*Journal of Cell and Molecular Research (2010) 2 (1), 15-28*

# Investigating protein features contribute to salt stability of halolysin proteins

Esmaeil Ebrahimie[1]*, Mansour Ebrahimi[2] and Narjes Rahpayma[3]

[1]*Department of Crop Production and Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran*
[2]*Bioinformatics Research Group, Green Research Center, Qom University, Qom, Iran*
[3]*Department of Crop Production and Plant Breeding, College of Agriculture, Shiraz University, Shiraz, Iran*

Received 4 July 2010 Accepted 6 September 2010

## Abstract

The study used various screening techniques, clustering, decision tree and generalized rule induction (association) (GRI) models and molecular phylogenic relationship to search for patterns of halophilicy and to find features contribute to halolysin salt stability. We found that Met was the sole N-terminal amino acid in halolysin proteins, whereas other amino acids found at that position of other proteases and termitase. Eighty-three protein features were shown to be important in feature selection modeling, and just one peer group with an anomaly index of 2.42 declined to 1.87 after being run using only important selected features. The depth of the trees generated by various decision tree models varied from 1 to 5 branches. Compared to datasets without feature selection the number of peer groups in clustering models was reduced significantly ($p<0.05$). In most decision tree models, the frequency of Gly - Gly was the most important feature for decision tree rule sets and this feature was used in antecedent to support the rules in most GRI association rules. Significant differences ($p < 0.001$) found in charged amino acids between halolysin and other proteins with more Asp and Glu in halolysin proteins, while more hydrophobic residues and aliphatic amino acids were found in other proteases.

***Keywords:*** bioinformatics, modeling, protein, halophilic, halolysin

## Introduction

*Halobacteria*, extremely halophilic red-pigmented bacteria, have been intensively studied during the past decades (Sumper, 1987; Oren, 1994; Kamekura, 1998; Mukohata et al., 1999; Joo and Kim, 2005), through which our understandings of various biological processes such as energy metabolism (Gonzalez-Hernandez and Pena, 2002), environmental response (Elevi Bardavid and Oren, 2008), gene regulation (Conover and Doolittle, 1990), and the Archaea l cell cycle (Cui et al., 2006) have been greatly increased. Their extraordinary ability to grow in hypertonic solution (above 300 g of NaCl per liter) and their potential ability to hydrolyze proteins are the main reasons for rapid increase in research in this field (Kristjansson et al., 1986). A microorganism corresponding to the description of *Halobacteria salinarum* was isolated from salted fish more than 80 years ago (Soppa, 2006). Since then, many haloArchaea l species have been isolated, which, after considerable renaming, are currently grouped into 25 genera. Several years ago, it was decided that the species *Halobacterium salinarum*, *Halobacterium halobium*, and *Halobcterium cutirubrum* are so similar that they should be regarded as strains of one species named Halobacterium salinarum. *Halobacterium salinarum* shows very high genetic variability that was attributed to the large number of insertion sequences (Yang et al., 2006).

A small percentage of proteins can tolerate salinity and dryness stress. The enzymes from extremely halophilic bacteria represent a fascinating example of adaptation. . These enzymes function in vivo and in vitro at ranges of 4 to 5 M NaCl and upon exposure to low salt densities they lose their activities very rapidly (Binbuga et al., 2007; Pesenti et al., 2008; Zhu et al., 2008). Recently, genes for a number of halophilic enzymes have been cloned, including dihydrofolate reductase from *Haloferax volcanii* (Fine et al., 2006), glutamate dehydrogenase from *Halobacterium salinarium* (Ingoldsby et al., 2005), and malate dehydrogenase from *Haloarcula marismortui* (Zaccai et al., 1986). The mechanism of halophilicity of these enzymes, however, has not been fully elucidated at the molecular level. It has been shown Glu243Arg, a mutant protein of the malate dehydrogenase, was more halophilic, and

*Corresponding author E-mail:
*ebrahimiet@shirazu.ac.ir*

required significantly higher concentrations of NaC1 or KCI for equivalent stability (Madern et al., 1995). Proteases are key enzymes in many processes important to the cell and are widely used in biotechnology and industry. Many representatives of the *Archaea* domain are extremophiles, thriving in conditions lethal to most cells. Thus, *Archaea* represent an important resource of enzymes, including proteases, in applied research as well as for basic enzymology. For applications requiring low water activity such as high salt or organic solvents, haloArchaea l and their enzymes have great potential to act as biocatalysts (Kamekura et al., 1992; De Castro et al., 2008).

