CLASSIFICATION OF PROTEINS USING SEQUENTIAL AND STRUCTURAL FEATURES

by

AYDIN ALBAYRAK

Submitted to the Graduate School of Engineering and Natural Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Sabanci University

SABANCI UNIVERSITY

July, 2011

CLASSIFICATION OF PROTEINS USING SEQUENTIAL AND STRUCTURAL FEATURES

Aydin Albayrak

Biological Sciences and Bioengineering, PhD Thesis, 2011

Thesis Supervisor: Assoc. Prof. Dr. Ugur Sezerman

Keywords: Protein Classification, Support Vector Machines, Classification-via-clustering, Thermostability, Protein Families, Sequential and Structural Features, Machine Learning, Relative Complexity Measure, Reduced Amino Acid Alphabets

ABSTRACT

Classification of proteins is an important process in many areas of bioinformatics research. In this thesis, we devised three different strategies to classify proteins with high accuracy that may have implications for function and attribute annotation. First, protein families were classified into different functional subtypes using a classification-via-clustering approach by using relative complexity measure with reduced amino acid alphabets (RAAA). The devised procedure does not require multiple alignment of sequences and produce high classification accuracies. Second, different fixed-length motif and RAAA combinations were used as features to represent proteins from different thermostability classes. A T-test based dimensionality reduction scheme was applied to reduce the number of features and those features were used to develop support vector machine classifiers. The devised procedure produced better results with less number of features than purely using native protein alphabet. Third, a non-homologous protein structure dataset containing hyperthermophilic, thermophilic, and mesophilic proteins was assembled *de novo*. Comprehensive statistical analyses of the dataset were carried out to highlight novel features correlated with increased thermostability and machine learning approaches were used to discriminate the proteins. For the first time, our results strongly indicate that combined sequential and structural features are better predictors of protein thermostability than purely sequential or structural features. Furthermore, the discrimination capability of machine learning models strongly depends on RAAAs.

PROTEINLERIN DIZISEL VE YAPISAL ÖZELLİKLERİNİN KULLANILARAK SINIFLANDIRILMASI

Aydın Albayrak
Biyoloji Bilimleri ve Biyomühendislik, Doktora Tezi, 2011
Tez Danışmanı: Doç. Dr. Uğur Sezerman

Anahtar Kelimeler: Proteinlerin Sınıflandırılması, LibSVM, Kümeleme ile Sınıflandırma, Sıcaklık Dayanıklılığı, Protein Aileleri, Dizisel ve Yapısal Özellikler, Bilgisayarlı öğrenme yöntemleri, Göreceli Zorluk Değeri, Sadeleştirilmiş Protein Alfabeleri

ÖZET

Proteinlerin sınıflandırılması biyoinformatik araştırmalarında kullanılan önemli bir yöntemdir. Bu tez de proteinlerin yüksek doğrulukuta sınıflandırılması için üç farklı yöntem geliştirilmiştir. İlk olarak, farklı yapısal alt türlere sahip protein aileleri kümeleme ile sınıflandırma yöntemi ile Göreceli Zorluk Değeri (GZD) ve Sadeleştirilmiş Protein Alfabeleri (SPA) kullanılarak sınıflandırılmıştır. Bu geliştirilen yöntem ile Çoklu Dizi Sıralama yöntemini kullanmaksızın yüksek doğrulukta sınıflandırma yapılması sağlanmıştır. İkinci olarak, sabit uzunluktaki dizi motifleri ve SPA kombinasyonları dizileri tanımlamada özellik olarak kullanılmış ve sıcaklığa karşı dirençleri farklı olan proteinler sınıflandırılmıştır. T-test ile hipotez sınaması yapılarak özellik sayısı azaltılmış ve bu seçilen özellikler kullanılarak Destek Vektör Sınıflandırıcıları geliştirilmiştir. Bu yöntem ile proteinler normal protein alfabesine kıyasla daha az özellik kullanılarak doğruluk değerleri yüksek sınıflandırma sonuçlar elde edilmiştir. Üçüncü olarak, aşırı sıcağa dayanıklı, normal sıcağa dayanıklı ve orta derecede sıcağa dayanıklı homolog olmayan proteinlerden oluşan yeni bir veri kümesi oluşturulmuştur. Daha sonra bu veri kümesi üzerinde proteinlerin sıçağa karşı dayanıklı

olmaları ile ilintili özelliklerini ayırt edebilmek için kapsamlı bir istatistiksel analiz yapılmış ve bilgisayarlı öğrenme yöntemleri kullanılarak proteinler sınıflandırılmıştır. Bu tez çalışması sonucunda yeni dizisel ve yapısal özelliklerin birlikte kullanılmasının proteinleri sıcağa karşı direncinin tahmin edilmesinde sadece dizisel yada yapısal özelliklerin kullanılmasından daha iyi sonuçlar alındığı gösterilmiştir. Ayrıca, proteinleri ayırmak için kullanılan bilgisayarlı öğrenme yöntemlerinin doğru sınıflandırma kapasitesinin kullanılan SPA'lere bağlı olduğu gösterilmiştir.

TABLE OF CONTENTS

FIGURES

TABLES

CHAPTER 1

1 INTRODUCTION

Classification of proteins is an important process in many areas of bioinformatics including drug target identification, drug design, protein family characterization, and protein annotation. Sequencing projects and high-throughput x-ray crystallography techniques have increased the number of novel proteins. Functional and structural proteomics techniques that have been used to correlate biological functions or structural motifs to specific proteins have led to the classification of a substantial number of proteins.

In the absence of experimental validation, similarity searches are routinely employed to transfer function or attribute of a known protein to a novel protein if the similarity is above a certain threshold. However, similarity searches do not necessarily perform well when similar proteins belong to different classes or families and significant mis-annotations can occur even at high sequence identity levels. In such cases, machine learning approaches can be used to predict the class of a novel protein using features derived from raw sequence or structure data.

In a biological context, classification of proteins refers to the determination of the class of a protein or the assignment of a protein into a predefined category based on the existence of certain similarities to other members of the same category. Proteins can be classified based on their structural components, catalytic function, cellular location, pH and optimum working temperature (T_{opt}).

Classification starts with the definition of a class and class properties that make it unique or different from other classes. Class boundaries may sometimes be difficult to establish due to following reasons: i) Class definition process is abstract in nature and does not represent underlying classes. ii) Established classes are not applicable to all proteins because of non-discovered classes. To eliminate boundary-related problems, a classification scheme may need to be updated with the availability of more data.

Previously, machine learning algorithms have been used in many classification problems particularly protein interaction prediction [1], cluster analysis of gene expression data [2], annotation of protein sequences by integration of different sources of information [3], automated function prediction [4], protein fold recognition and remote homology detection [5], SNP discovery [6], prediction of DNA binding proteins [7], and gene prediction in metagenomic fragments [8]. In many cases, classification with machine learning approaches provides simple and yet advantageous solutions over more traditional, laborious and sometimes error-prone means that employ protein similarity measures.

In classification, it is often interest to determine the class of a novel protein using features extracted from raw sequence or structure data rather than directly using the raw data. For example, a typical manual annotation of a novel protein can be carried out against a database which contains expert annotated proteins with other secondary attributes. The best match in the database can be used as a template and its properties may be transferred to the novel protein. The search would take the raw sequence information as input and find sequences that are similar to the given query sequence at a given similarity threshold.

However, in a machine learning framework, the same process may be carried out as follows: i) obtain representative sequences from the database, ii) extract features from these sequences such as number and kind of domains, motif, signal regions, length of proteins, and post-translational modification sites, iii) utilize machine learning classifiers to learn from this training data, and iv) generate a model that can be used to predict the class of a new sample by testing the model on it.

This thesis is organized into five chapters where Chapter 1 is a general introductory chapter and Chapter 5 is a general conclusion chapter. Chapter 2, 3, and 4 are organized as self-sufficient individual unit with their own Introduction, Methods, Results, and Conclusions sections. Each chapter is organized to address a different classification problem and provides novel classification strategies that outperform commonly utilized methods. In cases of regions of chapter overlaps, we refer those regions in that context, sometimes expanding on them without extensive references to previous chapters or previously cited references.

In Chapter 2, for the first time, a comprehensive set of different reduced amino acid alphabet (RAAA) and Relative Complexity Measure (RCM) combinations were

tested systematically to classify protein families into functional subtypes. The procedure developed in this chapter employs the alignment-free RCM algorithm. Utilization of RCM with RAAAs may be considered as an alternative or, in some way, a complementary strategy to the commonly used protein similarity comparison algorithm - *multiple sequence alignment* (MSA). The devised procedure is independent of manual expert handling that is generally required for consistent phylogenies and produces equal or better results in terms of accuracy than those achieved by MSA.

Chapter 3 introduces the classification of protein sequences into different thermostability classes using a combination of N-grams (subsequences of length n) and RAAAs, and a T-test based dimensionality reduction approach. Effects of different N-gram sizes and a larger repertoire of RAAAs on the classification of proteins are also examined along with the effects of T-test based dimensionality reduction scheme. The devised classification strategy can produce classification accuracies that are comparable or better than those achieved using native protein alphabet but with less number of features.

Chapter 4 is dedicated to comprehensive statistical analysis and classification of proteins from three different thermostability (a finer division of classes compared to Chapter 3) classes using novel and conventional sequence-based (sequential) and structure-based (structural) features. In the first part, a timeline of major computational and experimental research on protein thermostability was provided followed by the explication of major factors suggested for protein thermostabilization in a non-exhaustive manner. In the second part, a dataset has been assembled *de novo*; computer software were developed to extract novel sequential and structural features from raw protein sequence and structure data; comprehensive statistical analyses were carried out on each feature; and classification of proteins into different thermostability classes was carried out systematically using extracted features. In the third section, analyses of the significant features and classification results were carried out and compared to the accumulated knowledge in the literature to highlight differences and their implications.

In Chapter 5, important findings of this thesis are summarized along with remarks for future research topics.

1.1 References

1. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins* 2006, 63(3):490-500.
2. Jiang DX, Tang C, Zhang AD: Cluster analysis for gene expression data: A survey. *Ieee T Knowl Data En* 2004, 16(11):1370-1386.
3. Tetko IV, Rodchenkov IV, Walter MC, Rattei T, Mewes HW: Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information. *Bioinformatics* 2008, 24(5):621-628.
4. Friedberg I: Automated protein function prediction--the genomic challenge. *Brief Bioinform* 2006, 7(3):225-242.
5. Damoulas T, Girolami MA: Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics* 2008, 24(10):1264-1270.
6. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP: Application of machine learning in SNP discovery. *BMC Bioinformatics* 2006, 7:4.
7. Bhardwaj N, Langlois RE, Zhao G, Lu H: Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res* 2005, 33(20):6486-6493.
8. Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, Meinicke P: Gene prediction in metagenomic fragments: a large scale machine learning approach. *BMC Bioinformatics* 2008, 9:217.

CHAPTER 2

2 TREE-BASED CLASSIFICATION OF PROTEIN FAMILIES INTO FUNCTIONAL SUBTYPES USING RELATIVE COMPLEXITY MEASURE WITH REDUCED AMINO ACID ALPHABETS

2.1 Introduction

Proteins that evolve from a common ancestor can change functionality over time [1] and produce highly divergent protein families that can be divided into subfamilies with similar but distinct functions (i.e., functional subfamilies or subtypes) [2]. Identification of subfamilies using protein sequence information can be carried out using phylogenetic methods that can reveal the evolutionary relationship between proteins by clustering similar proteins together in a phylogenetic tree [3-5]. The most common method for identifying similarities in sequences through phylogenetic analysis starts with the construction of a multiple alignment of homologous sequences using a substitution matrix. Multiple alignment scores are then transformed into a distance matrix to construct a phylogenetic tree. Often the branching order of a phylogenetic tree exactly matches the known functional split between proteins [1] and branch lengths are proportional to the extent of evolutionary changes since the last common ancestor.

Multiple sequence alignment (MSA) is constructed using a scoring scheme which reward or penalize each substitution, insertion and deletion to get an optimum alignment of the given sequences. The quality of an MSA is connected to the chosen parameters that are entered manually and expert handling is almost always required to maintain alignment integrity by observing general trends in each protein family. As such different alignment parameters may yield different phylogenetic trees that are only as good as the MSA that the trees are derived from [6, 7].

Phylogenetic analysis is broadly divided into two groups of methods. Algorithms in the first group calculate a matrix representing the distance between each pair of sequences and then transform this matrix into a tree using a tree-clustering algorithm. Algorithms in the first category utilize various distance measures with different models to account for nucleotide or amino acid substitutions. In the second group, the tree that can best explain the observed sequences under the chosen evolutionary model is found by evaluating the fitness of different tree topologies [6, 8]. The second category can further be divided into two groups based on the optimality criterion used in tree evaluation: maximum parsimony and maximum likelihood. Under maximum parsimony [9], the preferred phylogenetic tree is the tree that requires the least evolutionary change to explain the observed data whereas under maximum likelihood [9, 10], it is the most probable tree under the chosen evolutionary assumption.

The prediction of subfamilies from protein MSAs have been carried out previously by comparing subfamily hidden Markov models, subfamily specific sequence profiles, analyzing positional entropies in an alignment, and ascending hierarchical method [4, 5, 11, 12]. All of these methods require an alignment of biological sequences that assume some sort of an evolutionary model. Computational complexity and the inherent ambiguity of the alignment cost criteria are two major problems in MSA along with controversial evolutionary models that are used to explain them.

A novel approach for phylogenetic analysis based on Relative Complexity Measure (RCM) of whole genomic sequences have been previously proposed by Otu *et al*, that eliminates the need for MSA and produces successful phylogenies on real and simulated datasets [8]. The algorithm employs Lempel-Ziv (LZ) complexity [13] and produces a score for each sequence pair that can be interpreted as the "closeness" of the sequence pairs. Unequal sequence length or different positioning of similar regions along sequences (such as different gene order in genomes) is not an issue as the method has been shown to handle both cases naturally. Moreover, RCM does not use any approximations and assumptions in calculating the distance between sequences. Therefore, RCM utilizes the information contained in sequences and requires no human intervention.

Application of RCM to genomic sequences for phylogenetic analysis was successfully carried out on various datasets containing genomic sequences [8, 14]. Moreover, Liu *et al* [15] extended this method further to integrate the hydropathy profile and a different LZ-based distance measure for phylogenetic analysis of protein sequences while Russell *et al* [16] integrated a merged amino acid alphabet containing 11 characters to represent all amino acids. The merged alphabet was used to reduce sequence complexity prior to calculating a pairwise distance measure to be used as a pairwise scoring function in determining the order with which sequences should be joined in a multiple sequence alignment problem.

Application of RCM to evaluate genomic sequences is relatively straight forward since RCM based on Lempel-Ziv complexity scores can capture each mutation in DNA sequences and register it as an increase in the complexity scores of compared sequences. However, substitution of one residue into another in proteins is tolerable as long as the substituted residue is not highly conserved and physicochemical and structural properties of the substituted and the native residues are not fundamentally different [17-19]. Employment of hydropathy-index-based grouping of residues is one way of a preprocessing requirement to capture only the mutations that would not be tolerated in a protein sequence since LZ algorithm is not capable of accounting for amino acid substitution frequencies and similarity scores. Hence, any application that uses RCM to generate a distance matrix of protein sequences should be linked to treating the sequence with a reduced amino acid alphabet (RAAA) prior to calculating their RCMs.

In this chapter, we systematically utilized RCM with different reduced amino acid alphabets and assessed RCM's potential in clustering protein families into functional subtypes based solely on sequence data using a tree-based classification algorithm. This method clustered seven well-characterized protein families into their functional subtypes with 92% - 100% accuracy.

2.2 Methods

2.2.1 Datasets

2.2.1.1 Simulated Dataset

Performance of RCM was tested on a simulated dataset that contains 10 randomly evolved protein sequences from a root sequence of length 500 by using INDELible V1.02 [20]. Simulated protein sequences were generated according to the following parameters:

1. JTT-dcmut [21] was chosen as the amino acid substitution model.
2. Power law insertion/deletion length distribution model with a=1.7 and maximum allowed insertion/deletion length of 500 were used.
3. Both insertion and deletion rates were set to the default parameter of 0.1 relative to average substitution rate of 1%.
4. Length of the root protein sequence was set to 500.
5. The rooted tree with 10 taxa that reflects the true phylogenetic evolution of the sequences was generated along with the true MSA from which the true tree was inferred.
6. The true MSA was then inputted into ClustalW2 [22] and the bootstrap tree was generated (1000 bootstrap trials, including positions with gaps, and correcting for multiple substitutions)

2.2.1.2 Protein Datasets

RCM was tested on seven protein datasets. Number of sequences, number of subfamilies, average length, standard deviation of sequence lengths and mean percent identities (PID) [23] of sequences for each family are summarized in Table 2.1. Protein sequences for mandelate racemases, crotonases, haloacid dehalogenases and vicinal oxygen chelates (VOC) were extracted from extensively curated Structure-Function Linkage Database which contains sets of subfamily grouping for a large set of protein families. SFLD contains protein families with a hierarchical classification scheme based on sequence, structure and conserved chemical reactions at the superfamily, subgroup,

and family levels [24]. Crotonases and haloacid dehalogenases were filtered such that subfamilies that contain less than 3 sequences or more than 200 sequences were removed to prevent sequence number bias and to reduce computational complexity. Unknown or unspecified amino acids were discarded (21, 22 and 10 occurrences in mandelate racemase, crotonase and VOC family, respectively). The protein sequences for acyl transferase (AT) domains and nucleotidyl cyclases were obtained from reference [25]. The protein sequences in the hard-to-align dataset that contains glycoside hydrolase family 2 (GH2) members were adapted from reference [3] .

Table 2.1 General Properties of the datasets
* Mean Percent Identity (μ PID) is the average of all pairwise sequence identities in a given family.

Family	# of sequences	# of subfamilies	μ Length	σ Length	μ PID*
Crotonases	467	13	332	87	21
Mandelate racemases	184	8	416	74	27
Vicinal oxygen chelates	309	18	294	108	14
Haloacid dehalogenases	195	14	303	137	12
Nucleotidyl cyclases	75	2	1059	200	21
Acyl transferases	177	2	290	12	41
GH2 hydrolases	33	4	872	160	15

2.2.2 Reduced Amino Acid Alphabets

Sequence space of proteins is redundant and generates only a limited number of folds, domains, and structures [26]. Various strategies have been devised that take a coarse-grained approach to account for the degeneracy of sequences by grouping similar amino acids together [17-19, 27-30]. Grouping is usually carried out based on structural and physiochemical similarities of amino acids [28]. Grouping of amino acids in sequence space can help develop prediction methods for various sequence determinants and decrease the amount of search space in procedures employed in directed evolution experiments [26, 31].

One of the finest examples is the reduction of amino acid alphabet into a binary code that is composed of characters representing polar and non-polar amino acid residues [27]. Grouping of amino acid residues has also been used extensively in

Hydrophobic-Polar (HP) lattice model to explain the hydrophobic collapse theory of protein folding [32].

A recent study was carried out by Peterson *et al* to test the performance of over 150 RAAAs on the sequence library from DALIpdb90 database and showed that RAAAs improves sensitivity and specificity in fold prediction between protein sequence pairs with high structural similarity and low sequence identity [33].

RAs have been integrated in many experimental and computational applications and have been known to produce superior results in certain computational biology domains. One of the most common use is undeniably the implicit usage of RA in a given multiple sequence alignment problem where a similarity matrix is employed to align sequences such that similar regions are aligned on top of each other. A good alignment is ensured as long as the residues in the aligned regions have similar properties based on the residue exchange matrix that is used to evaluate the fit of one residue with another.

We tested performances of six amino acid reduction schemes with 15 different levels of groupings to separate proteins into functional subfamilies (Table 2.2). These included three top performing RAAA (HSDM17, SDM12, GBMR4) from reference [33] and three random RAAA of size 4.

Table 2.2 Reduced Amino Acid Alphabets

* Substitution matrices for these reduced alphabets were obtained from reference [33]. [§] BL62 frequency counts were used to derive these substitution matrices using the formula outlined in reference [33]. [#]Gap opening/gap extension penalties used for MSAs in ClustalW2.

Scheme	Size	Matrix	Gaps[#]	Reference
ML*	4,8,10,15	BL50	12/2	[28]
EB[§]	13,11,9,8,5	BL62	11/1	[18]
HSDM*	17	HSDM	19/1	[29]
SDM*	12	SDM	7/1	[29]
GBMR*	4	BL62	11/1	[30]
RANDOM[§]	4,4,4	BL62	11/1	This study

2.2.3 Substitution Matrices

Amino acids that are within the same group in a RAAA are considered identical [33]. Substitution matrices that assign the same similarity score to each amino acid within the same group were obtained from reference [33]. For those RAAAs in the EB scheme and the three random RAAAs, new substitution matrices were created from BLOSUM62 frequency counts using the same procedure outlined in reference [33].

2.2.4 Lempel-Ziv Complexity

In this chapter, a normalized distance measure that was previously used for phylogenetic tree construction of whole genome sequences was employed. The distance measure was based on Lempel-Ziv [34] complexity and was known to accurately cluster all related genomic sequences under one branch of the tree [8].

Lempel-Ziv (LZ) complexity score of a sequence is obtained by counting the number of steps required to generate a copy of the primary sequence starting from a null state. At each step, an amino acid or a series of amino acids are copied from the subsequence that has been constructed thus far allowing for a single letter innovation. The number of steps needed to obtain the whole sequence is identified as the LZ-complexity score of the given sequence. The exhaustive library of a sequence is defined as the smallest number of distinct amino acid or amino acid combinations required to construct the sequence using a copying process described by Lempel and Ziv [34]. For example, the LZ-complexity of the simple sequence 'AAILNAIIANNL' would be obtained as shown in Table 2.3. Since seven steps are needed to generate the whole sequence, the LZ-complexity score for this sequence is 7. The LZ-complexity of a sequence 'X' compared to a sequence 'Y' is known as the RCM of 'X' with respect to 'Y'. This is the number of steps required to construct sequence 'X' beginning with 'Y' instead of a null sequence. Five different distance metrics have been suggested by Otu *et al* [8] , however, this work used only the following normalized distance metric that accounts for the differences in sequence lengths:

$$D_{XY} = \frac{c(XY) + c(YX) - c(X) - c(Y)}{\frac{c(XY) + c(YX)}{2}}$$

where c(XY) and c(YX) are RCM of X appended to Y and Y appended to X, respectively. Remaining four LZ-based distance measures defined by Out *et al* [8] performed slightly worse than the above distance (data not shown). Although in performance between five measures were not significant, we adopted the aforementioned distance for its ability to account for length variance.

Table 2.3 Exhaustive library construction and Lempel-Ziv complexity calculation

Sequence X = AAILNAIIANNL	
Exhaustive History	Complexity
A	1
AI	2
L	3
N	4
AII	5
AN	6
NL	7
$H_E(X)$	C(X)=7

2.2.5 Distance Matrix & Phylogenetic Tree

The relative complexity measure (RCM) for creation of the distance matrix was utilized as previously described [8]. Phylogenetic trees were generated from distance matrices using neighbor-joining [35] program of the phylogeny inference package, PHYLIP 3.68 [36]. Un-rooted trees were rooted with midpoint rooting by placing the root halfway between the two most distinct taxa. Midpoint-rooted trees were converted to cladograms (i.e., branch lengths are discarded) using the Retree program of PHYLIP package [36].

2.2.6 ClustalW2

Protein sequences in each family were aligned using ClustalW2 [22] for comparison with RCM. MSAs were performed using updated substitution matrices with gap extension and gap opening penalties provided in Table 2.2. Bootstrap analyses were carried out 100 times and trees containing bootstrap values were created using ClustalW2 with the neighbor-joining clustering algorithm. For convenience, MSAs that were carried out using ClustalW2 will be referred as the MSA or the MSA method for the rest of the article.

2.2.7 Tree Based Classification (TBC)

TBC algorithm [4] was used to check the accuracy of each tree in separating protein families into subfamilies. TBC divides a tree into disjoint subtrees and assigns a protein subfamily to a subtree that maximizes the number of true positives when the proportions of $fp/(tp+fp)$ and $fn/(tp+fn)$ are both equal to 0.5 for a given subtree, where *fp* is the number of false positives, *fn* is the number of false negatives and *tp* is the number of true positives. Above proportions correspond to the "maximal allowed contamination" level that minimizes the TBC error over the whole tree.

TBC requires a bifurcating tree of sequences in a protein family and an attribute file that contains expert curated assignment of each sequence to a particular subfamily. TBC accuracy (i.e., the percentage of correctly classified sequences) is the primary performance measure to evaluate the division of protein families into subtypes using the TBC algorithm. TBC accuracy is equal to 1- %TBC error where %TBC error is the total number of *fp*, *fn*, and *unclassified sequences* divided by the total number of sequences. For a detailed analysis of the TBC algorithm, refer to reference [4].

2.2.8 Protocol

The proposed algorithm operates on a set of sequences in FASTA format. After one of the alphabets given in Table 2.2 is applied to all the sequences in the dataset, RCMs are calculated and used to obtain the distance between each pair for the neighbor–joining clustering to create a phylogenetic tree. For each RAAA, a single tree based on RCM is generated and analyzed using TBC algorithm to determine how well it clusters different subfamilies under different branches of the tree.

For simulated dataset, three phylogenetic trees were compared: The true tree generated by INDELible, the bootstrap tree and the RCM tree. INDELible creates a true MSA of the simulated protein sequences. This alignment was used in ClustalW2 and bootstrapped 1000 times and the resulting tree was called the bootstrap tree. The third tree is the RCM tree that was generated by the proposed approach.

For seven protein datasets, first, the original fasta sequences were used to calculate RCMs and their associated RCM trees. Second, the original fasta sequences were re-coded using different RAAAs (Table 2.2) and the reduced sequences were used to calculate their RCMs and the associated RCM trees.

A similar procedure was applied to the phylogenetic trees using the MSA method. For each protein family, MSA was carried out using the corresponding substitution matrices and gap penalties provided in Table 2.2. MSA-based trees were created following bootstrap analysis (100 replicates) with ClustalW2.

Finally, for each family, a total of 16 phylogenetic trees (1 for 20-letter alphabet, 12 for RAAAs, and 3 for random RAAAs) for each method are generated and checked how well they separated families into subfamilies. A summary of the overall workflow is depicted in Figure 2.1.

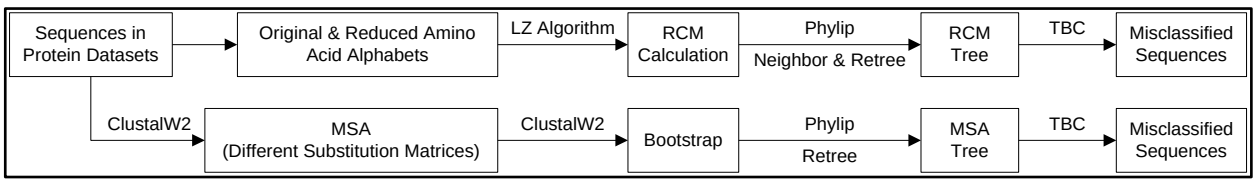

Figure 2.1 Overall workflow of the protocol

2.3 Results and Discussion

2.3.1 Simulated Dataset

Phylogenetic analysis of protein sequences has been intimately connected with MSA. A phylogenetic tree is generated from an evolutionary distance matrix using MSA of sequences. However, for real biological datasets, the true tree is rarely known. Therefore, protein sequence evolution was simulated to study the reliability of the RCM method. A simulated protein dataset containing 10 protein sequences was generated to show that RCM coupled with a RAAA can produce a phylogenetic tree (RCM tree) that is consistent with the true tree and the bootstrap tree. The true tree is produced by INDELible and is the original tree that reflects the evolution of 10 simulated sequences. On the other hand, the bootstrap tree is the tree that was produced by ClustalW2 using the true MSA implied by INDELible. The bootstrap tree is identical to the true tree and the bootstrap supports for all branches are high reflecting the consistency [37] in the branching. The RCM tree was produced by the alignment-free RCM approach. The RCM tree is identical to both the true tree and the bootstrap tree reflecting its potential for use in phylogenetic analysis of protein sequences. The tree topology of only one of the trees is shown in Figure 2.2 since they are all identical.

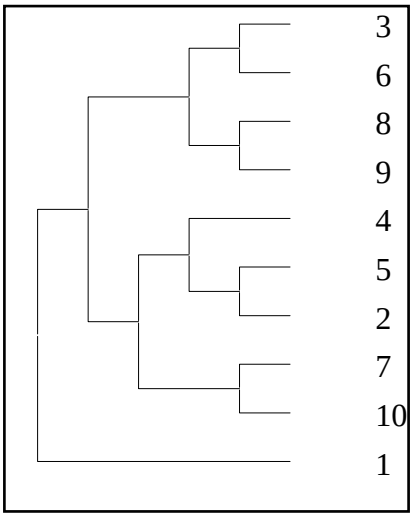

Figure 2.2 Tree topology of the simulated dataset

2.3.2 Performance of the RCM approach

We applied the RCM approach to seven protein datasets. RCM method showed an efficient division of protein families into subfamilies using RAAAs. Phylogenetic trees of the seven protein families using RCM approach are shown in Figure 2.3 for ML15 alphabet. Detailed comparison of RCM with MSA in terms of TBC accuracy, the number and percentage of TBC error for each RAAA and each dataset is provided in Appendix A.

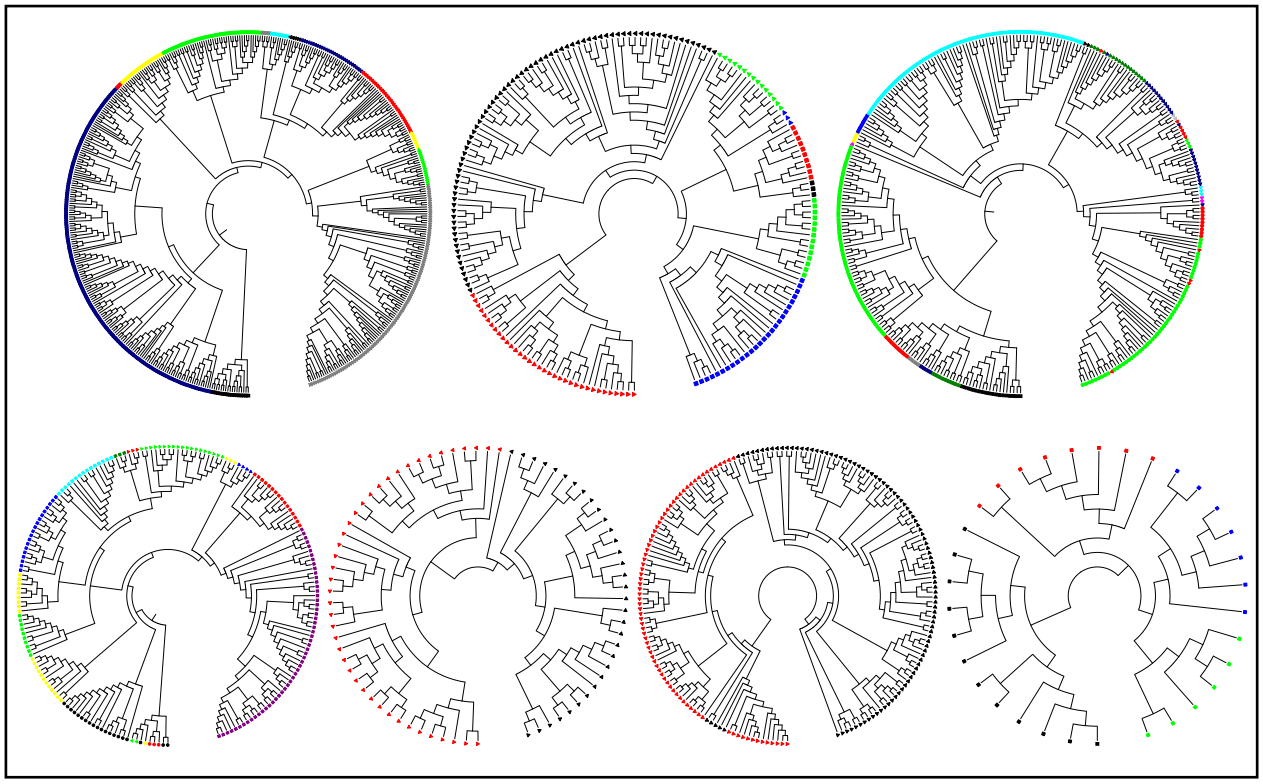

Figure 2.3 Phylogenetic trees of protein families
RCM trees were drawn using ML15 alphabet. For each family, the taxa corresponding to different subfamilies are colored differently. (A) Crotonases (B) Mandelate racemases (C) Vicinal oxygen chelates (D) Haloacid dehalogenase (E) Nucleotidyl cyclases (F) Acyl transferases (G) GH2 hydrolases

2.3.2.1 Crotonases

Members of crotonase family contain 467 protein sequences from 13 different subfamilies and catalyze diverse metabolic reactions with certain family members displaying dehalogenase, hydratase, and isomerase activities. TBC accuracy varied between 96.4% and 100% for RCM. The top performing RAAA with the smallest size was GBMR4 that resulted in 100% TBC accuracy. TBC accuracy was 100% for all RAAAs tested with MSA.

2.3.2.2 Mandelate Racemases

The mandelate racemase dataset contains 184 sequences that are assigned to 8 expert curated subfamilies. All mandelate racemases contain a conserved histidine, presumably acting as an active site base [38]. When the RCM approach was tested on mandelate racemases, all resulting trees showed correct assignment of functional subfamilies into 8 different clusters with 100% accuracy using all alphabets except GBMR4 that resulted in 96.7% TBC accuracy.

2.3.2.3 Vicinal oxygen chelates *(VOC)*

VOC family contains 309 sequences from 18 different subfamilies. The number of TBC accuracy varied between 77.7% and 92% for RCM and 81.9% to 91.3% for MSA. Members of VOC have an average sequence length of 294 amino acids and a mean PID of 14% (Table 2.1). The low PID and the highly divergent nature of this family make its subfamilies susceptible to misclassification more than other families based on sequence information alone. In this dataset, EB8 performed better than 20-letter alphabet (92.2% vs. 91.3%) with RCM while GBMR4, ML4, EB8, EB, EB13 and 20-letter alphabets resulted in 91.3% TBC errors with MSA.

2.3.2.4 Haloacid dehalogenases

Haloacid dehalogenases contains 195 sequences that belong to 14 different subfamilies. Haloacid dehalogenase family is similar to VOCs in its highly divergent nature based on the low mean PID (12%) that places the sequences in this family in the "twilight zone" to infer any relation between sequences based on sequence information alone. ML15 was the best performing RAAA for RCM with 96.9% accuracy (Table 2.4). The size of the best performing RAAA for this family is larger compared to other families hinting that highly divergent sequences may require larger alphabets with lower level of grouping.

