Protein Classification on the basis of thermal stability using Supervised Learning

Ankit Sharma - MT16121

Supervisor: Dr. Debajyoti Bera
Dr. Ganesh Bagler

Department of Computational Biology
Indraprastha Institute of Information Technology, Delhi

This dissertation is submitted for the degree of Master of Technology

May 2018
I would like to dedicate this thesis to my loving parents . . .
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Ankit Sharma - MT16121
May 2018
Acknowledgements

I sincerely acknowledge and thank my supervisors Dr. Debajyoti Bera (IIIT Delhi) and Dr. Ganesh Bagler (IIIT Delhi) for giving me the opportunity to work on this project and also for their supervision and valuable guidance during the course of this thesis. I would also like to thank my family for their consistent support, motivation and trust in my abilities.
Abstract

Species evolve by adapting to variable thermal conditions. The differences in the thermal stability of hyperthermophilic, thermophilic and mesophilic proteins arise partly due to their structural variations. The goal of this thesis is to identify structural features responsible for these variations using machines learning techniques that use features derived from the residue interaction graphs (RIG) and the amino acid sequences of proteins. For the RIG model, we studied the features linked to thermal stability which capture different notions of centrality, connection strength, weighted clustering coefficient and such. We evaluated them against a few features that were hitherto not studied in the context of thermal classification and demonstrated that the new features can significantly improve classification accuracy. We further improved the performance by using a histogram of centrality values as a feature vector instead of using a single statistic such as mean that has been the trend so far among researchers. We discovered that the histograms corresponding to edge betweenness centrality, current flow closeness centrality and 2-hop degree centrality lead to the best classification accuracy among the network-based features. We also investigated the state-of-the-art features based on amino acid sequences and proposed a new one using the amino acid tri-mers of a protein. For empirical evaluation, we investigated a set of 842 hyperthermophilic, 533 thermophilic and 2248 mesophilic proteins and compared our proposed features with the state-of-the-art features using commonly known classifiers such as SVM, ANN and random forest. We obtained an overall accuracy greater than 90% which is significantly better than what has been reported so far.
Table of contents

List of figures xiii

List of tables xv

1 Introduction 1
   1.1 Thermophilicity ............................................ 1
   1.2 Problem Statement ........................................ 1
   1.3 Related work .............................................. 2
   1.4 Key findings ............................................. 2

2 Data compilation 5
   2.1 Data curation ............................................. 5
       2.1.1 Data extraction .................................... 5
       2.1.2 Data preprocessing .................................. 6
       2.1.3 Finalizing the dataset for classification .......... 8

3 Methodology 11
   3.1 RIG models of protein structures ......................... 11
   3.2 Feature Extraction ....................................... 13
       3.2.1 Network properties as features .................... 14
       3.2.2 Sequence properties as features .................. 19
   3.3 Feature Selection ......................................... 20
   3.4 Training the Classifiers .................................. 21
   3.5 Predictions ................................................ 24

4 Results and Discussion 25
   4.1 Results ..................................................... 25
       4.1.1 Methodology 1 ........................................ 25
       4.1.2 Methodology 2 ........................................ 26
Table of contents

4.1.3 Methodology 3 ................................................. 27
4.1.4 Hold out feature comparison - Network features .......... 28
4.2 Discussion ......................................................... 29

References ......................................................... 33

Appendix A Thermophilic and Mesophilic organisms ........... 35
  A.1 Thermophilic Organisms .................................. 35
  A.2 Mesophilic Organisms .................................... 35
# List of figures

2.1 Cumulative distribution for thermal classes of proteins - Stage 1 dataset  ... 7
2.2 Accuracy comparison for Stage 1 Dataset and Stage 2 Dataset using Artificial Neural Network. .................................................. 8
2.3 Accuracy trend with increasing number of higher resolution structures  ... 8
2.4 Normalized Cumulative Distribution for Thermophilic-Mesophilic finalized dataset ................................................................. 9

3.1 A standard Machine Learning pipeline ........................................... 12
3.2 RIG models ............................................................................. 13
3.3 Comparison of methodologies for feature extraction ...................... 19
3.4 Feature selection process ............................................................. 20
3.5 SVM ..................................................................................... 21
3.6 SVM-Kernel in action ................................................................. 22
3.7 Artificial Neural Network architecture ........................................ 23
3.8 ReLU activation function ........................................................... 23

4.1 Performance comparison in among 3 methodologies using Support Vector Machine-SVM, Artificial Neural Network-ANN and Random forest.  ... 29
List of tables

2.1 Optimal growth temperature (OGT) ranges for different thermal classes. . . 5
2.2 Protein structure count at different data processing stages . . . . . . . . . . 9
3.1 All the network based features at a glance. . . . . . . . . . . . . . . . . . . 17
4.1 Results - Methodology 1 . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
4.2 Results - Methodology 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
4.3 Results - Methodology 3 . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
4.4 Hold out feature comparison for network features . . . . . . . . . . . . . . 29
Chapter 1

Introduction

1.1 Thermophilicity

Thermophilicity of a protein is a property associated with robust structure and function response of a protein at high temperatures. Unfavorable temperature conditions may lead to cleavage of bonds which play a vital role in holding a protein’s shape. Thermophilicity accounts for thermal stability of the protein and it is a non-linear property. On the basis of an organism’s ability to survive at different temperature conditions, organisms are classified into hyperthermophilic, thermophilic and mesophilic classes. The temperature ranges associated with hyperthermophilic, thermophilic and mesophilic class of proteins/organisms are >81°C, 42°C to 80°C and <41°C respectively [14]. Thermophilicity plays an important role in the survival of an organism, for instance, mesophilic proteins when subjected to a temperature of >41°C, may unfold resulting into its inactivation which may be lethal to the organism. Understanding thermophilicity can also provide significant insights into the structural compactness and molecular rigidity. Better understanding of structural and sequence-based correlates of thermal stability can be used to design proteins that would be stable at higher temperatures.

1.2 Problem Statement

The objective of this study was to identify structure- and sequence-based correlates of proteins that are critical for imparting stability at high temperatures. Towards this goal, supervised learning techniques were used to classification of proteins into thermophilic and mesophilic categories. Subsequently, key features were ascertained by ranking them according to their contribution towards classification accuracy.
1.3 Related work

Protein thermal stability has been an active area of research for a while. In order to pin down the factors that induce thermal stability, many researchers have used an exhaustive set of features and a wide spectrum of techniques. It has been inferred that thermal stability in proteins is a cumulative effect of a protein’s spatial organization as well as its primary sequence. Some of the factors responsible for holding the shape of a protein include hydrogen bonds, amino acid composition, di-peptide composition, salt bridges etc. These are discussed in more detail in the next chapter.

