupling Composition

A coupling pattern with amino acid type X and Z separated by d amino acids in a sequence is denoted by $[XdZ]$ ⁶². Based on statistical analysis on a set of mesophilic and thermophilic genomes, identify a set of 734 thermo-dominant coupling patterns and a set of 961 meso-dominant coupling patterns are identified. Let X^T (X^M respectively) denote the thermo-dominant (meso-dominant, respectively) coupling patterns $[XdZ]$, for all distance d and all type of amino acid Z . The thermo-coupling (meso-coupling, respectively) composition of a protein consists of 20 components representing the occurrence frequencies of thermo-dominant (meso-dominant) coupling patterns of X^T (X^M respectively) for each native amino acid X in it. The thermo-coupling composition and the meso-coupling composition are merged into a 40 dimensions coupling composition for the SVMs.

5.1.3.4 Experimental Environment Features (Temperature, pH)

As in the work of Capriotti *et al.*^{107,108}, the experimental environment features tempera-

ture and pH value in the $\Delta\Delta G$ detection are also considered.

5.1.4 Scoring and Performance

An important issue of optimizing SVMs is the assignment of model parameters, such as the penalty parameters and the kernel parameters of the RBF function. We use a 20-fold cross-validation on different sets of parameters for the model selection^{107,108}. In the cross-validation, the proteins of the same PDB ID are assigned to the same fold for proper non-redundancy. The performance indices of the predictor is overall accuracy, $Q2$, the coverage of class s , $Q(s)$, and the precision of class s , $P(s)$, where s is + or – for the sign of $\Delta\Delta G$ and Matthews' correlation coefficient, MCC .

For a class s , let T^s and F^s be the number of correct and incorrect predictions for mutants of $\Delta\Delta G$ sign s respectively. Then the overall accuracy

$$Q2 = \frac{T^+ + T^-}{T^+ + T^- + F^+ + F^-} \times 100\%. \quad (17)$$

Matthews's correlation coefficient

$$MCC = [T^+ \times T^- - F^+ \times F^-] / D, \quad (18)$$

where $D = [(T^+ + F^+) \times (T^+ + F^-) \times (T^- + F^+) \times (T^- + F^-)]^{1/2}$ is a normalization factor.

The coverage for class s is $Q(s) = T^s / [T^s + F^{\sim s}]$, where $\sim s$ is + if s is – and vice versa.

The precision for class s is $P(s) = T^s / [T^s + F^s]$.

5.2 Results and Discussion

Because our prediction method is based only on sequence information, we compare with

the counterpart in I-Mutant2.0. On the dataset S2048, our results show better *MCC* and better accuracy in stable mutants [$Q(+)$] than Capriotti's work^{107,108}. As shown in Table 4, *MCC* of our method is higher than that of Capriotti's by 0.05, which means that the overall performance is significantly improved. The performance is almost equal to *MCC* of Capriotti's prediction method with structural method¹⁰⁸. In our approach, the best accuracy for stable mutants, $Q(+)$ is 0.54. It is 9% accuracy improvement to I-Mutant2.0, and means the stable mutant prediction in our method is significant improved.

We also predict the same data set (S2048) without temperature or pH values information in our work (In table 4). The performance of prediction method without the 2 features drops drastically, which implies that the features temperature or pH values affect seriously the performance of prediction. However, that result is not available in any other publication of related works. In real application, the optimal temperature and pH value for stable mutant prediction are not always available. It would be more practical to have a prediction method for protein mutation stability change without temperature and pH value.

We also built SVMs using the sequence features, amino acid composition (20 features) or dipeptide (400 features), and experimental environment features, temperature and pH (2 features). Based on the result shown in Table 4, amino acid composition is efficient information for predicting protein mutation stability change. We can find the result of "AAC, T, pH" is very similar to the result of "I-Mutant2.0-seq". It means the sequence-based features in I-Mutant2.0 just include the information of amino acid compo-

sition, temperature and pH. The best performance of these SVMs is with MCC about 0.44 and $Q2$ about 0.78. However, the accuracy of stable mutant prediction $Q(+)$ are both smaller than 0.5. It means the result without features of significant coupling patterns is similar to the result of I-Mutant2.0. Nevertheless, the features of dipeptide are also adding distinct efficiency while adding 20 times of features than amino acid composition. The information of I-Mutant2.0-seq is just to add temperature and pH data in amino acid composition. Only features of significant coupling patterns are more useful than amino acid composition information, and coupling composition can improve stable mutation prediction.



Table 4. The performance of I-Mutant2.0 and SVM predictors based different combinations of features, coupling composition, amino acid composition, dipeptide composition, temperature and pH value.