2.3.2.5 Nucleotidyl cyclases

Nucleotidyl cyclase family has two functional subfamilies, adenylate and guanylate cyclases that correspond to use of the substrates ATP and GTP respectively. The nucleotidyl cyclase family with 33 adenylate cyclases and 42 guanylate cyclases was clustered into two distinct subfamilies with 100% accuracy using both methods and all RAAAs except EB5 and EB8 for RCM and ML4 and EB5 for MSA, all of which resulted in 98.7% accuracy (Table 2.4). Moreover, the clustering result for the nucleotidyl cyclases are in agreement with the result obtained previously by the MSA-dependent clustering algorithm that uses the residues with the highest evolutionary split statistic to split protein families into functional subfamilies [25].

2.3.2.6 Acyl transferases *(AT)*

The AT domains of Type I modular polyketide synthases are responsible for the substrate selection. Most incorporate either a C2 unit (malonyl-CoA substrate) or a C3 unit (methylmalonyl-CoA substrate). The choice of substrate can be deduced from the chemical structure of the polyketide product [25]. In the acyl transferase dataset, 99 of the 177 sequences use C2 units whereas 78 use C3 units as substrate.

Previously, Goldstein *et al* [25] used evolutionary split statistic and clustered the AT domains into 2 subfamilies with 2 false assignments for the 5 residue-long motif. The number of false assignments increased to 5 with increasing motif length (up to 30-residue long) suggesting that the utilization of a larger motif increases the noise and

error rate. As such, inclusion of only 5 residues (less noise) with high split statistics increases the assignment accuracy (5 vs. 2 false assignments).

A similar trend is observed in the case of RCM. While the TBC accuracy for AT domains was only 91% (15 false assignments) with the 20-letter alphabet (Table 2.4), the accuracy increased to 97% (5 false assignments) with the utilization of the ML4, ML8, EB9, ML10, EB11, SDM12, EB13, and HSDM17 alphabets. Furthermore, 4 of the 5 misclassified sequences using the above reduced alphabets are contained in the 2, 3 and 4 false assignments produced by the Goldstein *et al*'s approach using the 5,10 and 15 residue-long motifs, respectively. Although the accuracy was higher previously, it should be noted that the RCM approach did neither require an MSA of sequences nor any other sequence-based statistics. The accuracy was 97.2% for MSA using the top performing RAAAs. There was no immediate evidence suggesting a specific characteristic for incorrectly classified sequences.

2.3.2.7 Glycoside hydrolase family 2 *(GH2)*

The final dataset contains 33 members of the GH2 family with a $(\beta/\alpha)_8$ fold. The subfamilies and the number of sequences from each subfamily are β-galactosidases (6), β-mannosidases (12), β-glucuronidases (7) and exo-β-D-glucosaminidases (8). This dataset was used previously and chosen because it was cited as a "hard-to-align" dataset by classical alignment approaches [3]. The GH2 family was clustered into 4 functional subfamilies with 100% accuracy using ML4 and GBMR4 – the two top performing RAAAs – with RCM (Table 2.4). TBC accuracy was 100% for all RAAAs tested with MSA.

2.3.3 The effect of the size of the RAAA on clustering performance

The comparison of RCM with MSA in terms of TBC accuracy and the percentage of TBC error are summarized in Table 2.4 for the 20-letter alphabet and the top performing RAAA with the minimum size. In cases where two RAAAs of the same size give identical TBC results, both of them are reported. Three trends can be observed from the data in Table 2.4.

Table 2.4 TBC errors for top performing RAAA

TBC accuracy and percentage of TBC error are reported for the 20-letter alphabet and the top performing RAAA. If two RAAAs with the same size have identical TBC accuracies, both RAAAs are reported at the final row in Table 2.4. Bold entries correspond to top performers using RCM and MSA for the specified datasets

		Crotonases		Mandelate racemases		Vicinal oxygen chelates		Haloacid dehalogenases		Nucleotidyl cyclases		Acyl transferases		GH2 hydrolases	
		RCM	MSA	RCM	MSA	RCM	MSA	RCM	MSA	RCM	MSA	RCM	MSA	RCM	MSA
20 letter	Accuracy	100	100	100	100	91.6	91.3	93.3	99.5	100	100	91.5	97.2	87.9	100
	Error	0	0	0	0	8.4	8.7	6.7	0.5	0	0	8.5	2.8	12.1	0
Statistics for top performing RAAA	Accuracy	100	100	100	100	**92.2**	91.3	96.9	**99.5**	100	100	97.2	97.2	100	100
	Error	0	0	0	0	7.8	8.7	3.1	0.5	0	0	2.8	2.8	0	0
Top performing RAAAs	RAAA	GBMR4	ML4 GBMR4	ML4	GBMR4 ML4	EB8	GBMR4 ML4	ML15	ML8	ML4 GBMR4	GBMR4	ML4	ML4 GBMR4	ML4 GBMR4	ML4 GBMR4

First, for five of the seven families (crotonases, mandelate racemases, nucleotidyl cyclases, acyl transferases, and GH2 hydrolases), both methods perform equally well comparably. For VOC, RCM outperforms MSA while for haloacid dehalogenases, MSA slightly outperforms RCM. It is important to note that both VOCs and dehalogenases have the two lowest mean PIDs (12% vs. 14%) and low mean sequence lengths with large standard deviation. Low PID and low sequence length are two features in alignments that render inference of relationship based only on sequence information difficult. Nonetheless, TBC accuracies of both families with their respective top performing RAAAs are comparable to the results obtained from the protein families with higher mean PIDs and longer mean sequence lengths.

Second, either ML4 or GBMR4 is sufficient to obtain high TBC accuracy for all datasets except VOCs and haloacid dehalogenases. Indeed, apart from the aforementioned families, ML4 and GBMR4 can produce either identical or better results than all other alphabets using either RCM or MSA, implying that as little as an alphabet size of 4 would be sufficient to capture most of the sequence information that might yield considerable improvements in inferring relationship based on sequence information when both mean PID and the length of the aligned regions in an MSA is above a certain threshold.

Third, for the datasets with low mean PIDs and average sequence lengths, a larger RAAA size may be required to obtain identical or better results than the 20-letter alphabet using both RCM and MSA. This is especially evident with the RCM approach. While the minimum RAAA size of the top performer was 4 for 5 datasets that have relatively higher average sequence lengths and mean PIDs, it increases to 8 (EB8) for VOCs and 15 (ML15) for haloacid dehalogenases that have mean PIDs of 14% and 12%, respectively. Moreover, a subtle but a similar trend is also evident in the case of MSA. While the alphabet size of the top performer was 4 (GBMR4, ML4) for VOCs, it increased to 8 (ML8) for haloacid dehalogenases, implying that a larger RAAA size may perform better on sequences with lower sequence identities.

It is also interesting to note that the average TBC error for mandelate racemases, nucleotidyl cyclases and hydrolases with three random alphabets of size 4 varied between 0% and 15.6% for the MSA method. While the groupings of amino acids in the random alphabets do not have any physicochemical or structural significance that can justify this overall performance, the low percent TBC error may suggest that some

subfamilies of these protein families may be very tight with small distances between their sequences while larger distance between different subfamilies. This scenario coupled with the relatively longer sequences (top three families in terms of mean sequence length) within these families may generate sufficiently long aligned regions with enough informative sites that can result in a tree that correctly assigns subfamilies even the reduced alphabet groupings do not have any structural or biological meaning.

However, the trend of low TBC error is not apparent using RCM with random alphabets. TBC errors of different protein families using random RAAAs (average of three random alphabets) were significantly higher than TBC errors using biologically meaningful reduced alphabets for all the families except racemases and nucleotidyl cyclases, both of which overlap with the results obtained with MSA.

Performance of RCM approach with different RAAAs to cluster protein families into functional subfamilies is eminent. Yet, it must be noted that there is no uniformly superior algorithm for tree-based subfamily clustering and that simple protein similarity measures combined with hierarchical clustering produce trees with reasonable and often high accuracy. Furthermore, if much time has passed since the evolution of different subfamilies, then sequences may have diverged beyond the point where simple phylogenetic analysis cannot easily give a clear distinction of subfamilies.

2.4 Conclusions

The application of RCM in generating meaningful phylogenetic trees has been previously tested on genomic sequences and made RCM a good alternative to MSA-based phylogenetic analysis. However, integration of RCM to measure the closeness of protein sequences was simply problematic due to the lack and difficulty of accounting for amino acid substitutions. In this chapter, we introduced an RAAA-based approach as a preprocessing of protein sequences prior to calculating pairwise RCMs. Utilization of an RAAA that is consistent with the structure and function of the proteins or an RAAA that reflects the general trends in specific protein families under study can result in successful phylogenies that can cluster each protein superfamily into functional subfamilies.

In finding functional subtypes of a protein family, it is often of interest to find out if the mechanisms that manipulate a certain clustering are of evolutionary or functional origin. Although these two signals may be overlapping and hard to separate, RCM could be used to address this issue by finding differences in exhaustive histories in two sequences when they are concatenated. The "words" that result in an observed difference can then be analyzed and correlated to a functional and/or evolutionary origin. We believe future work can focus in this direction building on the current approach that does not attempt to trace back the origin of differentiating sequence signals but provides a powerful clustering method of protein families into functional subtypes without using multiple sequence alignment.

2.5 References

1. Wallace IM, Higgins DG: Supervised multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinformatics* 2007, 8:135.
2. Georgi B, Schultz J, Schliep A: Partially-supervised protein subclass discovery with simultaneous annotation of functional residues. *BMC Structural Biology* 2009, 9:68.
3. Kelil A, Wang S, Brzezinski R, Fleury A: CLUSS: clustering of protein sequences based on a new similarity measure. *BMC Bioinformatics* 2007, 8:286.
4. Lazareva-Ulitsky B, Diemer K, Thomas PD: On the quality of tree-based protein classification. *Bioinformatics* 2005, 21(9):1876-1890.
5. Wicker N, Perrin GR, Thierry JC, Poch O: Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol Biol Evol* 2001, 18(8):1435-1441.
6. Brocchieri L: Phylogenetic inferences from molecular sequences: review and critique. *Theoretical Population Biology* 2001, 59(1):27-40.
7. Baldauf SL: Phylogeny for the faint of heart: a tutorial. *Trends in Genetics* 2003, 19(6):345-351.
8. Otu HH, Sayood K: A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* 2003, 19(16):2122-2130.
9. Felsenstein J: Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 1981, 17(6):368-376.
10. Nei M: Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics* 1996, 30:371-403.
11. Hannenhalli SS, Russell RB: Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology* 2000, 303(1):61-76.
12. Brown DP, Krishnamurthy N, Sjolander K: Automated protein subfamily identification and classification. *PLoS Computational Biology* 2007, 3(8):e160.
13. Ziv J, Lempel A: A universal algorithm for sequential data compression. *IEEE Trans Inf Theory* 1977, 23:337-343.
14. Bastola DR, Otu HH, Doukas SE, Sayood K, Hinrichs SH, Iwen PC: Utilization of the relative complexity measure to construct a phylogenetic tree for fungi. *Mycol Res* 2004, 108(Pt 2):117-125.
15. Liu N, Wang T: Protein-based phylogenetic analysis by using hydropathy profile of amino acids. *FEBS Lett* 2006, 580(22):5321-5327.
16. Russell DJ, Otu HH, Sayood K: Grammar-based distance in progressive multiple sequence alignment. *BMC Bioinformatics* 2008, 9:306.
17. Wang J, Wang W: A computational approach to simplifying the protein folding alphabet. *Nature Structural Biology* 1999, 6(11):1033-1038.
18. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG: A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J* 2007, 36(8):1059-1069.
19. Li T, Fan K, Wang J, Wang W: Reduction of protein sequence complexity by residue grouping. *Protein Eng* 2003, 16(5):323-330.
20. Fletcher W, Yang Z: INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* 2009, 26(8):1879-1888.

21. Kosiol C, Goldman N: Different versions of the Dayhoff rate matrix. *Mol Biol Evol* 2005, 22(2):193-199.
22. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al*: Clustal W and Clustal X version 2.0. *Bioinformatics* 2007, 23(21):2947-2948.
23. Eddy SR: Profile hidden Markov models. *Bioinformatics* 1998, 14(9):755-763.
24. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC: Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry (Mosc)* 2006, 45(8):2545-2555.
25. Goldstein P, Zucko J, Vujaklija D, Krisko A, Hranueli D, Long PF, Etchebest C, Basrak B, Cullum J: Clustering of protein domains for functional and evolutionary studies. *BMC Bioinformatics* 2009, 10:335.
26. Strelets VB, Shindyalov IN, Lim HA: Analysis of peptides from known proteins: clusterization in sequence space. *J Mol Evol* 1994, 39(6):625-630.
27. Dill KA: Theory for the folding and stability of globular proteins. *Biochemistry* 1985, 24(6):1501-1509.
28. Murphy LR, Wallqvist A, Levy RM: Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng* 2000, 13(3):149-152.
29. Prlic A, Domingues FS, Sippl MJ: Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng* 2000, 13(8):545-550.
30. Solis AD, Rackovsky S: Optimized representations and maximal information in proteins. *Proteins* 2000, 38(2):149-164.
31. Munoz E, Deem MW: Amino acid alphabet size in protein evolution experiments: better to search a small library thoroughly or a large library sparsely? *Protein Eng Des Sel* 2008, 21(5):311-317.
32. Lau KF, Dill KA: A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 1989, 22(10):3986-3997.
33. Peterson EL, Kondev J, Theriot JA, Phillips R: Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics* 2009, 25(11):1356-1362.
34. Lempel A, Ziv J: On the Complexity of Finite Sequences. *IEEE Trans Inf Theory* 1976, 22(1):75-81.
35. Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987, 4(4):406-425.
36. Felsenstein J: PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989, 5:164-166.
37. Holmes S: Bootstrapping Phylogenetic Trees: Theory and Methods. *Stat Sci* 2003, 18(2):241-255.
38. Gerlt JA, Babbitt PC: Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* 2001, 70:209-246.

CHAPTER 3

3 DISCRIMINATION OF THERMOPHILIC AND MESOPHILIC PROTEINS USING REDUCED AMINO ACID ALPHABETS WITH N-GRAMS

3.1 Introduction

Proteins undertake many processes under physiological conditions that vary significantly for different organisms. Some of those conditions are considered extreme because the majority of proteins may not function properly due to increased irreversible unfolding rate under those conditions. Proteins have evolved to adapt to those conditions by making adjustments at different levels of the protein structural hierarchy. Currently, there is a growing interest to understand the mechanisms of adaptation to high temperatures by comparative analysis of proteins from heat-tolerant and heat-sensitive microorganisms. The mechanisms that result in an observed difference in thermostability of the proteins from such organisms can then be analyzed and used to design proteins with improved thermal properties and predict the thermostability class of a novel protein from its sequence or structure.

Microorganisms can be separated into four classes based on their optimum growth temperatures (T_{opt}): psychrophiles have T_{opt} of less than 15°C; mesophiles have T_{opt} in the range of 15 - 45°C; thermophiles have T_{opt} in the range of 45-80°C and hyperthermophiles with a T_{opt} above 80°C. Slightly different breakpoint regions for thermostability classes were also used in the literature. Throughout this article, a protein will be called mesophilic if it is from a mesophilic organism and thermophilic if it is from a thermophilic or hyperthermophilic organism.

Generally, proteins of mesophiles are considered as mesophilic and thermophiles as thermophilic. However, certain proteins that have been isolated from thermophiles are known to operate at temperatures that are well above the T_{opt} of their host organisms. For instance, *Pyrococcus furiosus* amylopullulanase is optimally active at

125°C, which is 27°C above the host organisms T_{opt} of 98°C [1]. The existence of such thermophilic proteins with elevated melting temperature (T_m) also has theoretical support from the equation, $T_m = 24.4 + 0.93\ T_{env}$ [2] that relates the T_m of a protein to the environmental temperature (T_{env}) of the host organism.

Current bioinformatics research on protein thermostability can be divided into two broad categories. In the first category, proteomic data from mesophiles and thermophiles are analyzed to discover discriminative patterns [3-13]. In the second category, homologous proteins from mesophiles and thermophiles are compared based on their sequential and structural features to understand specific underlying factors for the thermostabilization of the thermophilic homologs [5, 12, 14-18]. In general, the results of the first category can be used to understand generic properties of proteins from different thermostability classes. The results of the second category can be used to design mesophilic proteins with increased thermostability by mimicking the thermophilic homolog. A successful strategy for rational thermostable enzyme engineering should use a combination of these two approaches by observing the general trends conferring thermostability and simultaneously fine-tuning the protein based on its immediate homologous partners from thermophiles.

Rules obtained from comparison of non-homologous thermophilic and mesophilic proteins do not necessarily correlate well with the results of the comparison of homologous protein pairs and *vice versa*. For example, according to the study of Karshikoff and Ladenstein [19] and more recently Taylor and Vaisman [5], there is no significant difference in packing densities of non-homologous thermophilic and mesophilic proteins. Yet, an increase in the packing density due to an increase in Ile content was suggested by Britton *et al* [20] for the thermostabilization of *Pyrococcus furious* GDH compared to its mesophilic homolog from *Clostridium symbiosum*. In the next section, bioinformatical research examples on protein thermostability are summarized in a non-exhaustive manner.

Discrimination of proteins from different thermostability classes using sequence-based features was successfully carried out on various datasets and most of the results either overlap or encompass one another. For example, Gromiha *et al* [4] reported that the composition of charged residues Lys, Arg, Glu, Asp and hydrophobic residues Val, Ile are higher in thermophiles and Ala, Leu, Gln, Thr are higher in mesophiles based on the evaluation of the discriminative power of amino acid composition by using different

machine learning algorithms. Zeldovich *et al* [6] surveyed a total of 204 complete archaea and bacteria proteomes and showed that the total number of Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL) amino acids correlates well with the optimal growth temperature of the source organisms ranging from 10°C to 110°C. Kumar *et al* [15] performed a statistical analysis of 18 thermophilic and mesophilic protein homologs and reported that the number of salt-bridges and hydrogen bonds between side chains are increased in thermophiles. They have also shown that Arg and Tyr are more and Cys and Ser are less frequent in the thermophilic homologs. Yokota *et al* [21] also carried out a comparative statistical analysis on 94 mesophilic and thermophilic protein homologs and reported that the thermophilic proteins favor a higher frequency of Arg, Glu, Tyr and a lower frequency of Ala, Ser, Met and Gln residues at the protein surface. Taylor and Vaisman [5] tested various sequence based indices and Delaunay tessellation based descriptors. Delaunay tessellation of a protein structure refers to the representation of a protein where each amino acid is abstracted to a set of points (i.e., Cα atom coordinates) to generate non-overlapping, space-filling irregular tetrahedra that uniquely defines four nearest neighbor Cα atoms (i.e., four nearest-neighbor amino acid residues). They have shown that sequence-based indices such as IVYWREL and CvP bias (defined as the difference between charged, DEKR and polar, NQST residues [22]) are better discriminators of thermophilic and mesophilic proteins and the strongest contributors to thermostability is an increase in surface ion pairs and more hydrophobic protein core

Meanwhile, different studies have been devoted to grouping amino acids based on shared physicochemical and/or structural features [23-31]. A reduced amino acid alphabet (RAAA) contains different levels of amino acid grouping to account for the degeneracy of amino acid sequences which yield to only a limited number of folds, domains, and structures. RAAAs were used extensively in the Hydrophobic-Polar (HP) lattice model [31] to explain the hydrophobic collapse theory of protein folding and were shown to improve accuracy in fold prediction between protein sequence pairs with high structural similarity and low sequence identity [32].

In Chapter 2 [33], we have shown that RAAAs can be used to cluster protein families into functional subtypes with equal or better accuracy than the native alphabet. We also suggested that for the clustering of protein families with relatively high sequence similarity, a smaller size of RAAA may be sufficient to correctly cluster

protein sequences into corresponding subtypes with high accuracy, thereby enabling faster computation.

In this chapter, we systematically evaluated 65 different RAAAs with three different n-grams (subsequences of length n) combinations in the classification of protein sequences from thermophiles and mesophiles using support vector machines. A t-test based feature selection procedure was applied to reduce the number of features in a given feature vector. Classification using RAAAs with 1-grams and 2-grams gave better accuracies than with 3-grams. In most cases, a smaller RAAA size was sufficient to get the same level of accuracy as the native alphabet.

3.2 Methods

3.2.1 Datasets

Two different datasets were used in this study. Training and test sets were adapted from Gromiha *et al* [4]. The training set contains 1609 thermophilic and 3075 mesophilic sequences belonging to 9 and 15 organisms, respectively. Training set contains 8 protein sequences with unknown residues (ie, "X" residue). For those sequences, Uniprot database was checked to see if there exists an update for the unknown residues. If an update was available, it was incorporated into the sequences. For other sequences, unknown residues were simply discarded (a total of 5 residues).

The test set contains 707 protein sequences with 325 belonging to mesophilic *Xylella fastidosa* and 382 to thermophilic *Aquifex aeolicus*. Number of sequences, average length, standard deviation of sequence lengths, mean percent identities (μPID), and maximum pairwise identities of all sequences in these datasets are summarized in Table 3.1. μPID was calculated using the pairwise identity scores obtained from the result of Needleall many-to-many pairwise alignment script available in EMBOSS [34] suite and reported only for the test set. This is because μPID calculation requires summation of all pairwise sequence identities divided by the total number of such pairs. Calculation of μPID for the training set is rather impractical considering that there are 10,967,586 (4684*4683/2) possible pairwise alignments.

In addition to μPID values, we also report that no sequence pairs in any of the classes of the training or test datasets contain more than 50% sequence identity based on the results of the CD-HIT [35] sequence redundancy search algorithm.

It is a general practice to remove sequence redundancy at a predefined similarity threshold in many bioinformatical analyses. According to previous authors, no sequences in either the training set or the test set have more than 40% sequence identity. We checked, using CD-HIT suite, thermophilic training, mesophilic training, thermophilic test, and mesophilic test datasets and found that maximum sequence identities between any two sequences were indeed close to 40% for each of these individual datasets (see below). The slight differences in max identity values between the current and previously reported study may arise from different global alignment parameters used to determine pairwise sequence identities. Moreover, maximum

sequence identity between thermophilic sequences in the training and test set was 75% and between mesophilic sequences in the training set and test set was 76%.

Table 3.1 General properties of datasets

		# of sequences	μ length	σ length	Max % identity	μPID (%)
Training Set	Mesophilic	3075	339	225	40	--
	Thermophilic	1609	326	225	42	
Test Set	Mesophilic	325	358	209	47	8.40
	Thermophilic	382	349	204	50	

Furthermore, we calculated the maximum percent identities between the sequences of the test set and the training set for each class and reported these identities in Table 3.2. A 50% max identity cutoff would have eliminated only 36 sequences from thermophilic test set and 60 sequences from mesophilic test set.

Table 3.2 Maximum identity values between training and test sets

Dataset 1	Dataset 2	Max % Identity
Mesophilic Training	Mesophilic Test	76
Thermophilic Training	Thermophilic test	75

3.2.2 RAAA

We adopted the same approach as Peterson's [32] in naming the RAAAs. For a given RAAA, if a name is provided by the authors, it has also been used here; otherwise first letters of the names of first and last authors were used as abbreviations. The numerical value next to the letters of a RAAA corresponds to the size of the RAAA and only sizes larger than 10 were included in this work. The reason for the exclusion of smaller sized RAAAs was two-fold. First, μPID of the test set is very low which implies that each amino acid is highly informative. Using a small-size alphabet would mask the informative sites to the extent that no clear distinction can be made between sequences of different classes. In Chapter 2, we have also shown that using a larger RAAA size produces better accuracy for sequences with low μPID values. Second is the obvious

computational cost of generating feature vectors for sequences recoded with smaller-sized RAAAs and training LibSVM classifiers.

We also generated a random RAAA to determine whether RAAAs are biologically relevant and useful in classification or stochastic manifestations in a noisy data. A list of all RAAAs is provided in Table 3.3 while the amino acid groupings are provided in Appendix B.

Table 3.3 Reduced Amino Acid Alphabets

Alphabet	Size	Reference
Native	20	
Ab	10-19	[23]
Dssp	10-14	[30]
Eb	11, 13	[24]
Gbmr	10-14	[30]
Hsdm	10,12,14-17	[29]
Lr	10	[25]
Lwi	10-19	[26]
Lwni	10,11,14	[26]
Lzbl	10-16	[27]
Lzmj	10-16	[27]
Ml	10,15	[28]
Sdm	10-14	[29]
Random	10	This study

3.2.3 N-grams

N-grams are sequences of n amino acids in a sliding window over the length of the protein sequence [36]. In a biological context, n-grams where n is equal to 1, 2, and 3 correspond to amino acid, dipeptide and tripeptide compositions, respectively. Given the pentapeptide sequence "AYDIN", there is one count each of 2-grams AY, YD, DI, and IN. N-gram frequency is simply the number a particular n-gram divided by the total number of all n-grams in a given sequence. For example, frequencies of each of the above 2-grams would be 0.25 since there is one count for each 2-grams and there are a total of 4 such 2-grams. The formal definition of n-grams is given below.

Definition:

Given a sequence of N letters $S = s_1 s_2 ... s_N$ over the alphabet A, and n a positive integer, an n-gram of the sequence S is any subsequence $s_i ... s_{i+n-1}$ of n consecutive letters. There are N-n+1 such n-grams in S. For an alphabet A with |A| distinct letters, there are $|A|^n$ possible unique n-grams.

3.2.4 Curse of dimensionality

The curse of dimensionality refers to the problems associated with high dimensional feature space given a limited number of data samples. The problem can be illustrated as follows: let us assume that we have ten samples and one feature and the complete probability space of this feature is represented by the unit interval (0, 1), and each one of the 10 sample points equally represents 10% of the probability space (Table 3.4) .

Table 3.4 Probability space of ten samples with one feature

■	■	■	■	■	■	■	■	■	■
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

When a second feature with values in the same (0, 1) interval and represented by the same ten samples is added, 10 points on a two dimensional space representing a new probability space (Figure 3.1) are produced. Since the new space has 10 x 10 = 100 area units, each of the ten points now represents only 1% of the probability space. Therefore, 100 samples would be required for each point to represent the same 10% of the probability space that was represented by 10 points in only one dimension [37].

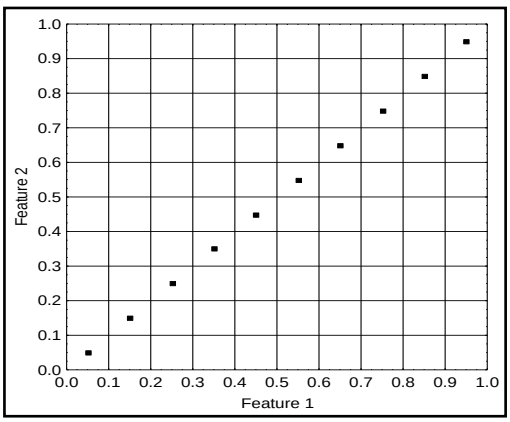

Figure 3.1 Probability space of ten samples with two features

Consequently, increasing the size of the feature space by adding more features reduces the coverage of the probability space thereby reducing accuracy. It is obvious that10^n samples would be required for an n-dimension problem. This is called the "curse of dimensionality" and places a practical limit above which additional features result in decreased accuracy as illustrated in Figure 3.2.

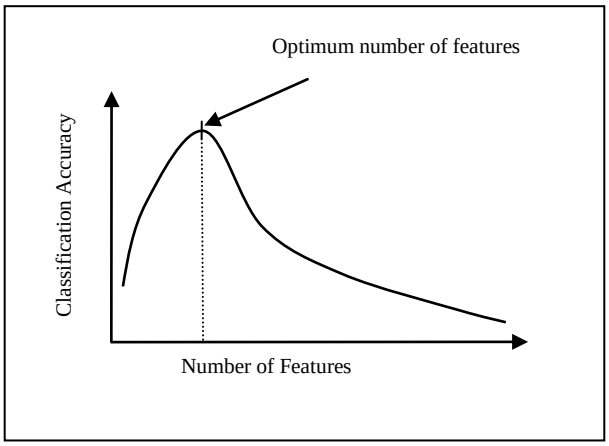

Figure 3.2 Effect of increasing feature size on classification accuracy

To minimize the effects of high dimensionality associated with the features (especially 2-grams and 3-grams) used in this chapter, we employed a dimensionality reduction or equivalently a feature selection procedure based on two-sided t-test as described in the next section.

3.2.5 T-test based feature reduction

Each protein sequence in the training set was transformed into a feature vector for each RAAA and n-gram combination. Two-sided t-test was performed at the 0.01 significance level. Dunn-Bonferroni correction was applied to the significance level to account for multiple comparisons by simply dividing the significance level by the size of the feature vector. For example, there are 20 features for the 20 letter native amino acid alphabet and the significance level would be set to $\alpha = 0.01/(2*20)$. The extra division by a factor of two was to account for the two sided t-test because according to the null-hypothesis, the mean of a given feature in thermophiles may be larger or smaller than the mean of the same feature in mesophiles.

3.2.6 SMOTE Sampling

Performance of machine learning algorithms is typically evaluated using predictive accuracy. However, this is not appropriate when the data is imbalanced and/or the costs of different errors vary markedly. As an example, consider the classification of a dataset with two classes 100 data points with 95% belonging to a negative class and 5% belonging to positive class. A simple default strategy of guessing the majority class would give a predictive accuracy of 95%. However, the nature of some applications requires a fairly high sensitivity for the detection of the minority class and allows for a small error rate in the majority class in order to achieve this. Simple predictive accuracy is clearly not appropriate in such situations.

Therefore, we used Synthetic Minority Over-sampling Technique (SMOTE) [38] to balance the size of the thermophilic and mesophilic protein classes in the training set. SMOTE, which is available in Weka [39] software, improves classifier performance by using a combination of over-sampling the minority class and under-sampling the majority class. In SMOTE, synthetic samples are created for the minority class as follows: Randomly select a sample from the minority class; find its nearest neighbor (or one of its *k* nearest neighbors); take the difference between the feature vector of the sample under consideration and its nearest neighbor; multiply the difference by a random number that is between 0 and1; and add it to the feature vector under consideration to create a synthetic sample.

3.2.7 Classification

3.2.7.1 Support vector machines (SVM)

An SVM machine performs classification by constructing an N-dimensional hyperplane to optimally separate the data into different categories. In SVM literature, a predictor variable is called an attribute, and a transformed attribute that is used to define the hyperplane is called a feature. A set of features that describes one case (i.e., a row of feature values) is called a feature vector. The goal of SVM is to find the optimal hyperplane that separates clusters of vector in such a way that cases that belong to one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyperplane are the support vectors.

3.2.7.2 Classification using LibSVM

Classification was carried out using WLSVM [40], a LibSVM [41] classifier interface for the widely distributed Weka (v3.6.3) [39] data mining software. The classifier was trained using five-fold cross validation on the normalized training set with RBF kernel-C-SVC, C=100, and ε=0.09 to generate a model.

In five-fold cross validation, the training set is randomly partitioned into five roughly equal-sized parts. Of the 5 parts, 4 parts are used as training data and the remaining single part is retained as the validation data for testing the model. The cross-validation process is then repeated 5 times, with each of the 5 parts used exactly once as the validation data. Although the performance of the classifier is evaluated using cross-validation, Weka outputs a model built from the full training set and that model is used to test on the normalized test set.

3.2.8 Performance Evaluation

Classifier performance was assessed by calculating sensitivity, specificity, accuracy, and area under the Receiver Operator Characteristic (ROC) curve (AUC) using the following equations;

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP are true positives (thermophilic proteins predicted as thermophilic); FN are false negatives (thermophilic proteins predicted as mesophilic); TN are true negatives (mesophilic proteins predicted as mesophilic) and FP are false positives (mesophilic proteins predicted as thermophilic).

In the current context, *sensitivity* refers to the number of correctly classified thermophilic proteins divided by the total number of thermophilic proteins; *specificity* is the number of correctly classified mesophilic proteins divided by the total number of mesophilic proteins; *accuracy* corresponds to the total number of correctly classified thermophilic and mesophilic proteins divided by the total number of thermophilic and mesophilic proteins.

AUC values was obtained using Weka [39] software. The top three performing RAAAs (with minimum alphabet size) in terms of classification accuracy were reported in Table 3.5. Classification results in terms of sensitivity, specificity, accuracy, and AUC for the test set with different n-grams and RAAAs were reported in Appendix C.

Table 3.5 Classification performance of the top three performing RAAAs
Top three performing RAAAs in terms of classification accuracy with the corresponding AUC, sensitivity and specificity values are reported for each n-grams

N-gram	RAAA	Features	Accuracy	AUC	Sensitivity	Specificity
Amino Acid (1-grams)	Hsdm16	13	**91.796**	0.960	0.921	0.914
	Lwi19	16	91.513	0.957	0.921	0.908
	Hsdm17	14	91.372	0.958	0.921	0.905
	Native	17	91.372	0.956	0.919	0.908
Dipeptide (2-grams)	Lwi18	158	**91.513**	0.965	0.906	0.926
	Hsdm17	141	91.089	0.962	0.893	0.932
	Ml15	120	90.806	0.955	0.898	0.920
	Native	190	90.806	0.965	0.887	0.932
Tripeptide (3-grams)	Sdm12	227	**88.826**	0.949	0.882	0.895
	Sdm11	220	88.543	0.952	0.882	0.889
	Sdm13	235	88.401	0.950	0.866	0.905
	Native	351	83.451	0.906	0.793	0.883

3.2.9 Protocol

After one of the alphabets given in Table 3.3 is applied to all the sequences in the training set, frequencies of 1-grams, 2-grams and 3-grams were calculated for each sequence. Features in an n-gram that are statistically significant were selected after performing a two-sided t-test on the "training set" and only those significant features were calculated for the test set. SMOTE sampling procedure was performed on the training set to balance the number of instances in each class using Weka [39]. A classification model for each RAAA and n-gram combination was generated by the LibSVM classifier using the training set. The classifier was tested on the test set using the model to determine how well it classified protein sequences to different thermostability classes. A summary of the overall workflow is also depicted in Figure 3.3.