Gromiha et al.[10] identified amino acid composition is one of the important aspects in classifying thermophilic and mesophilic proteins. The importance of residues was also established by Zhang et al.[15], and Wang et al.[6]. Ding et al.[7] have stated in their work that thermophilic proteins have higher average degree, higher connection strength and more compact structure than mesophilic proteins. Also, the thermophilic proteins are more rich in hubs than then mesophilic proteins. Gao et al. [9] used the residue interaction network (discussed in detail in the next chapter) to capture the three dimensional spatial information of a protein structure. These structural correlates were were used as features into machine learning algorithms. We aimed to identify key features towards classification of thermophilic and mesophilic proteins. Some of the algorithms that have also been proposed that implement mutation based thermo-stability prediction are as follows: Mutant 2.0 by Capriotti et al. [4], MU-pro by Cheng et al. [5], iPTREE-STAB by Huang et al. [12].

The main objective of this work is to improvise upon the existing approaches by proposing a new strategy which aims to capture the overall distribution of the features rather than banking upon a single value as the representation of the distribution of the feature. Along with it, we also focused on identification of critical network based features that play a major role in the classification.

1.4 Key findings

Subtle features in protein structures are expected to hold the information that differentiate thermophilic and mesophilic proteins. A battery of structure- and sequence-based features were used to classify such proteins. The key findings of this research work are as follows:

1. Following structural features obtained from the RIG model were found to be critical for thermal stability: degree, connection strength, weighted clustering coefficient, closeness centrality, degree centrality (1 hop and 2 hops), edge betweenness centrality, current flow closeness centrality, subgraph centrality and tri-peptide composition.
2. Machine learning models were able to achieve a classification accuracy of over 90%.

3. The top 3 network features found (by using hold out feature test) were: current flow closeness centrality, edge betweenness centrality and degree centrality (2 hops).
Chapter 2

Data compilation

2.1 Data curation

Data curation pipeline involved compilation of data of organisms belonging to different thermal classes followed by extraction of proteins for each organism. This was followed by filtering data using different protocols to sanitize it (data preprocessing). The dataset was finalized after validation using various experiments (finalizing the dataset).

2.1.1 Data extraction

First, we obtained the list of organisms [14] which had their “optimal growth temperature” (OGT) within the ranges specified in Table 2.1. These organisms were already classified into hyperthermophilic, thermophilic and mesophilic classes. All proteins corresponding to these organisms were extracted from Protein Data Bank (PDB) [1] in the form of PDB files and FASTA files. PDB files were used to construct a network model for every protein structure (Residue Interaction Graph) and subsequently were used for network-based feature extraction, whereas the FASTA files were used for extracting sequence-based features. The data comprised of 1248 hyperthermophilic, 716 thermophilic and 3386 mesophilic proteins.

<table>
<thead>
<tr>
<th>Thermal class</th>
<th>OGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyperthermophilic</td>
<td>&gt;81°C</td>
</tr>
<tr>
<td>Thermophilic</td>
<td>42°C to 80°C</td>
</tr>
<tr>
<td>Mesophilic</td>
<td>&lt;41°C</td>
</tr>
</tbody>
</table>

Table 2.1 Optimal growth temperature (OGT) ranges for different thermal classes.
2.1.2 Data preprocessing

Data preprocessing was done using the following protocol and the processed data after this was termed as Stage 1 Dataset.

- The raw data was pruned to remove structures to only keep proteins with acceptable resolution (\(\leq 2.5\) Å).
- In the presence of multiple structures for the same protein, the structure with higher molecular weight was considered. In case of multiple proteins after this step, the structure with the lowest resolution was kept.

After this protocol, Stage 1 dataset contained 842 hyperthermophilic, 533 thermophilic and 2248 mesophilic proteins. Intuitively, hyperthermophilic-mesophilic (HM) pair is more polar in comparison to thermophilic-mesophilic (TM) pair in terms of temperature ranges. As a result, it was expected that HM pair classification accuracy would be higher than TM pair classification accuracy. In order to validate this, supervised classification techniques were applied and are extensively explained in the next chapter under methodology 1. But, the results were not in accordance with this intuition. HM classification accuracy was computed along with the TM classification accuracy, and the results were 68.25% and 70.50% respectively for Stage 1 dataset.

The following was the possible rationale that that could explain the apparently contradictory results: Resolution plays a vital role in influencing the classification process. Theoretically, poorer the resolution the poorer is the error in capturing the exact co-ordinates of the amino-acid residues. Therefore, while randomly sampling randomly from the dataset, if the number of poor resolution structures are more, then it would have an adversely impact on the classification. This rationale was validated with the help of a series of experiments. In experiment 1, a cumulative distribution over resolution was obtained for hyperthermophilic (H), thermophilic (T) and mesophilic (M) proteins for Stage 1 Dataset and is illustrated in Fig. 2.1.

From Fig. 2.1 it is evident that the number of structures with poorer resolution are more in hyperthermophilic dataset in comparison to the thermophilic dataset. This implies that the H-M classification would be noisier when compared to T-M classification, if the rationale regarding the resolution holds. Furthermore, in experiment 2, a range of resolution (termed as optimal window) was to be identified wherein the number of structures would be comparable for hyperthermophilic and thermophilic proteins so as to account for the bias towards higher number of poor resolution structures in hyperthermophilic dataset. This optimal window is defined as 0 Å–1.9 Å and a matched number of data points were sampled. The classification
accuracy was expected to be comparable for structures present within the optimal window but higher than the experiment 1 in which the entire dataset was subjected to classification. The classification accuracy for H-M classification was found out to be 70.25% and for TM classification was found out to be 73.28%. The results are depicted in Fig. 2.2. The intuition of HM pair giving cleaner results couldn’t be validated and this might be due to the nature of the data. The processed data after this step is termed as Stage 2 Dataset.

In the final experiment, data was added in an incremental manner for thermophilic and mesophilic proteins by incorporating poorer resolution structures in every subsequent step. The step size was 50. that is from one step to the next one 50 more structures were added to the corpus in order of increasing resolution starting with 250 structures from each category. The accuracy trend is illustrated in Fig. 2.3. It is evident that as we incorporate structures with poorer resolution, the classification accuracy drops significantly. These experiments validate the suggested rational: poorer the resolution of the protein, worse would be the classification accuracy due to increased error in construction of RIG models and further propagation of errors into the computation of features.
2.1.3 Finalizing the dataset for classification

After all the experiments, we inferred that thermophilic dataset would be preferred data-set for classification against the mesophilic dataset instead of the hyperthermophilic dataset. Despite comparable accuracy we chose to use thermophilic-mesophilic dataset pairs for further experiments due to greater number of data points in the former which enables better training of the classifiers.
Thus, thermophilic-mesophilic Stage 2 dataset was finalized and the problem statement involved a binary classification challenge, wherein the objective was to design a robust classifier using network- and sequence-based features. Table 2.2 shows the statistics of protein structures at different stages of data processing. Fig. 2.4 shows the normalized cumulative distribution of the final thermophilic-mesophilic data with comparable number of structures from each class. This data is balanced compared to Fig. 2.1 with the two curves almost superimposed on each other. This suggests that the final processed data is unbiased and hence is suitable for the application of supervised learning techniques.

