Discrimination of thermophilic and mesophilic proteins using reduced amino acid alphabets with n-grams

Aydin Albayrak, Ugur O. Sezerman §

Biological Sciences and Bioengineering, Sabanci University, Orhanli, Tuzla, Istanbul, Turkey

§Corresponding author

Email addresses:

AA: [aydinalbay@su.sabanciuniv.edu](mailto:aydinalbay@su.sabanciuniv.edu)

UOS: [ugur@sabanciuniv.edu](mailto:ugur@sabanciuniv.edu)

# Abstract

Protein thermostabilization has been the focus of recent research due to growing interest in the production of enzymes that can operate at temperatures that are industrially beneficial. Understanding the determinants of thermostabilization at the level of sequence and structure are important to design such enzymes. A bioinformatical approach was used to determine the extent by which reduced amino acid alphabets (RAAA) with n-grams (subsequences of length n) that were subjected to a t-test-based feature selection procedure can be used to discriminate proteins from thermophiles and mesophiles. Classification performance of 65 different protein alphabets with 3 different n-gram sizes was systematically evaluated using support vector machines in a test set that contained 707 proteins from mesophilic *Xylella fastidosa* and thermophilic *Aquifex aeolicus*. A classification accuracy of 91.796% was achieved with Hsdm16 RAAA with 13 features: EK-ILV-ST-A-G-F-H-Q-N-R-M-W-Y. The t-test-based feature selection procedure reduced the classification time without significantly affecting classification accuracy. The overall combination of methods in this paper is useful and computationally fast for classifying protein sequences from thermophiles and mesophiles using sequence information alone.

**Keywords:** Amino acid composition, dipeptide, N-grams, reduced amino acid alphabets, statistically significant features, thermostability, tripeptide

# 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 (Topt): psychrophiles have Topt of less than 15°C; mesophiles have Topt in the range of 15 - 45°C; thermophiles have Topt in the range of 45-80°C and hyperthermophiles with a Topt 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 Topt of their host organisms. For instance, *Pyrococcus furiosus* amylopullulanase is optimally active at 125°C, which is 27°C above the host organisms Topt of 98°C [1]. The existence of such thermophilic proteins with elevated melting temperature (Tm) also has theoretical support from the equation, Tm = 24.4 + 0.93 Tenv [2] that relates the Tm of a protein to the environmental temperature (Tenv) 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 while the results of the second category can be used to design mesophilic proteins with increased thermostability by mimicking the thermophilic homolog.

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 (i.e., specific void volume) 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) [22]. They have shown that sequence-based indices such as IVYWREL and CvP bias (defined as the difference between charged, DEKR and polar, NQST residues [23]) 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 [5].

Meanwhile, different studies have been devoted to grouping amino acids based on shared physicochemical and/or structural features [24-32]. 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 [32] 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 [33].

In our previous work [34], we have shown that RAAAs can be used to cluster protein families into functional subtypes with equal or better accuracy than the native amino acid 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.

In this work, we systematically evaluated 65 different RAAAs with three different n-grams (subsequences of length n) in the classification of protein sequences from thermophiles and mesophiles using support vector machines. Classification using RAAAs with 1-grams and 2-grams resulted in better accuracies than with 3-grams. In most cases, a smaller RAAA size was sufficient to obtain the same level of accuracy as the native alphabet.

# Methods

## 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. 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 1. µPID was calculated using the pairwise identity scores obtained from the result of Needleall many-to-many pairwise alignment script available in EMBOSS [35] 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 [36] sequence redundancy search algorithm. 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%.

## RAAA

We adopted the same approach as Peterson's [33] 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. Previously, 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 2 while the amino acid groupings of all RAAAs are provided in Supplementary File 1.

## N-grams

N-grams are sequences of namino acids in a sliding window over the length of the protein sequence [37]. 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.

## T-test

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 α = 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.

## SMOTE Sampling

The training set was subjected to Synthetic Minority Over-sampling Technique (SMOTE) [38] to balance the size of the thermophilic and mesophilic protein classes. 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 [38]: 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.

## Classification

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.

## 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;

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 the minimum alphabet size) in terms of classification accuracy were reported in Table 3. Classification results in terms of sensitivity, specificity, accuracy and AUC for the test set with different n-grams and RAAAs were reported in Supplementary File 2.

## Protocol

After one of the alphabets given in Table 2 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 1.

# 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, 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.

**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.* [33] and was shown to improve classification accuracy in fold recognition prediction. The fact that the clusters of amino acids in the HSDM17 alphabet were also good predictors of protein thermostability in the current study may imply that the grouping of amino acids in this alphabet may reflect an evolutionary response to increased temperatures at the level of protein sequence.

Lwi18 was the top performing alphabet for 2-grams with 91.513% accuracy. The feature vector of Lwi18 alphabet has 158 significant features out of 324 (i.e., 182) 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. Overall, the t-test based feature selection resulted in 84-90% feature reduction for the top performing 3-grams.

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.

**Effect of RAAA size on classification accuracy**

Previously, 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 3-grams in Sdm11 alphabet has 1331 (i.e., 113) 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 (Supplementary File 2) 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 [37] 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-specific 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.

**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.

**Benchmark Results**

In Table 4, 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.

# Conclusions

It is possible to accurately discriminate proteins from thermophiles and mesophiles using RAAAs with n-grams. Classification accuracy of each RAAA usually decreases with increasing n-gram size and this decrease is especially more evident in 3-grams. Current approach of using RAAAs with different n-grams has produced better results with fewer features than the native alphabet in terms of accuracy. Our results also indicate that RAAAs can improve performance relative to full protein alphabet. Performing t-test to reduce the number of features in the training set decreases the compute time without significantly affecting classification accuracy and makes classification with 3-grams possible. Extensions of this work are currently underway that include compiling larger training and test sets with different levels of mean percent identities, generating organism-specific RAAAs, and separating thermostability classes by phyla.