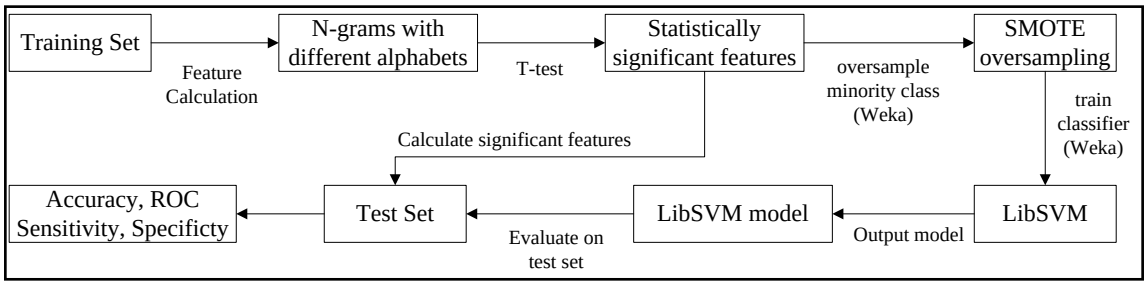

Figure 3.3 Overall workflow of the protocol

3.3 Results and Discussion

We have computed the reduced amino acid composition with three different n-gram sizes for thermophilic and mesophilic proteins. We have used a t-test based feature selection procedure to reduce the number of features that can be used to represent a protein sequence in feature space prior to generating a model using LibSVM classifier to predict the thermostability class of a protein. Based on the results reported in Table 3.5, it is clear that 1-grams are generally better predictors of thermostability than 2-grams and more so than 3-grams in terms of classification accuracy. In the following two sections, more in depth analysis was carried out to highlight the effects of n-gram and RAAA sizes on classification accuracy.

3.3.1 Effects of n-gram size on classification accuracy

The best discriminatory alphabet for 1-grams was Hsdm16 which showed 91.796% accuracy. The feature vector of this alphabet has only 13 features out of 16 possible features. The features that were included in this alphabet were [AGFHKMLNQRTWY]. K corresponds to negatively/positively-charged (EK) cluster; L corresponds to aliphatic (ILV) cluster and T corresponds to (ST) cluster. Lwi19 and Hsdm17 were the other top performers. Lwi19 contains 16 features which includes (IV) cluster whereas Hsdm17 contains 14 features which includes (EK) and (ILV) clusters. Hsdm17 can be derived from Hsdm16 by breaking the (ST) cluster and Lwi19 by breaking the (EK) and (ILV) clusters. Hsdm17, which has an accuracy as good as the native alphabet, was also one of the top three performers in the work of Peterson *et al* [32] and was shown to improve classification accuracy in fold recognition prediction.

Lwi18 was the top performing alphabet for 2-grams with 91.513% accuracy. The feature vector of this alphabet has 158 features out of 324 (i.e., 18^2) possible features. Lwi18 contains the clusters of aliphatic (IV) and aromatic (FY) residues. Hsdm17 and Ml15 were the other top performers. Ml15 contains aromatic (FY), positively-charged (KR) and aliphatic (ILVM) clusters. Classification accuracy of the native alphabet was 90.81%.

The best discriminatory alphabet for 3-grams was Sdm12 with 88.826% accuracy. Sdm11 and Sdm13 were the other top performers. There was a dramatic decrease in the number of features of 3-grams because only 13.1, 16.5 and 10.6% of all possible 3-grams were used for Sdm12, Sdm11, and Sdm13 alphabets, respectively.

In general, accuracy of a given RAAA decreases with increasing n-gram size. For 32 out of 64 RAAAs (excluding the random alphabet), 1-grams yield better accuracy than 2-grams and for 58 RAAAs 2-grams yield better accuracy than 3-grams. Decrease in accuracy for higher n-gram sizes is a weak manifestation of high dimensional feature space. Given a constant number of sequences, as the number of features or dimensions increase, the sparsity increases exponentially [42] and leads to redundancy in feature values (i.e., many features will have very similar values) and smaller distances between sequences [43]. This phenomenon makes it difficult to learn from the training set with limited number of sequences and leads poor classification performance. The lower accuracy of native alphabet with 3-grams compared to Sdm12 with 3-grams is a clear indication of negative effects of high dimensionality causing low classification accuracy for the native alphabet.

3.3.2 Effect of feature reduction through T-test on classification time

In a classification problem, it is often of interest to find the minimum number of features that can be used to separate a test dataset into corresponding classes with high sensitivity and specificity. However, increasing the number of features for a given dataset increases the classification accuracy up to an optimum number of features and then decreases. For all practical purposes, there is a trade-off between classification accuracy and the number of features. Too few features will not yield good classification accuracy as too many features.

In this current context, we used t-statistic to discard the features that are not statistically significant between thermophilic/mesophilic and hyperthermophilic/mesophilic protein sequences. The reduction in the feature space was more pronounced especially for the 3-grams composition. Without a feature reduction step, it would have required the calculation of 8000 (20^3) triplet frequencies and classification of the dataset using all the frequencies. However, for such a classification problem using LibSVM, the computation time is proportional to the square of the

feature space. In other words, doubling the number features for a given dataset would quadruple the computational time of the classification algorithm.

3.3.3 Effect of RAAA size on classification accuracy

In Chapter 2, we have shown that a smaller size alphabet is sufficient to obtain a classification accuracy that is identical or better than native alphabet in clustering protein families into functional subtypes. This trend was also observed in the classification of thermophilic and mesophilic proteins. For all three n-grams, the top performing RAAA gave better results than the native alphabet with less number of features. This trend is especially more pronounced with 3-grams since Sdm11 alphabet that produced the highest accuracy is an 11-sized alphabet. Using all features in Sdm11 alphabet would have meant that the feature space of the Sdm11 alphabet has 1331 features. However, based on t-test, only 227 features were used. Relatively smaller sizes of the top performing RAAAs in 3-grams may be attributable to the clustering of amino acids that make the feature vector less sparse compared to the native alphabet and avoid the negative effects of high dimensionality in feature space.

It is also interesting to note that the classification accuracy of the random alphabet was 76.09%. The grouping of amino acids in the random alphabet does not have any physicochemical or structural significance. Out of 10 different alphabets of size 10 used in 1-grams, Random10 produced the lowest accuracy compared to all other RAAAs. Moreover, in terms of accuracy, Random10 came amongst the lowest three for all three n-grams.

A recent study [36] revealed that particular n-grams are more abundant in certain organisms than others and may serve as proteomic signatures of those organisms. Organism preference for specific n-grams may indicate that organism- or protein family specific RAAAs may be prescribed that reflects the prevalent amino acid substitution preference in protein sequence space of an organism in a similar way that codon usage bias reflects genomic tRNA pool of an organism. Indeed, organism-specific RAAAs have not been addressed in the literature and require further research that may have implications for protein thermostabilization and protein function prediction.

3.3.4 Comparison with other methods

Gromiha *et al* [4] previously used different machine learning algorithms on the same test set and achieved overall accuracies of 91.3% and 89.7% with amino acid and dipeptide compositions, respectively. Current work can be considered as an extension to the work of Gromiha *et al* with the intension of decreasing the number of features that can be used to discriminate thermophilic and mesophilic proteins using RAAAs. To that end, accuracies of 91.796% and 91.513% were achieved using 1-grams with Hsdm16 alphabet and 2-grams with Lwi18 alphabet, respectively. The slight differences between accuracies of our works may be the result of using different machine learning algorithms and/or parameters. Nonetheless, performing t-test for feature selection prior to classification and utilizing RAAAs gave similar results to the previous work in terms of accuracy with fewer features.

3.3.5 Benchmark Results

In Table 3.6, computational times and accuracies of five runs of 5-fold cross validation on the training set are reported for native and Sdm12 alphabets with and without feature selection. Both alphabets with feature selection are computationally faster than without feature selection even though the classification accuracies did not change considerably. The reduction in computational time is especially more evident in 3-grams because without a feature selection step it is impossible to perform a 5-fold cross-validation using a PC clocked at 2.13 Ghz. Performing a feature selection step greatly reduced the computational times of 3-grams to the levels comparable to that of 2-grams for both alphabets.

Table 3.6 Benchmark results of 5-fold cross validation with and without feature selection through t-test

		With Feature Selection		Without Feature Selection	
Alphabet	N-gram	Time (*s*)	Accuracy	Time (*s*)	Accuracy
Native	1	84	89.901	90	90.286
	2	380	90.371	619	90.691
	3	264	85.781	--	--
Sdm12	1	57	86.187	77	87.019
	2	294	87.297	418	86.956
	3	512	85.973	--	--

Computational times and accuracies are reported as averages of 5 runs of five-fold cross-validation for each n-grams for the native alphabet and sdm12 RAAA with and without feature selection process. A personal computer with an Intel Celeron processor with 2.13 Ghz speed and 2GB RAM has been used for computations. 3-grams without feature selection could not be calculated due to computational limitations.

3.4 Conclusions

It is possible to accurately discriminate proteins from thermophiles and mesophiles using RAAAs with n-grams. Classification accuracy of RAAA usually decreases with increasing n-gram size and this decrease is more evident in 3-grams. Current approach of systematically using different RAAA and n-gram combinations has produced better results with fewer features than the native alphabet in terms of accuracy.

Our results also indicate that RAAAs can improve classification performance relative to native protein alphabet. Performing t-test to reduce the number of features in the training set also decreases the computational time significantly without significantly affecting classification accuracy and makes classification with 3-grams possible. A future avenue of research in this area may involve carrying out research in generating organism-specific RAAAs, and separating thermostability classes by phyla.

3.5 References

1. Brown SH, Kelly RM: Characterization of Amylolytic Enzymes, Having Both Alpha-1,4 and Alpha-1,6 Hydrolytic Activity, from the Thermophilic Archaea Pyrococcus-Furiosus and Thermococcus-Litoralis. *Appl Environ Microbiol* 1993, 59(8):2614-2621.
2. Gromiha MM, Oobatake M, Sarai A: Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophys Chem* 1999, 82(1):51-67.
3. Ding Y, Cai Y, Zhang G, Xu W: The influence of dipeptide composition on protein thermostability. *FEBS Lett* 2004, 569(1-3):284-288.
4. Gromiha MM, Suresh MX: Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins* 2008, 70(4):1274-1279.
5. Taylor TJ, Vaisman, II: Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol* 2010, 10 Suppl 1:S5.
6. Zeldovich KB, Berezovsky IN, Shakhnovich EI: Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 2007, 3(1):e5.
7. Zhang G, Li H, Gao J, Fang B: [Influence of amino acid and dipeptide composition on protein stability of piezophilic microbes]. *Wei Sheng Wu Xue Bao* 2009, 49(2):198-203.
8. Zhang GY, Fang BS: [A study on the discrimination of thermophilic and mesophilic proteins based on dipeptide composition]. *Sheng Wu Gong Cheng Xue Bao* 2006, 22(2):293-298.
9. Zhao W, Wang X, Deng R, Wang J, Zhou H: Discrimination of Thermostable and Thermophilic Lipases using Support Vector Machines. *Protein Pept Lett* 2011.
10. Kreil DP, Ouzounis CA: Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res* 2001, 29(7):1608-1615.
11. Singer GA, Hickey DA: Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* 2003, 317(1-2):39-47.
12. Cambillau C, Claverie JM: Structural and genomic correlates of hyperthermostability. *J Biol Chem* 2000, 275(42):32383-32386.
13. Zhang GY, Fang BS: Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. *Process Biochem* 2006, 41(8):1792-1798.
14. Ditursi MK, Kwon SJ, Reeder PJ, Dordick JS: Bioinformatics-driven, rational engineering of protein thermostability. *Protein Eng Des Sel* 2006, 19(11):517-524.
15. Kumar S, Tsai CJ, Nussinov R: Factors enhancing protein thermostability. *Protein Eng* 2000, 13(3):179-191.
16. Lehmann M, Loch C, Middendorf A, Studer D, Lassen SF, Pasamontes L, van Loon AP, Wyss M: The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng* 2002, 15(5):403-411.
17. Lehmann M, Pasamontes L, Lassen SF, Wyss M: The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta* 2000, 1543(2):408-415.

18. Szilagyi A, Zavodszky P: Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* 2000, 8(5):493-504.
19. Karshikoff A, Ladenstein R: Proteins from thermophilic and mesophilic organisms essentially do not differ in packing. *Protein Eng* 1998, 11(10):867-872.
20. Britton KL, Baker PJ, Borges KM, Engel PC, Pasquo A, Rice DW, Robb FT, Scandurra R, Stillman TJ, Yip KS: Insights into thermal stability from a comparison of the glutamate dehydrogenases from Pyrococcus furiosus and Thermococcus litoralis. *Eur J Biochem* 1995, 229(3):688-695.
21. Yokota K, Satou K, Ohki S: Comparative analysis of protein thermo stability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Sci Technol Adv Mater* 2006, 7(3):255-262.
22. Cambillau C, Claverie JM: Structural and genomic correlates of hyperthermostability. *J Biol Chem* 2000, 275(42):32383-32386.
23. Andersen CAF, Brunak S: Representation of protein-sequence information by amino acid subalphabets. *Ai Magazine* 2004, 25(1):97-104.
24. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG: A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J* 2007, 36(8):1059-1069.
25. Landès C, Risler J-L: Fast databank searching with a reduced amino-acid alphabet. *Comput Appl Biosci* 1994, 10(4):453-454.
26. Li T, Fan K, Wang J, Wang W: Reduction of protein sequence complexity by residue grouping. *Protein Eng* 2003, 16(5):323-330.
27. Liu X, Liu D, Qi J, Zheng WM: Simplified amino acid alphabets based on deviation of conditional probability from random background. *Phys Rev E Stat Nonlin Soft Matter Phys* 2002, 66(2 Pt 1):021906.
28. Murphy LR, Wallqvist A, Levy RM: Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng* 2000, 13(3):149-152.
29. Prlic A, Domingues FS, Sippl MJ: Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng* 2000, 13(8):545-550.
30. Solis AD, Rackovsky S: Optimized representations and maximal information in proteins. *Proteins* 2000, 38(2):149-164.
31. Lau KF, Dill KA: A Lattice Statistical-Mechanics Model of the Conformational and Sequence-Spaces of Proteins. *Macromolecules* 1989, 22(10):3986-3997.
32. Peterson EL, Kondev J, Theriot JA, Phillips R: Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. *Bioinformatics* 2009, 25(11):1356-1362.
33. Albayrak A, Otu HH, Sezerman UO: Clustering of protein families into functional subtypes using Relative Complexity Measure with reduced amino acid alphabets. *BMC Bioinformatics* 2010, 11:428.
34. Rice P, Longden I, Bleasby A: EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000, 16(6):276-277.
35. Huang Y, Niu B, Gao Y, Fu L, Li W: CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010, 26(5):680-682.
36. Osmanbeyoglu HU, Ganapathiraju MK: N-gram analysis of 970 microbial organisms reveals presence of biological language models. *BMC Bioinformatics* 2011, 12(1):12.

37. Annis C: Curse of Dimensionality.[http://www.statisticalengineering.com/curse_of_dimensionality.htm]
38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* 2002, 16:321-357.
39. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009, 11(1):10-18.
40. EL-Manzalawy Y, Honavar V: {WLSVM}: Integrating LibSVM into Weka Environment.[http://www.cs.iastate.edu/~yasser/wlsvm]
41. Chang C, Lin C: {LIBSVM}: a library for support vector machines.[http://www.csie.ntu.edu.tw/~cjlin/libsvm]
42. Silverman BW: Density estimation for statistics and data analysis. London ; New York: Chapman and Hall; 1986.
43. Verleysen M, François D: The Curse of Dimensionality in Data Mining and Time Series Prediction. In: *Computational Intelligence and Bioinspired Systems.* Edited by Cabestany J, Prieto A, Sandoval F, vol. 3512: Springer Berlin / Heidelberg; 2005: 85-125.

CHAPTER 4

4 STATISTICAL ANALYSIS AND CLASSIFICATION OF PROTEINS FROM DIFFERENT THERMOSTABILITY CLASSES USING SEQUENTIAL AND STRUCTURAL FEATURES

4.1 Introduction

Genome-sequencing initiatives and high-throughput X-ray crystallography technologies have increased the number of completely sequenced genomes, and proteomes to unprecedented levels. Currently, there are 2639 organisms with completely sequenced proteomes (mapped to genomes) containing approximately 5.2 million sequences. Approximately 92.4 % of those sequences have been automatically annotated for functional and structural domains using expert or computer generated rules. These rules are based on empirical and statistical evidence, literature review, and computational algorithms which can detect similarities between proteins, occurrence of structural motifs, domains and other important sites. Surprisingly, of the 5.2 million sequences, only 0.31% have corresponding three dimensional structures available in the Protein Databank (PDB) [1].

Sequencing initiatives and X-ray crystallography technologies generate raw sequence and structure data which need to be analyzed for biological significance and classified into categories for comparative studies. Understanding from such a wealth of biological data could have been a daunting task without the use of bioinformatical and computational tools, and biological databases that have been made available in the last two decades.

Existence of large amounts of biological data requires the undertaking of an equally challenging task of categorization of the data to mine for valuable biological information that is meaningful to researchers. Many different research groups supported

by the need of the scientific community have created databases to represent data in a categorical way facilitating comparative analysis.

After sequencing a novel protein or determining its X-ray structure, it is often interest to assign the protein to an appropriate category to study its biological significance, relevance to existing proteins, existence of catalytic residues, common motifs and domains, and its function within a catalytic or synthetic pathway. All these tasks are considered an integral part of the classification process which requires extraction of features from raw data.

Proteins perform a variety of functions in all living organisms under physiological conditions that vary significantly for different organisms. Environmental factors such as salinity, acidity, basicity and temperature are only some of the conditions that require fine adjustments at different level of protein structural hierarchy. Understanding mechanisms of adaptations to such conditions have both theoretical implications and practical applications. There are genomic, proteomic as well as extracellular components to adaptation and this chapter will focus on proteomic components.

In this chapter, we extracted a comprehensive set of sequential and structural features using protein sequence and structure, respectively by using computer software that was mainly developed in-house. Then, we systematically analyzed the extracted features for their statistical significance between hyperthermophiles and thermophiles compared to mesophiles. Finally, we carried out classification tasks using support vector machines to determine the extent by which those features can be used in a machine learning framework to predict the thermostability class of a protein.

In general, our results indicate that features that are based on RAAAs are better predictors of protein thermostability than Native protein alphabet and structural features. Structural features alone are not as good predictors of protein thermostability as sequential features. Combinations of structural and sequential features are better predictors than purely sequential or structural features.

4.1.1 Thermostability Classes

In Chapter 3, we defined the thermostability class of a protein as the one that was based on the optimum growth temperatures (T_{opt}) of the source organism: psychrophiles have T_{opt} of less than 15°C; mesophiles have T_{opt} in the range of 15 - 45°C; thermophiles have T_{opt} in the range of 45-80°C and hyperthermophiles with a Topt above 80°C. Moreover, in Chapter 3, we grouped hyperthermophilic and thermophilic proteins into one class, namely, thermophilic. However, in this chapter, we will consider hyperthermophiles and thermophiles as distinct classes and will compare proteins from these two classes to a control set of proteins from mesophiles.

4.1.2 Current Research on Thermostability

In Figure 4.1, we present the result of a literature search that reflects the increase in the number of journal articles that contains "Protein Thermostability" in the abstract section of journals that are indexed in PubMed database. It is evident that protein thermostability research has increased dramatically in the last 35 years. The increase in thermostability research is attributable to the need to develop proteins and enzymes with enhanced thermal properties that are demanded by a variety of industries. While thermostability research has been driven mainly by experimental study of thermophilic and mesophilic proteins in a case by case basis in the beginning, the advent of many computational tools and biological databases simultaneously increased bio-data mining related research that use high-throughput data to understand the factors effecting thermal adaptation and to generate classification models that can be used to detect the thermostability class of a protein (e.g., from metagenomic samples) that can be used for downstream computational or experimental analysis.

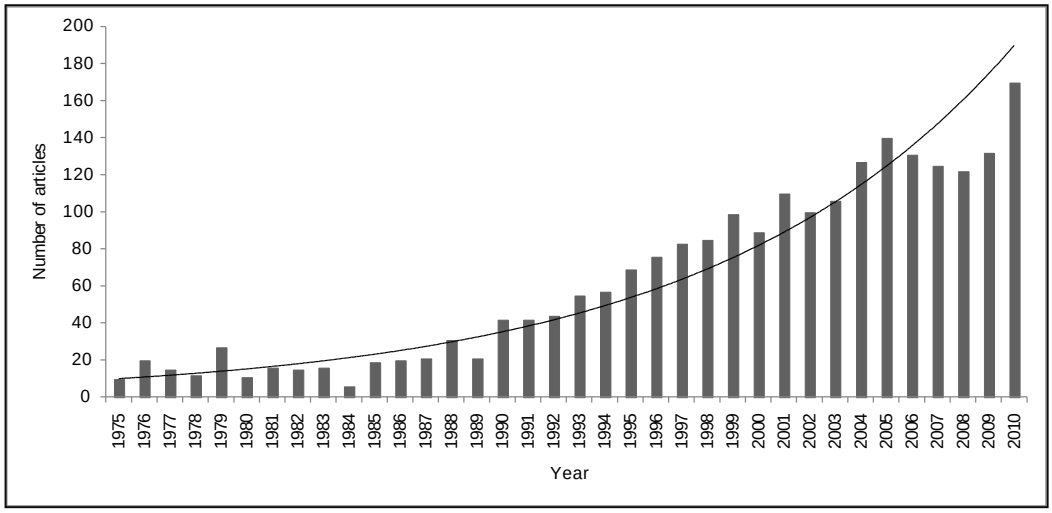

Figure 4.1 Number of articles related to protein thermostability in PubMed

4.1.3 Protein Structural Hierarchy

4.1.3.1 Amino Acids

Proteins are composed of amino acids linked through amide bonds (also called *peptide bond*). The peptide bonded polymer that forms the backbone of polypeptide structure is called the *α-chain* or *main chain*. The peptide bonds of the α-chain are rigid planar units formed by the dehydration reaction of the α-carboxyl of one amino acid with the α-amino group of another releasing one molecule of H_2O in the process. The carbonyl-amino amide bond has partial double bond character and possesses no rotational freedom [2].

The sequence and physiochemical properties of each amino acid ultimately determine protein structure, reactivity, and function. Each amino acid is composed of an amino group and a carboxyl group bound to a central carbon, called the Cα. Also bound to the Cα are a hydrogen atom and a side chain that is unique to each amino acid and contributes to the chemical properties of the protein. There are 20 common (also called *standard or primary*) amino acids found throughout nature, each containing a side chain with particular size, structure, charge, hydrogen bonding capacity, polarity, and reactivity. The side chains are not directly involved in the formation of the polypeptide backbone and are free to interact with their environment [2].

Amino acids may be grouped based on their side chain characteristics. There are seven amino acids that contain aliphatic side chains, which are relatively non-polar and hydrophobic in character: glycine, alanine, valine, leucine, isoleucine, methionine, and proline. Glycine (Gly) is the simplest amino acid with its side chain consisting of only a hydrogen atom. Alanine (Ala) possesses a single methyl group for its side chain. Valine (Val), leucine (Leu), and isoleucine (Ile) are slightly more complex with three or four carbon branched-chain constituents. Methionine (Met) contains a thioether ($-S-CH_3$) group at the terminus of its hydrocarbon chain. Proline (Pro) is actually the only *imino* acid and its side chain forms a ring structure with its α-amino group resulting in two covalent linkages to its Cα atom. Due to its unique structure, Pro often causes severe turns in a polypeptide chain and cannot be accommodated in normal α-helical structures, except at the ends where it may create a turning point for the chain [2].

Phenylalanine (Phe) and tryptophan (Trp) contain aromatic side chains that, like the aliphatic amino acids, are also relatively non-polar and hydrophobic. The presence of an accessible Trp in a protein is significant in that contributes more to its total absorption at 275‑280 nm on a mole-per-mole basis than any other amino acid. The Phe content, however, adds very little to the overall absorbance in this range.

All of the aliphatic and aromatic hydrophobic residues are usually located at the interior of protein molecules or in areas that interact with other non-polar structures. They usually form the hydrophobic core of proteins and are not readily accessible to water or other hydrophilic molecules.

There are four amino acids which have relatively polar side chains and are hydrophilic: asparagine (Asn), glutamine (Gln), threonine (Thr), and serine (Ser). They are usually found in hydrophilic regions of a protein molecule, especially at or near the surface where they can have favorable interactions with the surrounding hydrophilic environment.

There is also another group of amino acids that contain ionizable side chains and also hydrophilic in character: aspartic acid (Asp), glutamic acid (Glu), lysine (Lys), arginine (Arg), cysteine (Cys), histidine (His), and tyrosine (Tyr). Both Asp and Glu contain carboxylate groups with similar ionization properties as the C-terminal α-carboxylate. The theoretical pKa of the β-carboxyl of Asp (3.7-4.0) and the γ-carboxyl of Glu (4.2-4.5) are somewhat higher than the α-carboxyl groups at the C-terminal of a polypeptide chain (2.1-2.4). At pH values above their pKa, these groups are generally ionized to negatively charged carboxylates. Thus at physiological pH, they contribute to the overall negative charge of a protein [2].

Lys, Arginine, and His have ionizable amine containing side chains that, along with the N-terminal α-amine, contribute to a protein’s overall net positive charge. Lys contains an unbranched four-carbon chain terminating in a primary amine group. The theoretical pKa of ε-amine of Lys is around 9.3-9.5 and at pH values lower than the pKa of this group, Lys are generally protonated and possess a positive charge. At pH values greater than the pKa, Lys are unprotonated and contribute no net charge. Arg contains a strongly basic group on its side chain called a guanidino group. The ionization point of this residue is so high (pKa of 12.0) that it is virtually always protonated and carries a positive charge. The side chain of His is an imidazole ring that is potentially protonated at slightly acidic pH values (pKa of 6.7-7.1). Thus, at physiological pH, these residues

contribute to the overall net positive charge of an intact protein molecule. The amine containing side chains in Lysine, Arginine, and Histidine typically are located at the surface of proteins and can be involved in salt bridges through their interactions with the aspartic and glutamic acids [2].

Cys is the only amino acid containing a thiol group (-S-H). At physiological pH, this residue is normally protonated and possesses no charge. Ionization only occurs at high pH (pKa = 8.8-9.1) and results in a negatively charged thiolate group. The most important reaction of Cys residues in proteins is the formation of disulfide crosslinks with another Cys residue. Cys disulfides (also called *cystine or disulfide bridges*) often are key points in stabilizing protein structure and conformation. They frequently occur between polypeptide subunits, creating a covalent linkage to hold two chains together.

Cysteines are relatively hydrophobic due to the small electronegativity difference (i.e., 2.58 vs. 2.20) between the sulfur and hydrogen atoms and usually can be found within the core of a protein. For this reason, strong deforming agents may be needed to open up the protein core to fully reduce the disulfides of large proteins.

Tyrosine (Tyr) contains a phenolic side chain with a pK_a of about 9.7-10.1. Due to its aromatic character, Tyr is second only to Trp in contributing to a protein's overall absorptivity at 275-280 nm. Although the amino acid is only sparingly soluble in water (0.0453g/100g at 25°C), the ionizable nature of the phenolic group makes it often appear in hydrophilic regions of a protein [2].

4.1.3.2 Secondary and Tertiary Structures

Amino acids are linked through peptide bonds to form long polypeptide chains. The *primary* structure of protein molecules is simply the linear sequence of each amino acid residue along the α-chain. Each amino acid in the chain interacts with surrounding groups through various weak, noncovalent interactions and through its unique side chain functionalities. Noncovalent forces such as hydrogen bonding and ionic and hydrophobic interactions combine to create each protein's unique organization.

It is the sequence and types of amino acids and the way that they are folded that provides protein molecules with specific structure, activity, and function. Ionic charge, hydrogen bonding capability, and hydrophobicity are the major determinants for the resultant three-dimensional structure of protein molecules. The α-chain is twisted,

folded, and formed into globular structures, α-helices, and β-sheets based upon the side-chain amino acid sequence and weak intramolecular interactions such as hydrogen bonding between different parts of the peptide backbone.

Major secondary structures of proteins such as α-helices and β-sheets are held together solely through a network of hydrogen bonding created through the carbonyl oxygens of peptide bonds interacting with the hydrogen atoms of other peptide bonds. Other minor secondary structures can also be found in the proteins such as 3_{10} helix, π-helix, turns, and β-bridges.

In addition, negatively charged residues may become bonded to positively charged groups through ionic interactions. Non-polar side chains may attract other non-polar residues and form regions of hydrophobicity to the exclusion of water and other ionic groups. Occasionally, disulfide bonds also are found holding different regions of the polypeptide chain together. All of these forces combine to create the *secondary* structure of proteins, which is the way the polypeptide chain folds in local areas to form larger, sometimes periodic structures.

On a larger scale, the unique folding and structure of one complete polypeptide chain is termed the *tertiary* structure of protein molecules. The difference between local secondary structure and complete polypeptide tertiary structure is arbitrary and sometimes of little practical difference. Larger proteins often contain more than one polypeptide chain. These multi-subunit proteins have a more complex shape, but are still formed from the same forces that twist and fold the local polypeptide. The unique three-dimensional interaction between different polypeptides in multi-subunit proteins is called the *quaternary* structure. Subunits may be held together by noncovalent contacts, such as hydrophobic or ionic interactions, or by covalent disulfide bonds formed from the cysteine residue of one polypeptide chain being crosslinked to a cysteine sulfhydryl of another chain [2].

Thus, aside from the covalently polymerized α-chain itself, the majority of protein structure is determined by weaker, noncovalent interactions that potentially can be disturbed by environmental changes. It is for this reason that protein structure can be easily disrupted or denatured by fluctuations in pH, temperature, or by substances that can alter the structure of water, such as detergents.

4.1.4 Protein Data Bank (PDB)

First protein structure that was determined by protein X-ray crystallography was the structure of myoglobin, which gave the authors, Max Perutz and John Kendrew the Chemistry Nobel Prize in 1962. Since then, the number of proteins whose structures have been made publicly available grew exponentially over the last 50 years. Currently, high-throughput methods are routinely employed for the elucidation of protein structures through X-ray crystallography and NMR studies. Since the opening of PDB database in 1997 with only 6 structures, the number of X-ray structures with experimental data available has increased to a staggering 73000 as of April, 2011.

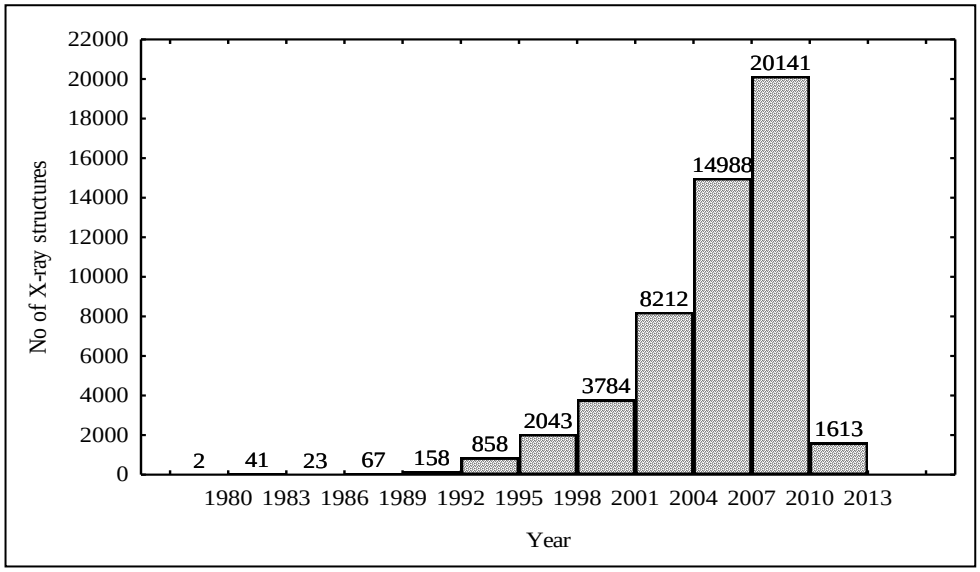

Figure 4.2 PDB X-ray structures deposited to RCSB PDB database

Quality of X-ray structures in terms of resolution has also increased in parallel to the number of structures enabling the comparative computational analysis of protein structures. Such analyses have improved our understanding of protein structures and led to the design and experimental validation of novel proteins with improved properties (May it be increased specificity, improved regioselectivity, acquired functionality and etc.) based on the comparison of different structures.

4.1.5 Mechanisms of Protein Thermostabilization

The hydrophobic effect is considered to be one of the dominant driving forces of protein folding driven by two factors: [3] 1) Hydrophobic groups prefer to avoid water and hydrophilic groups prefer to dissolve in the water. 2) Hydrophobicity drives the protein to a collapsed state from which the native structure is defined by the contribution of all types of non-covalent interactions (e.g., H bonds, ion pairs, and Van der Waals interactions). Dill reviewed the evidences supporting this theory and concluded that: (i) nonpolar solvents denature proteins; (ii) hydrophobic residues are typically sequestered into a core avoiding contact with water; and (iii) residues and hydrophobicity in the protein core are more strongly conserved and related to structure than any other type of residue (replacements of core hydrophobic residues are generally more disruptive than other types of substitutions). Given the central role of the hydrophobic effect in protein folding, it was easy to assume that the hydrophobic effect is also the major force responsible for protein stability [4].

The sequencing, structure, and mutagenesis information accumulated in the last 30 years confirm that hydrophobicity is, indeed, a main force in protein stability [4]. Two observations suggest that mesophilic, thermophilic and hyperthermophilic homologues have a common basic stability afforded by the conserved protein core: (i) hydrophobic interactions and core residues involved in secondary structures are more conserved than surface area features, and (ii) numerous stabilizing substitutions are found in solvent-exposed areas. The high level of similarity encountered in the core of mesophilic, thermophilic and hyperthermophilic protein homologues suggests that even mesophilic proteins are packed almost as efficiently as possible and that there is not much room left for stabilization inside the protein core. Stabilizing interactions in hyperthermophilic proteins are often found in the less conserved areas of the protein. Enough experimental evidence has been accumulated on thermostable proteins in recent years to conclude that no single mechanism is responsible for the remarkable stability of (hyper)thermophilic proteins and increased thermostability must be found, instead, in a small number of highly specific mutations that often do not obey any obvious traffic rules [4].

4.1.5.1 Amino acid composition

Protein amino acid composition is considered as one of the clearest manifestations of protein thermostability. Ponnuswamy *et al* [5] made one of the earliest systematic searches of amino acids that are more significant in protein thermostabilization using 30 protein sequences and about 65000 different amino acid combinations to find the best predictor of protein melting temperature. They have shown that some groups of residues that consist of polar-charged residues and nonpolar residues possessing high surrounding hydrophobicity stabilize proteins against temperature. Residue groups containing polar-uncharged residues destabilize the molecule against temperature, serine being the most destabilizing residue [5].