The list of organisms (Thermophilic and Mesophilic) is provided in the appendix. The complete finalized dataset is available here.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Hyperthermophilic</th>
<th>Thermophilic</th>
<th>Mesophilic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw data</td>
<td>1248</td>
<td>716</td>
<td>3386</td>
</tr>
<tr>
<td>Stage 1</td>
<td>842</td>
<td>533</td>
<td>2248</td>
</tr>
<tr>
<td>Stage 2</td>
<td>170</td>
<td>288</td>
<td>1089</td>
</tr>
</tbody>
</table>

Table 2.2 Protein structure count at different data processing stages
Chapter 3

Methodology

Beyond the data compilation and curation, in this chapter, we explain the methodology implemented for building the model, extracting features and performing the classification. Following is the structure of the chapter:

1. Building Residue Interaction Graph (RIG) models
2. Feature Extraction
3. Feature Selection
4. Training the Classifiers
5. Predictions

A standard machine learning pipeline is illustrated in Fig.3.1.

3.1 RIG models of protein structures

Thermophilicity can be attributed to a large number of factors within a protein structure. These factors also involve various types of weak interactions (non-covalent interactions) that play a vital role in retaining the folded structure of a protein. Some such factors have been summarized below:

- Hydrophobicity [13]: It is defined as the property of a molecule by the virtue of which it has a tendency to stay away from water molecules. Greater the hydrophobicity, higher is the resistance to thermal fluctuations.
• Shortening of loops [13] : Length of the loops is an important factor in determining the thermal behavior of a protein. Lesser the length of the loop, more compact and rigid is the structure.

• Hydrogen bonds [13] : It is a type of dipole-dipole interaction in between a hydrogen atom bound to a more electronegative atom like Nitrogen, Oxygen or Fluorine and another atom having a lone pair of electrons. It is stronger than the conventional dipole-dipole interaction but is weaker than ionic or covalent bonds. Greater the number of hydrogen bonds, stronger is the tendency to retain structural fold.

• Increased salt bridges [13] : A salt bridge involves two non-covalent interactions : a hydrogen bond and an ionic bond. These bridges greatly contribute to the stability of folded structure of a protein and if present in higher number, make the structure stronger.

In order to capture the impact of these factors on thermal stability, a graph theoretic model known as Residue Interaction Graph (RIG) was constructed for every protein. PDB file contains information of coordinates of atoms in the protein structure required for constructing a RIG. A weighted RIG was constructed by computing the euclidean distances among all α
carbon atoms. \( \alpha \) carbon atoms were chosen as they are consistently present in the backbone of all amino acids and present a good tangible geometric point representing an amino acid. If the euclidean distance between two \( \alpha \) carbon atoms is lesser than or equal to the cut-off radius then an edge exists in between those two \( \alpha \) carbon atoms with edge weight being equal to the difference between the cut-off radius and the obtained euclidean distance. The cut-off radius was set to be 6.5 Å. From previous studies it is known that the conclusions based on RIG models are not very sensitive to the cut-off and any value in the range of 5 Å to 10 Å meaningfully represents the structural details [9]. For the euclidean distance between two \( \alpha \) carbon atoms greater than 6.5 Å, the residues were considered to be non-interacting. Fig. 3.2 shows illustration of RIG models, one each from the thermophilic and the mesophilic class. RIG models of protein structures, obtained in this manner, were used for extracting network based features which were further incorporated in machine learning classifiers.

![RIG models](image)

Fig. 3.2 RIG visualization with threshold as 6.5 Å. Weights have been ignored for simplicity.

## 3.2 Feature Extraction

The features to be extracted were broadly classified into 2 categories:

1. **Network based features** : RIG model was used to extract network based features, and mainly include centrality based measures.

2. **Sequence based features** : Amino acid sequences were used to fetch sequence based features.

All features from these two categories are explained below.
3.2.1 Network properties as features

RIG models were exploited to obtain the network-based features. The aim was to consider those features which have been reported to have an impact on the compactness and rigidity of the structure. Centrality based features were the prime contributors along with other features. The list of features is as follows:

1. Degree
2. Connection Strength
3. Characteristic Path Length
4. Weighted Clustering Coefficient
5. Closeness Centrality
6. Edge Density
7. Degree Centrality (1 hop)
8. Degree Centrality (2 hops)
9. Edge Betweenness Centrality
10. Current-flow Closeness Centrality
11. Subgraph Centrality

Three types of methodologies were adapted in order to capture the thermal behavior of the proteins with the help of these features accurately. The main difference among these methodologies was the resolution at which these features were captured.

Methodology 1

Methodology 1 aimed at extracting the network based features wherein each and every feature would represent the averaged out value for every protein based RIG model. Therefore, a single value was used to represent the entire network and the nature of the distribution was ignored. The mathematical expressions along with a brief explanation are as follows:

1. **Average Degree**: The average degree of a network is defined as the sum of degree of all the nodes divided by the total number of nodes present in the network. Mathematically, average degree $= \frac{\sum_{i=1}^{N} d_i}{N}$ where $d_i$ is the degree of the $i^{th}$ node.
2. **Average Connection Strength**: The average connection [9] strength of a network is defined as the sum of connection strength of all the nodes divided by the total number of nodes present in the network. The connection strength of a node is defined as the sum over product of the number of edges in between that node and its immediate neighbors selected one at a time and the corresponding weight of that edge. Mathematically, average connection strength $= \frac{\sum_{i=1}^{N}cs_i}{N}$ where $cs_i = \sum_{j=1}^{N} a_{ij}w_{ij}$, $a$ is the adjacency matrix of the graph and $w$ is the weight matrix of the graph.

3. **Average Characteristic Path Length**: Average characteristic path length [9] is defined as the average of the shortest paths in between all the pair of nodes. Mathematically, average characteristic path length $= \frac{\sum_i\sum_j L_{ij}}{N(N-1)}$, where $L_{ij}$ is the shortest path in between node ‘i’ and node ‘j’.

4. **Average Weighted Clustering Coefficient**: Weighted Clustering Coefficient [9] is defined as the average of the measure of degree to which the nodes of the network tend to cluster together. Mathematically, average weighted clustering coefficient $= \frac{1}{N} (\sum_{i=1}^{N} \frac{1}{cs_i(a_i-1)} (\sum_{j,h} w_{ij}w_{ih}+wh 2a_{ih}a_{jh})).$ To compute the average it is added for all nodes and is divided by the total number of nodes.

5. **Average Closeness Centrality**: Closeness centrality [9] is defined as the degree up to which a node is near all other nodes of the network. Mathematically, average closeness centrality $= \frac{\sum_{i=1}^{N} cc_i}{N-1} \sum_{j\in S} distance(i,j)$, $S$ is the set of all nodes of the network.