Halolysin, a halophylic alkaline serine protease, has been extracted from *Archeaebacterium* and some other bacteria such as *Natrialba asiatica*, *Haloferax mediterranei*, *Natrialba magadii* and *Halobacterium sp. NRC-1* (Kamekura et al., 1992; Kamekura and Seno, 1993; Kamekura et al., 1996; De Castro et al., 2008). Halolysinfrom Halophilic archaeon is active at NaCl concentration of 4-4.5 M, loses its activity at salt concentration lower than 2M and is a very interesting sample of studying adaptation to harsh conditions (Feng and Yang, 2008; Strahl and Greie, 2008). The purpose of this study was to find the most important features contributing to these enzymes' ability to stand high concentration of salts and find other similar possible enzymes. Here we studied phylogenic relationship, feature selection, screening models, association models and statistical analyses among halolysin and other proteases extracted from few bacteria, fungi and plants in order to investigate features contributing to salt tolerance.

## Material and Methods

Nine halolysin sequences (A42605, AAG20619, AAV66536, BAA01049, BAA10958, CAP14928, NP_281139, P29143 and YP_001690274) were extracted from UniProt Knowledgebase (Swiss-Prot and TrEMBL). To find similar proteases, peptidases and termitase sequences, p29143 halolysin sequence was used to blast with available databases and 37 plant protease, 8 fungal proteases and 6 termitase were found and saved as FASTA format. To draw phylogenic tree, three software (CLCbio, MEGA4 and CLASTAL W) were used with similar parameters (i.e. Neighbor joining algorithm). Similar consensus sequences with 100% restrictions from alignment sequence with lower E value were chosen. Forty hundred and thirty nine protein features such as length, weight, isoelectric point, count and frequency of each element (carbon, nitrogen, sulphur, oxygen and hydrogen), count and frequency of each amino acid, count and frequency of negatively charged, positively charged, hydrophilic and hydrophobic residues, count and frequency of dipeptids, number of α-helix and β-strand and other secondary protein features were extracted.

To investigate protein features contributing to resistance of halolysin proteins to salty conditions and to compare them with other proteases and termitase studied in this paper, we divided dataset proteins into two groups: 1) T/F groups (T = halolysin proteins and F = other proteins; plant, bacterial and fungal proteases and termitase). 2) H/B/F/P/T groups (H = halolysin proteins, B = bacterial proteases, F = fungal proteases, P = plant proteases and T = termitase; comparing halolysin proteins with individual class of other proteins). The Protein name (either T/F or H/B/F/P/T) variable was set as the output variable and others as input variables. All features were classified as continuous variables, except the N-terminal amino acid, which was classified as categorical. A dataset of these protein features was imported into Clementine software (Clementine_NLV-11.1.0.95; Integral Solution, Ltd.).

Various decision tree algorithms were applied to the datasets to identify the most important features and find possible patterns that contribute to protein classes. These models allowed the development of classification systems that automatically included in their rules only the attributes important in making a decision. Attributes that did not contribute to the accuracy of the tree were ignored. This process yielded very useful information about the data and could be used to reduce the data to relevant fields only before training another learning technique, such as a neural network. As various algorithms were available for performing classification and segmentation analysis, and herein we used different decision tree and cluster analysis models. All models were run both with and without feature selection criteria to investigate the effects of the feature selection algorithm on other models behavior. All models run as previously described (Ebrahimi et al., 2009; Bijanzadeh et al., 2010; Ebrahimi and Ebrahimie, 2010).

### *Screening Models*

#### *Anomaly detection model*

This model was used to identify outliers or unusual cases in the data. Unlike other modeling methods that store rules about unusual cases,

anomaly detection models store information on what normal behavior looks like. This makes it possible to identify outliers even if they do not conform to any known pattern. While traditional methods of identifying outliers generally examine one or two variables at a time, anomaly detection can examine large numbers of fields to identify clusters or peer groups into which similar records fall. Each record then can be compared to others in its peer group to identify possible anomalies. The further away a case is from the normal center, the more likely it is to be unusual.