Tekaia *et al* [6] studied a set of 56 complete genomes and their predicted proteomes including significant numbers of representatives from the three domains of life and derived the following conclusions: First, thermophilic proteins display a relative abundance in Glu, which is more abundant in thermophiles at the expense of Gln. Second, in thermophilic species, the relative abundance in Glu (negative charge) is significantly correlated (Pearson correlation coefficient r=0.83 with P<0.0001), with the increase in the lumped Lys+Arg (positive charges) content. This correlation (absent in mesophiles) could be interpreted on a physicochemical basis, relevant to the thermostability of proteins. (4) Statistically significant differences are observed between the average lengths of thermophilic (283.0+/-5.8) versus mesophilic (340+/-9.4) genes and the "general" shortening of the primary sequences in thermophilic proteins may play a role in thermostability. (5) Considering various combinations of conservation properties (genes conserved exclusively in eukaryotes, in archaea, in bacteria, in combinations of two domains, etc.) correspondence analysis reveals a trend towards thermophilic-hyperthermophilic profiles for the most conserved subset of genes (ancient genes) [6].

Gromiha *et al* [7] reported that the composition of charged residues Lys, Arg, Glu and Asp and hydrophobic residues Val and Ile are higher in thermophiles. On the other hand, Ala, Leu, Glu and Thr are higher in mesophiles based on the evaluation of the discriminative power of amino acid composition by using different machine learning algorithms.

Gliakina *et al* [8] used a dataset of 392 homologous protein pairs from thermophilic and mesophilic organisms and found that proteins from thermophiles

contain more atom-atom contacts per residue in comparison with mesophilic homologues. They analyzed amino acid composition of interior, inaccessible for the solvent, and exterior amino acid residues of proteins from thermophilic and mesophilic organisms and concluded that exterior residues of proteins from thermophilic organisms contain residues such as Lys, Arg and Glu and smaller amino acids such as Ala, Asp, Asn. Gln, Ser, and Thr compared to mesophilic proteins. No significant difference could be detected for the amino acid compositions of interior regions of the considered proteins.

Kumar *et al* [9] performed a statistical analysis of 18 thermophilic and mesophilic protein homologs and reported that the number of salt-bridges and hydrogen bonds between side chains are increased. They have also reported that the frequency of Arg and Tyr is higher and Cys and Ser are lower in thermophiles.

Yokota *et al* [10] also carried out a comparative statistical analysis on 94 mesophilic and thermophilic protein homologs and reported that the thermophilic proteins favor a higher frequency of Arg, Glu, Tyr and a lower frequency of Ala, Ser, Met and Gln residues at the protein surface.

In a recent study, Zeldovich *et al* [11] performed an exhaustive enumeration of all possible sets of amino acids (1,048,574 such combinations, (2^{20}-2)) and surveyed a total of 204 complete archaea and bacteria proteomes and found that the total number of Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL) amino acids correlates well with the optimal growth temperature of the source organisms ranging from 10°C to 110°C with a correlation coefficient of 0.93. The IVYWREL set contains residues of all major types, aliphatic and aromatic hydrophobic (Ile, Val, Trp, Leu), polar (Tyr), and charged (Arg, Glu), both basic and acidic. They also argue that using exact statistical mechanical models of protein stability, the increase of the content of hydrophobic and charged amino acids can be quantitatively explained as a physical response to the requirement of enhanced thermostability, reflecting the positive and negative components of protein design.

In summary, all of these studies reveal that certain amino acids are more favored in thermophilic proteins and following conclusions can be drawn about the amino acid preference of thermophilic proteins: Thermophilic proteins are composed of amino acids that are on the opposite end of the hydrophobic scale. Thermophilic proteins have a higher percentage of both highly hydrophobic and highly hydrophilic (i.e., charged

amino acids) amino acids. Hydrophobic amino acids are allocated to the core of the proteins and the hydrophilic amino acids to the exterior surface. While a similar trend is also observed in mesophilic proteins, the mechanism by which the charged hydrophilic residues replacing the non-charged hydrophilic residues is unique to the thermophilic proteins.

4.1.5.2 Disulfide bridges

Disulfide bridges are formed through the coupling of thiol groups of two Cys residues (Figure 4.3). Disulfide bridges exert their stabilizing effects by reducing the entropy of the protein's unfolded state (denatured state). The bridge usually brings different parts of a polypeptide chain to close proximity and reduces the size of the allowable conformational space (entropic effect) [12]. Zhang *et al* [12] experimentally showed the effects of introducing multiple disulfide bridges. In that research, loop permutation analysis was carried out to vary the length of the region separating two Cys residues.

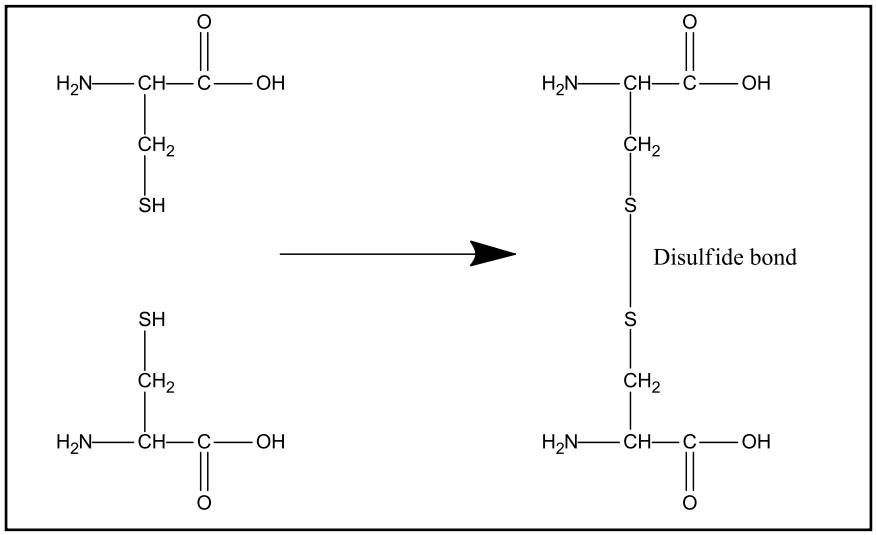

Figure 4.3 Disulfide bond formation

According to their results, the magnitude of the entropic effect of a disulfide bridge is proportional to the logarithm of the number of residues separating the two Cys residues involved in the formation of the bridge [12]. In other words, the higher the number of residues separating two Cys residues forming the disulfide bridge, the higher the magnitude of entropic stabilization.

On the other hand, Zavodsky *et al* [13] engineered *Cucurbita maxima* trypsin inhibitor-V variants containing multiple disulfide bridges and concluded that disulfide

bridges can stabilize not only the denatured state but also the native state of a protein, and differential stabilization of the two states causes either loss or gain in protein stability.

Wakarchuk *et al* [14] increased the thermostability of the xylanase from *Bacillus circulans* xylanase by introducing both intra- and intermolecular disulfide bridges through site-directed mutagenesis. The disulfide bridges that were engineered into the xylanase were mostly buried and, in the absence of protein denaturants, relatively insensitive to reduction by dithiothreitol (a strong reducing agent). All disulfide bond designs tested increased the thermostability of the xylanase, without enhancing the activity of the enzyme at elevated temperatures.

Khan and Deber [15] introduced a single Cys residue into transmembrane helical segment of a major coat protein of M13 bacteriophage and increased the thermostability of the coat protein by enabling the formation of a disulfide-bridged helical dimer in the hydrophobic transmembrane region.

Yamaguchi *et al* [16] engineered a disulfide bridge into the mesophilic *P. camembertii* lipase by observing that other homologous lipases from thermophilic *Rhizomucor miehei*, and *Humicola lanuginose* have a characteristic long disulfide bridge. While the introduction of the disulfide bridge increased the melting temperature of the mesophilic enzyme (51 to 63), it also decreased the optimal temperature for the catalytic activity of the enzyme implying an intrinsically unstable disulfide construct.

Imani *et al* [17] introduced a disulfide bridge into *Photinus pyralis* firefly luciferase and improved its thermal stability and specific activity (7.3-fold).

Matsumura *et al* [18] engineered T4 lysozyme (naturally a disulide-free enzyme) mutants containing one, two, and three disulfide bridges and showed that increase in melting temperature resulting from the individual disulphide bridges was approximately additive and the triple disulfide bridge variant had a melting temperature that was 23.4°C higher than the wild-type lysozyme.

4.1.5.3 Salt bridges

Proteins contain amino acids of opposite charges which may come consecutively, bring different parts of a single protein chain in close proximity to perform a catalytic function or bring different regions in a multi-chain protein. Salt bridges, also called ion-pairs, have been implicated to play significant roles in the thermostabilization of certain protein structures. They may exert their effects either through a single ion-pair or a network of ion-pairs.

Bogin *et al* [19] studied two highly homologous alcohol dehydrogenases, one from the mesophile *Clostridium beijerinckii* (CbADH) and the other from the extreme thermophile *Thermoanaerobacter brockii* (TbADH), suggested that in the thermophilic enzyme, an extra intrasubunit ion pair and a short ion-pair network at the intersubunit interface might contribute to the thermal stability of TbADH. Moreover, Bogin *et al* mutated structurally strategic residues of the mesophilic CbADH with the corresponding amino acids from TbADH and concluded that the amino acid substitutions in CbADH mutants enhanced the thermal stability of the mesophilic protein by reinforcing the quaternary structure of the enzyme through the formation of a new intrasubunit salt bridge and an extended network of intersubunit ion-pairs.

Tomazic and Klibanov [20] compared the half-lives of three *Bacillus* alpha-amylases at 90 °C and suggested that the increase in the half-lives in the series from *Bacillus amyloliquefaciens* to *Bacillus stearothermophilus* and *Bacillus licheniformis* (the difference in thermostability between the first and the third enzymes exceeds 2 orders of magnitude) is mainly attributable to the additional salt bridges involving a few specific Lys residues.

Matsutani *et al* [21] carried out a comparative genomics study of thermo-tolerant species and concluded that an increased Lys to Arg substitution in the salt bridges contributes to the thermotolerance of *Acetobacter tropicalis*.

Hendsch and Tidor [22] carried out a continuum electrostatic approach on 21 salt bridges in 9 protein X-ray crystal structures and found that the majority (17) of salt-bridges are electrostatically destabilizing due to a large, unfavorable desolvation contribution that was not fully compensated by favorable interactions within the salt bridge and between salt-bridge partners and other polar and charged groups in the folded protein. They also suggested that mutation of salt bridges, particularly those that are buried in hydrophobic regions can result in proteins with increased stability.

Kumar *et al* [23] also carried out continuum electrostatic calculations on a dataset of 222 non-equivalent salt bridges derived from 36 non-homologous high-resolution monomeric protein crystal structures and concluded that most of the salt bridges in their dataset are stabilizing, regardless of whether they are buried or exposed, isolated or networked, hydrogen bonded or non-hydrogen bonded. Moreover, one-third of the salt bridges in their dataset are buried in the protein core, with the remainder exposed to the solvent. The difference in the dielectric properties of water versus the hydrophobic protein interior cost buried salt bridges large desolvation penalties.

Kumar *et al* [24] assembled a dataset containing 18 non-redundant families of thermophilic and mesophilic proteins with each of the 18 families consisting of homologous thermophile-mesophile pairs. They observed that the number of salt bridges is increased in most of the thermophilic proteins. By comparing the salt bridges in the glutamate dehydrogenase from the hyperthermophilic *Pyrococcus furiosus* and the mesophilic *Clostridium symbiosum*, they concluded that while the salt-bridges in the former are highly stabilizing, they add only marginal stability to the mesophilic protein.

Karshikof and Ladenstein [25] suggested that the optimization of electrostatic interactions by increasing of the number of salt bridges is a driving force for enhancement of the thermotolerance of proteins from hyperthermophilic microorganisms and this feature is less evident in proteins from thermophilic organisms and is absent from mesophile-derived proteins.

Ge *et al* [26] concluded that the energy contribution of a salt bridge formed by two charged residues far apart in the primary sequence is higher than that of those formed between two very close ones based on the contribution of two conserved salt bridges to the stability free energy of the DNA-binding protein, Ssh10b, from the archaeon *Sulfolobusshibatae.*

4.1.5.4 Hydrophobic interactions

The statistical analysis carried out by Ikai *et al* [27] showed that the aliphatic index, the relative volume of a protein occupied by aliphatic side chains (e.g., Ala, Val, Leu and Ile) is significantly higher in thermophilic globular proteins than mesophilic proteins.

On the other hand, Merkler *et al* [28] argued that there exists only a relatively weak positive correlation between thermostability and aliphatic index by carrying out a study that only included slightly more than 20 enzymes (far less than the number to draw statistical significance with current standards) from closely mesophilic and thermophilic microorganisms.

Lu *et al* [29] used a larger dataset with 110 homologous sequences from mesophiles and thermophiles and claimed that the reason for thermophilic proteins having a higher aliphatic index is attributable to the higher Leu composition in thermophiles, and made the validity of aliphatic index as a positive indicator of thermostability equivocal.

4.1.5.5 Aromatic interactions

Proteins contain amino acids that are called aromatic because they contain ring structures with delocalized conjugated π systems which allow the movement of electrons over the entire ring structure providing resonance stabilization. Aromatic amino acids are phenylalanine, tyrosine, tryptophan and histidine. Although histidine (p*K*a=6.1) is an aromatic amino acid, the presence of positive charge at pH 7 complicates its interaction with other π systems and ions. The delocalized π electrons can assert their stabilizing effects through two modes of actions: π-π stacking and cation-π interactions.

4.1.5.5.1 π-π stacking

Stacking of two or more aromatic ring structures on top of each other is called π-π stacking. Some researchers [30] consider the increased stability of such stacks as yet another manifestation of strong Van der Waals forces that is attributable to an increase in interaction surface area while others [31] consider as a different stabilizing force that cannot only be explained by Van der Waals forces. There is not a consensus in the literature about the source of this interaction and or its strength as a stabilizer.

4.1.5.5.2 Cation-π interactions

Positively charged amino acids like Arg, Lys and His that are near an aromatic amino acid in certain orientations can be a stabilizing force due to cation-π interactions. The orientation of the participating partners is the most favorable when the positive charge is stabilized by the electron dense regions of a conjugated π system. The cation-π interaction is comparable in strength (ca. 2kcal/mol) to hydrogen bonding and can be a decisive intermolecular force depending on the physiological conditions. Possible interaction partners are Lys and Arg for cations and Tyr, Phe and Trp for π-systems.

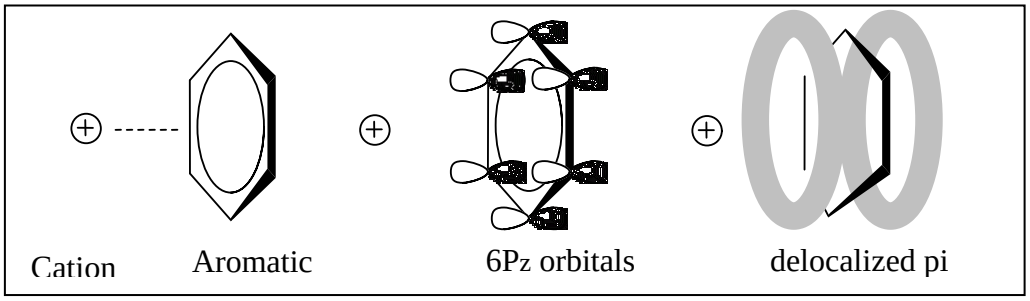

Figure 4.4 Different illustrations of a cation-pi interaction

Gallivan & Dougherty [32] reported results from a quantitative survey of cation-π (cation-pi) interactions in high-resolution structures obtained from the PDB database. Using an energy-based criterion for identifying significant sidechain interactions, they studied 593 protein structures with dissimiliar sequences. They found an average of one such interaction per 77 residues, with no significant effect of chain length, or multiple-chain vs. single chain structures. Arg was more likely than Lys to participate in a cation-π interaction, and the likelihood of aromatic sidechain participation was Trp > Tyr > Phe.

Moreover, they also found that over one quarter of all Trp's were involved in cation-π interactions, with the cation typically positioned over the 6-atom ring of Trp. Their study did not include His because of two different modes of action that depends on His residue's protonation state; it could participate either as a cation or as a π-system. Lys and Arg were assumed always to be protonated and hence cationic.

Based on their findings, Gallivan and Dougherty [32] concluded "When a cationic sidechain is near an aromatic sidechain, the geometry is biased toward one that would experience a favorable cation-π interaction", and "cation-π interactions should be considered alongside the more conventional hydrogen bonds, salt bridges, and hydrophobic effects in any analysis of protein structure".

Chakravarty and Varadarajan [33] showed that cation-π interactions, estimated to be twice as strong as ion-pairs, are significantly enriched in thermophiles.

Folch *et al* [34] carried out *in silico* analyses of protein thermostability using statistical residue-residue potentials and derived the following conclusions: Thermostabilizing interactions include salt bridges and cation-π interactions (especially those involving arginine), aromatic interactions, and H-bonds between negatively charged and some aromatic residues. H-bonds between two polar non-charged residues or between a polar non-charged residue and a negatively charged residue are relatively less stabilizing at high temperatures. It is necessary to consider both repulsive and attractive interactions in overall thermostabilization, as the degree of repulsion may also vary with increasing temperature [34].

4.1.5.6 Structural rigidity

The B-factor (also called B-value, Debye–Waller factor, or temperature factor) is used to measure local flexibility (mobility) of residues. B-factor values are reported from experimental atomic-resolution structures. They quantify the decrease of intensity in diffraction due to the dynamic disorder caused by the temperature-dependent vibration of the atoms and the static disorder related to orientation of the protein molecule. High values indicate higher mobility of residues in crystal structures. B-factor values of Cα atoms are commonly used to represent motion of the backbone and depend on a number of other factors such as the overall resolution of the protein structure,

crystal contacts, and applied refinement procedures. As a result, they are usually normalized [20, 21].

The distribution of B-factor values along a protein sequence reflects flexibility and dynamics of the underlying structure. For instance, protein core is usually characterized by low B-factor values since it should be well packed to provide rigidity for the entire structure. At the same time, surface would usually include some flexible regions which would have high B-factor values. The reason is that the protein interacts with other molecules, which requires certain degree of structural flexibility.

Parthasarathy and Murthy [35] carried out an analysis of B values reported in high-resolution X-ray crystal structures of mesophilic and thermophilic proteins and concluded that Ser and Thr have lesser flexibility in thermophiles than in mesophiles; the proportion of Glu and Lys in high B value regions of thermophiles is higher and that of Ser and Thr is lower; and the dispersion of B values within spheres at Cα atoms is similar in mesophiles and thermophiles.

Jochens *et al* [36] increased the thermostability (without compromising specific activity) of various *Pseudomonas fluorescens* esterase variants up to 9 degrees compared to wild type by generating site-saturation libraries targeting surface positions on the basis of B-factor iterative test principle, a method that was developed to aid in the design of "small, but smart" mutant libraries.

The knowledge of B values was used in prediction of protein flexibility [37, 38], analysis of protein thermal stability [35, 39] and active sites [40-42], correlating the side chain mobility with protein conformation [43, 44], and prediction of protein-protein binding sites [45].

4.1.5.7 Dipole Stabilization

Due to the presence of a significant number of charged residues at neutral pH, proteins are macro-zwitterions whose electrostatic properties are important for their stability and function. Many proteins have surface patches of positive or negative potential that might be important for their function. Such regions are indicative of an excess of net positive or negative charge and/or a significant imbalance in the spatial distribution of the charges or, in other words, of a large dipole moment. Many examples of proteins with large net charges or dipole moments have been reported.

Eijsink *et al* [46] increased the thermostability of a *Bacillus subtilis* neutral protease up to 1.2 degrees by replacing the Lys residue at the N-terminal with Ser or Asp. Substitutions improved the electrostatic interactions by introducing favorable residues at the end of α-helices.

Nicholson *et al* [47] constructed two stabilizing mutations, T109D and N116D, in phage T4 lysozyme that showed a pH-dependent increase in thermal stability due to the interaction of the aspartic acids with the α-helix dipole. They also showed that the mutant N116D did not show enhanced stability due to a favorable salt-bridge interaction but rather an interaction with the alpha-helix dipole.

4.1.6 Reduced Amino Acid Alphabets

In Chapter 3, we have systematically shown that n-gram-RAAA combinations can be used to discriminate proteins from different thermostability classes using less number of features than a native alphabet. In this chapter, we expand the number of features used in Chapter 3 by incorporating other sequential and structural features. However, we excluded dipeptide and tripeptide compositions as possible feature sets because they are not as good discriminators of thermostability as the amino acid composition in the classification of proteins using support vector machines.

4.2 Methods

4.2.1 Dataset acquisition

We assembled a protein data set *de novo* that contains 2022 proteins with an x-ray structure of good quality (resolution 2.5 Å) and known T_{opt} of the source organism. Only monomeric proteins were included in the data set, as identified by the Protein Quaternary Structure server. The dataset contains proteins from mesophilic, thermophilic and hyperthermophilic organisms and was used to test the significance of different sequential and structural features in hyperthermophilic and thermophilic proteins compared to a control set of mesophilic proteins. The dataset was assembled according to the protocol that was outlined below:

1) Prokaryotic Growth Temperature database (PGTdb) [48]was used to download the names of the source organisms that belong to three distinct classes based on the optimal growth temperature (T_{opt}) of the source organism:
   a) Hyperthermophilic T_{opt} >80°C
   b) Thermophilic 45°C<T_{opt} <80°C
   c) Mesophilic 15°C<T_{opt}<45°C
2) Taxonomy identification numbers (Taxids) that correspond to the above organisms were downloaded from the National Center for Biotechnology Information (NCBI) Taxonomy Homepage.
3) PDB database was searched using the names and Taxids of the source organisms along with the following two criteria:
   a) Proteins that do not contain any modified residues
   b) Proteins that contain only a single chain (i.e., monomeric proteins)
4) The above search resulted in a total of 6505 protein structures.
5) Further refinement of the PDB structures was carried out using the PDB culling server, PISCES, with the following criteria. Only the PDB entries satisfying the following criteria were kept and all others were discarded.
   a) Maximum percent identity 90%
   b) Maximum resolution 2.5Å
   c) Maximum R-value 0.3Å
   d) Minimum chain length 40 amino acids

e) Maximum chain length 10000 amino acids
f) PDB structures obtained by X-ray crystallography

6) The culling procedure further reduced the size of the structure-based dataset to a total of 2087.
7) PDB structures containing inserted amino acids, fragments, or unknown residues were also eliminated (32 total).
8) PDB structures annotated as membrane proteins according to Structural Classification of Proteins (SCOP) database were also discarded (33 total)
9) The final dataset contains 2022 PDB structures.
10) All further analyses were carried out using these 2022 PDB structures which henceforth will be called as *Structure-Based Dataset* (SB). Number of sequences in SB dataset is presented in a pie-chart in Figure 4.5. PDBids and the corresponding thermostability classes are provided in Appendix D
11) Three different datasets were generated from the SB dataset:
    a) Hyperthermophile-Mesophile (HM) dataset contains only hyperthermophilic and mesophilic proteins
    b) Thermophile-Mesophile (TM) dataset contains only thermophilic and mesophilic proteins
    c) Hyperthermophile/Thermophile-Mesophile ((HT)M) dataset contains two classes: Hyperthermophilic and thermophilic proteins are combined into a single class that is called (HT) and proteins from mesophilic organisms form the M class.

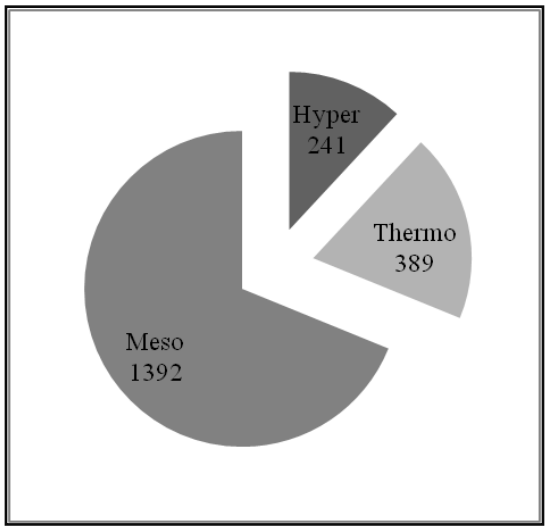

Figure 4.5 Distribution of the number of sequences to different classes in SB dataset

4.2.2 Software development

In this chapter, we calculated one of the most comprehensive set of features using either protein sequence or structure as input. In certain cases, there exists webservers or standalone computer software for the calculation of such features (such as disulfide bridges) using a single protein sequence or structure. However, batch extraction of such features using such servers is not only computationally time consuming but also laborious and not viable for more than a handful of training samples.

Therefore, in many cases, we developed either computer software for batch processing of input data to output a numerical value for a particular feature (feature extraction) or developed so called “wrapper scripts” to parse such features from a remote server and formatted them to be used for downstream analysis. To that end, we developed more than 120 different python scripts for data manipulation, feature extraction, sequence and structure parsing, batch processing and plotting. These scripts will be made publicly available for the use of scientific community.

Unless otherwise stated, all sequence and structure derived features except cation-pi interactions and dipole related features have been calculated using computer software developed in-house using Python programming language and Biopython (v1.53) [49] module for Python v2.5. All boxplots were generated using Matplotlib v1.0.1 [50] module.

4.2.3 Sequential features

Sequential features refer to the features that are extracted from protein primary sequence only. In this chapter, both sequential and structural features are organized into feature sets. For example, in the "Amino acid composition" feature set, there are a total of 20 features with each feature representing the composition of a particular amino acid. Moreover, each RAAA-alphabet size combination is a different feature set. For example, Ab10 feature set is different from Ab11 feature set and they contain 10 and 11 features, respectively. A list of all sequential feature sets along with the number of features in each set is provided in Table 4.1. In the next sections, calculation of features will be explained briefly.

Table 4.1 Sequential feature sets that were used in this study

Feature Set	**# of features**
Amino acid composition (Native)	20
Basics	8
Ab10 – Ab19	145
Dssp10 – Dssp14	60
Eb11, Eb13	24
Gbmr10 – Gbmr14	60
Hsdm10, Hsdm12, Hsdm14 – Hsdm17	121
Lr10	10
Lwi10 – Lwi19	145
Lwni10, Lwni11, Lwni14	35
Lzbl10 – Lzbl16	91
Lzmj10 – Lzmj16	91
Ml10, Ml15	25
Sdm10-Sdm14	60
Total of 65 Feature sets	Total of 895 features

4.2.3.1 Amino acid composition

Amino acid composition is calculated by counting the number of each amino acid and dividing by the total number of amino acids (i.e., protein length). Only the composition of 20 standard amino acids is calculated and reported for a given sequence.

4.2.3.2 Basic Indices

4.2.3.2.1 Aromaticity

Aromaticity is calculated using the aromaticity value of a protein according to Lobry [51]. It is total number of Phe, Trp, and Tyr residues divided by the total number of residues in a protein sequence.

4.2.3.2.2 Helix, sheet, turn propensity

Propensity values are calculated by counting the total number of residues which are more likely to be included in a given secondary structural element (SSE) and dividing by protein sequence length. Residues that are more likely to be in each of these structures are provided in Table 4.2

Table 4.2 Secondary structure propensity

Secondary structure propensity	Amino acids
Helix	V, I, Y, F, W, L
Turn	N, P, G, S
Sheet	E, M, A, L

4.2.3.2.3 Grand average of hydrophobicity (Gravy)

Gravy index is an estimate of the overall hydrophobicity of the protein. Each amino acid has a hydrophobicity score that ranges between -4.6 and 4.6 with negative and positive values indicating hydrophilic and hydrophobic residues, respectively. Gravy index of each amino acid is provided in Figure 6. Gravy is calculated by taking the average of all hydrophobicity scores in a given protein sequence according to the hydrophobicity values provided by Kyte-Dolittle [52].

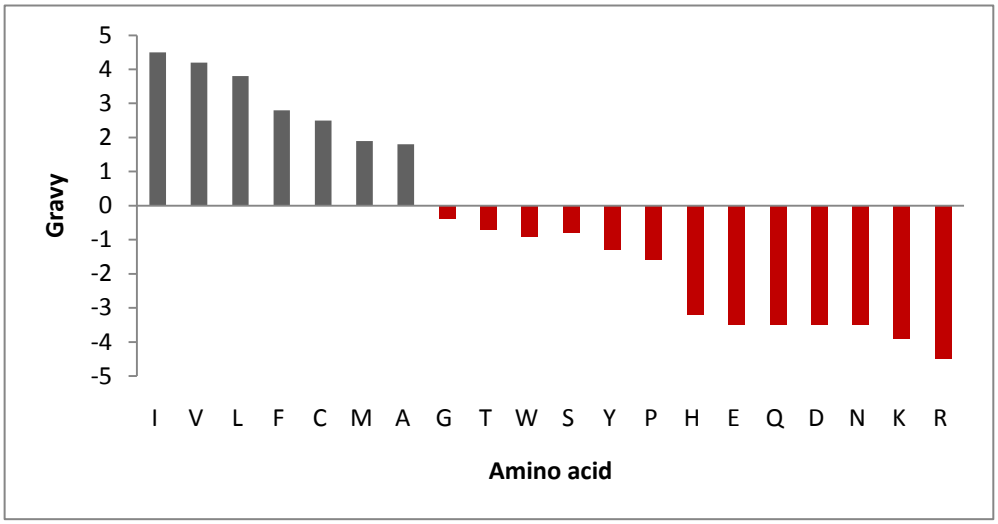

Figure 4.6 Hydrophobicity values according to Kyte-Dolittle scale

4.2.3.2.4 IVYWREL

IVYWREL corresponds to the total number of IVYWREL residues in a protein normalized by the length of the protein sequence.

4.2.3.2.5 Instability Index

Instability index is calculated according to Guruprasad *et al* [53]. For a given protein, the summation of dipeptide instability weight values is normalized by the length of the protein sequence.

4.2.3.2.6 Isoelectric Point (pI)

pI is the pH value at which net charge of the protein is equal to 0. Theoretical pI value of the protein is calculated using Isoelectric Point function in Biopython package [49]

4.2.3.2.7 Molecular Weight (MW)

MW for each protein is calculated using molecular weight function in Biopython package [49]

4.2.3.2.8 Protein Length

Protein length is simply the total number of amino acids in a protein sequence.

4.2.3.3 Reduced amino acid composition

Reduced amino acid composition is calculated using the RAAAs (except random alphabet) provided in Chapter 3 by counting the number of a particular reduced amino acid and dividing by the length of the protein sequence.

4.2.4 Structural features

All structural features were extracted from the PDB file of the corresponding protein. A list of all features (alphabetically organized) along with the number of features in each feature set is provided in Table 4.

Table 4.3 Structural features obtained from protein structure

Feature Set	# of features
Amino acid composition in SSs	140
B-values of SSs	7
Cation-π interactions and related features	13
Dipole related features	10
Disulfide bridges and related features	3
Salt bridge related features	14
Secondary structure content	7
Hinge and loop related features	4
Secondary structure embedded sequence alphabet	180
Total	378

4.2.4.1 Cation-π interactions

Two different approaches exist in the literature to define cation-π interactions. In the first approach, a cation-π pair is considered interacting if the distance between them

is less than 6 Å. This is a distance-based approach and is employed in Protein Explorer. In the second approach, only energetically significant cation-π pairs are considered *interacting*. This approach is employed by Gallivan and Dougherty [32] and extensively tested.

The number of cation-π interactions resulting from the energy-based and distance based approaches may be different. Distance-based method employed in Protein Explorer may result an overestimation of the number of cation-π interactions. In some cases, a cation is within the requisite 6Å of an aromatic sidechain, but the interaction would in fact be energetically insignificant due to the suboptimal orientation of the cation with respect to the aromatic ring. In other cases, the requirement for three alternate carbon atoms in aromatic rings is met by carbons from different residues. The latter type of incorrect results is usually obvious because a ring will be shown with no proximal cation, or *vice versa*.

Determination of the extent of cation-π interactions was carried out using the CaPTURE program [32] that uses an energy-based calculation of the favorable interactions between cationic arginine and lysine residues with the aromatic sidechains of the tyrosine, phenylalanine and tryptophan residues. All cation-π related features in this feature set are provided in Figure 4.7. Energetically the most significant cation-π pairs correspond to the number of interacting pairs with an electrostatic energy, E(es), of less than -2 kcal/mol or -2kcal/mol < E(es) < -1kcal/mol and E(vdw) <= -1kcal/mol [32].

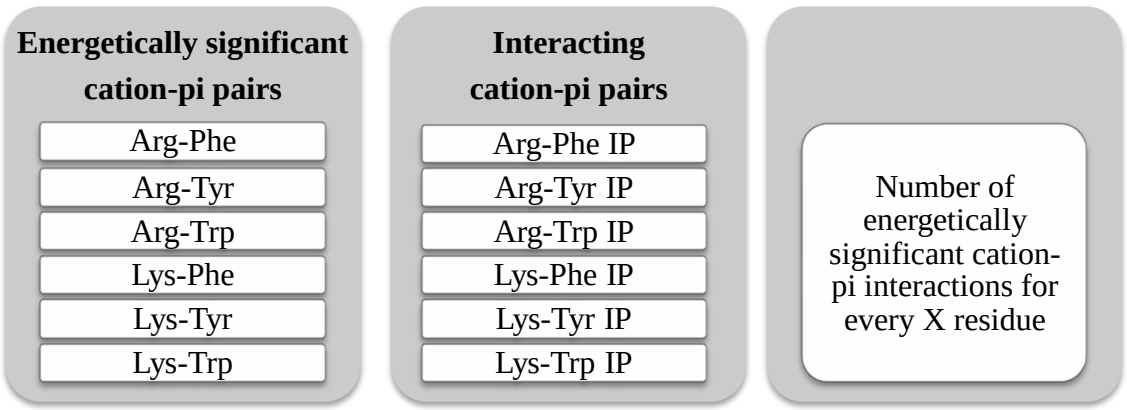

Figure 4.7 Cation-pi related features

4.2.4.2 B-values of SSs

Protein B-values were extracted from PDB file and each B-factor was normalized according to the formula:

$$B_{i,normalized} = \frac{B_i - B_{avg}}{\sigma}$$

where B_{avg} is the average of the B-factor values of a given structure, B_i is the B-value of i^{th} residue, and σ is the corresponding standard deviation of all residues in protein. Normalized B-value of each residue that folds into one of 7 secondary structural elements (SS) defined by STRIDE [54] were summed and divided by the total number of residues assuming that particular SSE.

4.2.4.3 Secondary structure content

Secondary structure content is defined as the percentage of total number of amino acids in a particular secondary structural element in a protein. For proteins with X-ray structures and corresponding PDB file, secondary structure content is calculated according to the formula below:

$$Count_x = \frac{x}{L}$$

where x = α-helix, β-strand, coil, 3_{10} helix, π-helix, turns, and β-bridges according to the 7-states defined by STRIDE [54], $Count_x$ denotes the percentage of residues assuming secondary structure of type x, and L is the length of the protein chain. The secondary structure content encapsulates the bulk (protein-wide) information concerning secondary structure without the knowledge of which residues assume a particular secondary structure (see section 4.2.3.7 for the amino acid composition in a particular secondary structure). This information is useful to characterize an overall type of the protein fold, such as those defined in the SCOP [55] and CATH [56] databases.