6. **Edge Density**: It is defined as the number of edges of the network divided by the number of nodes of the network.

7. **Average Degree Centrality (1 hop)**: It is defined for every node of a network and corresponds to the fraction of nodes that particular node is connected to. To compute the average it is added for all nodes and is divided by the total number of nodes.

8. **Average Degree Centrality (2 hops)**: Degree centrality (2 hops) is defined as the degree centrality with the hop radii set to 2, i.e. while computing the centrality, neighbors up to 2 hops are considered for every node. To compute the average it is added for all nodes and is divided by the total number of nodes.

9. **Average Edge Betweenness Centrality**: Betweenness centrality for an edge [2] is the sum of the fraction of all-pairs shortest paths that pass through that edge. Mathematically, edge betweenness centrality for an edge $e = \sum_{s,t\in V} \frac{\sigma(s,t|e)}{\sigma(s,t)}$, where $V$ is the set
of vertices, $\sigma(s,t)$ is the number of shortest (s,t) paths and $\sigma(s,t|e)$ is the number of those paths passing through e. To compute the average it is added for all edges and is divided by the total number of edges.

10. **Average Current-flow Closeness Centrality**: Current-flow closeness centrality [3] is variant of closeness centrality based on effective resistance between nodes in a network. This metric is also known as information centrality. Mathematically, current flow closeness centrality for a node $n = \left[ \frac{1}{N} \sum_j \frac{1}{I_{ij}} \right]^{-1}$, where $I_{ij}^{-1} = (B^{-1})_{ii} + (B^{-1})_{jj} - 2(B^{-1})_{ij}$ and $B = D - A + J$ where D is the diagonal matrix of the degree of each node, A is the adjacency matrix of the network and J is a matrix with all elements equal to 1. To compute the average it is added for all nodes and is divided by the total number of nodes.

11. **Average Subgraph Centrality**: Subgraph centrality [8] for a node ‘n’ is defined as the sum of weighted closed walks of all lengths starting and ending at node ‘n’. The weights decrease with path length and each closed walk is associated with a connected subgraph. It can be found out using a spectral decomposition of the adjacency matrix, Mathematically, subgraph centrality for a node ‘n’, $S(n) = \sum_{j=1}^{n} (v_j)^2 e^{\lambda_j}$ where $v_j$ is an eigen vector of the adjacency matrix A of G corresponding to the eigen value $\lambda_j$. To compute the average it is added for all nodes and is divided by the total number of nodes.

Thereby, 11 network based feature values are extracted for every protein ‘RIG model’.

The above mentioned network features are summarized in table 3.1.

**Methodology 2**

In order to overcome the loss of information because of taking the average value of the distribution as the feature in the previous scheme, a more robust and elaborative feature extraction method was used. Under this, a new concept of ‘feature histogram’ was used. Feature bucket extraction was done by the following steps:

1. For a particular feature, its value was computed for all the nodes of the network.

2. Maximum value of the feature was extracted from amongst all the nodes.

3. All the values obtained in Step 1 were divided by the maximum value extracted in Step 2, so that the values correspond to the range in between 0 and 1.
### Table 3.1 All the network based features at a glance.

<table>
<thead>
<tr>
<th>Network feature</th>
<th>Feature type</th>
<th>References</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>Node based</td>
<td>Gao et. al [9]</td>
<td>Previously applied to RIG models</td>
</tr>
<tr>
<td>Connection Strength</td>
<td>Node based</td>
<td>Gao et. al [9]</td>
<td>Previously applied to RIG models</td>
</tr>
<tr>
<td>Characteristic path length</td>
<td>Node based</td>
<td>Gao et. al [9]</td>
<td>Previously applied to RIG models</td>
</tr>
<tr>
<td>Weighted Clustering Coefficient</td>
<td>Node based</td>
<td>Gao et. al [9]</td>
<td>Previously applied to RIG models</td>
</tr>
<tr>
<td>Closeness Centrality</td>
<td>Node based</td>
<td>Gao et. al [9]</td>
<td>Previously applied to RIG models</td>
</tr>
<tr>
<td>Edge density</td>
<td>Edge based</td>
<td>-</td>
<td>Previously applied to biological networks</td>
</tr>
<tr>
<td>Degree centrality (1 hop)</td>
<td>Node based</td>
<td>-</td>
<td>Applied to complex networks but not to biological networks</td>
</tr>
<tr>
<td>Degree centrality (2 hops)</td>
<td>Node based</td>
<td>-</td>
<td>Proposed by us</td>
</tr>
<tr>
<td>Edge betweenness centrality</td>
<td>Node based</td>
<td>Brandes et. al [2]</td>
<td>Applied to complex social networks but not to biological networks</td>
</tr>
<tr>
<td>Subgraph centrality</td>
<td>Node based</td>
<td>Estrada et. al [8]</td>
<td>Previously applied to RIG models</td>
</tr>
</tbody>
</table>
4. The range in between 0 and 1 was binned using a bin size of 0.1; and 10 such bins were obtained.

5. Each bin is populated with number of nodes having value that lies in that range. For instance, bin 1 contains the number of nodes that have the feature value in between 0 and 0.1, bin 2 contains the number of nodes that have the feature value in between 0.1 and 0.2 and so on.

6. This bin count was then divided by the total number of nodes present in the network so as to normalize the feature across protein RIG models of different sizes.

Each feature hence gets split into 10 feature buckets/bins and these 10 ‘feature buckets/bins’ are collectively used for training and prediction.

**Methodology 3**

The third methodology is a hybrid approach wherein the resolution at which the features are picked is a combination of methodology 1 and methodology 2. It takes into account the distribution of feature values for all nodes, but at the same time does not work over individual node values. The strategy is listed as follows:

1. For a particular feature, its value was computed for all the nodes of the network.

2. Maximum value of the feature was extracted from amongst all the nodes.

3. All the values obtained in Step 1 were divided by the maximum value extracted in Step 2, so that the values correspond to the range in between 0 and 1.

4. The range in between 0 and 1 was binned using a bin size of 0.1 and 10 such bins were obtained.

5. Each bin was populated with the nodes having value that lies in that range. For instance, bin 1 contains the nodes that have the feature value in between 0 and 0.1, bin 2 contains the nodes that have the feature value in between 0.1 and 0.2 and so on.

6. Average value of the feature was computed for every bin wherein the total sum of value of all nodes in that bin is divided by the total number of nodes of the network.

In this strategy as well each feature gets split into 10 feature buckets/bins and these 10 feature buckets/bins were collectively used for training and prediction. This is strategy a balanced and an optimum strategy for extracting network-based features. The granularity attains a
balance where the features are neither too coarse grained as they were in methodology 1 nor too fine grained as they were in methodology 2. This is illustrated in Fig. 3.3. The only disadvantage of methodology 2 and 3 is the fact that all 10 buckets/bins must used together so as to use the feature for classification.