***Feature selection algorithm***

The feature selection algorithm was applied to identify the attributes having a strong correlation with the thermostability of enzymes. The algorithm considers one attribute at a time to determine how well each predictor alone predicts the target variable. The important value for each variable is then calculated as (1–*p*), where *p* is the *p* value of the appropriate test of association between the candidate predictor and the target variable. The association test for the categorized output variables differs from the test for continuous variables. In our study, when the target value was categorical (as in our datasets), *p* values based on the F statistic were used. The idea was to perform a one-way ANOVA F test for each predictor; otherwise, the *p* value was based on the asymptotic t distribution of a transformation of the Pearson correlation coefficient. Other models, such as likelihood-ratio chi-square (also tests for target-predictor independence), Cramer's V (a measure of association based on Pearson's chi-square statistic), and Lambda (a measure of association that reflects the proportional reduction in error when the variable is used to predict the target value) were conducted to check the possible effects of calculation on feature selection criteria. The predictors were then labeled as important, marginal, and unimportant, with values > 0.95, between 0.95 and 0.90, and < 0.90.

***Clustering Models***

***K-Means***

The K-Means model can be used to cluster data into distinct groups when the content of the groups is unknown. Unlike most learning methods in Clementine, K-Means models do not use a target field. This type of learning, with no target field, is called unsupervised learning. Instead of trying to predict an outcome, K-Means tries to uncover patterns in the set of input fields. Records are grouped so that those which are within a group or a cluster tend to be similar to each other, whereas those which are in different groups are dissimilar. K-Means works by defining a set of starting cluster centers derived from the data. It then assigns each record to the cluster to which it is most similar based on the record's input field values. After all cases have been assigned, the cluster centers are updated to reflect the new set of those records assigned to each cluster. The records are then checked again to see whether they should be reassigned to a different cluster, and the record assignment/cluster iteration process continues until either the maximum number of iterations is reached or the change between one iteration and the next fails to exceed a specified threshold.

***Two-Step cluster***

The Two-Step cluster model is a two-step clustering method. The first step makes a single pass through the data, during which it compresses the raw input data into a manageable set of subclusters. The second step uses a hierarchical clustering method to progressively merge the subclusters into larger and larger clusters, without requiring another pass through the data. Hierarchical clustering has the advantage of not requiring the number of clusters to be selected ahead of time. Many hierarchical clustering methods start with individual records as starting clusters and merge them recursively to produce ever-larger clusters.

***Decision Tree Models***

***Classification and regression tree (C & RT)***

This model uses recursive partitioning to split the training records into segments by minimizing the impurity at each step. A node is considered pure if 100% of cases in the node fall into a specific category of the target field.

***CHAID***

This method generates decision trees using chi-square statistics to identify optimal splits. Unlike the C&RT and QUEST models, CHAID can generate non-binary trees that means some splits can have more than two branches.

***Exhaustive CHAID***

This model is a modification of CHAID thatmore thoroughly examins all possible splits, but it takes longer to compute.

***QUEST***

The QUEST model provides a binary

classification method to builde decision trees. It is designed to reduce the processing time required for large C & RT analyses while also reducing the tendency are performed in classification tree methods to favor predictors that allow more splits.

***C5.0***

The C5.0 model builds either a decision tree or a rule set. The model works by splitting the samples based on the field providing the maximum information gained at each level. The target field must be categorical. Multiple splits into more than two subgroups are allowed.

***Association Model***

The generalized rule induction (GRI) model discovers association rules in the data. GRI extracts a set of rules from the existing data, pulling them out the rules with the highest information content. Information content is measured using an index that takes both the generality (support) and accuracy (confidence) of rules into account.

Statistical analyses; general linear model comparisons (pairwise comparisons with Tukey test and confidence level of 95.0% were done by the SPSS software (version 13, Michigan, USA).