4.2.4.4 Salt-bridge related features

Different criteria exist in the literature for the definition of a salt bridge. In the current study, the criterion for determining salt bridges is that the distance between any of the two carboxyl oxygen atoms on the side chain of Glu or Asp and nitrogen atoms on the side chain of Arg or Lys is within 4.00 Å. Histidine is excluded as a potential partner in a salt-bridge or ion pair due to the fact that it is very sensitive to pH changes in the physiological range (The R group of histidine has 10% probability to become positively charged at pH = 7, but the probability increases to 50% at pH = 6. Thus, histidine is very sensitive to pH change in the physiological range). All possible combinations of salt-bridge forming pairs are provided in Figure 4.8. No_of_IPs is the number of total ion-pairs, IPs_per_residue is total ion pair divided by the length of protein sequence; Arg-Glu_per_residue is the number of Arg-Glu ion-pairs divided by protein length; Arg_involved_in pairs is the number of Arg involved in an ion pair and etc.

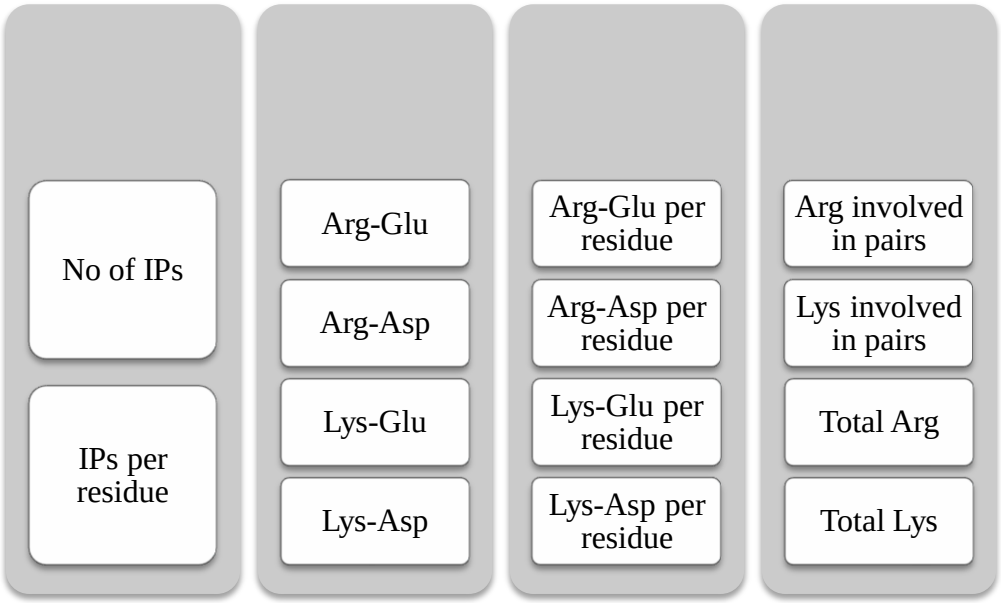

Figure 4.8 Salt-bridge related features

4.2.4.5 Dipole related features

Electrostatic properties can play a significant role in affecting the properties and activities of proteins, for example influencing how and where various substrates, inhibitors, cofactors, and other proteins bind. If proteins have a large net charge or dipole, this effect might be particularly significant. While the precise electrostatic potential about a protein involves a detailed and complex calculation and interpretation, one can often get a first clue by examining two very simple overall properties, the net charge and the dipole moment, and how the latter lines up in comparison with key structural features of the protein.

Actually, since proteins in general are not electrically neutral, one should really speak in terms of a position-dependent first moment of charge distribution. However, we chose to refer to this quantity as a "dipole moment", because this is how the concept is best recognized by the general scientific community. Although this calculated dipole moment is only a rough approximation, due to several simplifying assumptions made, comparisons of calculated dipoles among different proteins can be meaningful if they are all calculated the same way, e.g., if the same degree of ionization of the residue sidechains, the same atom charges, and same centering of the protein within the coordinate system are used.

We wanted to know if such dipole moments are actually unusually large or small for proteins from different thermostability classes. We therefore utilized Protein Dipole Moment Server [57] server calculated the following dipole related features: Number of atoms, number of residues, positive residues, negative residues, net molecular charge, net molecular charge per atom, net molecular dipole, overall molecular dipole moment in Debyes, net molecular dipole moment per atom, quadrapole, and mean protein radius.

4.2.4.6 Disulfide related features

For calculating disulfide related features, we utilized a distance based approach. A disulfide bridge or disulfide bond is defined between two residues if the distance between the partnering sulfur atoms is less than 2.3Å. In addition to the number of disulfide bridges, we also included total number of Cys residues and free Cys residues in this feature set.

4.2.4.7 Amino acid content in secondary structures (aa_content_in_ss)

Amino acid content in secondary structures is calculated for each amino acid by finding the secondary structure of that amino acid and summing the number of all such amino acids and dividing by the length of the protein sequence. For example, H_Ala as a feature corresponds to the percentage of Ala residues (amongst all residues) assuming α-helical secondary structure in the protein. Since there are 20 amino acids and 7 secondary structures (see section 4.2.4.3), there are a total of 140 features in this feature set.

4.2.4.8 Secondary structure embedded sequence alphabet (SSESA)

SSESA corresponds to a structural sequence alphabet that is composed of 180 triplets. First letter in each triplet carries secondary structure information of each amino acid in a particular protein under study and can take one of three different states according to STRIDE definition: H: α-helix, S: β-sheet and L: Loop. Second letter in each triplet corresponds to relative solvent accessibility (RSA) of each amino acid and can take one of three values: B: Buried, P: Partially buried, and E: Exposed. Third letter of each triplet corresponds to one of 20 different standard amino acids. There are a total of such 180 SSESAs (3x3x20=180) and composition of each of this SSESA was normalized by the length of the secondary structural element of that triplet.

RSA is calculated by finding the solvent exposed surface area (SESA) of each amino acid in a particular protein structure using STRIDE and normalizing by the maximum surface area of that particular amino acid according to standard values provided in reference [58]. An amino acid is considered Buried if RSA < 0.09, Partially-buried if 0.09 < RSA < 0.36, and Exposed if RSA > 0.36.

4.2.4.9 Hinge region related features

HingeProt [59] program was used for predicting rigid parts of proteins and the flexible hinge regions connecting them in the native topology of protein chains. HingeProt utilizes two elastic network (EN) models: Gaussian Network Model (GNM) and Anisotropic Network models (ANM).

HingeProt takes a protein PDB structure as input and outputs a list of rigid parts and hinge regions for the two slowest modes. HingeProt also outputs a list of short flexible fragments for the two slowest modes which correspond to rigid segments with less than 15 amino acid residues.

A python *HingeProt* wrapper script was developed to parse the number of rigid fragments in slow mode 1 and slow mode 2, and number of short flexible fragments in slow mode 1 and slow mode 2. For each protein PDB structure, hinge-region related values were normalized by protein sequence length.

4.2.5 Kolmogorov-Smirnov Test

The SB dataset contains relatively fewer number of training instances compared to the sequence based dataset that was used in Chapter 3. Therefore, an initial test of normality was carried to determine whether most of the features in a feature set are normally distributed using Shapiro-Wilk test [60]. It was found that the distribution of most of the features did not follow a normal distribution. Therefore, we performed two-sided two-sample Kolmogorov-Smirnov (KS) non-parametric statistical significance test. KS test does not assume that the underlying distributions are normally distributed. KS test results showed that for many features the underlying distribution is significantly different between hyperthermophiles and thermophiles compared to the control set of mesophiles. After performing KS test, the value at 50^{th} percentile (i.e., median) of a particular feature in each thermostability class was used as an indicator of the central tendency of the value of the feature and the interpretation of KS test results were carried out based on median value rather than mean and standard deviation.

For each feature, following additional descriptive statistics were calculated: hyperthermophile mean, mesophile mean, hyperthermophile standard deviation, mesophile standard deviation, value at 25^{th}, 50^{th}, and 75^{th} percentile for hyperthermophiles and mesophiles, minimum and maximum values of the feature in hyperthermophiles and mesophiles, KS-statistic of hyperthermophiles with respect to mesophiles, p-value of the KS-statistic, a binary code where 1 implies statistical significance at the level of 0.01 for the two-sided KS-test, and a nominal value of either OVER or UNDER which implies hyperthermophile median is higher or lower than mesophile median, respectively. Same descriptive statistics were also calculated for TM dataset.

4.2.6 Boxplots

For each significant feature in a feature set, boxplots were generated using Matplotlib module to visualize the distribution of that feature. In a boxplot, 25th percentile, 50th percentile, 75th percentile, mean (shown with an ☆) and outliers (+ signs) can be identified as demonstrated in Figure 4.9.

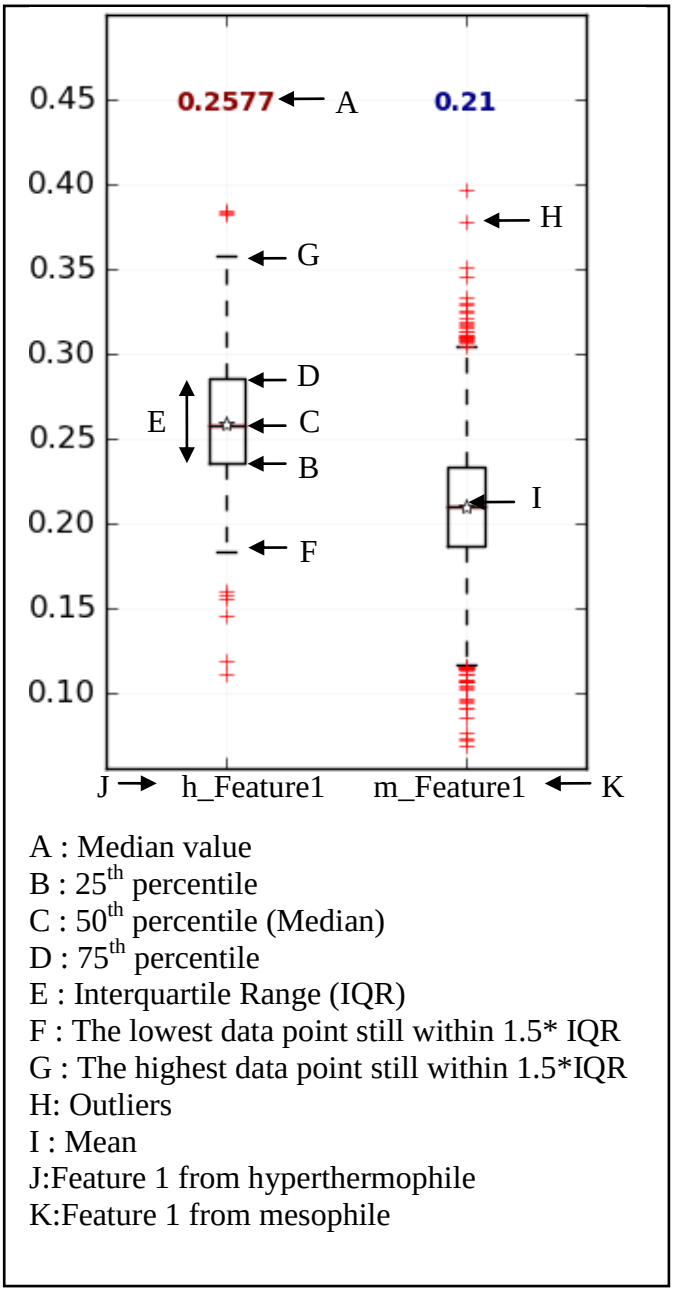

Figure 4.9 Boxplot example of a hypothetical feature in HM dataset

4.2.7 Classification

Fasta or PDB files were downloaded from RCSB PDB database. Feature extraction was carried out on each dataset using either dedicated web-servers or written in-house python scripts to extract sequence or structure related features.

Classification was carried out using WLSVM [61], a LibSVM [62] classifier interface for the widely distributed Weka (v3.6.3) [63] data mining software. For each feature set and dataset (HM or TM), the dataset was split randomly with 80 % of the data used for training and the remaining 20% for testing. A weight that is inversely proportional to the class size was assigned to each class to account for unbalanced class sizes during the training phase of the LibSVM classifier. The classifier was trained on the normalized training set with the parameters set to RBF kernel-C-SVC, C=100, and ε=0.09 to generate a model and tested on the test set. After each testing, sensitivity, specificity, accuracy and AUC values were recorded. The 80-20 split was carried out 100 times or cycles where a different random seed was used for splitting at each cycle. After 100 cycles of testing, accuracy values were averaged for each feature set and reported as average accuracy. An overview of the workflow is also provided in Figure 4.10.

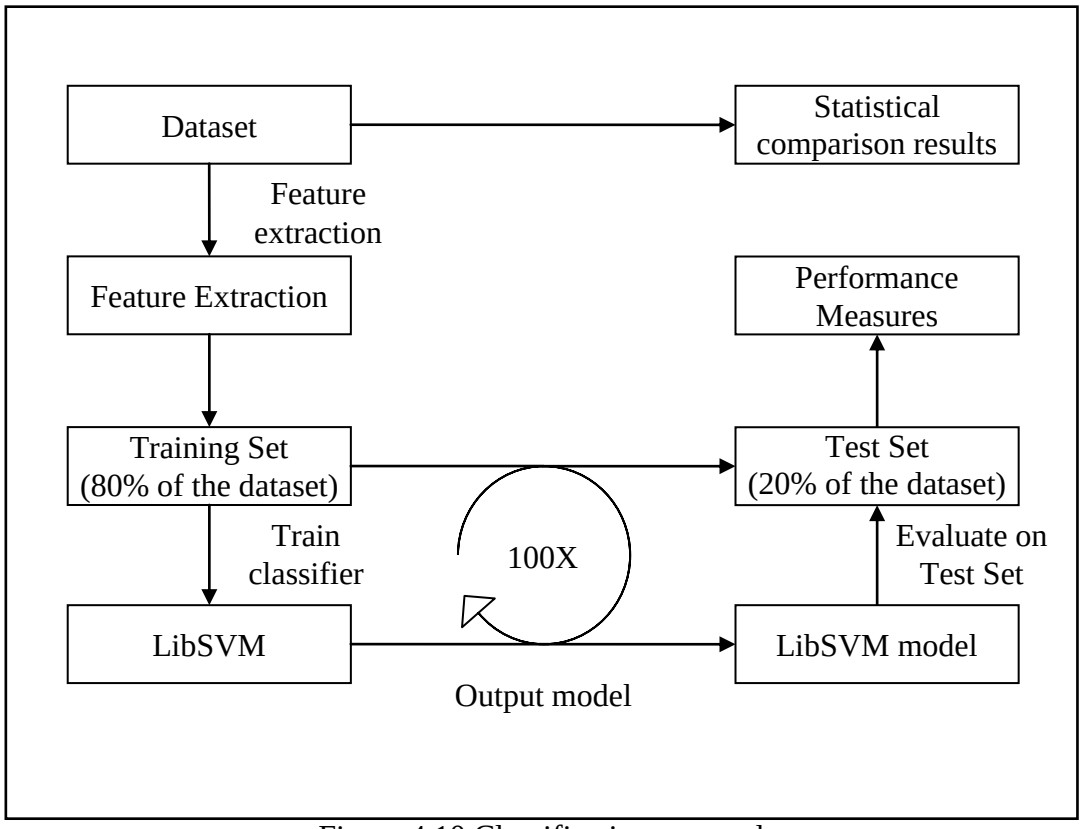

Figure 4.10 Classification protocol

4.2.8 Performance measures

The performance of machine learning algorithms is typically evaluated by a confusion matrix as illustrated in Table 4.4 (for the TM dataset). The columns are the Predicted classes and the rows are the Actual classes. In the confusion matrix of the TM dataset, TP is the number of true positives (thermophilic proteins predicted as thermophilic); FN is the number of are false negatives (thermophilic proteins predicted as mesophilic); TN is the number of true negatives (mesophilic proteins predicted as mesophilic) and FP is the number of false positives (mesophilic proteins predicted as thermophilic). In the confusion matrix of the HM dataset, TP is the number of true positives (hyperthermophilic proteins predicted as hyperthermophilic); FN is the number of are false negatives (hyperthermophilic proteins predicted as mesophilic); TN is the number of true negatives (mesophilic proteins predicted as mesophilic) and FP is the number of false positives (mesophilic proteins predicted as hyperthermophilic).

Table 4.4 Confusion Matrix of the TM dataset

		Predicted	
		Thermo	**Meso**
Actual	**Thermo**	TP	FN
	Meso	FP	TN

Five different statistics were used as performance measures to evaluate the discriminative power of the each feature set using LibSVM classifier in classifying proteins into different thermostability classes: Sensitivity (TP Rate), Specificity (FP Rate), Accuracy, and AUC. Classifier performance was assessed using the following equations;

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

AUC values were obtained using Weka [63] software. All classification results are provided in Appendix E.

4.3 Results

4.3.1 Statistically significant features

In the next two sections, we will present the results of the KS test in a systematic manner and use the median value of a feature to elaborate on the central tendency of that feature in hyperthermophiles or thermophiles compared to mesophiles. To eliminate confusion due to long sentence constructs and to improve readability of the text, we will simply say "feature X is higher in hyperthermophiles compared to mesophiles" when we really mean that "median value of X is higher in hyperthermophiles compared to the median value of X in mesophiles."

4.3.1.1 Basic Features

Amongst basic features, *IVYWREL* index is the most significant difference between either hyperthermophiles or thermophiles and mesophiles with p-values of 1.17E-77 and 3.46E-54 in HM and TM datasets, respectively. The *IVYWREL* (Figure 4.11) index is higher in both hyperthermophiles and thermophiles than mesophiles. This result is in broad agreement with the previously published results of Taylor *et al* [64] and Zeldovich *et al [65]*.

Helix propensity, and *Turn* are other features with significant differences in HM and TM datasets (boxplots not shown). Interestingly, Helix propensity is higher in hyperthermophiles and thermophiles than mesophiles; and Turn propensity is lower in hyperthermophiles and thermophiles than mesophiles. Lower Turn propensity in thermostable proteins implies that the percentage of residues with propensity to be part of a Turn is lower in more thermostable proteins. In other words, loops are shortened in more thermostable proteins.

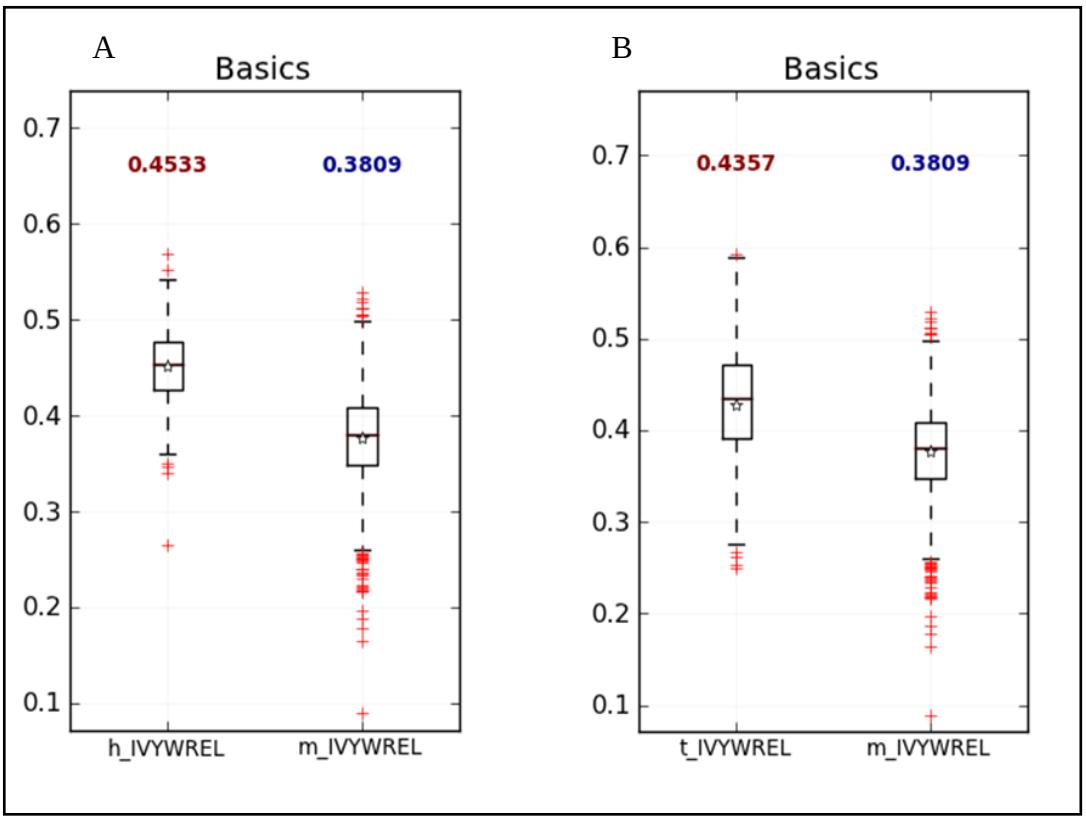

Figure 4.11 Boxplots of IVYWREL index in HM and TM datasets. The

4.3.1.2 Amino acid composition

According to KS test results, distributions of 13 amino acids are different between hyperthermophiles and thermophiles compared to mesophiles. In Figure 4.12, boxplots of only the three most significant amino acid compositions were shown for HM dataset. Glu (E) and Lys (K) residues are significantly higher in hyperthermophiles with p-values of 3.88E-54 and 5.86E-46, respectively, at the expense of thermolabile Gln (Q) residue which is significantly lower in hyperthermophiles with a p-value of 5.61E-58.

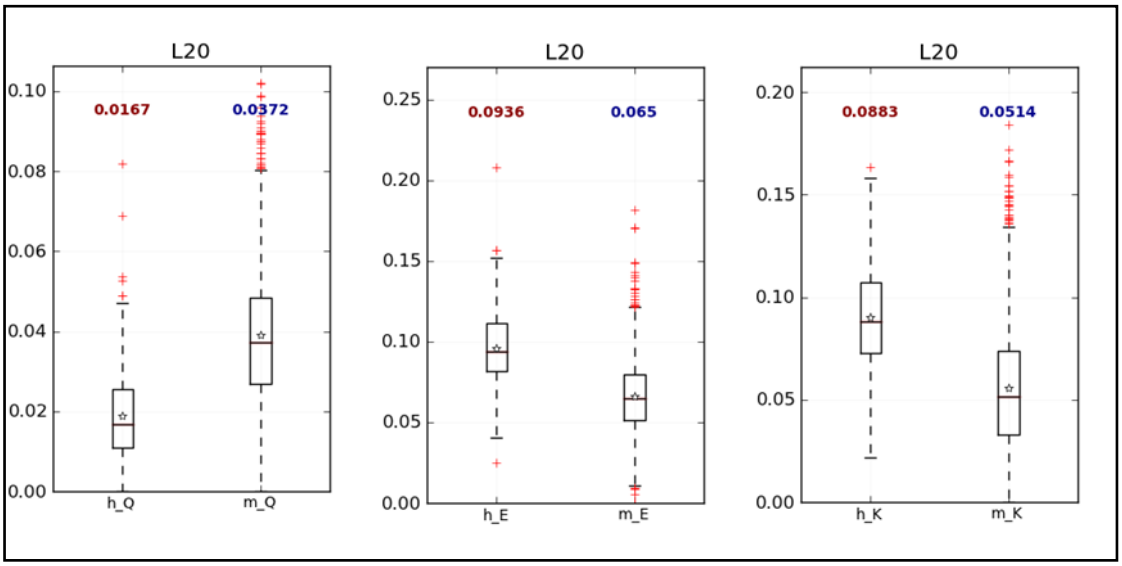

Figure 4.12 Boxplots of three most significant amino acids in HM dataset

In TM dataset, Glu (E) residue (Figure 4.13) is significantly higher in thermophiles with a p-value of 2.02E-40. On the other hand, compositions of Gln (Q) and Asp (D) are lower in hyperthermophiles than mesophiles.

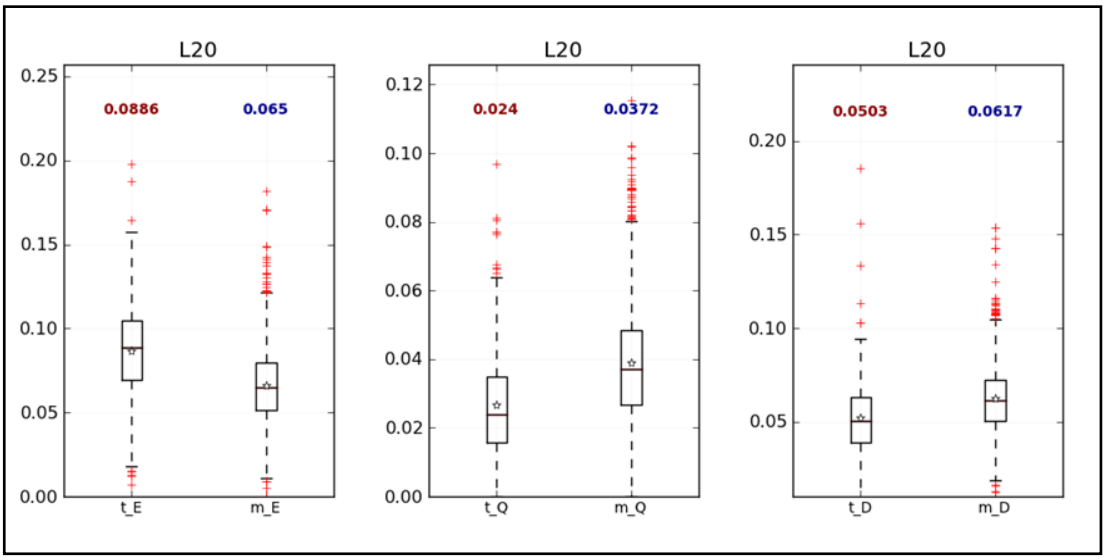

Figure 4.13 Boxplots of three most significant amino acids in TM dataset

4.3.1.3 Reduced amino acid composition

We have utilized 63 RAAAs and analyzed each one of the clusters (amino acid groupings) in these alphabets on a case by case basis for statistical significance. We have already elaborated on the implications of different RAAAs on the tree-based classification of protein families in Chapter 2 and on the classification of proteins from different thermostability classes using different RAAA and n-gram size combinations in Chapter 3. Therefore, in this chapter, we sorted all features in all RAAAs based on their p-values and report the general characteristics of only top three RAAA clusters.

The K cluster (Figure 4.14) which contains the grouping of charged residues EKR is same in Sdm11, Sdm12, Sdm13, and Sdm14 alphabets and is the most significant feature between hyperthermophiles and mesophiles in HM dataset. This grouping of residues is consistent with the finding in the previous section that E and K residues are higher in hyperthermophiles with the only addition of R residue to this duo.

Although R residue as a feature in the amino acid composition feature set had a p-value of 0.0004, it was not counted as a significant difference between hyperthermophiles and mesophiles, because the p-value was still higher than the effective two sided α-level of 0.00025 (after correcting for multiple testing and two-sided test).

Similar to the K cluster, the T clusters in Sdm alphabets of sizes 11, 12, 13, and 14 are equivalent and reflect the grouping of QST residues. This cluster is the second most significant difference between hyperthermophiles and mesophiles. The lump sum composition of QST residues (i.e., T cluster) is significantly lower in hyperthermophiles than mesophiles with median values of 10.07% and 14.89%, respectively (Figure 4.14).

The A cluster of Gbmr14 has the grouping of EKAFILV residues which correspond to the residues that are on the opposite end of the hydrophobicity scale. Gbmr14 A cluster contains positively charged EK residues and hydrophobic AFILV residues. This cluster is the third most significant difference (amongst RAAA features) between hyperthermophiles and mesophiles. The percentage of amino acids that make up the A cluster is higher in hyperthermophiles than mesophiles with 54.2% and 46.8% of all residues, respectively (Figure 4.14).

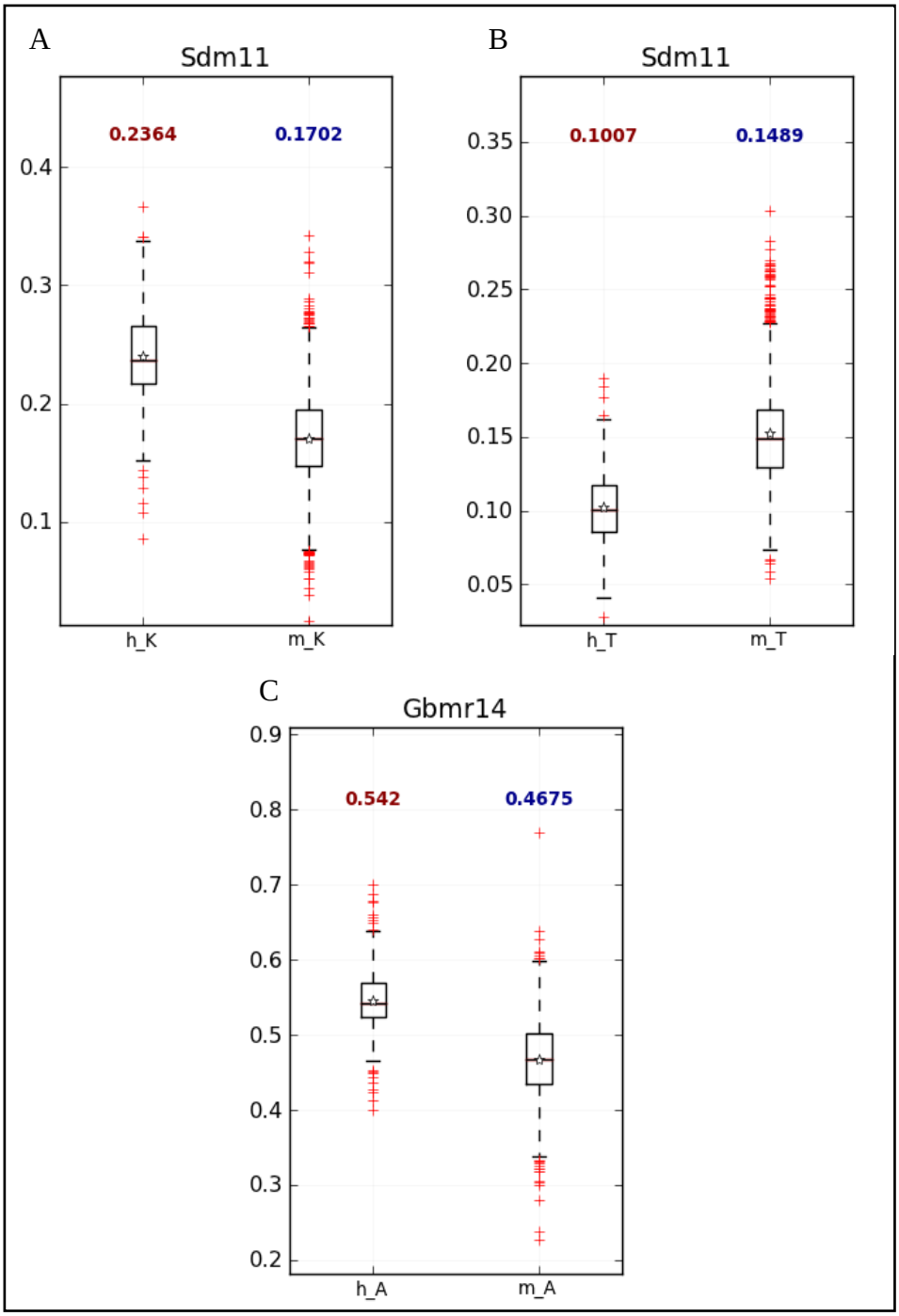

Figure 4.14 Boxplots of K and T clusters in Sdm11 alphabet and A cluster in Gbmr14 alphabet for HM dataset

4.3.1.4 Cation-Pi

Lys-Tyr and Lys-Phe interacting pairs (IP) were the only two significant features between hyperthermophiles and mesophiles (Figure 4.15). Number of Lys–Tyr or Lys–Phe IPs in hyperthermophiles is twice their values in mesophiles.

It has been suggested that several properties of Arg residues make them more suitable to higher temperatures than Lys residues: Arg has a reduced chemical reactivity due to the high pKa and resonance stabilization of its guanido group which has more surface enabling more favorable interactions. Arg residue contains one fewer methylene group than Lys residue and has less surface area for unfavorable interactions with the solvent [4].

According to the study by Folch *et al* [34], cation-π interactions involving especially those involving Arg residues are more thermostabilizing. Yet, our results indicate otherwise. The number of cation-π interactions involving Lys residues is higher in hyperthermophiles than mesophiles. Moreover and no significant difference in the number of interactions involving Arg residues has been observed between hyperthermophiles and mesophiles. These results suggest that if an increased cation-π interactions involving Arg residues is indeed stabilizing; this mechanism is not universal in either hyperthermophiles or thermophiles. No significant feature was found in the TM dataset using cation-pi feature set.

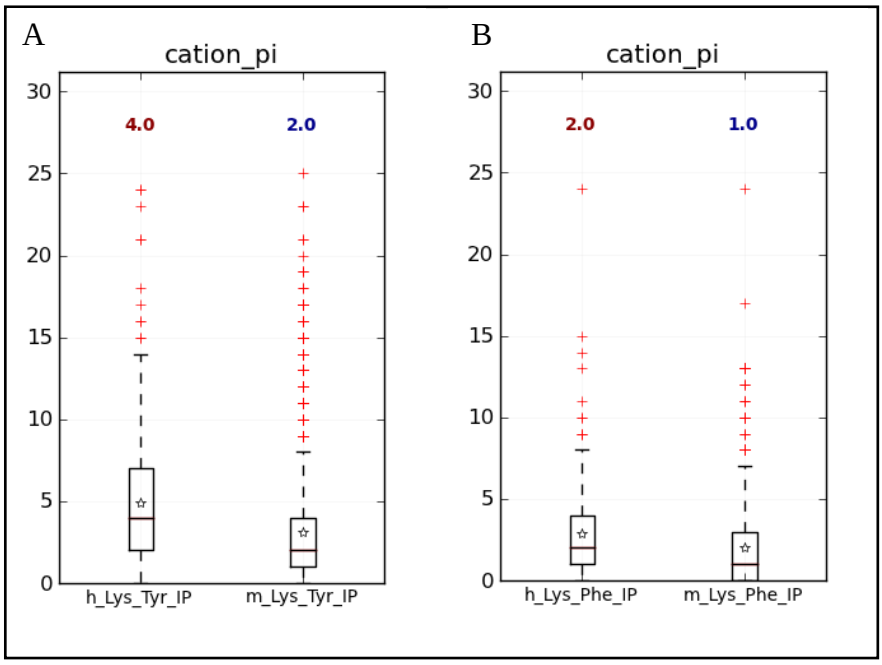

Figure 4.15 Boxplots of Lys- Tyr and Lys-Phe interacting pairs in HM dataset

4.3.1.5 Dipole related features

No obvious correlations were observed between net charges and dipole moments, or of either of them with the number of atoms, the number of residues or the mean radius; nor was there any obvious relationship between dipole or quadropole and thermostability. Even among proteins with negative charges or dipoles, there was, in most cases, no clear pattern. The two exceptions are *charges* or *charge per atom* values, which are more negative in mesophilic proteins.