Method	Feature number	Prediction result					
		$Q2$	$P(+)$	$P(-)$	$Q(+)$	$Q(-)$	MCC
I-Mutant2.0-str	43	0.80	0.73	0.56	0.83	0.91	0.51
I-Mutant2.0-seq	42	0.77	0.69	0.79	0.46	0.91	0.42
CC, T, pH	42	0.79	0.69	0.82	0.54	0.90	0.47
AAC, CC, T, pH	62	0.78	0.64	0.82	0.55	0.87	0.44
CC	40	0.69	0.48	0.81	0.59	0.73	0.30
CC, T	41	0.77	0.65	0.80	0.48	0.89	0.41
CC, pH	41	0.77	0.68	0.79	0.43	0.91	0.4
AAC, T, pH	22	0.78	0.69	0.81	0.48	0.91	0.44
AAC, DC, T, pH	422	0.78	0.68	0.81	0.5	0.9	0.44

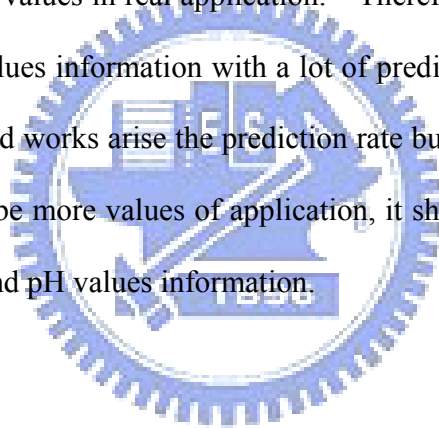
CC: Coupling Composition, AAC: Amino Acid Composition,

DC: Dipeptide Composition, T: Temperature

5.3 Conclusion

In real protein engineering, the prediction ability of stable proteins is the most concerned. However, due to less stable mutation data for the training set, it is much difficult to predict stable mutants. Our result shows that prediction method using SVM based on coupling composition can make a significant improvement in $Q(+)$.

Our result supports that temperature or pH values information are of crucial role in prediction of the same data set (S2048). However, it is difficult to find the protein optimal temperature and pH values in real application. Therefore, we consider the diverse of temperature and pH values information with a lot of prediction hint. Those information are just for that related works arise the prediction rate but not useful. We suggest if future related works will be more values of application, it should compare the same data set without temperature and pH values information.



Chapter 6 TheCUP: A Web Tool for Protein Thermostability

At present, it is lack of practical tools in identifying possible mutation sites for enhancing thermostability enhancement. As a result, experimentalists still rely on the strategy of random mutation to improve protein thermostability. Here we present a tool that can suggest possible mutation sites for increasing protein thermostability based on thermal coupling sequence pattern of protein sequences. The *protein thermostability profiles using Thermal CoUpling Patterns* (TheCUP) provides sequence level thermostability analysis based on essential thermostability coupling-pattern profiles. These profiles can be used to identify potential mutation site for thermostability improvement. TheCUP provides a rational approach to improve protein thermostability and should be complementary to the usual experimental approaches based on random mutations.

We present a web server that will suggest potential mutation sites from protein sequences that will increase protein thermostability. This web server, referred to as the *protein thermostability profile using Thermal CoUpling Patterns* (TheCUP), generates thermostability profile based on the sequence coupling patterns⁶². The sequence-coupling patterns are defined as any two types of amino acids separated by one or more amino acids. Liang and co-workers⁶² have developed the statistical analysis approach to identify the sequence-coupling patterns from the thermophilic and mesophilic microbial genomes. They found that the thermophiles and mesophiles have their own preferred sequence coupling patterns and that these sequence-coupling patterns are useful in distinguishing thermophilic and mesophilic sequences⁶². For example, the sequence-coupling pattern

EXXXK, where X represents any amino acid type, is the preferred sequence-coupling pattern (p -value $< 10^{-8}$) in the thermophilic microbial genomes. We will refer to those coupling patterns preferred by thermophiles as the thermo-patterns, and those preferred by mesophiles as the meso-patterns. Here, we derive the thermostability profiles from these sequence-coupling patterns to identify the locally stabilized regions of protein sequences.

6.1 Methods and Implementation

6.1.1 Thermostability Profile

Thermostability profiles are defined in equations 13 and 14. The thermostability profile of this protein sequence is given as $T_i, M_i, T_i - M_i, i = 1, \dots, n$.

6.1.2 Input and Output Format

The web page of the TheCUP web server is shown in Figure 18. The users can either paste one sequence in the FASTA format or 2 sequences for the analysis of potential mutation sites for thermostability improvement upload a structural file in the PDB format. If the user enters 1 sequence, theCUP will return three thermostability profiles – the T profile, the M profile and the $T - M$ profile (Figure 19). The potential mutation sites, which appear at the local minima of the $T - M$ profile, are marked by \times sign. If the user intends to compare 2 sequences, TheCUP will perform sequence alignment of these sequences and plot their thermostability profiles (Figure 20). The user can compare difference in local stability between these two sequences. This feature will be particularly useful when the experimentalist intends to improve the thermostability of one

sequence based on other homologous sequence.