## Results

More than 72% (155) of proteins studied here were bacterial proteases while 17.21% (37), 4.19% (9), 3.72% (8) and 2.79% (6) were plant proteases, halolysins, fungal proteases and termitase. The average length, weight, isoelectric point, and aliphatic indices of proteins studied here were 573.27 ± 260.91, 60.95 ± 28.30, 6.68 ± 1.59, and 81.40 ± 7.80 (mean ± SD). The average counts of sulphur, carbon, nitrogen, oxygen, and hydrogen were 12.86, 2695.85, 733.94, 852.92, and 4196.46, respectively, and the average counts of hydrophobic, hydrophilic, and other residues were 292.84 ± 134.88, 169.84 ± 75.40, and 110.58 ± 61.90 (mean ± SD). The frequencies of hydrogen, carbon, oxygen, nitrogen, and sulphur in all enzymes were 0.494 ± 0.005, 0.317 ± 0.003, 0.087 ± 0.003, 0.101 ± 0.005 and 0.002 ± 0.001, and the frequencies of hydrophobic, hydrophilic, other, negatively, and positively residues were 0.509 ± 0.048, 0.303 ± 0.051, 0.188 ± 0.033, 56.55 ± 36.38, and 43.66 ± 25.14, respectively. The frequencies of amino acids ranged from a low amount of 0.0001 ± 0.00001 for Ile, Asp and Gln to a high amount of 0.176 ± 0.024 for Ala.

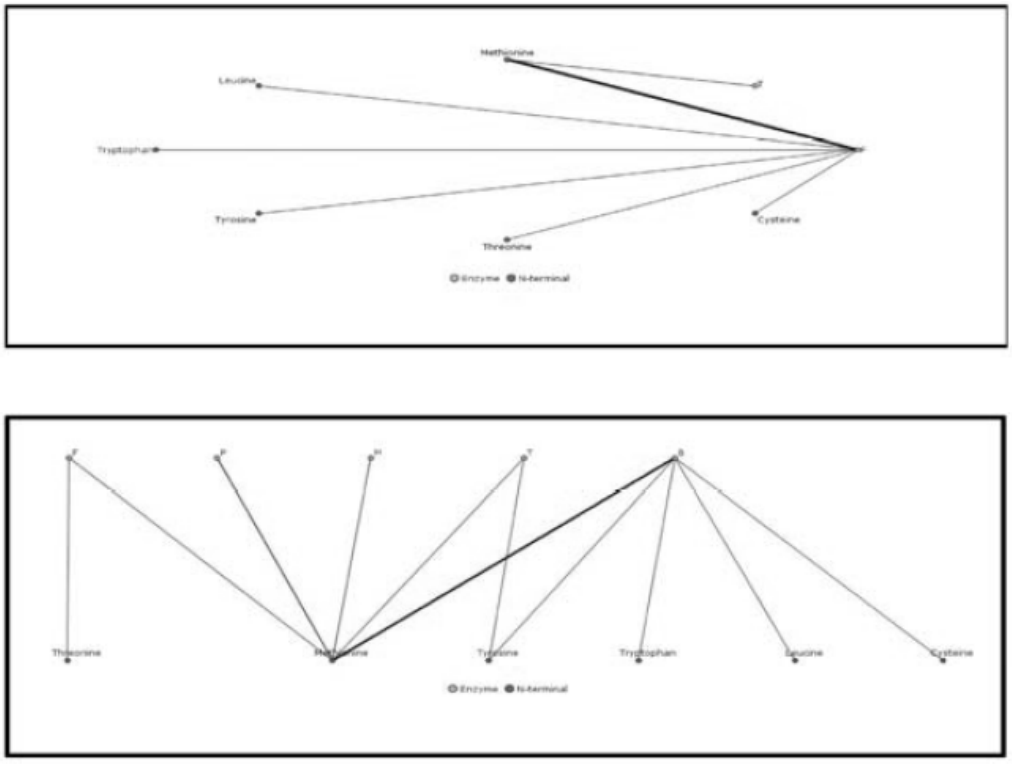


**Figure 1.** Web graph of N-terminal amino acids in a) T (halolysin proteins) and F (bacterial, Fungi and plant proteases and termitase), and b) bacterial proteases (B), fungal proteases (F), halolysin proteins H), plant proteases (p) and termitase (T) groups, thicker lines showing higher incidences of amino acids.