C*harge per atom* (i.e., difference between positive charges and negative charges divided by the total number of atoms) and *charge* is higher in hyperthermophiles and thermophiles than mesophiles (Figure 4.16). The *charge per atom* value is 3.75 and 1.5 times higher in mesophiles (more negative in mesophiles) than hyperthermophiles and thermophiles, respectively whereas the *charge* value is 3 and 1.5 times higher in mesophiles (more negative in mesophiles) than hyperthermophiles and thermophiles, respectively. These findings imply that there is a net charge imbalance in mesophilic proteins.

According to Tekaia *et al* [6], increased Glu concentration in more thermostable proteins is correlated with an increase in the lumped pool Lys + Arg content. It appears that the ultimate result of this correlation manifests itself by giving more thermostable proteins a more balanced charge distribution over the entire structure, a property that is absent in mesophilic proteins. While mesophilic proteins have also ionazible residues, the extent that these residues are compensated through favorable ionic interactions is below the level seen in more thermostable proteins.

In addition to the *positive charge* feature whose value is higher in hyperthermophiles than mesophiles (data not shown), no other significant differences were observed in HM or TM datasets.

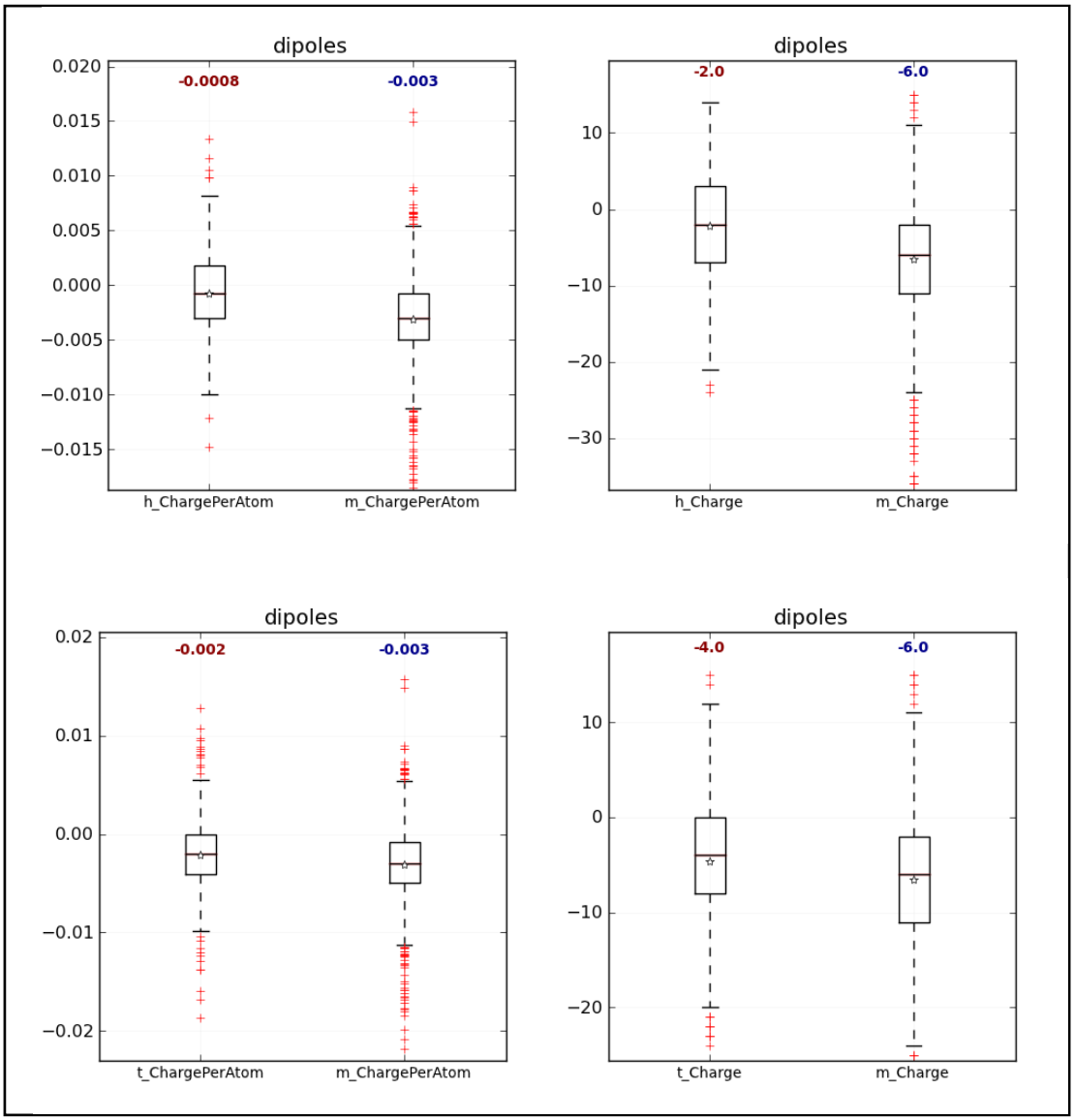

Figure 4.16 Boxplots of significant dipole related features

4.3.1.6 Salt bridges

Out of all possible ionic interactions, Lys-Glu is more prevalent (2.5 times) in hyperthermophiles than mesophiles with a p-value of 1.61E-23 (Figure 4.17-A). Similarly, Lys-Glu interactions per residue (Figure 4.17-B) is significantly higher in hyperthermophiles than mesophiles (0.0216 vs 0.0086) which implies that the increased number of Lys-Glu interaction in hyperthermophiles is not a result of longer hyperthermophilic proteins. Roughly 1 out of 4 Lys residues is involved in an ionic interaction in proteins from all three thermostability classes (Data not shown). Number of Lys residues involved in an ionic interaction (Figure 4.17-C) and total Lys (Figure 4.17-D) content are higher in hyperthermophiles than mesophiles.

Previous studies have pointed that stabilization at high temperatures is stronger for salt-bridges involving Arg residues than Lys due to the fact that Lys residue is longer and posses more rotational freedom compared to Arg [4, 34]. Surprisingly, while our statistical significance test has not selected Arg as significant between hyperthermophiles and mesophiles, salt-bridges involving Arg interacting with Glu showed a less pronounced statistical significance in TM dataset (Data not shown) immediately followed by even lesser significant salt-bridges involving Lys interacting with Glu residue. It appears that even salt-bridges involving Arg is more stabilizing in high temperatures, this mechanism is not universal because Lys is more preferred than Arg in more thermostable hyperthermophilic organisms.

Kumar and Nussinov [23] indicated that buried salt bridges prefer Arg over Lys, while exposed salt bridges prefer Lys over Arg. Since we have not made a distinction of salt-bridges based on their location, it is difficult to compare our results with the results of Kumar and Nussinov. However, an indirect conclusion of our results that is in line with the previous published results may suggest that our dataset has more proteins involving surface exposed salt-bridges containing Lys residues.

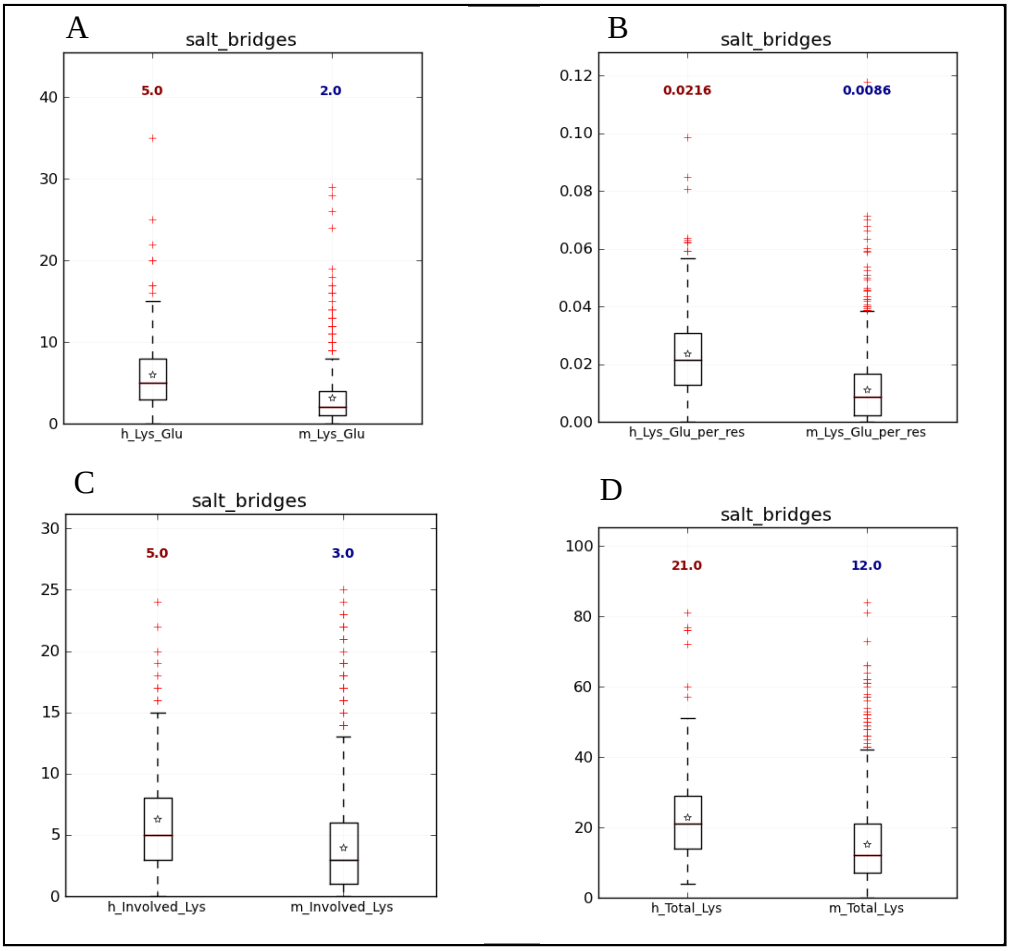

Figure 4.17 Boxplots of significant salt-bridge related features in HM dataset A) Lys-Glu interactions B) Lys-Glu interactions per residue D) Number of Lys involved in an ionic interaction C) Total number of Lys

4.3.1.7 Amino acid content in SSs

Amino acid content in SSs is a variation of amino acid composition that is tuned to reflect the preference of each amino acid to be in a particular secondary structure. For example, Ala in α-helix as a feature in this feature set indicates what percent of Ala is found in α-helices. There are a total of 30 and 12 significant features from this feature set in HM and TM datasets, respectively. Top two significant features in terms of lowest p-value in both datasets are provided in . While there are less Gln residues in α-helices of hyperthermophiles and thermophiles compared to mesophiles, there are more Lys residues in α-helices of hyperthermophiles and more Glu residues in thermophiles compared to mesophiles.

Table 4.5 Most significant feature in aa_content_in_ss feature set

			Median (%)	
	aa_content_in_ss	**p-value**	**hyper/thermo**	**meso**
HM	H_GLN	1.82E-34	0.78	1.62
	H_LYS	9.65E-31	3.76	1.79
TM	H_GLU	3.03E-16	4.18	2.94
	H_GLN	2.80E-15	1.11	1.62

4.3.1.8 Secondary structure content

In this feature set, contents of 7 different secondary structures as defined by STRIDE were tested. Based on KS-test results, only *Turn* content was significantly different between hyperthermophiles and mesophiles. *Turns* of hyperthermophiles are composed of less number of amino acids than the turns of mesophiles with median values of 17% and 19%, respectively (Figure 4.18). This finding is also in agreement with the previous results [33] obtained by homologous pair comparisons.

According to the results of Chakravarty and Varadarajan [33] based on a dataset of 900 mesophilic and 300 thermophilic protein single chains, there is an approximate decrease of 1% in the overall loop content and a corresponding increase in helical content in thermophiles. While our results indicate that the difference between median loop content of hyperthermophiles and mesophiles is approximately 2 percentage points, we could not find such a significant increase in α-helix content in either thermophiles or hyperthermophiles. This may be due to the fact that we used a finer definition of secondary structure content where α-helices were considered separate from 3_{10} and π-helices. No other significant difference was observed for other SSEs.

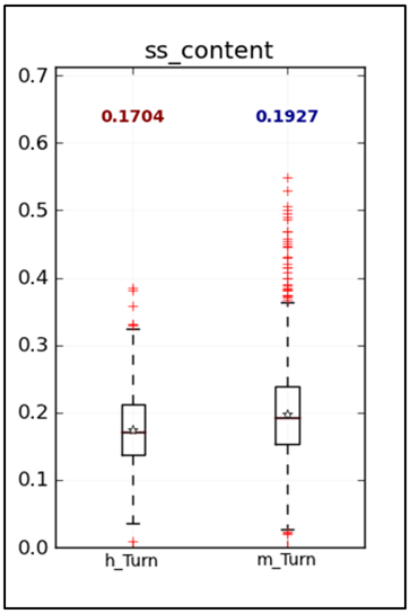

Figure 4.18 Boxplot of Turn content in HM dataset

4.3.2 Classification Results

4.3.2.1 Accuracy

A perfect method of classification would have an accuracy of 100% resulting from the correct identification of all true positives and true negatives. Our classification results indicate that highest average accuracies are achieved using mostly sequential features (Table 4.6). Classification accuracies that were obtained using structural features are slightly lower than those obtained using sequential features in both HM and TM datasets.

Generally, proteins in HM dataset are classified with higher average accuracies than the proteins in TM dataset. However, this is an expected outcome of this study because hyperthermophilic, thermophilic and mesophilic proteins are by definition on a linear temperature scale. The further the T_{opt} of an organism from the reference point of mesophilic temperatures, the easier it is to separate its proteins from the proteins of mesophiles.

Top three sequential feature sets in terms of accuracy are Native, Lwi19, and Ab19 alphabets for HM dataset and Ab19, Native and Lzmj11 alphabets for TM dataset, respectively. Native alphabet is the top performer for HM dataset with 92.5% accuracy while Ab19 alphabet is for TM dataset with 82.35% accuracy. Lwi19 feature set contains only the grouping of aliphatic IV residues and Ab19 the grouping of aromatic FY residues.

Amino_acid_content_in_ss is the top performing feature set amongst the structural feature sets for both HM and TM datasets. It contains the distribution of amino acids into 7 different secondary structural elements. While the ss_content as a feature set produced 57.19% and 54.55% average accuracies for HM and TM datasets respectively, the inclusion of amino acid distribution information in the secondary structural elements increased the accuracies to 91.39% and 77.80%. The increase in accuracy in both datasets is a major improvement over using only ss_content.

Moreover, slightly lower accuracy obtained using aa_content_in_ss compared to sequential feature sets may be attributable to the effects of high dimensionality. The accuracy may be increased simply by performing a feature selection procedure and classifying only with those selected features.

The top performing sequential and structural feature sets for the (HT)M dataset were Ab19 alphabet and aa_content_in_ss with 84.73% and 82.15% accuracies, respectively. Furthermore, the feature set that contains Native alphabet and all structural features including hinge-related features and secondary structure embedded sequence alphabets gave a classification accuracy of 83.21% for the (HT)M dataset.

On the other hand, the average classification accuracy using only the statistically significant features that were extracted from Native alphabet and all structural features gave the highest accuracy of 85.1% for the (HT)M dataset. Prior to significance testing, there were a total of 214 features. Statistical significance testing based on KS test at the confidence level of 0.01 reduced the size of the feature vector to 101 features and simultaneously resulted in higher classification accuracy.

Classification results using (HT)M dataset which contains thermostable proteins from hyperthermophiles and thermophiles as a single group indicate that using statistically significant features of the combined sequential and structural feature sets is better predictor of protein thermostability than purely sequential or structural features.

Table 4.6 Top performing sequential, structural, and combined feature sets in terms of average accuracy

* Only significant features were used in this classification scheme

Dataset	Feature Set	Avg. Acc.
	Sequential	
HM	Native	**92.95**
	Lwi19	92.81
	Ab19	92.72
TM	Ab19	**82.35**
	Native	82.33
	Lzmj11	82.10
	Structural	
HM	aa_content_in_ss	**91.39**
	dipole_related	80.97
	salt_bridges	79.91
TM	aa_content_in_ss	**77.80**
	salt_bridges	72.09
	dipole_related	65.41
	Sequential	
(HT)M	Ab19	**84.73**
	Native	84.55
	Lwi19	84.44
	Structural	
(HT)M	aa_content_in_ss	**82.15**
	salt_bridges	73.11
	dipole_related	71.74
	Combined Sequential and Structural	
(HT)M	*Native + all structural features	**85.10**
(HT)M	Native + all structural features	83.21

4.3.2.2 Effect of alphabet size on classification accuracy

In chapter 2 and 3, we have shown that a smaller size alphabet is sufficient for the classification of proteins with identical or better than accuracies than the native alphabet. A similar trend was also observed in the classification of thermophilic and mesophilic protein structures. For both datasets, there were 2 RAAAs amongst the top performers in terms of accuracy. However, the sizes of top performing alphabets were larger compared to our results in Chapter 2. This is due to the fact that HM and TM datasets contain proteins from a wide range of organisms and such a diversity of source organisms imply an equally diverse repertoire of sequences which may not be separated into different thermostability classes with smaller and coarser RAAA sizes.

4.4 Conclusions

In this chapter, we carried out one of the most comprehensive statistical analyses of proteins from different thermostability classes using conventional and novel sequential and structural features. We also generated classification models using those features to predict the thermostability class of a protein. Finding features to predict the class (e.g., enzyme function, protein family, SCOP, or CATH class) of a protein has been carried out by different research groups. In such undertakings, the aim is to achieve high accuracy, sensitivity, and specificity. Once a model with high predictive accuracy is generated, it can be used to predict the class of a newly sequenced protein or resolved protein structure and further downstream computational and experimental analysis can be carried out in a more informed manner.

In some cases, even a good predictive model cannot replace the solid results that are obtained through extensive experimental validation simply due to the fact that a novel protein may come from a different pool of sequences where different features may be involved in thermostabilization or the class labels may not be sufficient. In other words, a predictive model may be too coarse and finer divisions into more classes may be necessary to reflect the biological complexity of the system or a continuous value such as T_{opt} may need to be assigned and the problem needs to be addressed in the realm of principal component analysis. However, this is not to say that highly predictive features of protein thermostability or activity cannot be used in a manner to design proteins by improving their properties based on those features. Various studies have already been cited in the literature that takes either a statistical or machine learning approach to improve biological property, function, and activity by using sequential or structural features (refer to Chapter 1).

In general, our results indicate that sequential features are superior in terms of accuracy in both HM, TM and (HT)M datasets than structural features in a machine learning framework. This may be due to a multitude of reasons such as low dimensionality of structural feature vectors, non-universality of thermal adaptation mechanisms which yield low accuracy for individual structural feature sets.

To address this problem, a combined feature set that contains Native alphabet and all structural feature sets were created; KS-test based dimensionality reduction scheme was applied; and increased classification accuracy was achieved.

However, the point that structural features alone are not good discriminators of protein thermostability has been raised previously [64]. In their study, Taylor *et al* [64] concluded that combined sequential and structural features are only slightly better discriminators of protein thermostability than either sequential or structural features alone.

Surprisingly, while sequential features are --in general-- better predictors of protein thermostability than structural features alone, addition of structural features may or may not increase prediction accuracy. The uncertainty can be expunged by striking a balance between the antagonistic effects of two different phenomena: High dimensional feature space vs. inclusion of informative features. While increasing the size of the feature space for a limited number of samples may increase the classification accuracy up to a certain practical limit (see Chapter 3, Figure 3.2), it is difficult to ensure that the total number of features used is not well above the level to obtain optimum classification accuracy. Therefore, using only statistical significant features of the combined feature set gave better classification accuracy than using either sequential or structural feature sets.

On the other hand, KS-test results indicate that the distributions of many sequential and structural features are significantly different in hyperthermophiles and thermophiles compared to the control set of mesophiles. While some of our results overlap with the previously published results, there were also sources of separation. For example, while salt-bridges involving Arg were believed to be more thermostabilizing, our results indicate that in fact salt-bridges involving Lys residues were more significant in more thermostable hyperthermophilic proteins.

While we could extract a diverse set of significant sequential and structural features from both HM and TM datasets, utilization of some of these feature sets, mostly structural, did not translate to high classification accuracies, especially in TM dataset. We believe that other structural factors such as hydrogen bonding pattern, defining protein exterior or interior for the calculation of structural features, metal binding capacity, and post-translational modifications can be included future studies. Further research may also be built upon the effects of these factors and systematic combination of novel sequential and structural features on protein thermostability prediction.

4.5 References

1. Berman H, Henrick K, Nakamura H: Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 2003, 10(12):980.
2. Hermanson GT: Bioconjugate techniques, 2nd edn. Amsterdam ; Boston: Elsevier Academic Press; 2008.
3. Dill KA: Dominant forces in protein folding. *Biochemistry* 1990, 29(31):7133-7155.
4. Vieille C, Zeikus GJ: Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 2001, 65(1):1-43.
5. Ponnuswamy PK, Muthusamy R, Manavalan P: Amino acid composition and thermal stability of proteins. *International Journal of Biological Macromolecules* 1982, 4(3):186-190.
6. Tekaia F, Yeramian E, Dujon B: Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 2002, 297(1-2):51-60.
7. Gromiha MM, Suresh MX: Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. *Proteins* 2008, 70(4):1274-1279.
8. Gliakina AV, Lobanov AV, Galzitskaia OV: [Search for structural factors responsible for the stability of proteins from thermophilic organisms]. *Mol Biol (Mosk)* 2007, 41(4):681-687.
9. Kumar S, Tsai CJ, Nussinov R: Factors enhancing protein thermostability. *Protein Eng* 2000, 13(3):179-191.
10. Yokota K, Satou K, Ohki S: Comparative analysis of protein thermo stability: Differences in amino acid content and substitution at the surfaces and in the core regions of thermophilic and mesophilic proteins. *Sci Technol Adv Mater* 2006, 7(3):255-262.
11. Zeldovich KB, Berezovsky IN, Shakhnovich EI: Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 2007, 3(1):e5.
12. Zhang T, Bertelsen E, Alber T: Entropic effects of disulphide bonds on protein stability. *Nat Struct Biol* 1994, 1(7):434-438.
13. Zavodszky M, Chen CW, Huang JK, Zolkiewski M, Wen L, Krishnamoorthi R: Disulfide bond effects on protein stability: designed variants of Cucurbita maxima trypsin inhibitor-V. *Protein Sci* 2001, 10(1):149-160.
14. Wakarchuk WW, Sung WL, Campbell RL, Cunningham A, Watson DC, Yaguchi M: Thermostabilization of the Bacillus circulans xylanase by the introduction of disulfide bonds. *Protein Eng* 1994, 7(11):1379-1386.
15. Khan AR, Deber CM: An engineered disulfide bridge in the transmembrane region of phage M13 coat protein stabilizes the alpha-helical dimer. *Biochem Biophys Res Commun* 1995, 206(1):230-237.
16. Yamaguchi S, Takeuchi K, Mase T, Oikawa K, McMullen T, Derewenda U, McElhaney RN, Kay CM, Derewenda ZS: The consequences of engineering an extra disulfide bond in the Penicillium camembertii mono- and diglyceride specific lipase. *Protein Eng* 1996, 9(9):789-795.

17. Imani M, Hosseinkhani S, Ahmadian S, Nazari M: Design and introduction of a disulfide bridge in firefly luciferase: increase of thermostability and decrease of pH sensitivity. *Photochem Photobiol Sci* 2010, 9(8):1167-1177.
18. Matsumura M, Signor G, Matthews BW: Substantial increase of protein stability by multiple disulphide bonds. *Nature* 1989, 342(6247):291-293.
19. Bogin O, Levin I, Hacham Y, Tel-Or S, Peretz M, Frolow F, Burstein Y: Structural basis for the enhanced thermal stability of alcohol dehydrogenase mutants from the mesophilic bacterium Clostridium beijerinckii: contribution of salt bridging. *Protein Sci* 2002, 11(11):2561-2574.
20. Tomazic SJ, Klibanov AM: Why is one Bacillus alpha-amylase more resistant against irreversible thermoinactivation than another? *J Biol Chem* 1988, 263(7):3092-3096.
21. Matsutani M, Hirakawa H, Nishikura M, Soemphol W, Ali IA, Yakushi T, Matsushita K: Increased number of Arginine-based salt bridges contributes to the thermotolerance of thermotolerant acetic acid bacteria, Acetobacter tropicalis SKU1100. *Biochem Biophys Res Commun* 2011, 409(1):120-124.
22. Hendsch ZS, Tidor B: Do Salt Bridges Stabilize Proteins - a Continuum Electrostatic Analysis. *Protein Sci* 1994, 3(2):211-226.
23. Kumar S, Nussinov R: Salt bridge stability in monomeric proteins. *J Mol Biol* 1999, 293(5):1241-1255.
24. Kumar S, Tsai CJ, Ma BY, Nussinov R: Contribution of salt bridges toward protein thermostability. *J Biomol Struct Dyn* 2000:79-85.
25. Karshikoff A, Ladenstein R: Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads. *Trends Biochem Sci* 2001, 26(9):550-556.
26. Ge M, Xia XY, Pan XM: Salt bridges in the hyperthermophilic protein Ssh10b are resilient to temperature increases. *J Biol Chem* 2008, 283(46):31690-31696.
27. Ikai A: Thermostability and aliphatic index of globular proteins. *J Biochem* 1980, 88(6):1895-1898.
28. Merkler DJ, Farrington GK, Wedler FC: Protein thermostability. Correlations between calculated macroscopic parameters and growth temperature for closely related thermophilic and mesophilic bacilli. *Int J Pept Protein Res* 1981, 18(5):430-442.
29. Lu B, Wang G, Huang P: [A comparison of amino acid composition of proteins from thermophiles and mesophiles]. *Wei Sheng Wu Xue Bao* 1998, 38(1):20-25.
30. Grimme S: Do Special Noncovalent π–π Stacking Interactions Really Exist? *Angewandte Chemie International Edition* 2008, 47(18):3430-3434.
31. Wang L, Sun N, Terzyan S, Zhang X, Benson DR: A histidine/tryptophan pi-stacking interaction stabilizes the heme-independent folding core of microsomal apocytochrome b5 relative to that of mitochondrial apocytochrome b5. *Biochemistry (Mosc)* 2006, 45(46):13750-13759.
32. Gallivan JP, Dougherty DA: Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A* 1999, 96(17):9459-9464.
33. Chakravarty S, Varadarajan R: Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. *Biochemistry (Mosc)* 2002, 41(25):8152-8161.
34. Folch B, Dehouck Y, Rooman M: Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophysical Journal* 2010, 98(4):667-677.

35. Parthasarathy S, Murthy MR: Protein thermal stability: insights from atomic displacement parameters (B values). *Protein Eng* 2000, 13(1):9-13.
36. Jochens H, Aerts D, Bornscheuer UT: Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Eng Des Sel* 2010, 23(12):903-909.
37. Karplus PA, Schulz GE: Prediction of Chain Flexibility in Proteins - a Tool for the Selection of Peptide Antigens. *Naturwissenschaften* 1985, 72(4):212-213.
38. Vihinen M, Torkkila E, Riikonen P: Accuracy of protein flexibility predictions. *Proteins* 1994, 19(2):141-149.
39. Vihinen M: Relationship of protein flexibility to thermostability. *Protein Eng* 1987, 1(6):477-480.
40. Carugo O, Argos P: Accessibility to internal cavities and ligand binding sites monitored by protein crystallographic thermal factors. *Proteins* 1998, 31(2):201-213.
41. Mohan S, Sinha N, Smith-Gill SJ: Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophys J* 2003, 85(5):3221-3236.
42. Yuan Z, Zhao J, Wang ZX: Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng* 2003, 16(2):109-114.
43. Carugo O, Argos P: Correlation between side chain mobility and conformation in protein structures. *Protein Eng* 1997, 10(7):777-787.
44. Eyal E, Najmanovich R, Edelman M, Sobolev V: Protein side-chain rearrangement in regions of point mutations. *Proteins* 2003, 50(2):272-282.
45. Chung JL, Wang W, Bourne PE: Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* 2006, 62(3):630-640.
46. Eijsink VG, Vriend G, van den Burg B, van der Zee JR, Venema G: Increasing the thermostability of a neutral protease by replacing positively charged amino acids in the N-terminal turn of alpha-helices. *Protein Eng* 1992, 5(2):165-170.
47. Nicholson H, Anderson DE, Dao-pin S, Matthews BW: Analysis of the interaction between charged side chains and the alpha-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* 1991, 30(41):9816-9828.
48. Huang SL, Wu LC, Liang HK, Pan KT, Horng JT, Ko MT: PGTdb: a database providing growth temperatures of prokaryotes. *Bioinformatics* 2004, 20(2):276-278.
49. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al*: Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009, 25(11):1422-1423.
50. Hunter JD: Matplotlib: A 2D graphics environment. *Comput Sci Eng* 2007, 9(3):90-95.
51. Lobry JR, Gautier C: Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res* 1994, 22(15):3174-3180.
52. Kyte J, Doolittle RF: A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982, 157(1):105-132.
53. Guruprasad K, Reddy BV, Pandit MW: Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo

stability of a protein from its primary sequence. *Protein Eng* 1990, 4(2):155-161.
54. Heinig M, Frishman D: STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Research* 2004, 32(Web Server issue):W500-502.
55. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995, 247(4):536-540.
56. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: CATH--a hierarchic classification of protein domain structures. *Structure* 1997, 5(8):1093-1108.
57. Felder CE, Prilusky J, Silman I, Sussman JL: A server and database for dipole moments of proteins. *Nucleic Acids Res* 2007, 35(Web Server issue):W512-521.
58. Samanta U, Bahadur RP, Chakrabarti P: Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng* 2002, 15(8):659-667.
59. Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T: HingeProt: automated prediction of hinges in protein structures. *Proteins* 2008, 70(4):1219-1227.
60. Shapiro SS, Wilk MB: An analysis of variance test for normality (complete samples). *Biometrika* 1965, 52(3-4):591-611.
61. EL-Manzalawy Y, Honavar V: {WLSVM}: Integrating LibSVM into Weka Environment.[http://www.cs.iastate.edu/~yasser/wlsvm]
62. Chang C, Lin C: {LIBSVM}: a library for support vector machines.[http://www.csie.ntu.edu.tw/~cjlin/libsvm]
63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: The WEKA data mining software: an update. *SIGKDD Explor Newsl* 2009, 11(1):10-18.
64. Taylor TJ, Vaisman, II: Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol* 2010, 10 Suppl 1:S5.
65. Zeldovich KB, Berezovsky IN, Shakhnovich EI: Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Computational Biology* 2007, 3(1):e5.

CHAPTER 5

5 CONCLUSIONS

Proteins are complex biological molecules that perform a tremendous variety of functions in cells under diverse physiological conditions. Proteins can be classified based on their function, structural motifs that they possess, cellular location or adaptation to an external variable such as temperature, pH or salinity. All the diversity that is present in protein structure and function is encoded in protein sequence which determines many weak non-covalent interactions that ultimately determines a protein's structure, function or response to an external variable. Bridging the gap between protein sequence and structure or function is an active area of research in Bioinformatics that will enable us to design proteins that are optimized for biotechnological processes or provide the means to determine a target attribute of a novel protein (e.g., structural class, pH and temperature stability, important binding sites, half life and etc.) with high accuracy.

To that end, we introduced three different strategies to systematically classify proteins with high accuracy using novel and conventional sequential and structural features. In Chapter 2, we introduced an RAAA-based approach as a preprocessing of protein sequences prior to calculating RCMs. The devised procedure may be used as an alternative to the widely used MSA method for the classification of proteins into functional subtypes. The procedure does not require expert handling of the data and is computationally fast.

Utilization of an RAAA that is consistent with the structure and function of the proteins or an RAAA that reflects the general trends in specific protein families under study can result in successful phylogenies that can cluster each protein superfamily into functional subtypes.

Moreover, RCM can also be used to find the underlying sequential differences in exhaustive histories in two sequences when they are concatenated. The "words" that result in an observed difference can then be analyzed and correlated to a functional

and/or evolutionary origin. We believe future work can focus in this direction building on the current approach that does not attempt to trace back the origin of differentiating sequence signals but provides a powerful classification-via-clustering method of protein families into functional subtypes without using multiple sequence alignment.

In Chapter 3, we systematically tested the potential of using different n-gram and RAAA combinations to classify proteins from thermophiles and mesophiles in a machine learning framework. Our results indicate that classification accuracy usually decreases with increasing n-gram sizes for a given RAAA. Classification using the procedure outlined in Chapter 3 has produced better results with fewer features than the native alphabet in terms of accuracy. Our results also indicate that RAAAs can improve classification performance relative to standard protein alphabet. Furthermore, performing t-test to reduce the size of the feature space decreases computational time without significantly affecting classification accuracy and makes classification with 3-grams possible. A future avenue of research in this area may involve carrying out research in generating organism-specific RAAAs, and separating thermostability classes by phyla.

In Chapter 4, we carried out one of the most comprehensive statistical analyses of different sequential and structural features and generated classification models using those features to predict thermostability class of a protein. Finding features that correlate well with the thermostability class of a protein can be used to understand the evolutionary response to high environmental temperatures and further downstream computational and experimental analysis can be carried out in a more informed manner.

In general, our results indicate that combined sequential and structural features are better predictors of protein thermostability than using purely sequential or structural features. Furthermore, the fact that structural features alone are not as good predictors as sequential features may be due to a multitude of reasons such as i) low dimensionality of structural feature vectors, ii) heterogeneity of the structural features where each structural feature set is not a good discriminator but rather combinations of feature sets may need to be tested for their predictive accuracy, and iii) low coverage of structural features in each protein structure.