**Hold out feature comparison**

Hold out feature comparison was also performed to get a sense of relative more important features wherein every feature was dropped turn by turn. The corresponding drop in the accuracy marks the importance of the feature dropped. This is explained in detail in the next chapter.

![Fig. 3.3 Comparison of feature extraction methodologies in terms of feature value granularity.](image)

### 3.2.2 Sequence properties as features

Each protein has an amino acid sequence associated with it which is a significant contributor to the thermal stability. Sequences were downloaded in the form of FASTA files and following sequence based features extracted:

1. **Amino acid composition**: Amino acid composition is defined as the number of amino acid residues of a particular type divided by the total length of the sequence. As a result, 20 amino acid compositions were obtained for 20 different amino acids. Some of the amino acids are believed to have a significant impact on thermal stability of proteins.

   Mathematically, amino acid composition for an amino acid = \( \frac{n}{N} \) where \( n \) is the number of occurrences of that particular amino acid and \( N \) is the total number of amino acids in the protein.
2. **Di-peptide composition**: It is defined as the ration of number of di-peptides of each type in a protein and the total number of di-peptides.

Mathematically, di-peptide composition = \( \frac{ndp}{n_{ndp}} \) where ndp is the total number of di-peptides of a particular type in the amino acid sequence and \( n_{ndp} \) is the number of all possible di-peptides obtainable from the amino acid sequence.

3. **Tri-peptide composition**: It is defined as the number of tri-peptides of each type in a protein divided by the total number of tri-peptides.

Mathematically, tri-peptide composition = \( \frac{n_{tp}}{n_{ntp}} \) where \( n_{tp} \) is the total number of tri-peptides of a particular type in the amino acid sequence and \( n_{ntp} \) is the number of all possible tri-peptides obtainable from the amino acid sequence.

### 3.3 Feature Selection

An exhaustive feature selection strategy was implemented through which all possible subsets of features were generated. After subset generation, classifiers were selected from amongst Support Vector Machine, Artificial Neural Network and Random Forest. The classifier was trained using 10 folds cross validation and the hyper-parameters are optimized. Feature subset ranking was computed after this step and the subset with the highest accuracy in maximum folds (along with no over-fitting) was reported as the selected set of features. This strategy was applied to all 3 methodologies with the only difference that as the features subjected to feature selection must be independent, thereby, in methodology 2 and 3, all the bins corresponding to a particular feature were either picked or dropped together. This process has been illustrated in Fig. 3.4. This is a brute-force feature selection technique and gives the best results. On the other hand, due to its exhaustive nature it is slow when large number of features are used. Weka’s ClassifierSubsetEval [11] function provides an implementation for the same.

![Fig. 3.4 The exhaustive process of feature selection.](image)
3.4 Training the Classifiers

The classifiers chosen for this problem statement included Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest. A SVM is a discriminative classifier which works on the principle of identification of a hyper-plane that would separate the classes (thermophilic and mesophilic) better (Fig. 3.5). One basic metric of identifying the appropriate hyper-plane is maximizing the distance between the nearest data points. This distance is termed as ‘margin’ and on the basis of margin, SVM’s can be made to function as a ‘hard margin classifier’ or a ‘soft margin classifier’. A hard margin classifier does not allow mis-classifications and is a lot stricter when compared to a soft margin classifier that allows some mis-classifications. SVM’s can also be used to ignore outliers and hence is a robust classifier.

If the data is not separable using a linear hyper-plane, then SVM’s can be made to make use of kernels to perform classification (Fig. 3.5). Kernels are mathematical functions that take low dimensional input and return a high dimensional output.

![Support Vector Machine classifier with an optimal hyper-plane separating the two classes.](image)

From amongst the advanced classifiers, Artificial Neural Network and Random Forest were chosen. ANN consists of hidden layers as shown in Fig. 3.7. It was chosen because of the fact that an ANN progressively improves the classification by the virtue of back-propagation in which the weights are readjusted after computing the difference in between the ground truth and the predicted label. Stepwise framework of an ANN is as follows:

1. Initially assign random weights to the connections.
Fig. 3.6 Non-linear classification using a kernel function which helps in defining a decision surface.

2. Input nodes as well the connections in between the input nodes and the hidden nodes is used to find the rate of activation of the hidden nodes.

3. Using the activation rate obtained in the above step, and the connections to the output nodes, the activation rates of the output nodes are obtained.

4. Output error is computed and the connections in between the hidden nodes and the output nodes are re calibrated.

5. Using the error and the weights, error is cascaded down to the hidden nodes.

6. Weights in between the hidden nodes and the input nodes are re calibrated.

7. This process is repeated until the algorithm converges.

8. The final weights are used for prediction.

ReLU (Rectified Linear Unit) shown in Fig.3.8 is used as the activation function in ANN because of its advantages over sigmoid and tanh like reduced likelihood of the gradient to vanish. Mathematically, $y = \max(0; z); z = Wx + b$ where $'W'$ represents the weight and $'b'$ represents the bias.

**Random Forest** is a bootstrapping technique based upon decision tree model. This algorithm generates a forest of decision trees and then predicts the result which has got the maximum consensus. Steps for creating a random forest are as follows:

1. Select $'k'$ features randomly from a set of $'l'$ total features.

2. From the $'k'$ features, compute the root node $'r'$ using best split.
3.4 Training the Classifiers

Fig. 3.7 Architecture of Artificial Neural Network (ANN).

Fig. 3.8 ReLU activation function, $y = \max(0, z)$ where $z = Wx + b$. 
3. Split the root node into child nodes.

4. Repeat Steps 1 to 3 until 'm' nodes have been reached (threshold limit).

5. Construct the forest by repeating Steps 1 to 4 and generate 't' such trees.

6. In order to make predictions, use the trained model and predict the outcome with the maximum votes from amongst all 't' decision trees.

3.5 Predictions

The trained classifier was then used to make predictions for the unseen test dataset. The predicted labels were then compared against the true labels and accuracy was computed. Accuracy is defined as the percentage of correctly classified samples from the entire test set. Mathematically,

\[
\text{Accuracy} = \frac{x}{y} \times 100, \quad \text{where} \quad x = \text{number of correctly predicted labels from the test set and} \\
y = \text{total number of labels in the test set.}
\]
Chapter 4

Results and Discussion

4.1 Results

After performing all the experiments described in the previous chapter, results obtained with all three methodologies were compared. Fig. a plot is presented in Fig 4.1.