In 97.21% of proteins, the N-terminal amino acid was Met and in 0.93% of proteins the same position was occupied by Tyr. In 0.47%, the last amino acid was Cys, Leu, Thr and Trp. The average non-reduced Cys extinction coefficient at 280 nm was 71367.07 ± 31759, non-reduced Cys absorption was 1.21 ± 0.28, the reduced Cys extinction coefficient was 71109.49 ± 31629.72, and the reduced Cys absorption was 1.20 ± 0.28 (mean ± SD). Figure 1 is a web graph that illustrates the strength of the relationship between N-terminal amino acids and halophilic properties of proteins. Met exhibited a strong relationship with all proteins (a thicker line shows a stronger relationship). Met was the only N-terminal amino acid found in halolysin proteins, whereas Cys, Leu, Tyr, Try and Thr were found at N-terminal position of other proteases and termitase proteins. When halolysin proteins were compared with individual classes of other proteases and termitase, Met exhibits a strong relationship with all proteins and was the only N-terminal amino acid found in halolysin proteins, whereas Tyr and Thr were found at N-terminal position of fungal proteases and termitase proteins and Tyr, Try, Cys and Leu were found at the N-terminal in bacterial proteases.

The results showed that halolysin proteins can be inserted in a separate phylum between eukaryotes (plants and fungi) and bacteria, called Archea (figure 2). Some bacterial proteases such as thermophilic proteases [Q45670 (b118), EDL64549 (b147), ZP-01860436 (b149), YP-002603898(b108) and YP-002603888] showed close relationship with halolysin proteins. Plant proteases from pherphion family with EEF49096 (Tripeptidyle peptidase II, putative) and some bacterial proteases are classified in a separate group. According to figure 2, plants proteins are located at the top of phylogenic tree while fungi proteins with other bacteria proteases such as CAD85094 (b129) and CAD43134 (b50) are put near the top of the tree, confirming their place as eukaryotes. The results of protein blast showed that some parts of the proteins are conserved in all proteins studied here (E value 0). These conserved proteins have been known as putative, pattern formation or hypothetical proteins with a common amino acid sequence of Sec 7; this central region serves as exchange factor.

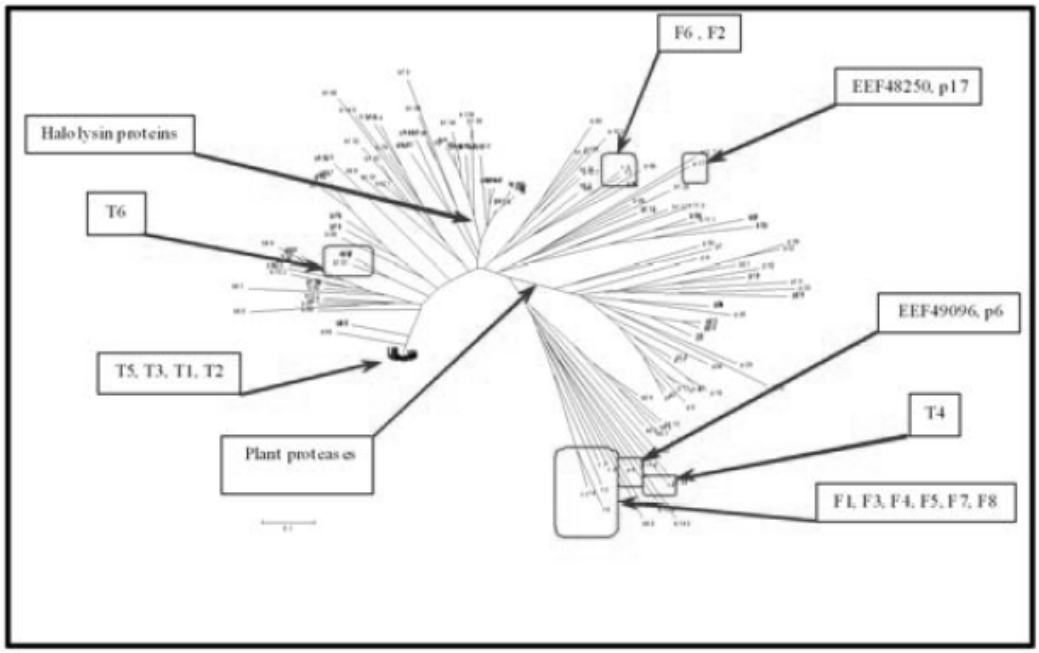


**Figure 2.** Phylogenic tree generated by MEGA4 software, showing halolysin proteins position regarding to other proteases and termitase (f: Fungi protease, T: Termitase, P: Plant protease, H: Halolysine).