In addition, we have shown that thermostable proteins have both sequence and structure based preferences based on statistical significance testing on each feature between the proteins of different thermostability classes. While some of our results are

in agreement with the previously published results, sources of separation were also borne out which require further studies in the area of protein thermostability with larger datasets. The fact that many structural feature sets that contain highly significant features did not even result 100% accuracies, implies that there is not a universal set of features that works for the thermostabilization of all proteins but rather a combination of different mechanisms may be determining the delicate balance of protein thermostability and unfolding at high temperatures.

We have seen that protein thermostability is a phenomenon that is complicated by high level of sequence and structure similarity between proteins of different thermostability classes, the lack of theoretical knowledge about the temperature dependence of the interactions that stabilize protein structures, and the multitude of ways that can be used to achieve thermostability. While numerous studies comparing homologous proteins or proteins belonging to organisms from different thermostability classes indicate that there is a series of thermostability-influencing factors, many such factors do not seem to be universal [1].

Lately, it was suggested that there are two kinds of adaptation to high temperatures: a structure adaptation undergone by proteins coming from archaea, and a sequence adaptation undergone by proteins coming from mesophilic organisms that have been transferred to extremely hot environments [1]. To test this hypothesis, proteins from different thermostability that are also separated by phyla can be generated and their sequential and structural features can be extracted and classified.

More recently, structural alphabets have emerged in the literature that have been obtained by converting 3D protein structures to the corresponding 1D structural letter sequences (i.e., structural sequences). They have been implicated to produce superior results in metal- or ligand-binding site discovery [2], 1D motif detection methods with structural alphabets to discover locally conserved protein structural motifs [3, 4], classification of proteins that belong to distinct folds of CATH [5].

Structural motif discovery methods based on protein structural alphabets can be applied to any set of proteins with known 3D structures. These new alphabets are timely considering the increasing number of structures for proteins with unknown function that are being solved from structural genomics initiatives. For such proteins, which share no significant sequence homology to proteins of known function, the presence of a structural motif that maps to a specific protein function in the structure would suggest

likely active/binding sites and a particular biological function. Moreover, the effects of such structural motifs on protein thermostability can also be investigated.

One of the mechanisms of protein thermostabilization that we have not touched upon in this thesis is the metal binding properties. It has been suggested that metal binding may also be a significant determinant in protein thermostabilization. The fact that metal binding regions of a protein structure can be found using structural protein alphabets makes it all easier to determine whether a correlation exist between metal binding capacity of proteins and their thermostability classes.

5.1 References

1. Folch B, Dehouck Y, Rooman M: Thermo- and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophys J* 2010, 98(4):667-677.
2. Dudev M, Lim C: Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. *BMC Bioinformatics* 2007, 8:106.
3. Ku SY, Hu YJ: Protein structure search and local structure characterization. *BMC Bioinformatics* 2008, 9:349.
4. Ku SY, Hu YJ: Discovery of structural motifs using protein structural alphabets and 1D motif-finding methods. *Adv Exp Med Biol* 2010, 680:117-123.
5. Le Q, Pollastri G, Koehl P: Structural alphabets for protein structure classification: a comparison study. *J Mol Biol* 2009, 387(2):431-450.

APPENDIX A

TBC errors for all protein families and all RAAAs

			Crotonases	
			467 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	100	100
ML15	Accuracy	%	100	100
ML10	Accuracy	%	100	100
ML8	Accuracy	%	96.4	100
ML4	Accuracy	%	99.6	100
EB13	Accuracy	%	100	100
EB11	Accuracy	%	100	100
EB9	Accuracy	%	99.8	100
EB8	Accuracy	%	100	100
EB5	Accuracy	%	98.5	100
HSDM17	Accuracy	%	100	100
SDM12	Accuracy	%	100	100
GBMR4	Accuracy	%	100	100
RANDOM4[§]	Accuracy	%	76.8	98.9

			Mandale racemases	
			184 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	100	100
ML15	Accuracy	%	100	100
ML10	Accuracy	%	100	100
ML8	Accuracy	%	100	100
ML4	Accuracy	%	100	100
EB13	Accuracy	%	100	100
EB11	Accuracy	%	100	100
EB9	Accuracy	%	100	100
EB8	Accuracy	%	100	100
EB5	Accuracy	%	100	100
HSDM17	Accuracy	%	100	100
SDM12	Accuracy	%	100	100
GBMR4	Accuracy	%	96.7	100
RANDOM4[§]	Accuracy	%	96.9	100

			Vicinal oxygen chelates	
			309 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	91.6	91.3
ML15	Accuracy	%	91.6	81.9
ML10	Accuracy	%	91.9	88.3
ML8	Accuracy	%	89.6	89
ML4	Accuracy	%	86.1	91.3
EB13	Accuracy	%	91.6	91.3
EB11	Accuracy	%	90.9	82.8
EB9	Accuracy	%	90.9	91.3
EB8	Accuracy	%	92.2	91.3
EB5	Accuracy	%	77.7	90.6
HSDM17	Accuracy	%	89	88.7
SDM12	Accuracy	%	89	90.3
GBMR4	Accuracy	%	91.3	91.3
RANDOM4[§]	Accuracy	%	66.8	84.4

			Haloacid dehalogenases	
			195 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	93.3	99.5
ML15	Accuracy	%	96.9	99
ML10	Accuracy	%	91.3	99.5
ML8	Accuracy	%	92.8	99.5
ML4	Accuracy	%	92.8	98.5
EB13	Accuracy	%	93.3	99
EB11	Accuracy	%	93.3	99.5
EB9	Accuracy	%	92.8	99
EB8	Accuracy	%	92.8	98.5
EB5	Accuracy	%	88.2	97.4
HSDM17	Accuracy	%	93.8	99
SDM12	Accuracy	%	96.4	99
GBMR4	Accuracy	%	94.4	99
RANDOM4[§]	Accuracy	%	60.3	93.7

			Nucleotidyl cyclases	
			75 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	100	100
ML15	Accuracy	%	100	100
ML10	Accuracy	%	100	100
ML8	Accuracy	%	100	100
ML4	Accuracy	%	100	98.7
EB13	Accuracy	%	100	100
EB11	Accuracy	%	100	100
EB9	Accuracy	%	100	100
EB8	Accuracy	%	98.7	100
EB5	Accuracy	%	98.7	98.7
HSDM17	Accuracy	%	100	100
SDM12	Accuracy	%	100	100
GBMR4	Accuracy	%	100	100
RANDOM4[§]	Accuracy	%	97.3	97.8

			Acyl transferases	
			177 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	91.5	97.2
ML15	Accuracy	%	96.6	97.2
ML10	Accuracy	%	97.2	96.6
ML8	Accuracy	%	97.2	97.2
ML4	Accuracy	%	97.2	97.2
EB13	Accuracy	%	97.2	97.2
EB11	Accuracy	%	97.2	97.2
EB9	Accuracy	%	97.2	97.2
EB8	Accuracy	%	96.6	97.2
EB5	Accuracy	%	91.5	97.2
HSDM17	Accuracy	%	97.2	96.6
SDM12	Accuracy	%	97.2	97.2
GBMR4	Accuracy	%	87.6	97.2
RANDOM4[§]	Accuracy	%	91.5	95.5

			GH2 hydrolases	
			33 sequences	
			RCM	MSA
Reduced Alphabet				
20*	Accuracy	%	87.9	100
ML15	Accuracy	%	97	100
ML10	Accuracy	%	87.9	100
ML8	Accuracy	%	100	100
ML4	Accuracy	%	100	100
EB13	Accuracy	%	87.9	100
EB11	Accuracy	%	97	100
EB9	Accuracy	%	97	100
EB8	Accuracy	%	100	100
EB5	Accuracy	%	97	100
HSDM17	Accuracy	%	63.6	100
SDM12	Accuracy	%	87.9	100
GBMR4	Accuracy	%	100	100
RANDOM4[§]	Accuracy	%	62.6	99

APPENDIX B

RAAA groupings

Ab10 = [A:A E:EKQR D:DNS G:G I:IV H:CH L:ML P:P T:T W:FWY]
Ab11 = [A:A E:EKQR D:D G:G I:IV H:CH L:ML P:P S:NS T:T W:FWY]
Ab12 = [A:A C:C E:EKQR D:D G:G I:IV H:H L:ML P:P S:NS T:T W:FWY]
Ab13 = [A:A C:C E:E D:D G:G I:IV H:H K:KQR L:ML P:P S:NS T:T W:FWY]
Ab14 = [A:A C:C E:E D:D G:G I:IV H:H K:KQR L:ML N:N P:P S:S T:T W:FWY]
Ab15 = [A:A C:C E:E D:D G:G I:IV H:H K:KQR L:ML N:N P:P S:S T:T W:W Y:FY]
Ab16 = [A:A C:C E:E D:D G:G I:IV H:H K:KQR M:M L:L N:N P:P S:S T:T W:W Y:FY]
Ab17 = [A:A C:C E:E D:D G:G I:IV H:H K:K M:M L:L N:N P:P S:S R:QR T:T W:W Y:FY]
Ab18 = [A:A C:C E:E D:D G:G I:I H:H K:K M:M L:L N:N P:P S:S R:QR T:T W:W V:V Y:FY]
Ab19 = [A:A C:C E:E D:D G:G I:I H:H K:K M:M L:L N:N Q:Q P:P S:S R:R T:T W:W V:V Y:FY]

Dssp10 = [A:AM C:C E:EKQR D:DNS G:GP F:F I:IV H:HT L:LY W:W]
Dssp11 = [A:AEKQR C:C D:DNS G:GP F:F I:IV H:H L:ML T:T W:W Y:Y]
Dssp12 = [A:AEKQR C:C D:DNS G:GP F:F I:I H:H L:ML T:T W:W V:V Y:Y]
Dssp13 = [A:AEKQR C:C D:DNS G:GP F:F I:I H:H M:M L:L T:T W:W V:V Y:Y]
Dssp14 = [A:A C:C D:EDKQ G:GNPS F:F I:I H:H M:M L:L R:R T:T W:W V:V Y:Y]

Et11 = [A:A C:C E:EKQR G:G F:FWY I:IV H:HS L:ML N:DN P:P T:T]
Et13 = [A:A C:C E:E G:G F:FWY I:IV H:HS M:M L:L N:DN Q:KQR P:P T:T]

Gbmr10 = [A:AEFIKMLQRWV C:C D:D G:G H:H N:N P:P S:S T:T Y:Y]
Gbmr11 = [A:AEFIKMLQRV C:C D:D G:G H:H N:N P:P S:S T:T W:W Y:Y]
Gbmr12 = [A:AEFIKMLQV C:C D:D G:G H:H N:N P:P S:S R:R T:T W:W Y:Y]
Gbmr13 = [A:AEFIKMLV C:C D:D G:G H:H N:N Q:Q P:P S:S R:R T:T W:W Y:Y]
Gbmr14 = [A:AEFIKLV C:C D:D G:G H:H M:M N:N Q:Q P:P S:S R:R T:T W:W Y:Y]

Hsdm10 = [A:A C:C D:DN G:G H:H L:FIMLV P:P T:EKQSRT W:W Y:Y]
Hsdm12 = [A:A C:C D:DN G:G H:H M:M L:FILV P:P R:R T:EKQST W:W Y:Y]
Hsdm14 = [A:A C:C D:DN G:G F:F H:H K:EKQ M:M L:ILV P:P R:R T:ST W:W Y:Y]
Hsdm15 = [A:A C:C D:D G:G F:F H:H K:EKQ M:M L:ILV N:N P:P R:R T:ST W:W Y:Y]
Hsdm16 = [A:A C:C D:D G:G F:F H:H K:EK M:M L:ILV N:N Q:Q P:P R:R T:ST W:W Y:Y]
Hsdm17 = [A:A C:C D:D G:G F:F H:H K:EK M:M L:ILV N:N Q:Q P:P S:S R:R T:T W:W Y:Y]

Native20 = [A:A C:C E:E D:D G:G F:F I:I H:H K:K M:M L:L N:N Q:Q P:P S:S R:R T:T W:W V:V Y:Y]

Lr10 = [A:AST C:C D:EDN G:G F:FY I:IMLV H:H K:KQR P:P W:W]

Lwi10 = [C:AC D:ED G:G F:FWY I:IV M:ML N:HN Q:KQR P:P T:ST]
Lwi11 = [C:AC D:D G:G F:FWY I:IV M:ML N:HN Q:EQ P:P R:KR T:ST]
Lwi12 = [A:A C:C D:D G:G F:FWY I:IV M:ML N:HN Q:EQ P:P R:KR T:ST]
Lwi13 = [A:A C:C D:D G:G F:FWY I:IV M:ML N:HN Q:EQ P:P S:S R:KR T:T]
Lwi14 = [A:A C:C D:D G:G F:FWY I:IV K:K M:ML N:HN Q:EQ P:P S:S R:R T:T]
Lwi15 = [A:A C:C D:D G:G F:FWY I:IV H:H K:K M:ML N:N Q:EQ P:P S:S R:R T:T]
Lwi16 = [A:A C:C D:D G:G F:FY I:IV H:H K:K M:ML N:N Q:EQ P:P S:S R:R T:T W:W]
Lwi17 = [A:A C:C E:E D:D G:G F:FY I:IV H:H K:K M:ML N:N Q:Q P:P S:S R:R T:T W:W]
Lwi18 = [A:A C:C E:E D:D G:G F:FY I:IV H:H K:K M:M L:L N:N Q:Q P:P S:S R:R T:T W:W]
Lwi19 = [A:A C:C E:E D:D G:G F:F I:IV H:H K:K M:M L:L N:N Q:Q P:P S:S R:R T:T W:W Y:Y]

Lwni10 = [A:AST C:C G:G F:FWY I:IV M:ML N:HN Q:EDQ P:P R:KR]
Lwni11 = [A:A C:C G:G F:FWY I:IV M:ML N:HN Q:EDQ P:P R:KR T:ST]
Lwni14 = [A:A C:C D:D G:G F:FWY I:IV H:H M:ML N:N Q:EQ P:P S:S R:KR T:T]
Lzbl10 = [A:A C:C D:D G:G F:FWY I:IMV L:L P:P S:HNST R:EKQR]
Lzbl11 = [A:A C:C E:E D:D G:G F:FWY I:IMV L:L P:P S:HNST R:KQR]
Lzbl12 = [A:A C:C E:E D:D G:G F:FWY I:IMV H:HKQR L:L N:N P:P S:ST]
Lzbl13 = [A:A C:C E:E D:D G:G F:F I:IMV H:HKQR L:L N:N P:P S:ST W:WY]
Lzbl14 = [A:A C:C E:E D:D G:G F:F I:IMV H:HKQR L:L N:N P:P S:S T:T W:WY]
Lzbl15 = [A:A C:C E:E D:D G:G F:F I:IMV H:H L:L N:N P:P S:S R:KQR T:T W:WY]
Lzbl16 = [A:A C:C E:E D:D G:G F:F I:IMV H:H L:L N:N P:P S:S R:KQR T:T W:W Y:Y]

Lzmj10 = [A:A C:C E:E D:D G:G I:ILV M:FM R:KR T:NST W:HQPWY]
Lzmj11 = [A:A C:C E:E D:D G:G I:IL M:FM R:KR T:NST W:HQPWY V:V]
Lzmj12 = [A:A C:C E:E D:D G:G I:IL M:FM N:N R:KR T:ST W:HQPWY V:V]
Lzmj13 = [A:A C:C E:E D:D G:G I:IL M:FM N:N S:S R:KR T:T W:HQPWY V:V]
Lzmj14 = [A:A C:C E:E D:D G:G I:I M:FM L:L N:N S:S R:KR T:T W:HQPWY V:V]
Lzmj15 = [A:A C:C E:E D:D G:G I:IL H:H M:FM N:N P:P S:S R:KR T:T W:QWY V:V]
Lzmj16 = [A:A C:C E:E D:D G:G I:I H:H M:FM L:L N:N P:P S:S R:KR T:T W:QWY V:V]

Ml10 = [A:A C:C E:EDNQ G:G F:FWY H:H K:KR L:IMLV P:P S:ST]
Ml15 = [A:A C:C E:E D:D G:G F:FY H:H K:KR L:IMLV N:N Q:Q P:P S:S T:T W:W]

Random10 = [A:AC D:EDM F:F I:I N:N Q:GQ R:R T:KTY W:LSW V:HPV]

Sdm10 = [A:A C:C D:DN G:G H:H L:IMLV P:P T:EKQSRT W:W Y:FY]
Sdm11 = [A:A C:C D:DN G:G H:H K:EKR L:IMLV P:P T:QST W:W Y:FY]
Sdm12 = [A:A C:C D:D G:G H:H K:EKR L:IMLV N:N P:P T:QST W:W Y:FY]
Sdm13 = [A:A C:C D:D G:G F:F H:H K:EKR L:IMLV N:N P:P T:QST W:W Y:Y]
Sdm14 = [A:A C:C D:D G:G F:F H:H K:EKR M:M L:ILV N:N P:P T:QST W:W Y:Y]

APPENDIX C

Classification performance in terms of sensitivity, specificity, accuracy, and AUC for all RAAAs and n-grams.

RAAA	N-grams	Accuracy	AUC	Sensitivity	Specificity
Ab10	1	84.5827	0.929	0.916	0.763
Ab10	2	84.017	0.926	0.895	0.775
Ab10	3	83.7341	0.904	0.83	0.846
Ab11	1	84.017	0.909	0.906	0.763
Ab11	2	86.1386	0.925	0.903	0.812
Ab11	3	82.0368	0.899	0.83	0.809
Ab12	1	83.7341	0.906	0.906	0.757
Ab12	2	86.1386	0.922	0.898	0.818
Ab12	3	82.6025	0.898	0.835	0.815
Ab13	1	87.8359	0.939	0.929	0.818
Ab13	2	88.5431	0.938	0.893	0.877
Ab13	3	83.7341	0.906	0.827	0.849
Ab14	1	86.5629	0.931	0.914	0.809
Ab14	2	85.1485	0.924	0.851	0.852
Ab14	3	81.471	0.887	0.785	0.849
Ab15	1	86.4215	0.933	0.914	0.806
Ab15	2	86.4215	0.931	0.869	0.858
Ab15	3	81.6124	0.887	0.785	0.852
Ab16	1	86.4215	0.932	0.908	0.812
Ab16	2	86.4215	0.931	0.861	0.868
Ab16	3	83.4512	0.891	0.812	0.862
Ab17	1	89.6747	0.955	0.921	0.868
Ab17	2	89.3918	0.962	0.89	0.898
Ab17	3	86.5629	0.914	0.838	0.898
Ab18	1	89.3918	0.954	0.927	0.855
Ab18	2	88.5431	0.955	0.88	0.892
Ab18	3	83.7341	0.902	0.817	0.862
Ab19	1	89.3918	0.951	0.911	0.874
Ab19	2	90.8062	0.961	0.895	0.923
Ab19	3	82.1782	0.905	0.772	0.88
Dssp10	1	83.5926	0.92	0.901	0.76
Dssp10	2	82.6025	0.916	0.877	0.766
Dssp10	3	82.3197	0.897	0.832	0.812
Dssp11	1	82.8854	0.896	0.887	0.76
Dssp11	2	82.3197	0.885	0.866	0.772
Dssp11	3	78.3593	0.855	0.785	0.782
Dssp12	1	81.471	0.887	0.887	0.729
Dssp12	2	83.1683	0.892	0.872	0.785
Dssp12	3	76.9448	0.841	0.77	0.769
Dssp13	1	81.6124	0.89	0.885	0.735
Dssp13	2	81.7539	0.895	0.874	0.751
Dssp13	3	76.5205	0.829	0.772	0.757
Dssp14	1	72.5601	0.813	0.764	0.68
Dssp14	2	75.389	0.83	0.759	0.748
Dssp14	3	69.4484	0.767	0.675	0.717

Et11	1	84.4413	0.918	0.916	0.76
Et11	2	84.2999	0.93	0.88	0.8
Et11	3	83.0269	0.901	0.84	0.818
Et13	1	88.2603	0.947	0.924	0.834
Et13	2	87.2702	0.939	0.882	0.862
Et13	3	86.1386	0.922	0.859	0.865
Gbmr10	1	79.4908	0.86	0.866	0.711
Gbmr10	2	77.6521	0.851	0.84	0.702
Gbmr10	3	76.3791	0.837	0.801	0.72
Gbmr11	1	79.4908	0.866	0.869	0.708
Gbmr11	2	78.3593	0.856	0.851	0.705
Gbmr11	3	76.3791	0.845	0.783	0.742
Gbmr12	1	76.0962	0.821	0.83	0.68
Gbmr12	2	75.389	0.821	0.838	0.655
Gbmr12	3	73.2673	0.804	0.788	0.668
Gbmr13	1	86.2801	0.922	0.877	0.846
Gbmr13	2	85.7143	0.928	0.859	0.855
Gbmr13	3	82.8854	0.904	0.806	0.855
Gbmr14	1	86.7044	0.922	0.893	0.837
Gbmr14	2	87.553	0.934	0.887	0.862
Gbmr14	3	83.0269	0.905	0.809	0.855
Hsdm10	1	83.5926	0.912	0.877	0.788
Hsdm10	2	81.6124	0.892	0.861	0.763
Hsdm10	3	81.471	0.89	0.809	0.822
Hsdm12	1	76.662	0.835	0.804	0.723
Hsdm12	2	76.3791	0.835	0.83	0.686
Hsdm12	3	74.6818	0.807	0.749	0.745
Hsdm14	1	81.3296	0.904	0.88	0.735
Hsdm14	2	79.7737	0.886	0.859	0.726
Hsdm14	3	80.3395	0.875	0.822	0.782
Hsdm15	1	81.0467	0.901	0.877	0.732
Hsdm15	2	80.7638	0.883	0.848	0.76
Hsdm15	3	78.5007	0.867	0.806	0.76
Hsdm16	1	91.7963	0.96	0.921	0.914
Hsdm16	2	90.6648	0.963	0.893	0.923
Hsdm16	3	87.9774	0.949	0.838	0.929
Hsdm17	1	91.372	0.958	0.921	0.905
Hsdm17	2	91.0891	0.962	0.893	0.932
Hsdm17	3	88.1188	0.945	0.835	0.935
Lr10	1	85.1485	0.926	0.89	0.806
Lr10	2	79.3494	0.884	0.819	0.763
Lr10	3	81.1881	0.885	0.793	0.834
Lwi10	1	85.0071	0.922	0.903	0.788
Lwi10	2	83.5926	0.911	0.885	0.778
Lwi10	3	82.3197	0.891	0.801	0.849
Lwi11	1	82.4611	0.91	0.903	0.732
Lwi11	2	82.744	0.911	0.869	0.778
Lwi11	3	80.7638	0.884	0.835	0.775
Lwi12	1	81.8953	0.905	0.903	0.72
Lwi12	2	82.1782	0.9	0.866	0.769
Lwi12	3	79.0665	0.878	0.809	0.769
Lwi13	1	81.0467	0.893	0.898	0.708
Lwi13	2	83.3098	0.902	0.861	0.8
Lwi13	3	79.9151	0.868	0.817	0.778
Lwi14	1	80.7638	0.888	0.893	0.708

Lwi14	2	82.1782	0.903	0.84	0.8
Lwi14	3	78.925	0.865	0.806	0.769
Lwi15	1	80.4809	0.888	0.893	0.702
Lwi15	2	83.0269	0.91	0.853	0.803
Lwi15	3	79.7737	0.859	0.798	0.797
Lwi16	1	80.3395	0.889	0.89	0.702
Lwi16	2	82.0368	0.904	0.848	0.788
Lwi16	3	79.6322	0.851	0.793	0.8
Lwi17	1	89.3918	0.951	0.901	0.886
Lwi17	2	90.8062	0.964	0.887	0.932
Lwi17	3	86.1386	0.926	0.814	0.917
Lwi18	1	89.5332	0.95	0.911	0.877
Lwi18	2	91.5134	0.965	0.906	0.926
Lwi18	3	83.7341	0.914	0.777	0.908
Lwi19	1	91.5134	0.957	0.921	0.908
Lwi19	2	90.8062	0.964	0.887	0.932
Lwi19	3	83.5926	0.913	0.791	0.889
Lwni10	1	83.1683	0.908	0.908	0.742
Lwni10	2	83.4512	0.914	0.885	0.775
Lwni10	3	82.1782	0.895	0.812	0.834
Lwni11	1	81.8953	0.905	0.903	0.72
Lwni11	2	82.6025	0.908	0.874	0.769
Lwni11	3	80.9052	0.886	0.822	0.794
Lwni14	1	80.7638	0.891	0.893	0.708
Lwni14	2	83.3098	0.91	0.877	0.782
Lwni14	3	81.7539	0.864	0.814	0.822
Lzbl10	1	83.1683	0.903	0.887	0.766
Lzbl10	2	83.1683	0.915	0.864	0.794
Lzbl10	3	82.4611	0.894	0.814	0.837
Lzbl11	1	87.4116	0.937	0.921	0.818
Lzbl11	2	86.5629	0.931	0.885	0.843
Lzbl11	3	84.2999	0.908	0.798	0.895
Lzbl12	1	85.29	0.918	0.906	0.791
Lzbl12	2	85.4314	0.925	0.864	0.843
Lzbl12	3	81.0467	0.895	0.767	0.862
Lzbl13	1	87.2702	0.931	0.911	0.828
Lzbl13	2	84.8656	0.923	0.853	0.843
Lzbl13	3	81.6124	0.888	0.764	0.877
Lzbl14	1	85.0071	0.92	0.89	0.803
Lzbl14	2	84.8656	0.924	0.84	0.858
Lzbl14	3	81.1881	0.884	0.793	0.834
Lzbl15	1	87.4116	0.937	0.893	0.852
Lzbl15	2	84.5827	0.922	0.83	0.865
Lzbl15	3	81.471	0.884	0.777	0.858
Lzbl16	1	88.5431	0.943	0.911	0.855
Lzbl16	2	84.7242	0.926	0.827	0.871
Lzbl16	3	82.1782	0.885	0.78	0.871
Lzmj10	1	84.7242	0.933	0.901	0.785
Lzmj10	2	85.29	0.937	0.874	0.828
Lzmj10	3	85.5728	0.928	0.83	0.886
Lzmj11	1	84.4413	0.932	0.906	0.772
Lzmj11	2	85.1485	0.926	0.872	0.828
Lzmj11	3	85.1485	0.926	0.832	0.874
Lzmj12	1	84.5827	0.93	0.908	0.772
Lzmj12	2	85.0071	0.922	0.885	0.809

Lzmj12	3	83.7341	0.916	0.806	0.874
Lzmj13	1	82.6025	0.921	0.893	0.748
Lzmj13	2	84.017	0.915	0.864	0.812
Lzmj13	3	81.7539	0.904	0.791	0.849
Lzmj14	1	84.5827	0.923	0.893	0.791
Lzmj14	2	87.2702	0.928	0.893	0.849
Lzmj14	3	81.471	0.895	0.767	0.871
Lzmj15	1	84.017	0.921	0.911	0.757
Lzmj15	2	84.4413	0.918	0.848	0.84
Lzmj15	3	83.0269	0.898	0.809	0.855
Lzmj16	1	85.0071	0.924	0.901	0.791
Lzmj16	2	85.9972	0.934	0.869	0.849
Lzmj16	3	81.3296	0.888	0.775	0.858
Ml10	1	81.3296	0.902	0.898	0.714
Ml10	2	79.7737	0.872	0.861	0.723
Ml10	3	79.2079	0.866	0.812	0.769
Ml15	1	90.099	0.954	0.911	0.889
Ml15	2	90.8062	0.955	0.898	0.92
Ml15	3	86.9873	0.932	0.817	0.932
Native20	1	91.372	0.956	0.919	0.908
Native20	2	90.8062	0.965	0.887	0.932
Native20	3	83.4512	0.906	0.793	0.883
Ra10	1	75.8133	0.828	0.793	0.717
Ra10	2	76.0962	0.84	0.764	0.757
Ra10	3	73.9745	0.812	0.657	0.837
Sdm10	1	83.1683	0.899	0.895	0.757
Sdm10	2	79.7737	0.88	0.853	0.732
Sdm10	3	78.925	0.864	0.804	0.772
Sdm11	1	89.6747	0.96	0.94	0.846
Sdm11	2	87.4116	0.955	0.921	0.818
Sdm11	3	88.5431	0.952	0.882	0.889
Sdm12	1	88.2603	0.95	0.929	0.828
Sdm12	2	88.6846	0.957	0.924	0.843
Sdm12	3	88.826	0.949	0.882	0.895
Sdm13	1	89.2504	0.959	0.927	0.852
Sdm13	2	88.9675	0.956	0.927	0.846
Sdm13	3	88.4017	0.95	0.866	0.905
Sdm14	1	89.5332	0.959	0.932	0.852
Sdm14	2	88.826	0.954	0.914	0.858
Sdm14	3	86.9873	0.946	0.853	0.889

APPENDIX D

PDBids of hyperthermophilic, thermophilic and mesophilic proteins

Hyperthermophilic

1EU8_A,1GC5_A,2F5T_X,1MXG_A,1C3P_A,1M1H_A,1OZ9_A,1T6C_A,1UDS_A,1ULZ_A,1XRF_A,
1YFT_A,1ZJR_A,2AU3_A,2DBO_A,2E8B_A,2E8E_A,2EG2_A,2EHS_A,2EJB_A,2EWV_A,2J5A_A,
2OQO_A,2PCL_A,2R75_1,2RGX_A,2YV5_A,2YVW_A,2YWE_A,2ZXY_A,3BN0_A,3ECC_A,3FTD_A,
3HJL_A,1IK6_A,1XQO_A,1UXT_A,1G8A_A,1IM5_A,1IQV_A,1IU9_A,1J3A_A,1L2L_A,1RI7_A,
1UB9_A,1UKU_A,1UZ5_A,1V30_A,1V33_A,1V3W_A,1V43_A,1V6T_A,1V77_A,1V7R_A,1VCT_A,
1VDX_A,1VFF_A,1WKB_A,1WN2_A,1WR2_A,1WY0_A,1WZU_A,1X0M_A,1X0T_A,1X3L_A,1X54_A,
1Y0Y_A,2CWE_A,2CWP_A,2CZW_A,2D59_A,2D7J_A,2D7U_A,2DEJ_A,2DT4_A,2DUL_A,2E1B_A,
2E3U_A,2EVB_A,2HD9_A,2HOQ_A,2OWF_A,2P8T_A,2PHC_B,2WZN_A,2WR8_A,2YX0_A,2YXL_A,
2YZQ_A,2ZUM_A,2ZZF_A,3A25_A,3AF5_A,3BPP_A,3CG3_A,3D79_A,3HPD_A,3IGH_X,3MFY_A,
1A8L_A,1BRF_A,1DQ3_A,1G3Q_A,1IZ4_A,1J24_A,1JG1_A,1KA2_A,1NNH_A,1PVV_A,1RYQ_A,
1SGW_A,1TWL_A,1UA4_A,1VJK_A,1X7N_A,1XHC_A,1XI6_A,1YBZ_A,1Z26_A,1ZD0_A,2CWR_A,
2I0X_A,2JGU_A,2PK8_A,2QS0_A,2ZJ8_A,3CAX_A,3E70_C,3I4H_X,1EZW_A,1GPJ_A,1QLM_A,
2X7M_A,3C19_A,3F8T_A,3M7D_A,1TUA_A,1WTA_A,2CXC_A,2CYA_A,2DVK_A,2EPJ_A,2FC3_A,
2H9U_A,2YV2_A,3A04_A,3A32_A,3HA9_A,1A53_A,1EH9_A,1H0Y_A,1HH1_A,1IM4_A,1NVG_A,
1OB9_A,1OXX_K,1R7J_A,1S5J_A,1VST_A,1Z6A_A,2DFL_A,2H6E_A,2HIV_A,2I6J_A,2IVY_A,
2PLF_A,2Q18_X,2QTF_A,2RDI_A,2W0M_A,2X7B_A,3F3X_A,3FHG_A,3HHV_A,1B74_A,1COJ_A,
1A76_A,1G6H_A,1G8S_A,1H72_C,1HYE_A,1I2A_A,1KKH_A,1NH9_A,1QWG_A,1S0U_A,1T5J_A,
1XBI_A,2AEU_A,2B0A_A,2C4E_A,2EFV_A,2EIF_A,2EJ9_A,2HXD_A,2P5D_A,2PKP_A,2R7K_A,
2VAP_A,2VBU_A,2YV1_A,2YWJ_A,2YWX_A,2YX5_A,2YZL_A,2Z2U_A,2ZP1_A,3A27_A,3AJD_A,
3CFZ_A,3EWG_A,3F47_A,3GIA_A,3GMI_A,3GRU_A,3KH5_A,3KPC_A,3M5F_A,1IO2_A,1MGT_A,
2CW7_A,2CZR_A,2Z1E_A,2Z2Z_A,1KK1_A,1MKH_A,1ML4_A,1YK4_A,1YR7_A,1YZ7_A,2BFW_A,
2HL0_A,2IVN_A,2JJQ_A,2QEN_A,2V7F_A,3BK7_A,3MB5_A,1W41_A,2BO1_A,1TGO_A