TheCUP: protein thermostability profile using Thermal CoUpling Pattern

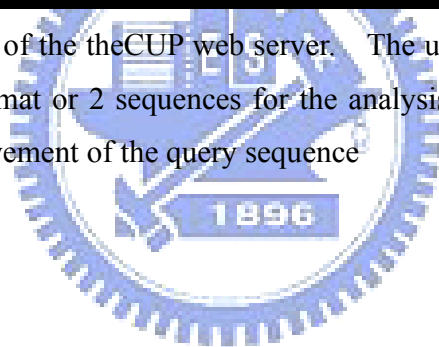
Paste the query sequences in FASTA format below

```
>RNaseHI-EC  
MLKQVEIFTDGSLGNPGGYGAILRYRGREKTF SAGYTRTTNNRMELMAAIVALEALKEHCEVILSTD  
SQYVRQGITQWIHNWKKRGWKTADKKPVKNVDLWQRLDAALGQHQIKWEWVKGHAGHPENERCDELARAA  
AMNPTLEDGTGYQVEV
```

Suggested Mutation Sites to Improve Thermostability

Comparison of Two Sequences

Figure 18. The web page of the theCUP web server. The users can either paste one sequence in the FASTA format or 2 sequences for the analysis of potential mutation sites for thermostability improvement of the query sequence



TheCUP: protein thermostability profile using Thermal CoUpling Pattern

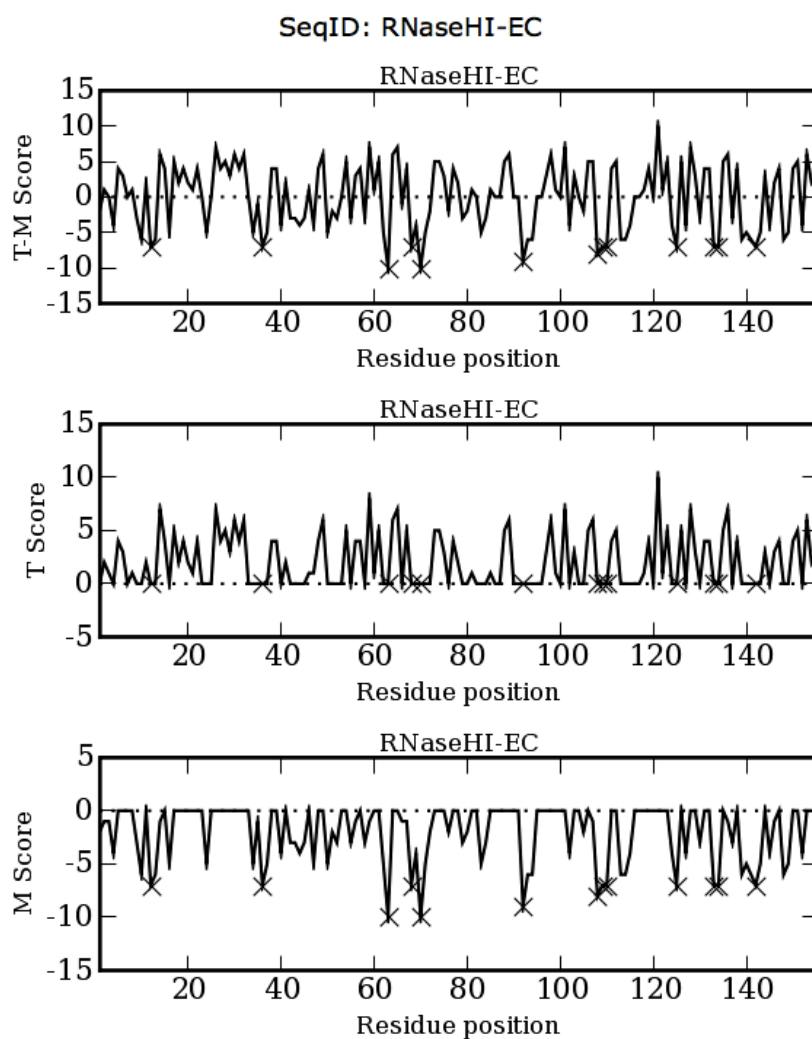


Figure 19. TheCUP will return for the query sequence three thermostability profiles – T profile, M profile and $T-M$ profile. The potential mutation sites, which appear at the local minima of the $T-M$ profile, are marked by \times sign.

TheCUP: protein thermostability profile using Thermal CoUpling Pattern

1st SeqID (Green): RNaseHI-EC
2nd SeqID (Blue): RNaseHI-R4R5R6R7
Comparison (Black): 1st-2nd