When feature selection model applied on dataset of protein features compared halolysin with other proteins (T/F groups), 83 of 215 features were ranked as important ($p > 0.95$) in contribution to halolysin ability to stand harsh conditions (table 1) and 15 features were found to be marginal ($0.90 < p > 0.95$). When the halolysin was compared with each individual protein classes (H/B/F/P/T), 176 out of 215 features were ranked as important and 15 features as marginal. Each time, a node was generated with just important features and was used whenever it was necessary to run all other models on feature selection dataset (as mentioned in Materials and Methods).

**Table 1:** Results of feature selection on important and marginal features contributing to the optimum temperature of proteins

| No | Field | Value | Rank | No | Field | Value | Rank |
|---|---|---|---|---|---|---|---|
| **1** | Freq. of Gly-Gly | 1.0 | Important | **50** | Freq. of Phe-Ala | 0.991 | Important |
| **2** | Freq. of Ala –Asp | 1.0 | Important | **51** | Freq. of Ala - Cys | 0.99 | Important |
| **3** | Freq. of Gly-Asp | 1.0 | Important | **52** | Freq. of His- Asp | 0.989 | Important |
| **4** | Freq. of Asp-Pro | 1.0 | Important | **53** | Mature peptide | 0.989 | Important |
| **5** | Freq. of Glu-Leu | 1.0 | Important | **54** | Freq. of Glu-Val | 0.989 | Important |
| **6** | Freq. of Aspartic Acid | 1.0 | Important | **55** | Freq. of Cys-Ala | 0.988 | Important |
| **7** | Freq. of Cys-Trp | 1.0 | Important | **56** | Freq. of His- Glu | 0.988 | Important |
| **8** | Freq. of Gly-Arg | 1.0 | Important | **57** | Freq. of Glu-Lys | 0.987 | Important |
| **9** | Freq. Negatively Charged | 1.0 | Important | **58** | Freq. of Gly-Thr | 0.987 | Important |
| **10** | Freq. of Gly | 1.0 | Important | **59** | Active site | 0.986 | Important |
| **11** | Freq. of Glu-Tyr | 1.0 | Important | **60** | Freq. of Glu-Ser | 0.985 | Important |
| **12** | Freq. of Asp-Leu | 1.0 | Important | **61** | Freq. of Met | 0.984 | Important |
| **13** | Gene | 1.0 | Important | **62** | Freq. of His- Lys | 0.984 | Important |
| **14** | Freq. of Ile | 1.0 | Important | **63** | Freq. of Phe-His | 0.984 | Important |
| **15** | Freq. of Asp-Asp | 1.0 | Important | **64** | Freq. of lie-Lys | 0.983 | Important |
| **16** | Freq. of Asp-Gly | 1.0 | Important | **65** | Count of Phe | 0.983 | Important |
| **17** | Freq. of Asp-Glu | 1.0 | Important | **66** | Freq. of Cys-Pro | 0.982 | Important |
| **18** | Freq. of Ala -Lys | 1.0 | Important | **67** | Count of Ile | 0.981 | Important |
| **19** | Isoelectric point | 1.0 | Important | **68** | Freq. of Phe-Trp | 0.981 | Important |
| **20** | Freq. Positively Charged | 1.0 | Important | **69** | Freq. of Ala -lle | 0.98 | Important |
| **21** | Freq. of Phe | 1.0 | Important | **70** | Freq. of Glu-Gly | 0.977 | Important |
| **22** | Freq. of lie-Arg | 1.0 | Important | **71** | Freq. of Gly-Phe | 0.977 | Important |
| **23** | Freq. of Lys | 1.0 | Important | **72** | Freq. of Ala -Pro | 0.976 | Important |
| **24** | Freq. of Glu-Gln | 1.0 | Important | **73** | Positively Charged residues | 0.976 | Important |
| **25** | Freq. of Glu-Pro | 1.0 | Important | **74** | Freq. of Ser | 0.975 | Important |
| **26** | Freq. of Glu-lle | 1.0 | Important | **75** | Freq. of Gly-His | 0.975 | Important |