4.1.1 Methodology 1

In methodology 1, the feature under consideration was computed for each and every node of the residue interaction graph. Average value for that feature was calculated over all the nodes of the network and was used as a dimension for that protein. This process was repeated for every protein. When fed to the machine learning classifiers (post hyper-parameter tuning), the results averaged over 10 runs for 10 folds cross-validation were as per Table 4.1. The network features selected from amongst the superset of all the features mentioned in the previous chapter were:

1. Degree centrality
2. Two hop degree centrality
3. Edge betweenness centrality
4. Current flow closeness centrality

The amino acids for which the composition was chosen were:

1. Alanine - ala - A
2. Glutamine - gln - Q
3. Glutamic acid - glu - E
4. Histidine - his - H
5. Threonine - thr - T

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Prediction accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>83.03%</td>
<td>83.1%</td>
<td>83.0%</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>81.11%</td>
<td>81.0%</td>
<td>81.11%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>83.48%</td>
<td>83.4%</td>
<td>83.0%</td>
</tr>
</tbody>
</table>

Table 4.1 Results - Methodology 1

4.1.2 Methodology 2

In methodology 2, the feature values were normalized for each and every node. Bins were created within the range 0–1 and the fraction of total nodes having its normalized feature value within those bins was computed. These binned feature vectors were treated as dimensions. As the features are expected to be independent in machine learning, bins for any feature were either selected in totality or were all dropped. When fed to the machine learning classifiers (post hyper-parameter tuning), the results averaged over 10 runs for 10 folds cross-validation were as per Table 4.2. The network features selected from amongst the superset of all the features mentioned in the previous chapter were:

1. Degree
2. Connection strength
3. Clustering coefficient
4. Closeness centrality
5. Two hop degree centrality
6. Edge betweenness centrality
7. Current flow closeness centrality
8. Sub graph centrality

The amino acids for which the composition was chosen were:
4.1 Results

1. Alanine - ala - A
2. Glutamine - gln - Q
3. Glutamic acid - glu - E
4. Histidine - his - H
5. Threonine - thr - T

Along with all these features, di-peptide composition for proteins was chosen over tri-peptide composition.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Prediction accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>80.6%</td>
<td>80.6%</td>
<td>80.6%</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>86.28%</td>
<td>86.4%</td>
<td>86.3%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>87.5%</td>
<td>87.9%</td>
<td>87.5%</td>
</tr>
</tbody>
</table>

Table 4.2 Results - Methodology 2

4.1.3 Methodology 3

In methodology 3, the feature values were normalized for each and every node. Bins were created within the range 0 - 1 and the average value for nodes having its normalized feature value within those bins was computed. These binned feature vectors were treated as dimensions. Just like in methodology 2, bins for any feature were either selected in totality or were all dropped. When fed to the machine learning classifiers (post hyper-parameter tuning), the results averaged over 10 runs for 10 folds cross-validation were as per Table 4.2. The network features selected from amongst the superset of all the features mentioned in the previous chapter were:

1. Degree
2. Connection strength
3. Clustering coefficient
4. Closeness centrality
5. One hop degree centrality
6. Two hop degree centrality
7. Edge betweenness centrality

8. Current flow closeness centrality

9. Sub graph centrality

The amino acids for which the composition was chosen were:

1. Alanine - ala - A
2. Glutamine - gln - Q
3. Glutamic acid - glu - E
4. Histidine - his - H
5. Threonine - thr - T

Along with all these features, tri-peptide composition for proteins was chosen over di-peptide composition.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Prediction accuracy</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machine</td>
<td>81.4%</td>
<td>81.4%</td>
<td>81.6%</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>88.6%</td>
<td>88.5%</td>
<td>88.0%</td>
</tr>
<tr>
<td>Random Forest</td>
<td>90.8%</td>
<td>90.9%</td>
<td>90.8%</td>
</tr>
</tbody>
</table>

Table 4.3 Results - Methodology 3

Parallel experiments were performed and the work done by Gao. et al [9] was replicated over the curated dataset. We couldn’t implement our methodology over their dataset because their dataset is not accessible. The accuracy achieved was 84.3% with SVM as the classifier.

4.1.4 Hold out feature comparison - Network features

In order to infer about the importance of network features, hold out feature comparison was performed. In this strategy, accuracy is recomputed by leaving out one feature vector completely from the selected feature vector set. The drop in the accuracy signifies the contribution of the feature in the classification. Higher is the accuracy drop, greater is the importance of that feature. This was repeated for every feature and the results are tabulated as per Table 4.4.
4.2 Discussion

<table>
<thead>
<tr>
<th>Held out feature</th>
<th>Prediction accuracy</th>
<th>Dropped accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>87.32%</td>
<td>3.48%</td>
</tr>
<tr>
<td>Connection strength</td>
<td>87.28%</td>
<td>3.52%</td>
</tr>
<tr>
<td>Weighted clustering coefficient</td>
<td>88.19%</td>
<td>2.61%</td>
</tr>
<tr>
<td>Closeness centrality</td>
<td>87.15%</td>
<td>3.65%</td>
</tr>
<tr>
<td>Degree centrality (1 hop)</td>
<td>87.5%</td>
<td>3.3%</td>
</tr>
<tr>
<td>Degree centrality (2 hop)</td>
<td>86.84%</td>
<td>3.96%</td>
</tr>
<tr>
<td>Edge betweenness centrality</td>
<td>86.74%</td>
<td>4.06%</td>
</tr>
<tr>
<td>Current flow closeness</td>
<td>86.19%</td>
<td>4.61%</td>
</tr>
<tr>
<td>Subgraph centrality</td>
<td>87.5%</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

Table 4.4 Hold out feature comparison for network features

![Table 4.4 Hold out feature comparison for network features](image)

Fig. 4.1 Performance comparison in among 3 methodologies using Support Vector Machine-SVM, Artificial Neural Network-ANN and Random forest.

4.2 Discussion

Starting with the methodologies, the most obvious difference lies in the granularity at which features have been extracted. Methodology 1 deals with the features at a coarse grained level. Methodology 2 looks at the feature vector at a very fine grained level where each and every node retains its identity in the feature vector space. Methodology 3 strikes a balance in terms of granularity and is neither too coarse grained nor too fine grained. When the features are very coarse grained, information loss is encountered as the distribution of the feature value within the network is ignored and the mean value is extracted. This would lead to poorer
classification accuracy as evident from table 4.1. The classifier is not fed with refined feature vector set and hence the classification is noisy.

In methodology 2, the feature vector contains too much information as it is very fine grained. This makes it difficult for the classifier to carry out the classification. This is quite evident from the results of SVM where we were not able to achieve more than 81% accuracy. SVM being a simple classifier working over support vectors struggles greatly and thereby complex classifiers like ANN and random forest were used. The property of ANN to use the feedback to readjust the weights boosts the classification accuracy for fine grained feature vectors. The accuracy achieved was 86.28%. Random forests gave slightly better but comparable result with classification accuracy reaching 87.5%.