# Acknowledgements

Aydin Albayrak would like to thank Cem Meydan for his sincere efforts in answering questions about writing many in-house python scripts that made many calculations possible and Michael Gromiha for kindly providing the datasets. The authors also would like to thank Murat Cokol, Gokhan Demirkan and Stuart James Lucas for proofreading the initial manuscript; and three anonymous reviewers for critical feedback.

# 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): 2614-21.

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): 284-8.

4. Gromiha MM, Suresh MX. Discrimination of mesophilic and thermophilic proteins using machine learning algorithms. Proteins 2008; 70(4): 1274-9.

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): 293-8.

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): 1608-15.

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): 32383-6.

13. Zhang GY, Fang BS. Application of amino acid distribution along the sequence for discriminating mesophilic and thermophilic proteins. Process Biochemistry 2006; 41(8): 1792-8.

14. Ditursi MK, Kwon SJ, Reeder PJ, Dordick JS. Bioinformatics-driven, rational engineering of protein thermostability. Protein Eng Des Sel 2006; 19(11): 517-24.

15. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng 2000; 13(3): 179-91.

16. Lehmann M, Loch C, Middendorf A, et al. The consensus concept for thermostability engineering of proteins: further proof of concept. Protein Eng 2002; 15(5): 403-11.

17. Lehmann M, Pasamontes L, Lassen SF, Wyss M. The consensus concept for thermostability engineering of proteins. Biochim Biophys Acta 2000; 1543(2): 408-15.

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): 867-72.

20. Britton KL, Baker PJ, Borges KM, et al. Insights into thermal stability from a comparison of the glutamate dehydrogenases from Pyrococcus furiosus and Thermococcus litoralis. Eur J Biochem 1995; 229(3): 688-95.

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): 255-62.

22. Singh RK, Tropsha A, Vaisman, II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996; 3(2): 213-21.

23. Cambillau C, Claverie JM. Structural and genomic correlates of hyperthermostability. J Biol Chem 2000; 275(42): 32383-6.

24. Andersen CAF, Brunak S. Representation of protein-sequence information by amino acid subalphabets. Ai Magazine 2004; 25(1): 97-104.

25. 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): 1059-69.

26. Landès C, Risler J-L. Fast databank searching with a reduced amino-acid alphabet. Comput Appl Biosci 1994; 10(4): 453-4.

27. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003; 16(5): 323-30.

28. 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.

29. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000; 13(3): 149-52.

30. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Eng 2000; 13(8): 545-50.

31. Solis AD, Rackovsky S. Optimized representations and maximal information in proteins. Proteins 2000; 38(2): 149-64.

32. Lau KF, Dill KA. A Lattice Statistical-Mechanics Model of the Conformational and Sequence-Spaces of Proteins. Macromolecules 1989; 22(10): 3986-97.

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): 1356-62.

34. 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.

35. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000; 16(6): 276-7.

36. 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): 680-2.

37. Osmanbeyoglu HU, Ganapathiraju MK. N-gram analysis of 970 microbial organisms reveals presence of biological language models. BMC Bioinformatics 2011; 12(1): 12.

38. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002; 16: 321-57.

39. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. SIGKDD Explor Newsl 2009; 11(1): 10-8.

40. EL-Manzalawy Y, Honavar V. {WLSVM}: Integrating LibSVM into Weka Environment. 2005; Available from: http://www.cs.iastate.edu/~yasser/wlsvm.

41. Chang C, Lin C. {LIBSVM}: a library for support vector machines. 2001; Available from: 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: Cabestany J, Prieto A, Sandoval F, editors. Computational Intelligence and Bioinspired Systems: Springer Berlin / Heidelberg; 2005. p. 85-125.

**Fig. (1).** Overall workflow of the protocol

Refer to the Protocol section under Methods for a detailed explanation of the workflow.

Thermostability_protocol_overview.emf

**Table 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 |

**Table 2.** Reduced Amino Acid Alphabets

|  |  |  |
| --- | --- | --- |
| **Alphabet** | **Size** | **Reference** |
| Native | 20 |  |
| Ab | 10-19 | [24] |
| Dssp | 10-14 | [31] |
| Eb | 11, 13 | [25] |
| Gbmr | 10-14 | [31] |
| Hsdm | 10,12,14-17 | [30] |
| Lr | 10 | [26] |
| Lwi | 10-19 | [27] |
| Lwni | 10,11,14 | [27] |
| Lzbl | 10-16 | [28] |
| Lzmj | 10-16 | [28] |
| Ml | 10,15 | [29] |
| Sdm | 10-14 | [30] |
| Random | 10 | This study |

**Table 3.** 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** | Hsdm16 | 13 | 91.796 | 0.960 | 0.921 | 0.914 |
| **(1-grams)** | 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** | Lwi18 | 158 | 91.513 | 0.965 | 0.906 | 0.926 |
| **(2-grams)** | 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** | Sdm12 | 227 | 88.826 | 0.949 | 0.882 | 0.895 |
| **(3-grams)** | 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 |

**Table 4.** 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.

Supplementary File 1: RAAA groupings and statistically significant n-grams in the training set

For each RAAA, first line is the amino acid grouping and the next three lines correspond to significant 1-grams, 2-grams and 3-grams, respectively.

Supplementary File 2: Classification results for all RAAAs and n-grams

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