| **27** | Freq. of Asp-Gln | 1.0 | Important | **76** | Freq. of Gly-Ala | 0.973 | Important |
| **28** | Freq. of Ala -Thr | 1.0 | Important | **77** | Freq. of Ala -Met | 0.971 | Important |
| **29** | CDS | 1.0 | Important | **78** | Freq. of His- Cys | 0.965 | Important |
| **30** | Freq. of Cys-Glu | 1.0 | Important | **79** | Freq. of lie-Phe | 0.963 | Important |
| **31** | Freq. of Asp-Arg | 1.0 | Important | **80** | Freq. of Asp-Ala | 0.962 | Important |
| **32** | Freq. of lie-lie | 1.0 | Important | **81** | Freq. of sulphur | 0.958 | Important |
| **33** | Freq. of Ala -Ser | 0.999 | Important | **82** | Freq. of Gly-Pro | 0.957 | Important |
| **34** | Freq. of Asp-His | 0.999 | Important | **83** | Freq. of Gly-Lys | 0.953 | Important |
| **35** | Freq. of Glu | 0.999 | Important | **84** | Count of Asp | 0.944 | Marginal |
| **36** | Freq. of Asp-Lys | 0.999 | Important | **85** | Count of Met | 0.943 | Marginal |
| **37** | Freq. of Asp-Phe | 0.999 | Important | **86** | Freq. of lie-Met | 0.94 | Marginal |
| **38** | Freq. of Phe-Asn | 0.998 | Important | **87** | Count of His | 0.933 | Marginal |
| **39** | Count of Lysine | 0.998 | Important | **88** | Freq. of Ala -Tyr | 0.93 | Marginal |
| **40** | Freq. of Phe-Glu | 0.997 | Important | **89** | Freq. of His- Ser | 0.927 | Marginal |
| **41** | Freq. of Glu-Glu | 0.996 | Important | **90** | Count of Beta-strand | 0.926 | Marginal |
| **42** | Freq. of Glu-Thr | 0.995 | Important | **91** | Freq. of Tryp | 0.924 | Marginal |
| **43** | Freq. of Phe-Phe | 0.995 | Important | **92** | Freq. of lie-Leu | 0.922 | Marginal |
| **44** | Freq. of Asp-Thr | 0.994 | Important | **93** | Freq. of Cys-Ser | 0.92 | Marginal |
| **45** | Freq. of His- Leu | 0.993 | Important | **94** | Freq. of Ala –Glu | 0.919 | Marginal |
| **46** | Freq. of Phe-lie | 0.993 | Important | **95** | Freq. of Glu-Cys | 0.918 | Marginal |
| **47** | Freq. of His | 0.992 | Important | **96** | Freq. of Phe-Lys | 0.912 | Marginal |
| **48** | Freq. of Gly-Asn | 0.992 | Important | **97** | Freq. of Glu-Asp | 0.908 | Marginal |
| **49** | Freq. of Gly-Ser | 0.992 | Important | | | | |

When the anomaly detection model was used on T/F groups, the records were divided into just one peer groups with an anomaly index cutoff of 2.42 and 3 records of this peer group of 215 records were found to be anomalies. When the models were applied using feature selection criteria, one peer groups with an anomaly index cutoff of 1.92 was found. When the model was used on H/B/F/P/T groups, one peer group with three records and anomaly index of 2.41 and 1.87 for dataset with or without feature selection filtering was found, respectively.

When the K-Means model was applied on T/F groups, the records were put into 5 groups or clusters (46, 14, 90, 10 and 55). When the model was applied on dataset with feature selection filtering, again five clusters with 58, 56, 21, 26 and 54 records were generated. When the halolysin was compared with each individual class of proteins, (H/B/F/P/T groups), 47 of the records were put into the first cluster and 14, 89, 10, and 55 records were put into the second, third, fourth, and fifth clusters, respectively. When the K-Means model was applied on the dataset with the feature selection filtering, again five clusters were generated, with 56, 3, 12, 77, and 67 records in each cluster.