Using methodology 3, we were able to capture the features at a balanced granularity scale. The results are a reflection of this as in methodology 3, SVM gave 81.4% accuracy which is greater than SVM’s accuracy in methodology 2. This can be attributed to the fact that the simpler feature vectors makes it easier for the simpler classifiers to carry out the classification. The results were also better for complex classifiers like ANN and random forest with 88.6% and 90.8% classification accuracy respectively.

**Comparison with the existing work:** The results obtained by using feature histogram were much better as compared to dealing with gross averages[9] as per results from methodology 1 and methodology 3. This is illustrated by the difference in the classification accuracies and therefore, the feature extraction strategy proposed by us is better.

Binning feature vectors by using methodology 3 serves as a better feature extraction strategy as compared to methodology 1 which has been used extensively by researchers over the past few years. Another important goal of our work was to find important network features that attribute towards thermal stability. Hold out feature comparison shows that the most significant drop in accuracy was for edge betweenness centrality, current flow closeness centrality and degree centrality. This marks the importance of these features as the top 3 most important network features for thermo stability. The reason for these features being important is explained below:

- **Edge betweenness centrality** - Edge betweenness centrality is defined for every interaction of the RIG model and it quantifies the frequency of occurrence of an interaction in the shortest paths of the RIG model. Information in a network is believed to flow from the shortest/low resistance paths. Hence, an interaction being important on the basis of edge betweenness centrality value, signifies that as higher information content flows through that interaction, thereby, that interaction is important for that RIG model to retain its function. In our problem statement, this essentially translates to the importance of a connection in between two α carbon atoms, being responsible
for thermal stability. In a RIG model, edge/interactions in between nodes/Cα atoms helps in fold retention of the protein. The edge betweenness feature value in the RIG model essentially means that the average frequency of the interactions to appear in the shortest paths is relatively higher which translates to robustness of the tertiary protein structure. This invariant behavior towards extreme conditions, as captured by edge betweenness centrality is causal to thermal stability of a protein.

• Current flow closeness centrality - Current flow closeness centrality deals with the other aspect of information flow through a RIG model. As per closeness centrality, the RIG is thought of as a network with electrical resistors and is defined for every Cα atom of the network. Interactions represent the resistors and the Cα atoms represent the junction in between resistors. Two Cα atoms are considered close when the resistance in between them is low. The interaction weights could be used to represent resistance or conductance depending upon the problem statement. This centrality measure is also known as information centrality and is based on the Laplacian of a graph. Laplacian refers to the spectral signature of a graph which captures the structural density which is a prime contributor to thermal stability.

• Degree centrality(2 hops) - It measures the number of interactions incident on a Cα atom and the interactions are considered to be incident if they are found in its 2 hop neighborhood. It is more important than 1 hop degree centrality because it spans the interactions over a greater reach within a RIG. Higher values of degree centrality implies greater importance of the corresponding Cα atom. In a way, it also captures compactness, therefore, is of relatively more importance than other features.

Some of the other relevant network features included closeness centrality, connection strength and degree. Closeness centrality looks from the perspective of weights associated with the interactions within the RIG. Higher the closeness centrality feature value, greater is the rigidity of the protein structure. Connection strength captures a sense of structural rigidity and is therefore deemed useful in context of protein fold retention. Degree being a very simple feature captures the significance of the Cα atom. Greater value of the degree feature of the RIG signifies stronger protein fold retention.

From amongst the remaining features, weighted clustering coefficient led to the least drop in accuracy. This implies that clustering coefficient in itself might not be a significant contributor to thermal stability but certainly adds value to the classification process. As per our problem statement, it quantifies and captures the frequency of the neighborhood of a Cα atom resembling a dense cluster. Greater number of clusters in a RIG implies greater rigidity
as it becomes very difficult to dis-orient a network/RIG with a large number of densely packed sub graphs/motifs.

As far as sequence based features are concerned, both di-peptide and tri-peptide compositions (k-mer based composition with \( k = 2 \) and \( k = 3 \) respectively) could be used for classification along with amino acid composition. Tri-peptide composition is believed to capture slightly more information as it looks at an amino acid triplet at a time. This concept of k-mers was extended up to \( k = 5 \), but feature selection became very compute intensive due to the large number of k-mers that are obtained and hence this experiment was aborted. Amino acid composition does not contribute much to the classification process as it works over the presence/absence of the amino acids within the protein. It considers all k-mers with \( k = 1 \). Relating it to the problem statement, it can be concluded that only the presence/absence of a particular amino acid does not fully dictate thermal stability, but also the neighboring amino acids matter.

Hence, we would like to conclude by stating that feature extraction strategy greatly affects the classification process and hence granularity of features should be kept in mind. The proposed strategy evidently works well as demonstrated in this thesis. Our results suggest that three features, namely, edge betweenness centrality, current flow closeness centrality and 2 hops degree centrality as the top 3 key structural correlates of thermal stability and these correlates could be easily linked to compactness and structural rigidity which imparts thermal stability to the protein.
References


# Appendix A

**Thermophilic and Mesophilic organisms**

The list of organisms along with their Optimal Growth Temperature (OGT) from which the proteins were extracted in a categorical manner is as follows:

## A.1 Thermophilic Organisms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Organism Name</th>
<th>OGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>mth</td>
<td>Methanothermobacter thermautotrophicus</td>
<td>70</td>
</tr>
<tr>
<td>tv o</td>
<td>Thermoplasma volcanium</td>
<td>60</td>
</tr>
<tr>
<td>tac</td>
<td>Thermoplasma acidophilum</td>
<td>59</td>
</tr>
<tr>
<td>lac</td>
<td>Lactobacillus acidophilus NCFM</td>
<td>45</td>
</tr>
<tr>
<td>sto</td>
<td>Sulfolobus tokodaii</td>
<td>80</td>
</tr>
<tr>
<td>sai</td>
<td>Sulfolobus acidocaldarius DSM 639</td>
<td>75</td>
</tr>
<tr>
<td>pto</td>
<td>Picrophilus torridus</td>
<td>60</td>
</tr>
<tr>
<td>t te</td>
<td>Thermoanaerobacter tengcongensis</td>
<td>75</td>
</tr>
<tr>
<td>tfu</td>
<td>Thermobifida fusca</td>
<td>50</td>
</tr>
</tbody>
</table>