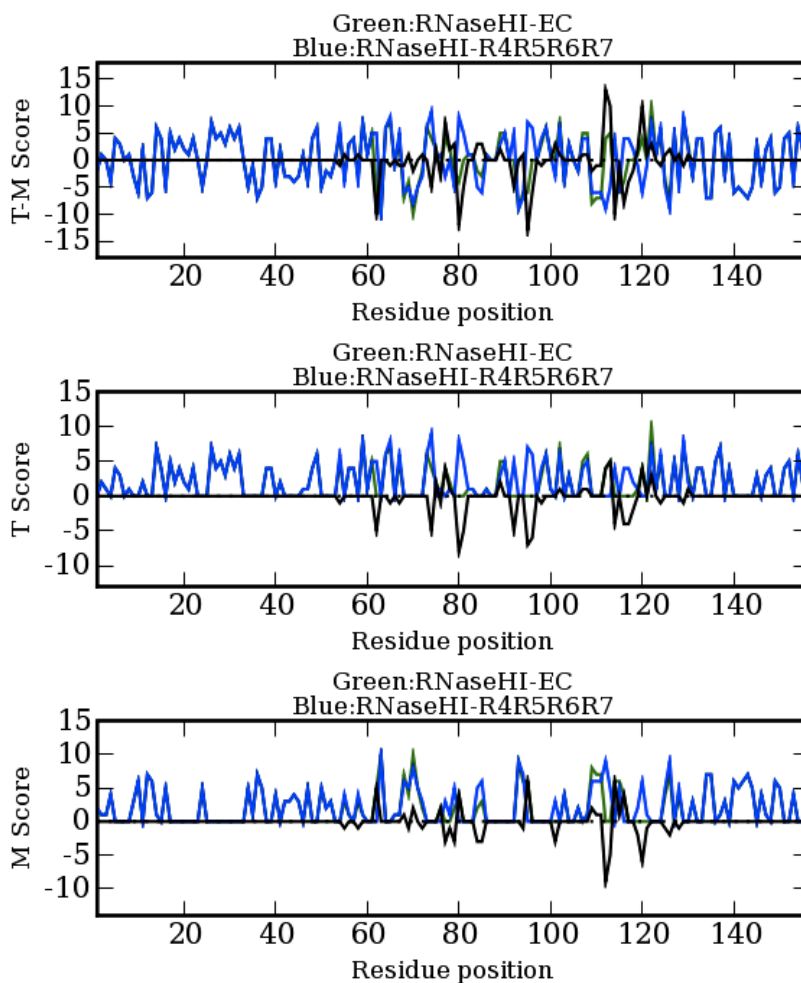


Figure 20. If the user inputs 2 sequences, TheCUP will perform sequence alignment of these sequences and plot their thermostability profiles (one in black and another in blue). This function is particularly useful when the experimentalist intends to improve the thermostability of one sequences based on other homologous sequence.

6.2 Conclusion

We have developed a web server called TheCUP to predict potential mutational sites for thermostability enhancement. Based on the findings that thermophilic and mesophilic microbes prefer different types of sequence-coupling patterns, TheCUP generates the thermostability profiles of the query sequences. The local minima of the $T - M$ profiles suggest local unstable regions and, hence, are the candidates for mutations that may help enhance protein thermostability. We believe that TheCUP will become a useful tool in rationally designing thermal stable sequences and should be complementary to the usual experimental design based on random mutations.



Chapter 7 Conclusion and Future Works

In recent years, many structure features of thermophilic enzymes are published in experimental evidences, but there are no single remarkable feature of protein thermostability can be concluded. Therefore, people have no obvious typical rule used for protein thermostability engineering to the present. As no single remarkable feature of protein thermostability, we try to make an integration of numerical methods to generate a single pattern profile. We have developed statistics based methods including sequence derived structural entropy (SDSE), and amino acid coupling patterns, to evaluate protein thermostability.

SDSE is a technique to compute structural entropy directly from protein sequences. We explored the possibility of using structural entropy to identify residues involved in thermal stabilization of various protein families. Examples include methanococcal adenylate kinase, Ribonuclease HI and holocytochrome c551. Our results show that the positions of the largest structural entropy differences between wild type and mutant usually coincide with the residues relevant to thermostability. We also observed a good linear relationship between the average structural entropy and the melting temperatures for adenylate kinase and its chimeric constructs. To validate this linear relationship, we compiled a large dataset comprised of 1,153 sequences and found that most protein families still display similar linear relationships. Our results suggest that the multitude of interactions involved in thermal stabilization may be generalized into the tendency of proteins to maintain local structural conservation. The linear relationship between structural entropy and protein thermostability should be useful in the study of protein thermal stabilization.

Amino acid coupling pattern is defined as any 2 types of amino acids separated by 1 or

more amino acids. Using this approach, we also construct the ρ profiles for the coupling patterns. The ρ value gives a measure of the relative occurrence of a coupling pattern in thermophiles compared with mesophiles. We study the amino acid coupling sequence patterns for a data set comprising 74 mesophilic and 15 thermophilic genomes, and we found that thermophiles and mesophiles exhibit significant bias in their amino acid coupling patterns. We showed that such bias is mainly due to temperature adaptation instead of species or GC content variations. Though no single outstanding coupling pattern can adequately account for protein thermostability, we can use a group of amino acid coupling patterns having strong statistical significance (p values $< 10^{-7}$) to distinguish between thermophilic and mesophilic proteins. We found a good correlation between the optimal growth temperatures of the genomes and the occurrences of the coupling patterns (the correlation coefficient is 0.89). Furthermore, we can separate the thermophilic proteins from their mesophilic orthologs using the amino acid coupling patterns. These results may be useful in the study of the enhanced stability of proteins from thermophiles—especially when structural information is scarce.