Thermophilic

1BQC_A,1W4X_A,2BMM_A,2BOG_X,3MCR_A,1KMZ_A,1A8Q_A,1BRT_A,1MGR_A,1T2I_A,1SQC_A,
1URG_A,2HM7_A,3CIV_A,3GZK_A,1JM1_A,3CRV_A,1Y80_A,2V9V_A,1RHC_A,1LBU_A,1GKU_B,
1HYQ_A,1I39_A,1LJO_A,1M2K_A,1NFJ_A,1P1L_A,1R89_A,1RWZ_A,1SAU_A,1SR8_A,1SW2_A,
1TSF_A,1W9H_A,1Y8A_A,1YOY_A,1YQE_A,1Z0W_A,1ZAR_A,2B2H_A,2B3M_A,2FZ4_A,2FZL_A,
2HC8_A,2I5H_A,2OEB_A,2ONS_A,2OO2_A,2QVO_A,2XDH_A,3CNU_A,3GDE_A,1RO2_A,1E3P_A,
2ZE9_A,3H4X_A,1D9X_A,3FPL_A,1ESW_A,1H98_A,1KU3_A,1LS1_A,1TAQ_A,2AUJ_D,2FXQ_A,
2HPI_A,2IYL_D,3A38_A,1OBR_A,1THM_A,1UH4_A,2ZYO_A,1BH6_A,1GBG_A,1NRF_A,1SCA_A,
1VJS_A,2JDC_A,2JEN_A,1ZXX_A,1WST_A,1ANU_A,1CEO_A,1CLC_A,1H6X_A,1JJF_A,1KWF_A,
1UT9_A,1UXX_X,2BM3_A,2C26_A,2C71_A,2E4T_A,2HQ1_A,2V3G_A,2W47_A,2WAO_A,2WNX_A,
2WZ8_A,3JTE_A,3JWG_A,1A32_A,1C7Q_A,1CYG_A,1DZ3_A,1H2E_A,1HVX_A,1I6M_A,1JQ5_A,
1L8N_A,1LQY_A,1O98_A,1PHP_A,1PJR_A,1PKP_A,1QHO_A,1R85_A,1RL6_A,1SDO_A,1SFS_A,
1T8H_A,1TIF_A,1TIG_A,1TQH_A,1U84_A,1WHI_A,1XWL_A,1XWM_A,1Y50_A,1ZIN_A,2D0B_A,
2E28_A,2HBB_A,2OGT_A,2PVU_A,2RA1_A,2TS1_A,2VIA_A,3CU9_A,4PFK_A,1A7W_A,1HTA_A,
1JDL_A,1QHQ_A,2AAN_A,3CMN_A,1AM2_A,3GMV_X,2SFA_A,2Z5L_A,1O9H_A,1AD2_A,1C4O_A,
1C52_A,1EH1_A,1H4V_B,1IOM_A,1IPA_A,1IQ0_A,1IUH_A,1IUK_A,1IYZ_A,1J09_A,1J27_A,
1J33_A,1KWG_A,1LOU_A,1NOX_A,1NZA_A,1OD6_A,1OI7_A,1RSS_A,1SRV_A,1TFE_A,1TUO_A,
1UB0_A,1UDX_A,1UEK_A,1UFY_A,1UG6_A,1UI0_A,1UJ5_A,1UJP_A,1ULR_A,1US5_A,1USM_A,
1V2D_A,1V2X_A,1V70_A,1V7Q_A,1V8Y_A,1V93_A,1VBI_A,1VC2_A,1VCO_A,1VD6_A,1VE1_A,
1VE4_A,1WCV_1,1WCW_A,1WDI_A,1WF3_A,1WJ9_A,1WJX_A,1WKA_A,1WLU_A,1WNZ_A,1WUB_A,
1X3O_A,1XAA_A,1YOA_A,1ZPW_X,2BM0_A,2C78_A,2CU2_A,2CUL_A,2CVE_A,2D2E_A,2D3Y_A,
2D4P_A,2D5B_A,2D8E_A,2DFA_A,2DP9_A,2DY1_A,2DYI_A,2E7U_A,2EF0_A,2EF4_A,2EKP_A,
2GH9_A,2GXQ_A,2H0A_A,2IHR_1,2IIH_A,2J07_A,2NXC_A,2P5Y_A,2PRD_A,2QHS_A,2V0C_A,
2WDC_A,2YSK_A,2YVP_A,2YVY_A,2YW4_A,2YWD_A,2YYK_A,2YZV_A,2YZY_A,2Z08_A,2Z0X_A,
2Z0Z_A,2Z1A_A,2ZDB_A,2ZO4_A,2ZQE_A,2ZVC_A,2ZXR_A,3A3U_A,3A9I_A,3ACD_A,3BYR_A,

3BZG_A,3CGM_A,3CM0_A,3DMG_A,3FUU_A,3G5R_A,3I31_A,2HKJ_A,1OD3_A,1UY4_A,1J8M_F,
2X41_A,1NQ6_A,1ASS_A,1E0R_B,1MTZ_A,1NIG_A,1NOG_A,1RLH_A,1RLK_A,1Z1W_A,2ART_A,
2FSJ_A,2I1O_A,2OHE_A,2QI2_A,2VSF_A,3CTA_A,2CC6_A,2JAF_A,2G3I_A,1W2L_A,3CP5_A,
3H31_A,1QRE_A,1DD5_A,1DMG_A,1E4F_T,1GJW_A,1GUI_A,1HH2_P,1I5D_A,1I8A_A,1IN4_A,
1J5P_A,1J5T_A,1JCF_A,1K9V_F,1KQ3_A,1L9G_A,1LKV_X,1NSJ_A,1NV9_A,1O0X_A,1O1Z_A,
1O4V_A,1O50_A,1O54_A,1O5L_A,1O5Z_A,1O6D_A,1OH4_A,1P1M_A,1RIY_A,1SUM_B,1THF_D,
1TMY_A,1TQG_A,1TZV_A,1V4I_A,1VJW_A,1VL1_A,1VLC_A,1VLV_A,1VMB_A,1VMJ_A,1VPE_A,
1VPK_A,1VPL_A,1VPQ_A,1VQT_A,1WOS_A,1XKR_A,1YD0_A,1Z9F_A,2AMU_A,2B5D_X,2C2A_A,
2E54_A,2EJC_A,2ESH_A,2FN8_A,2FNC_A,2G1U_A,2G36_A,2H2I_A,2H2W_A,2HP7_A,2HS3_A,
2J71_A,2NRR_A,2NSC_A,2O7I_A,2P2W_A,2P61_A,2W6R_A,2YYZ_A,2Z4S_A,3A0S_A,3A0U_A,
3A0X_A,3A1T_A,3AFH_A,3CIN_A,3CNL_A,3DCM_X,3DGF_C,3FRN_A,3H38_A,3HR8_A,3HYI_A,
1DOI_A,3EEH_A,1IQZ_A,3DNZ_A

Mesophilic

1RWH_A,1IRO_A,3C8Y_A,1B1B_A,1DF7_A,1DQY_A,1EYE_A,1F0N_A,1G19_A,1GR0_A,1GSI_A,
1H05_A,1I9G_A,1LMI_A,1LU4_A,1MQE_A,1N40_A,1N8I_A,1NH8_A,1NWA_A,1P0H_A,1PZS_A,
1RWL_A,1S8N_A,1SXV_A,1TFU_A,1TPY_A,1U2P_A,1U5H_A,1UOZ_A,1W66_A,1WQG_A,1YK9_A,
1YM3_A,1YWF_A,1Z9W_A,1ZA0_A,1ZNW_A,2A11_A,2A15_A,2A84_A,2BM6_A,2BYO_A,2C2Y_A,
2C7G_A,2CDN_A,2CGQ_A,2CIB_A,2CJG_A,2CM1_A,2E3J_A,2FEZ_A,2FK8_A,2FR2_A,2FSX_A,
2FWV_A,2G2D_A,2H7M_A,2HH7_A,2I1U_A,2I6Y_A,2IN0_A,2IRX_A,2IYV_A,2JA2_A,2JAY_A,
2JEK_A,2O03_A,2O0B_A,2O2I_A,2OQR_A,2PFC_A,2PKN_A,2Q3D_A,2QC3_A,2QHF_A,2UUQ_A,
2VKL_A,2VOR_A,2WGE_A,2WM5_A,2WU8_A,2Z2I_A,2Z99_A,2ZQ5_A,2ZYL_A,3AEZ_A,3B4W_A,
3BZQ_A,3DK5_A,3E26_A,3E3U_A,3EE4_A,3ELF_A,3EO5_A,3F61_A,3FVE_A,3G1M_A,3PL1_A,
3HEM_A,3HGB_A,3HT5_A,3HZA_A,3HZU_A,3I7T_A,3IB7_A,3IG0_A,3IJF_X,3IOS_A,3IU7_A,
3IVY_A,3K1D_A,3L60_A,3M6B_A,3MD0_A,3NE3_B,1RGV_A,1W7O_A,1ET7_A,1PAZ_A,1RK6_A,
1KEH_A,1CPN_A,2CC1_A,3JSC_A,1XDW_A,3IOG_A,2QQ9_A,3CP3_A,3HR6_A,1A8D_A,1Z7H_A,
2QYZ_A,1FT5_A,2JE3_A,3B9W_A,3CYM_A,1B2V_A,1C7S_A,1EDQ_A,1SAT_A,1WM1_A,2QUA_A,
1RTQ_A,1V7W_A,1K51_A,1MHX_A,2ZW1_A,1E4I_A,1JWQ_A,1W0N_A,2O9P_A,1KRQ_A,1W55_A,
2CIC_A,2FGS_A,2P2V_A,2QCO_A,2WQQ_A,2WY4_A,3BFP_A,3D6L_A,3E13_X,3FNR_A,1A7J_A,
1CXC_A,1EG2_A,1MZY_A,2FWT_A,2NWF_A,1D9Y_A,2G5D_A,3D2M_A,1GU3_A,1V6Y_A,2BVY_A,
3CUI_A,1AH7_A,1MQO_A,1NPC_A,1QS2_A,1UOK_A,1VEM_A,1WS0_A,1YWQ_A,2P8B_A,2PTD_A,
2UYR_X,2VDJ_A,2X2O_A,3BVS_A,3FCE_A,3N17_A,1C75_A,1EAR_A,1H6T_A,1H6U_A,1O6T_A,
1XEU_A,2I2C_A,2PLC_A,3BZ5_A,3FI7_A,3K2T_A,1G2R_A,1G97_A,1N7O_A,1OX0_A,1PSZ_A,
1QME_A,1YDF_A,2A9O_A,2AIE_P,2B06_A,2BG1_A,2BIB_A,2C1I_A,2DPM_A,2FI0_A,2FI1_A,
2FPH_X,2HQ0_A,2I0M_A,2J22_A,2J44_A,2JKB_A,2OI2_A,2QF5_A,2V05_A,2W91_A,2WAE_A,
2WJE_A,2WMF_A,2WW5_A,2WW8_A,2X9Y_A,2XD3_A,3GON_A,3GUV_A,3IM8_A,3KR9_A,3MMS_A,
1J2Z_A,1KLX_A,1MW7_A,1OUV_A,1UM8_A,1XNH_A,1ZUH_A,2BMV_A,2BO3_A,2BQX_A,2C4W_A,
2EW5_A,2I9I_A,2QGH_A,2QV3_A,2WLT_A,3GUQ_A,3H1G_A,3HVM_A,3K1H_A,3MYD_A,3SSI_A,
1BED_A,1MC4_A,1W0P_A,1XEZ_A,1XFK_A,1XMX_A,1YC9_A,1YG2_A,1ZU0_A,2G7E_A,2GU1_A,
2HAF_A,2I6V_A,2JHQ_A,2OXN_A,2V1L_A,3C38_A,3E9A_A,3ESS_A,3GBG_A,3LV8_A,3N28_A,
3NVS_A,1NPS_A,2GKG_A,1EDT_A,1JAK_A,1D2T_A,3BLD_A,1NTH_A,3EZX_A,1YS1_X,1ZOD_A,
2PIA_A,1LYV_A,1P9H_A,1XVX_A,1XVY_A,2BHO_A,2J0P_A,2JD9_A,2UVJ_A,2V8I_A,3H7Z_A,
1COT_A,2OV0_A,2R31_A,3D4T_A,3M97_X,1AVK_A,1PJ5_A,1ULV_A,1CBY_A,1W99_A,2C9K_A,
2RCI_A,3DHA_A,3GMS_A,3HNR_A,1Z8O_A,2FR1_A,2JJN_A,3EL6_A,1Q8C_A,1CB8_A,1DBO_A,
1K8T_A,1RZ2_A,1XP3_A,1YQY_A,2A1Y_A,2C5S_A,2H3G_X,2IFY_A,2J13_A,2J9R_A,2VU5_A,
2WAG_A,2Z5W_A,3CUX_A,3DAT_A,3DD6_A,3ERV_A,3FBQ_A,3HMC_A,3LQN_A,3M3H_A,1ZHV_A,
2D2J_A,2H8O_A,2HLY_A,2NWH_A,2PNW_A,2RH3_A,2ZE7_A,3AFL_A,3DB9_A,3IPC_A,3DON_A,
1G41_A,1J7G_A,1JJV_A,1JNI_A,1JOS_A,1JOV_A,1MXI_A,1NNF_A,1OZA_A,1P77_A,1T3B_A,
1UAL_A,1YNM_A,2AUD_A,2GKE_A,2RL1_A,2V0H_A,3A3J_A,3B50_A,3BAC_A,3EMI_A,3ET4_A,
3M73_A,2HRO_A,1EI5_A,2PVQ_A,1TKJ_A,1AKL_A,1DXH_A,1EB7_A,1EX9_A,1G6A_A,1H70_A,
1HZU_A,1K0I_A,1K2Y_X,1KO3_A,1L7L_A,1LRY_A,1MDL_A,1MZB_A,1NF9_A,1OVP_A,1PEA_A,
1R6M_A,1RTT_A,1RTV_A,1S7I_A,1SB8_A,1SK7_A,1TP6_A,1TU9_A,1U4G_A,1W4T_A,1X6Z_A,
1XKW_A,1Y0N_A,1Z6N_A,1ZK7_A,1ZWL_A,2AZP_A,2CFU_A,2FBH_A,2FBI_A,2FBQ_A,2FD5_A,
2FGO_A,2FTP_A,2GJL_A,2HL7_A,2IAH_A,2NS6_A,2OZ6_A,2PST_X,2QDX_A,2R79_A,2V3A_A,
2VAW_A,2VQD_A,2VW8_A,2W6L_A,2WOQ_A,2WZX_A,2X3N_A,2X4G_A,2XU2_A,2ZWS_A,3B40_A,

3BZC_A,3C96_A,3CRR_A,3DWO_X,3EAT_X,3FSA_A,3H6J_A,3HV8_A,3JZZ_A,3KH7_A,3KKW_A,
3LLB_A,3MOK_A,3NYT_A,451C_A,3BNJ_A,1G5A_A,1BQG_A,1GKM_A,1KV9_A,1LVL_A,1MXS_A,
1P4C_A,1Q6Z_A,1V72_A,1WVF_A,1YIQ_A,1ZI8_A,2AZQ_A,2IMF_A,2QPZ_A,2R14_A,2ZWU_A,
3A8U_X,3CPO_A,3EF6_A,3FMX_X,3KYF_A,3L5L_A,3N9T_A,1Q0R_A,1QZZ_A,1G94_A,1H12_A,
2ZOO_A,1DUR_A,1A8P_A,1ATG_A,1CC5_A,1DPB_A,1H4K_X,1H9K_A,1LRV_A,1P90_A,1RW4_A,
7FD1_A,1VLB_A,1WAD_A,1Z1N_X,2DSX_A,2WFB_A,2AIO_A,2ECF_A,1GIR_A,2BF6_A,2BJF_A,
2J1A_A,2QUO_A,2V72_A,2VCC_A,2VMH_A,2VNR_A,2VO8_A,2W1N_A,2WXU_A,3GFO_A,1BAM_A,
1E43_A,1H6L_A,1SUP_A,1YVS_A,1BQB_A,1CV8_A,1D2P_A,1DHN_A,1DUA_A,1DYQ_A,1ENF_A,
1EY4_A,1GHP_A,1HSK_A,1JF8_A,1JIL_A,1LRZ_A,1N67_A,1OJQ_A,1P4X_A,1P99_A,1QTF_A,
1QWY_A,1QWZ_A,1QXY_A,1UNS_A,1XAG_A,1XG8_A,1XXG_A,1YN4_A,1ZJC_A,2AP3_A,2F68_X,
2INR_A,2JG6_A,2O85_A,2O8L_A,2OKT_A,2Q8Q_A,2QMT_A,2RDG_A,2REU_A,2SAK_A,2W5Q_A,
2W9H_A,2WQD_A,2X75_A,2X7I_A,2XCQ_A,2Z6F_A,2Z8L_A,2ZCO_A,2ZKL_A,3B2N_A,3BCI_A,
3BL6_A,3DOA_A,3EA6_A,3F3M_A,3FYM_A,3GNS_A,3H04_A,3HZS_A,3IM9_A,3JSN_A,3KI9_A,
3KSH_A,3LHS_A,3NIZ_A,3SEB_A,3TSS_A,1EW0_A,1QKK_A,2Q88_A,3JU2_A,1NFP_A,1OAL_A,
3DDY_A,2B8I_A,2B78_A,2HCU_A,2NQ5_A,2RK5_A,2W3Z_A,2ZIC_A,3BJV_A,3EXT_A,3IOX_A,
3K8U_A,1JTA_A,1K7I_A,1NOF_A,1O88_A,1RU4_A,1G5Z_A,1P4P_A,1SUU_A,2HKD_A,2I5V_O,
3K9G_A,1GWE_A,3IF5_A,2QR3_A,3CMG_A,3E8V_A,1A3C_A,1A6F_A,1B02_A,1BLE_A,1BOW_A,
1F7L_A,1GPR_A,1GSK_A,1H99_A,1ISP_A,1JBG_A,1M7V_A,1MDB_A,1MVO_A,1NAT_A,1NC5_A,
1NE8_A,1NG6_A,1OYG_A,1OYS_A,1PUJ_A,1Q2Y_A,1QAM_A,1QE3_A,1QG8_A,1QR0_A,1R0U_A,
1RLJ_A,1SVI_A,1T9H_A,1TF5_A,1TVL_A,1TWU_A,1UA7_A,1UV4_A,1UX8_A,1W53_A,1W5D_A,
1XC3_A,1XD7_A,1XDZ_A,1Z91_A,1ZCH_A,2B18_A,2B4L_A,2CB9_A,2D1V_A,2F5C_A,2FGT_A,
2FQT_A,2G0C_A,2GKO_A,2GU3_A,2H1V_A,2I5M_X,2NX2_A,2O04_A,2P3S_A,2P5K_A,2V09_A,
2VD2_A,2VSQ_A,2VXY_A,2W1R_A,2WHK_A,2WI8_A,2X2B_A,2ZUY_A,3BR8_A,3BS8_A,3C71_A,
3C7F_A,3D30_A,3DKV_A,3E7W_A,3EHG_A,3F6P_A,3GHA_A,3I9Z_A,3MIX_A,3N54_B,3THI_A,
1B0Y_A,3EUN_A,1NML_A,2BH4_X,3I9X_A,1SJW_A,1A8S_A,1JOI_A,1QZ9_A,1UK8_A,1XUB_A,
1YIO_A,2AQJ_A,2QZ6_A,2V7K_A,3DGB_A,3G63_A,3NOV_A,1KNG_A,1XJ3_A,2J4X_A,2YXP_X,
1BGV_A,1KBL_A,3GF3_A,1CXY_A,1HPI_A,2B0T_A,2DAP_A,2NLO_A,2NSF_A,2P3H_A,3FH2_A,
3FNN_A,3H5T_A,3LMD_A,2VOW_A,3EWK_A,1H9A_A,2FX5_A,2Q3W_A,1JX6_A,2HJE_A,3B9E_A,
1VEI_A,2GEK_A,2HW2_A,2JFR_A,2VKS_A,2X1M_A,2ZR9_A,3DG6_A,3GWM_A,3MOY_A,3NWO_A,
3NXS_A,1LTZ_A,1D4D_A,1M1Q_A,1SP3_A,2BBE_A,2E4L_A,2GOU_A,3NJK_A,1B87_A,2JFU_A,
1PCH_A,2GPR_A,1JFX_A,1NM2_A,1ODO_A,1S1F_A,1T00_A,1VZW_A,2DKK_A,2EWT_A,2HYJ_A,
2OG5_A,2VFR_A,2WDS_A,2X4L_A,2ZCX_A,3EXM_A,3ID7_A,3KB9_A,3NF2_A,1J84_A,1T71_A,
1TD6_A,1A0P_A,1A2J_A,1A5T_A,1AB4_A,1AHN_A,1AJ2_A,1AKO_A,1AMF_A,1AOB_A,1AOP_A,
1AQT_A,1ARS_A,1B0A_A,1B4E_A,1B63_A,1B6R_A,1B7E_A,1B8X_A,1BBU_A,1BDO_A,1BIA_A,
1BS0_A,1BYI_A,1C5K_A,1CEI_A,1CHU_A,1CKE_A,1CTF_A,1CTT_A,1CUK_A,1CYX_A,1D6M_A,
1DB3_A,1DD9_A,1DDI_A,1DI6_A,1DKQ_A,1DPE_A,1DRW_A,1DVO_A,1E1O_A,1E2X_A,1E4C_P,
1E59_A,1E5K_A,1E6U_A,1E9F_A,1EEH_A,1EF9_A,1EFD_N,1EJ0_A,1EKR_A,1EUW_A,1EW4_A,
1F00_I,1F3Z_A,1FDR_A,1FL2_A,1FSF_A,1G6S_A,1G79_A,1G7V_A,1GEW_A,1GN0_A,1GRJ_A,
1GS5_A,1GSA_A,1GSO_A,1GTK_A,1GVH_A,1GVP_A,1GXQ_A,1GXU_A,1GYN_A,1H16_A,1H3D_A,
1H75_A,1HNJ_A,1HP1_A,1HQ0_A,1HW7_A,1HZ4_A,1HZT_A,1I2K_A,1I52_A,1I6A_A,1I6P_A,
1ID0_A,1IHU_A,1IOW_A,1ISE_A,1IVN_A,1IWL_A,1IWN_A,1IXH_A,1J2A_A,1JBK_A,1JF9_A,
1JGS_A,1JHC_A,1JHG_A,1JL1_A,1JPD_X,1JQN_A,1JR7_A,1JSX_A,1JY8_A,1JYE_A,1JYH_A,
1K6K_A,1K92_A,1KFN_A,1KID_A,1KNW_A,1KON_A,1KS9_A,1KV7_A,1L6P_A,1LN4_A,1LV7_A,
1M40_A,1M65_A,1M6T_A,1MJC_A,1ML8_A,1MLA_A,1MOQ_A,1MSK_A,1MUG_A,1MUL_A,1MUN_A,
1MW9_X,1MXA_A,1N8N_A,1NIJ_A,1NKD_A,1NNX_A,1NQK_A,1NYL_A,1NZJ_A,1OAP_A,1OLT_A,
1ONS_A,1OPC_A,1OPD_A,1OYY_A,1PB1_A,1PDO_A,1PF5_A,1PII_A,1POT_A,1PQE_A,1PS9_A,
1Q09_A,1Q0L_A,1Q2L_A,1Q3B_A,1Q6U_A,1Q6Y_A,1Q8I_A,1QF5_A,1QHL_A,1QOY_A,1QSA_A,
1QTW_A,1QUS_A,1QXX_A,1QZM_A,1R3F_A,1R62_A,1R6W_A,1R9L_A,1RA0_A,1RI6_A,1RKD_A,
1RLR_A,1RNL_A,1RPJ_A,1S3C_A,1S7C_A,1SCZ_A,1SDI_A,1SFE_A,1SI7_A,1SIG_A,1SQG_A,
1SUR_A,1T8K_A,1TDJ_A,1TJD_A,1TKE_A,1TOL_A,1TRB_A,1TT8_A,1TUV_A,1TXL_A,1U2K_A,
1U7U_A,1U94_A,1U9P_A,1UDC_A,1UJ8_A,1UJC_A,1USG_A,1UUF_A,1UWF_A,1UXY_A,1V9F_A,
1VB3_A,1VBV_A,1VI7_A,1VSR_A,1W8G_A,1WAU_A,1WDN_A,1WL6_A,1WXI_A,1X09_A,1X6J_A,
1XEO_A,1XVU_A,1XWY_A,1Y2F_A,1Y79_1,1YFE_A,1YLJ_A,1YOE_A,1YRW_A,1YSP_A,1YSQ_A,
1YT3_A,1Z15_A,1Z5P_A,1ZK5_A,1ZMR_A,1ZVU_A,1ZYL_A,1ZZM_A,2A0B_A,2ABK_A,2AE0_X,

2ALX_A,2ANB_A,2ARA_A,2ASR_A,2AU7_A,2B0C_A,2B1K_A,2BJV_A,2BLL_A,2BTD_A,2BUE_A,
2C4N_A,2CMD_A,2DBN_A,2DH5_A,2DJH_A,2DRI_A,2EA9_A,2EX2_A,2F1C_X,2F1N_A,2FDJ_A,
2FVY_A,2FWH_A,2FWM_X,2G7O_A,2GAR_A,2GGC_A,2GNK_A,2GUI_A,2GUS_A,2H09_A,2H8E_A,
2HG2_A,2HNH_A,2HQ2_A,2I88_A,2IOR_A,2IW1_A,2IX0_A,2IY9_A,2J0W_A,2J6G_A,2J7L_A,
2JFN_A,2JG0_A,2NSH_A,2NUL_A,2O90_A,2O9G_A,2OBL_A,2OLR_A,2OML_A,2P0B_A,2P5Z_X,
2P67_A,2PA3_A,2PII_A,2PMK_A,2PQX_A,2PTH_A,2PYU_A,2QCP_X,2QDF_A,2QFL_A,2QIA_A,
2QOP_A,2QXF_A,2QY9_A,2QZS_A,2RH2_A,2TIR_A,2TPT_A,2UYT_A,2V9L_A,2VGD_A,2VK2_A,
2VKE_A,2W21_A,2WJR_A,2WKX_A,2XE1_A,2XFD_A,2YXN_A,2Z1J_A,2Z98_A,2ZCU_A,2ZQ7_A,
3A2Z_A,3A6T_A,3A7L_A,3A7R_A,3B34_A,3B44_A,3B8J_A,3BEC_A,3BKF_A,3BQW_A,3BXY_A,
3BY8_A,3BZM_A,3BZS_A,3C5A_A,3C8F_A,3CDI_A,3CHY_A,3CLA_A,3CP2_A,3CUZ_A,3D1R_A,
3DAU_A,3DJL_A,3DXY_A,3E2Q_A,3ERS_X,3ESQ_A,3EYE_A,3EZ7_A,3F85_A,3FEW_X,3FRH_A,
3FWM_A,3FZG_A,3G7E_A,3G7U_A,3GA8_A,3GP6_A,3GR5_A,3GWI_A,3GX0_A,3GZH_A,3H4R_A,
3H9C_A,3HFI_A,3HJH_A,3HLR_A,3HO9_A,3HVV_A,3HXW_A,3HYF_A,3I87_A,3I9W_A,3ID1_A,
3ID4_A,3IP0_A,3KJT_A,3KQJ_A,3L1L_A,3VUB_A,4EUG_A,4TMK_A,8ABP_A,2WHL_A,7A3H_A,
1DFX_A,1DGJ_A,1DUW_A,1I77_A,1UP9_A,2A3M_A,2NAP_A,3KAP_A,1FD9_A,2IM9_A,2OO3_A,
2WZF_A,2WZG_A,3AAP_A,3I0O_A,3I47_A,1HUF_A,1JL5_A,1R6F_A,1Z21_A,2JLI_A,2X55_A,
3FWW_A,3GSE_A,3HID_A,3JTZ_A,3L92_A,3LXY_A,3N4J_A,3NRS_A,2FDN_A,1CPQ_A,1DMR_A,
1G8P_A,1UWM_A,2BGI_A,2JK1_A,2WC1_A,1EPW_A,1ZB7_A,2A8A_A,2FPQ_A,2J3X_A,2QN0_A,
2VU9_A,2VXR_A,3BON_A,3FUQ_A,3IRD_A,1ESC_A,1GCY_A,2CY8_A,1K0F_A,1XS5_A,2FQX_A,
2V84_A,1CO6_A,1J77_A,1OJT_A,1R1M_A,1RV9_A,1SS9_A,2A0J_A,2FY6_A,2GW8_A,2JC4_A,
2JC5_A,2OLS_A,2VQ2_A,2WLC_A,2ZDR_A,3A2S_X,3BQH_A,3HZ8_A,1AN8_A,1DLJ_A,1ET9_A,
1SU0_B,1Y08_A,1YS9_A,1Z0P_A,2C3F_A,2NX8_A,2OHG_A,2OS3_A,2OZE_A,2QGZ_A,2WB3_A,
2WH7_A,2WLU_A,3EIF_A,3FN7_A,3HH8_A,2BS5_A,2CHH_A,3EOJ_A,1NQZ_A,1SJY_A,1VH2_A,
1XP8_A,2A1V_A,2BHU_A,2BOO_A,2C2J_A,2C2Q_A,2C2U_A,2G40_A,2HZ7_A,2IMR_A,2NVO_A,
2O5V_A,2O9C_A,2VPA_A,3BT5_A,3E1S_A,3GG7_A,1EDG_A,1G43_A,1G9G_A,1IA6_A,2V4V_A,
3C2C_A,3I45_A,1NH1_A,1GCI_A,1DAB_A,1RWR_A,1Y9U_A,2PFZ_A,3EFM_A,3F2V_A,1P9P_A,
1Y9L_A,3CYV_A,3DR3_A,3GY0_A,1CGT_A,1D3C_A,1ITX_A,1QGI_A,1W3U_A,1XNB_A,2C81_A,
2J66_A,1OK0_A,2OLN_A,3H0O_A,1CWV_A,2H7O_A,2UYO_A,2GLK_A,1F1S_A,1YWM_A,3PHS_A,
2ZZR_A,3EPR_A,3EVN_A,3FAW_A,2FR7_A,3DM8_A,3FG2_P,3HUI_A,3LMO_A,2DKH_A,3FDD_A,
8CHO_A,1LNS_A,1PIE_A,2G0D_A,2IYO_A,2PBG_A,2WF7_A,2WQF_A,3F8C_A,3IAN_A,3L6G_A,
1EOK_A,1PGS_A,2EBN_A,2IXA_A,3IAJ_A,5NUL_A,2I1Q_A,1QHX_A,1R6D_A,2C7X_A,3I3L_A,
1IS1_A,3A57_A,3CFY_A,3I9Y_A,1BOO_A,1HN0_A

APPENDIX E

Average accuracy (AA) values for all feature sets

HM Dataset	
Feature Set	**Avg. Acc.**
Native	**92.95**
Lwi19	92.81
Ab19	92.72
Lwi18	92.70
Lwi17	92.32
Hsdm16	92.18
Hsdm17	92.06
Ml15	91.53
aa_content_in_ss	91.39
Lwi15	90.85
Lwi16	90.72
Lwi11	90.64
Ab18	90.33
Lwi14	90.28
Sdm14	90.25
Ab17	90.25
Lwni14	90.24
Lwi12	90.23
Lzmj16	90.21
Sdm12	90.15
Lwi13	90.03
Sdm11	89.98
Lwni11	89.89
Sdm13	89.85
Lzmj14	89.70
Lwni10	89.55
Gbmr13	89.35
Gbmr14	89.31
Ml10	89.25
Lzmj13	89.11
Lzmj12	89.04
Lzmj11	88.97
Ab16	88.93
Lzmj15	88.74
Et13	88.70
Lzmj10	88.60
Ab11	88.57
Ab15	88.56

TM Dataset	
Feature Set	**Avg. Acc.**
Ab19	**82.35**
Native	82.33
Lzmj11	82.10
Sdm11	82.09
Lwi19	82.09
Lwi18	82.08
Lzmj14	81.98
Lwi17	81.97
Ml15	81.93
Lzmj12	81.93
Sdm12	81.79
Sdm14	81.77
Lzmj10	81.75
Lzmj13	81.73
Sdm13	81.67
Lzmj15	81.63
Lzmj16	81.63
Dssp10	81.61
Hsdm16	81.51
Hsdm17	81.37
Gbmr14	81.06
Lwi14	80.97
Lwi15	80.78
Lwi16	80.75
Gbmr13	80.69
Lwi12	80.56
Ab18	80.48
Lwi11	80.35
Lwi13	80.34
Lwni14	80.26
Ab15	80.22
Ab17	80.20
Ab16	80.11
Et13	80.11
Hsdm14	80.02
Et11	80.01
Hsdm15	79.93
Ab13	79.91

(HT)M Dataset	
Feature Set	**Avg. Acc.**
*all features	**85.10**
Ab19	84.73
Native20	84.55
Lwi18	84.44
Lwi19	84.44
Lwi17	84.40
Ml15	84.12
Hsdm16	83.95
Hsdm17	83.91
Sdm11	83.77
Sdm12	83.68
Sdm13	83.48
Sdm14	83.41
Lzmj14	83.31
Gbmr13	83.28
all features	83.21
Lzmj12	83.19
Gbmr14	83.15
Lzmj13	83.14
Lzmj16	83.12
Lzmj11	83.10
Lzmj10	83.07
Lwi12	82.98
Lwi11	82.97
Lwi14	82.93
Lwi15	82.91
ssesa related	82.81
Lwi13	82.76
Lwi16	82.75
Lwni14	82.72
Lzmj15	82.53
Dssp10	82.53
Lwni11	82.39
Ab18	82.29
Ab17	82.29
aa_cont_ss	82.15
Et13	82.14
Ab14	81.98

Ab13	88.52	Ab14	79.85	Ab13	81.98
Ab14	88.37	Lzbl13	79.83	Ab10	81.87
Ab10	88.36	Dssp11	79.78	Ml10	81.85
Dssp10	88.36	Dssp12	79.75	Lwni10	81.71
Ab12	88.33	Lzbl11	79.71	Ab16	81.71
Dssp13	87.95	Ab12	79.64	Ab15	81.66
Et11	87.91	Lzbl12	79.64	Ab11	81.57
Lzbl15	87.69	Lzbl16	79.54	Et11	81.57
Lzbl11	87.69	Ab11	79.49	Ab12	81.54
Lzbl16	87.66	Ab10	79.48	Lzbl11	81.53
Dssp12	87.64	Dssp13	79.47	Lzbl15	81.14
Lzbl10	87.30	Lzbl15	79.44	Lzbl10	81.09
Hsdm15	87.20	Lzbl14	79.43	Hsdm14	81.09
Lzbl12	87.09	Lwni11	79.41	Hsdm15	81.07
Hsdm14	86.97	Lzbl10	78.94	Dssp13	81.04
Lzbl14	86.91	Ml10	78.86	Lzbl12	81.00
Dssp11	86.90	Lwni10	78.74	Dssp12	80.90
Lzbl13	86.80	Basics	78.33	Dssp11	80.84
Lwi10	86.69	Gbmr12	77.87	Lzbl14	80.83
Dssp14	85.73	Dssp14	77.85	Lzbl13	80.80
Basics	85.64	aa_content_in_ss	77.80	Lzbl16	80.77
Hsdm12	84.01	Gbmr11	77.60	Basics	80.01
Lr10	82.50	Gbmr10	77.54	Lwi10	79.43
Hsdm10	82.05	Hsdm12	77.38	Dssp14	78.41
Sdm10	81.74	Lwi10	76.80	Hsdm12	77.93
Gbmr12	81.73	Sdm10	76.68	Gbmr12	77.43
Gbmr11	81.52	Hsdm10	75.93	Gbmr11	77.30
dipoles	80.97	salt_bridges	72.09	Sdm10	76.98
Gbmr10	80.49	Lr10	72.00	Gbmr10	76.89
salt_bridges	79.91	dipoles	65.41	Hsdm10	76.63
cation_pi	71.04	cation_pi	63.39	Lr10	75.00
ss_content	57.19	ss_content	54.55	salt_bridges	73.11
bfactors_in_ss	47.17	disulfides	50.18	dipoles	71.74
disulfides	38.70	bfactors_in_ss	47.71	cation_pi	61.57
				hinge related	57.53
				ss_content	53.75
				disulfides	51.00
				bfactors_ss	38.06

Shaded cells correspond to feature sets that contain only structural features. *Only statistically significant (based on KS test) features of the combined sequential and structural feature set were included.