## A.2 Mesophilic Organisms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Organism Name</th>
<th>OGT</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmp</td>
<td>Methanococcus maripaludis S2</td>
<td>40</td>
</tr>
<tr>
<td>m ba</td>
<td>Methanosarcina barkeri Fusaro</td>
<td>37</td>
</tr>
<tr>
<td>t de</td>
<td>Treponema denticola</td>
<td>37</td>
</tr>
<tr>
<td>mac</td>
<td>Methanosarcina acetivorans</td>
<td>40</td>
</tr>
<tr>
<td>spy</td>
<td>Streptococcus pyogenes M1 GAS (serotype M1)</td>
<td>27</td>
</tr>
</tbody>
</table>
Thermophilic and Mesophilic organisms

sps  Streptococcus pyogenes SSI-1 (serotype M3)  27
spm  Streptococcus pyogenes MGAS8232 (serotype M18)  27
spg  Streptococcus pyogenes MGAS315 (serotype M3)  27
sag  Streptococcus agalactiae 2603 (serotype V)  37
san  Streptococcus agalactiae NEM316 (serotype III)  37
sak  Streptococcus agalactiae A909 (serotype Ia)  37
hpy  Helicobacter pylori 26695  37
hpj  Helicobacter pylori J99  37
spb  Streptococcus pyogenes MGAS6180 (serotype M28)  27
stc  Streptococcus thermophilus CNRZ1066  37
stl  Streptococcus thermophilus LMG18311  37
spz  Streptococcus pyogenes MGAS5005 (serotype M1)  27
mma  Methanosarcina mazei Go1  40
spa  Streptococcus pyogenes MGAS10394 (serotype M6)  27
spn  Streptococcus pneumoniae TIGR4 (virulent serotype 4)  37
spr  Streptococcus pneumoniae R6 (avirulent)  37
cte  Chlorobium tepidum  35
ser  Staphylococcus epidermidis RP62A  37
cjr  Campylobacter jejuni RM1221  45
cje  Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819  45
saa  Staphylococcus aureus subsp. aureus USA300-FPR3757 (CA-MRSA)  37
wsu  Wolinella succinogenes  35
smu  Streptococcus mutans UA159  37
hhe  Helicobacter hepaticus  37
sac  Staphylococcus aureus subsp. aureus COL (MRSA)  37
sep  Staphylococcus epidermidis ATCC 12228  37
mle  Mycobacterium leprae TN  37
sab  Staphylococcus aureus RF122  37
hit  Haemophilus influenzae 86-028NP (nontypeable)  37
lla  Lactococcus lactis subsp. lactis II1403  30
sha  Staphylococcus haemolyticus JCSC1435  40
nme  Neisseria meningitidis MC58 (serogroup B)  37
nma  Neisseria meningitidis Z2491 (serogroup A)  37
hin  Haemophilus influenzae Rd KW20 (serotype d)  37
zmo  Zymomonas mobilis subsp. mobilis ZM4  30
A.2 Mesophilic Organisms

sam  Staphylococcus aureus subsp. aureus MW2 (CA-MRSA) 37
sau  Staphylococcus aureus subsp. aureus N315 (MRSA/VSSA) 37
sav  Staphylococcus aureus subsp. aureus Mu50 (MRSA/VISA) 37
pac  Propionibacterium acnes KPA171202 32
mca  Methylcoccus capsulatus 37
lin  Listeria innocua 25
sar  Staphylococcus aureus subsp. aureus MRSA252 (MRSA) 37
sas  Staphylococcus aureus subsp. aureus MSSA476 (MSSA) 37
dra  Deinococcus radiodurans 35
lmf  Listeria monocytogenes F2365 37
ssp  Staphylococcus saprophyticus 35
dde  Desulfovibrio alaskensis 37
neu  Nitrosomonas europaea 30
pmu  Pasteurella multocida subsp. multocida Pm70 37
cac  Clostridium acetobutylicum ATCC 824 37
bmf  Brucella abortus 2308 37
cgb  Corynebacterium glutamicum ATCC 13032 (Bielefeld) 37
cgl  Corynebacterium glutamicum ATCC 13032 (Kyowa Hakko) 37
noc  Nitrosococcus oceanii 20
son  Shewanella oneidensis 30
vch  Vibrio cholerae O1 biovar El Tor N16961 37
mtc  Mycobacterium tuberculosis CDC1551 37
bce  Bacillus cereus ATCC 14579 30
bsu  Bacillus subtilis subsp. subtilis 168 37
vfi  Aliivibrio fischeri ES114 30
rsp  Rhodobacter sphaeroides 2.4.1 40
bca  Bacillus cereus ATCC 10987 30
pae  Pseudomonas aeruginosa PAO1 37
ppu  Pseudomonas putida KT2440 30
xac  Xanthomonas citri pv. citri 306 30
xcv  Xanthomonas campestris pv. vesicatoria 39
pst  Pseudomonas syringae pv. tomato DC3000 30
ypk  Yersinia pestis KIM10+ (biovar Mediaevalis) 37
xcb  Xanthomonas campestris pv. campestris 8004 39
bme  Brucella melitensis bv. 1 16M 37
xcc  Xanthomonas campestris pv. campestris ATCC 33913 39
<table>
<thead>
<tr>
<th>Code</th>
<th>Organism Name</th>
<th>Strain/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>psb</td>
<td>Pseudomonas syringae pv. syringae B728a</td>
<td>30</td>
</tr>
<tr>
<td>atu</td>
<td>Agrobacterium fabrum</td>
<td>28</td>
</tr>
<tr>
<td>ypm</td>
<td>Yersinia pestis 91001 (biovar Microtus)</td>
<td>37</td>
</tr>
<tr>
<td>pfo</td>
<td>Pseudomonas fluorescens Pf0-1</td>
<td>30</td>
</tr>
<tr>
<td>ecc</td>
<td>Escherichia coli O6:K2:H1 CFT073 (UPEC)</td>
<td>37</td>
</tr>
<tr>
<td>bcz</td>
<td>Bacillus cereus E33L</td>
<td>30</td>
</tr>
<tr>
<td>ype</td>
<td>Yersinia pestis CO92 (biovar Orientalis)</td>
<td>37</td>
</tr>
<tr>
<td>ecj</td>
<td>Escherichia coli K-12 W3110</td>
<td>37</td>
</tr>
<tr>
<td>psp</td>
<td>Pseudomonas savastanoi pv. phaseolicola 1448A</td>
<td>30</td>
</tr>
<tr>
<td>rso</td>
<td>Ralstonia solanacearum GMI1000</td>
<td>30</td>
</tr>
<tr>
<td>eci</td>
<td>Escherichia coli O18:K1:H7 UTI89 (UPEC)</td>
<td>37</td>
</tr>
<tr>
<td>pfl</td>
<td>Pseudomonas protegens Pf-5</td>
<td>30</td>
</tr>
<tr>
<td>reu</td>
<td>Ralstonia eutropha JMP134</td>
<td>30</td>
</tr>
<tr>
<td>rpa</td>
<td>Rhodopseudomonas palustris CGA009</td>
<td>37</td>
</tr>
<tr>
<td>yps</td>
<td>Yersinia pseudotuberculosis IP 32953 (serotype I)</td>
<td>37</td>
</tr>
<tr>
<td>ece</td>
<td>Escherichia coli O157:H7 EDL933 (EHEC)</td>
<td>37</td>
</tr>
<tr>
<td>eco</td>
<td>Escherichia coli K-12 MG1655</td>
<td>37</td>
</tr>
</tbody>
</table>