When we develop these methods to distinguish thermophilic and mesophilic proteins in sequence, it means people can use these methods for enzyme thermostability engineering when there is only sequence data. Both SDSE and amino acid coupling patterns methods integrate the complex features and interpret the protein thermostability level in local sequence. Furthermore, our developing can construct a possible strategy for thermostability protein engineering. SDSE and coupling pattern profiles are easier to apply in any case of protein thermostability engineering.

The prediction methods may make huge impact if real experimental confirm is available.

In the future, we will try to confirm the performance in real site-directed mutagenesis experiments. We believe that can also verify the prediction accuracy directly.

Furthermore, the training data set is unbalanced, because most of the mutation data reported is success data. There are lacks of fail (unstable) results that are also useful in proofing prediction. The fail data is needed for balancing the training data set

Moreover, we hope to do high-throughput experiment systematically to solve the basic problems, like the temperature limitation of enzyme, or integrating super stable essential genes in hyperthermophilic microbial, and creating new limitation of life.....maybe can create any incredible application. In addition, the mystery of life origin maybe will be solved one day.



8. References

1. Watanabe K, Suzuki Y. Protein thermostabilization by proline substitutions. *J Mol Catal B: Enzy* 1998;4:167-180.
2. Bommarius AS, Broering JM, Chaparro-Riggers JF, Polizzi KM. High-throughput screening for enhanced protein stability. *Curr Opin Biotechnol* 2006;17(6):606-610.
3. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 2001;65(1):1-43.
4. Hao J, Berry A. A thermostable variant of fructose bisphosphate aldolase constructed by directed evolution also shows increased stability in organic solvents. *Protein Eng Des Sel* 2004;17(9):689-697.
5. Sadeghi M, Naderi-Manesh H, Zarrabi M, Ranjbar B. Effective factors in thermostability of thermophilic proteins. *Biophys Chem* 2006;119(3):256-270.
6. Berezovsky IN, Shakhnovich EI. Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* 2005;102(36):12742-12747.
7. Vieille C, Hess JM, Kelly RM, Zeikus JG. xylA cloning and sequencing and biochemical characterization of xylose isomerase from *Thermotoga neapolitana*. *Appl Environ Microbiol* 1995;61(5):1867-1875.
8. Russell RJ, Ferguson JM, Hough DW, Danson MJ, Taylor GL. The crystal structure of citrate synthase from the hyperthermophilic archaeon *pyrococcus furiosus* at 1.9 Å resolution. *Biochemistry* 1997;36(33):9983-9994.
9. Bauer MW, Kelly RM. The family I beta-glucosidases from *Pyrococcus furiosus* and *Agrobacterium faecalis* share a common catalytic mechanism. *Biochemistry* 1998;37(49):17170-17178.
10. Chakravarty S, Varadarajan R. Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry* 2002;41(25):8152-8161.
11. Lim JH, Yu YG, Han YS, Cho S, Ahn BY, Kim SH, Cho Y. The crystal structure of an Fe-superoxide dismutase from the hyperthermophile *Aquifex pyrophilus* at 1.9 Å resolution: structural basis for thermostability. *J Mol Biol* 1997;270(2):259-274.
12. Chang C, Park BC, Lee DS, Suh SW. Crystal structures of thermostable xylose isomerases from *Thermus caldophilus* and *Thermus thermophilus*: possible structural determinants of thermostability. *J Mol Biol* 1999;288(4):623-634.
13. Szilagyí A, Zavodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure Fold Des* 2000;8(5):493-504.
14. Hasegawa J, Shimahara H, Mizutani M, Uchiyama S, Arai H, Ishii M, Kobayashi Y, Ferguson SJ, Sambongi Y, Igarashi Y. Stabilization of *Pseudomonas aeruginosa* cytochrome c(551) by systematic amino acid substitutions based on the structure of thermophilic *Hydrogenobacter thermophilus* cytochrome c(552). *J Biol Chem* 1999;274(53):37533-37537.
15. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. *Protein*

- Eng 2000;13(3):179-191.
16. Burley SK, Petsko GA. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* 1985;229(4708):23-28.
 17. Dong G, Vieille C, Savchenko A, Zeikus JG. Cloning, sequencing, and expression of the gene encoding extracellular alpha-amylase from *Pyrococcus furiosus* and biochemical characterization of the recombinant enzyme. *Appl Environ Microbiol* 1997;63(9):3569-3576.
 18. Teplyakov AV, Kuranova IP, Harutyunyan EH, Vainshtein BK, Frommel C, Hohne WE, Wilson KS. Crystal structure of thermitase at 1.4 Å resolution. *J Mol Biol* 1990;214(1):261-279.
 19. Ishikawa K, Okumura M, Katayanagi K, Kimura S, Kanaya S, Nakamura H, Morikawa K. Crystal structure of ribonuclease H from *Thermus thermophilus* HB8 refined at 2.8 Å resolution. *J Mol Biol* 1993;230(2):529-542.
 20. Serrano L, Bycroft M, Fersht AR. Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles. *J Mol Biol* 1991;218(2):465-475.
 21. Dougherty DA. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* 1996;271(5246):163-168.
 22. Mozo-Villarias A, Cedano J, Querol E. Hydrophobicity density profiles to predict thermal stability enhancement in proteins. *Protein J* 2006;25(7-8):529-535.
 23. Hurley JH, Baase WA, Matthews BW. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol* 1992;224(4):1143-1159.
 24. Rosato V, Pucello N, Giuliano G. Evidence for cysteine clustering in thermophilic proteomes. *Trends Genet* 2002;18:278-281.
 25. Thompson MJ, Eisenberg D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. *J Mol Biol* 1999;290(2):595-604.
 26. Haney PJ, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proc Natl Acad Sci U S A* 1999;96(7):3578-3583.
 27. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. *J Biol Chem* 2000;275(42):32383-32386.
 28. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Mol Biol Evol* 1999;16(12):1785-1790.
 29. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 2007;3(1):e5.
 30. Perutz MF, Raidt H. Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. *Nature* 1975;255(5505):256-259.
 31. Argos P, Rossman MG, Grau UM, Zuber H, Frank G, Tratschin JD. Thermal stability and protein structure. *Biochemistry* 1979;18(25):5698-5703.
 32. Jaenicke R, Zavodszky P. Proteins under extreme physical conditions. *FEBS Lett* 1990;268(2):344-349.
 33. Jaenicke R. Protein stability and molecular adaptation to extreme conditions. *Eur J Biochem* 1991;202(3):715-728.
 34. Querol E, Perez-Pons JA, Mozo-Villarias A. Analysis of protein conformational characteristics related to thermostability. *Protein Eng* 1996;9(3):265-271.

35. Vogt G, Woell S, Argos P. Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol* 1997;269(4):631-643.
36. Jaenicke R, Bohm G. The stability of proteins in extreme environments. *Curr Opin Struct Biol* 1998;8(6):738-748.
37. Dill KA. Dominant forces in protein folding. *Biochemistry* 1990;29(31):7133-7155.
38. Pace CN, Shirley BA, McNutt M, Gajiwala K. Forces contributing to the conformational stability of proteins. *Faseb J* 1996;10(1):75-83.
39. Matsumura M, Yasumura S, Aiba S. Cumulative effect of intragenic amino-acid replacements on the thermostability of a protein. *Nature* 1986;323(6086):356-358.
40. Pantoliano MW, Whitlow M, Wood JF, Dodd SW, Hardman KD, Rollence ML, Bryan PN. Large increases in general stability for subtilisin BPN' through incremental changes in the free energy of unfolding. *Biochemistry* 1989;28(18):7205-7213.
41. Serrano L, Day AG, Fersht AR. Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol* 1993;233(2):305-312.
42. Shih P, Kirsch JF. Design and structural analysis of an engineered thermostable chicken lysozyme. *Protein Sci* 1995;4(10):2063-2072.
43. Vetriani C, Maeder DL, Tolliday N, Yip KS, Stillman TJ, Britton KL, Rice DW, Klump HH, Robb FT. Protein thermostability above 100 degreesC: a key role for ionic interactions. *Proc Natl Acad Sci U S A* 1998;95(21):12300-12305.
44. Saven JG. Combinatorial protein design. *Curr Opin Struct Biol* 2002;12(4):453-458.
45. Mozo-Villiaría A, Querol E. Theoretical analysis and computational predictions of protein thermostability. *Current Bioinformatics* 2006;1:25-32.
46. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A. ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. *Nucleic Acids Res* 2002;30(1):301-302.
47. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* 2002;320(2):369-387.
48. Kwasigroch JM, Gilis D, Dehouck Y, Rooman M. PoPMuSiC, rationally designing point mutations in protein structures. *Bioinformatics* 2002;18(12):1701-1702.
49. Capriotti E, Fariselli P, Casadio R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* 2004;20 Suppl 1:i63-68.
50. Bordner AJ, Abagyan RA. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins* 2004;57(2):400-413.
51. Hoppe C, Schomburg D. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Sci* 2005;14(10):2682-2692.
52. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 2006;62(4):1125-1132.
53. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res* 2006;34(Web Server is-

- sue):W239-242.
54. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. *Proteins* 2007;66(1):41-52.
 55. Magyar C, Gromiha MM, Pujadas G, Tusnady GE, Simon I. SRide: a server for identifying stabilizing residues in proteins. *Nucleic Acids Res* 2005;33(Web Server issue):W303-305.
 56. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;22:2577-2637.
 57. Nakashima H, Fukuchi S, K. N. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *J Biochem* 2003;133:507-513.
 58. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* 2001;29:1608-1615.
 59. La D, Silver M, Edgar RC, Livesay DR. Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins. *Biochemistry* 2003;42(30):8988-8998.
 60. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. *FEBS Lett* 2000;470(1):65-69.
 61. Mizuguchi K, Sele M, Cubellis MV. Environment specific substitution tables for thermophilic proteins. *BMC Bioinformatics* 2007;8 Suppl 1:S15.
 62. Liang HK, Huang CM, Ko MT, Hwang JK. Amino acid coupling patterns in thermophilic proteins. *Proteins* 2005;59(1):58-63.
 63. Matsui I, Harata K. Implication for buried polar contacts and ion pairs in hyperthermostable enzymes. *Febs J* 2007;274(16):4012-4022.
 64. Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry* 1990;29(9):2403-2408.
 65. Kumar S, Nussinov R. How do thermophilic proteins deal with heat? *Cell Mol Life Sci* 2001;58(9):1216-1233.
 66. Kumar S, Ma B, Tsai CJ, Nussinov R. Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins* 2000;38(4):368-383.
 67. Yip KS, Stillman TJ, Britton KL, Artymiuk PJ, Baker PJ, Sedelnikova SE, Engel PC, Pasquo A, Chiaraluce R, Consalvi V. The structure of *Pyrococcus furiosus* glutamate dehydrogenase reveals a key role for ion-pair networks in maintaining enzyme stability at extreme temperatures. *Structure* 1995;3(11):1147-1158.
 68. Shirley BA, Stanssens P, Hahn U, Pace CN. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* 1992;31(3):725-732.
 69. Tanner JJ, Hecht RM, Krause KL. Determinants of enzyme thermostability observed in the molecular structure of *Thermus aquaticus* D-glyceraldehyde-3-phosphate dehydrogenase at 25 Angstroms Resolution. *Biochemistry* 1996;35(8):2597-2609.

70. Gallivan JP, Dougherty DA. Cation- π interactions in structural biology. *Proc Natl Acad Sci U S A* 1999;96(17):9459-9464.
71. Pace CN. Contribution of the hydrophobic effect to globular protein stability. *J Mol Biol* 1992;226(1):29-35.
72. Haney PJ, Stees M, Konisky J. Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the genus *Methanococcus*. *J Biol Chem* 1999;274(40):28453-28458.
73. Kirino H, Aoki M, Aoshima M, Hayashi Y, Ohba M, Yamagishi A, Wakagi T, Oshima T. Hydrophobic interaction at the subunit interface contributes to the thermostability of 3-isopropylmalate dehydrogenase from an extreme thermophile, *Thermus thermophilus*. *Eur J Biochem* 1994;220(1):275-281.
74. Britton KL, Baker PJ, Borges KMM, Engel PC, Pasquo A, Rice DW, Robb FT, Scandurra R, Stillman TJ, Yip KSP. Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis*. *Eur J Biochem* 1995;229:688-695.
75. Zhu W, Sandman K, Lee GE, Reeve JN, Summers MF. NMR structure and comparison of the archaeal histone HfOB from the mesophile *Methanobacterium formicicum* with HMfB from the hyperthermophile *Methanothermobacter fervidus*. *Biochemistry* 1998;37(30):10573-10580.
76. Li WT, Grayling RA, Sandman K, Edmondson S, Shriver JW, Reeve JN. Thermodynamic stability of archaeal histones. *Biochemistry* 1998;37(30):10563-10572.
77. Matsumura M, Signor G, Matthews BW. Substantial increase of protein stability by multiple disulphide bonds. *Nature* 1989;342(6247):291-293.
78. Choi IG, Bang WG, Kim SH, Yu YG. Extremely thermostable serine-type protease from *Aquifex pyrophilus*. Molecular cloning, expression, and characterization. *J Biol Chem* 1999;274(2):881-888.
79. Volkin DB, Klibanov AM. Thermal destruction processes in proteins involving cystine residues. *J Biol Chem* 1987;262(7):2945-2950.
80. Auerbach G, Ostendorp R, Prade L, Korndorfer I, Dams T, Huber R, Jaenicke R. Lactate dehydrogenase from the hyperthermophilic bacterium *thermotoga maritima*: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic protein stabilization. *Structure* 1998;6(6):769-781.
81. Suhre K, Claverie JM. Genomic correlates of hyperthermostability, an update. *J Biol Chem* 2003;278(19):17198-17202.
82. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Research* 2000;28(1):254-256.
83. Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A* 1994;91(25):12091-12095.
84. Criswell AR, Bae E, Stec B, Konisky J, Phillips GN, Jr. Structures of thermophilic and mesophilic adenylate kinases from the genus *Methanococcus*. *J Mol Biol* 2003;330(5):1087-1099.
85. Kanaya S, Itaya M. Expression, purification, and characterization of a recombinant ribonuclease H from *Thermus thermophilus* HB8. *J Biol Chem* 1992;267(14):10184-10192.

86. Sayle R, Bissel A. RasMol: A program for fast realistic rendering of molecular structures with shadows. Proceedings of the 10th Eurographics UK'92 Conference 1992.
87. Machius M, Declerck N, Huber R, Wiegand G. Kinetic stabilization of *Bacillus licheniformis* alpha-amylase through introduction of hydrophobic residues at the surface. *J Biol Chem* 2003;278(13):11546-11553.
88. Kimura S, Nakamura H, Hashimoto T, Oobatake M, Kanaya S. Stabilization of *Escherichia coli* ribonuclease HI by strategic replacement of amino acid residues with those from the thermophilic counterpart. *J Biol Chem* 1992;267(30):21535-21542.
89. Tanaka A, Flanagan J, Sturtevant JM. Thermal unfolding of staphylococcal nuclease and several mutant forms thereof studied by differential scanning calorimetry. *Protein Sci* 1993;2(4):567-576.
90. Brown BM, Sauer RT. Tolerance of Arc repressor to multiple-alanine substitutions. *Proc Natl Acad Sci U S A* 1999;96(5):1983-1988.
91. Strop P, Mayo SL. Contribution of surface salt bridges to protein stability. *Biochemistry* 2000;39(6):1251-1255.
92. Culajay JF, Blaber SI, Khurana A, Blaber M. Thermodynamic characterization of mutants of human fibroblast growth factor 1 with an increased physiological half-life. *Biochemistry* 2000;39(24):7153-7158.
93. Georgette D, Damien B, Blaise V, Depiereux E, Uversky VN, Gerday C, Feller G. Structural and functional adaptations to extreme temperatures in psychrophilic, mesophilic, and thermophilic DNA ligases. *J Biol Chem* 2003;278(39):37015-37023.
94. Bogin O, Peretz M, Hacham Y, Korkhin Y, Frolow F, Kalb AJ, Burstein Y. Enhanced thermal stability of *Clostridium beijerinckii* alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic *Thermoanaerobacter brockii* alcohol dehydrogenase. *Protein Sci* 1998;7(5):1156-1163.
95. Kawamura S, Abe Y, Ueda T, Masumoto K, Imoto T, Yamasaki N, Kimura M. Investigation of the structural basis for thermostability of DNA-binding protein HU from *Bacillus stearothermophilus*. *J Biol Chem* 1998;273(32):19982-19987.
96. Northey JG, Di Nardo AA, Davidson AR. Hydrophobic core packing in the SH3 domain folding transition state. *Nat Struct Biol* 2002;9(2):126-130.
97. Perl D, Mueller U, Heinemann U, Schmid FX. Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat Struct Biol* 2000;7(5):380-383.
98. Martin A, Kather I, Schmid FX. Origins of the high stability of an in vitro-selected cold-shock protein. *J Mol Biol* 2002;318(5):1341-1349.
99. Dalhus B, Saarinen M, Sauer UH, Eklund P, Johansson K, Karlsson A, Ramaswamy S, Bjork A, Synstad B, Naterstad K, Sirevag R, Eklund H. Structural basis for thermophilic protein stability: structures of thermophilic and mesophilic malate dehydrogenases. *J Mol Biol* 2002;318(3):707-721.
100. Yano JK, Blasco F, Li H, Schmid RD, Henne A, Poulos TL. Preliminary characterization and crystal structure of a thermostable cytochrome P450 from *Thermus thermophilus*. *J Biol Chem* 2003;278(1):608-616.

101. Jiang X, Kowalski J, Kelly JW. Increasing protein stability using a rational approach combining sequence homology and structural alignment: Stabilizing the WW domain. *Protein Sci* 2001;10(7):1454-1465.
102. Yu MH, Weissman JS, Kim PS. Contribution of individual side-chains to the stability of BPTI examined by alanine-scanning mutagenesis. *J Mol Biol* 1995;249(2):388-397.
103. Kuroda Y, Kim PS. Folding of bovine pancreatic trypsin inhibitor (BPTI) variants in which almost half the residues are alanine. *J Mol Biol* 2000;298(3):493-501.
104. Jermutus L, Tessier M, Pasamontes L, van Loon AP, Lehmann M. Structure-based chimeric enzymes as an alternative to directed enzyme evolution: phytase as a test case. *J Biotechnol* 2001;85(1):15-24.
105. Compiani M, Fariselli P, Martelli PL, Casadio R. An entropy criterion to detect minimally frustrated intermediates in native proteins. *Proc Natl Acad Sci U S A* 1998;95(16):9290-9294.
106. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;278(5338):631-637.
107. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res* 2005;33(Web Server issue):W306-310.
108. Capriotti E, Fariselli P, Calabrese R, Casadio R. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* 2005;21 Suppl 2:ii54-ii58.
109. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 1995.
110. Yu CS, Lin CJ, Hwang JK. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Sci* 2004;13(5):1402-1406.
111. Chang C-C, Lin C-J. LIBSVM v. 2.81: a library for support vector machines. (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>); 2005.