Two-Step method clustered records (from T/F groups) into two groups with 52 and 159 records in each cluster, and three clusters (with 109, 52 and 54 records in each cluster) were created for the filtered dataset using feature selection criteria. Two clusters (52 and 195 records and 163 and 52 records) were created with or without the feature selection filtering; when the model applied on H/B/F/P/T groups.

When halolysin proteins (T group) were compared with other proteases and termitase (F group), the C5.0 model generated a decision tree with a depth of 2 and cross-validation of 98.1 ± 0.8. The most important feature used to build the tree was the frequency of oxygen. If the value of this feature was equal to or less than 0.111, the proteins fell into F category (bacterial, fungal and plant proteases and termitase); otherwise they were put into the T category. In this category, if the frequency of Tyr was equal to or less than 0.036, they were placed in the F subgroup; otherwise they were put into the T subgroup (halolysin proteins). When a 10-fold cross-validation was applied to the same dataset, again a tree with a depth of 2 and cross-validation of 97.6.1 ± 1.1 was created. The same protein features and values were used to create tree branches. When the same models were applied to datasets using the feature selection filtering, a tree with the same depth (2) and cross-validation of 96.3 ± 1.1 and 89.1 ± 1.0 were generated for C5.0 and C5.0 with a 10-fold cross-validation, respectively. The frequency of Glu-Leu features were used to create the first branch (value < 0.007 in F mode and > 0.007 T Mode); in T mode if the frequency of Gly was equal to or less than 0.121 they were put in F mode (proteases and termitase); otherwise they were in T mode (halolysin proteins).

When the H/B/F/P/T dataset was used, the C5.0 model generated a decision tree with a depth of 5 and cross-validation of 86.9 ± 1.7. The most important feature used to build the tree was the count of sulphur. If the value of this feature was equal to or less than 18, the proteins fell into the bacterial proteases category; otherwise they were put into the plant proteases category. In the bacterial proteases subgroup, the frequency of Glu-Ser was used to create the next tree branches, with < 0.009 as the bacterial protein mode and > 0.009 as the halolysin protein mode. In the plant proteases subgroup, if the value for the frequency of other residues was equal to or less than 0.164, they were placed in the fungal proteases subgroup; otherwise they were put into the plant proteases subgroup. When a 10-fold cross-validation was applied to the same dataset, again a tree with a depth of 5 and cross-validation of 85.5 ± 1.5 was created. The same protein features and values were used to create tree branches. When the same models were applied to datasets using feature selection filtering, a tree with a depth of 4 and cross-validation of 87.5 ± 2.2 and 86.1 ± 2.5 were generated for C5.0 and C5.0 with 10-fold cross-validation. The same protein features were used to create the first and second subgroups.

In the C&RT node, a tree with a depth of 1 was created, and the most important feature used to build the tree was the frequency of Gly - Gly (value < 0.026 for the F mode and > 0.026 for the T mode (the halolysin protein). The same results were obtained when the feature selection was selected. When the halolysin was compared with each individual class of other proteins (H/B/F/P/T groups), a tree with a depth of 4 was created, and the most important feature used to build the tree was the count of sulphur (value < 18.5 for bacterial and > 18.5 for plant proteases). The frequency of Gly - Gly was used to create the second level for the first subgroups (< 0.026 for bacterial and > 0.026 for halolysin proteins) and the frequency of Glu (< 0.032 for plant and >0.032 for bacterial proteases). The same results were obtained when a feature selection was used.

In the Quest modeling, a tree with a depth of 2 was generated, and the frequency of Gly - Gly (with a value equal to or less than 0.021) was used

to create the first tree branches (the F mode) and the frequency of Ala-Lys was used to generate the next subgroup (< 0 for the halolysin protein and > 0 for other proteases and termitase). The same results occurred when a feature selection filtering was applied. When H/B/F/P/T groups compared, a tree with a depth of 2 was generated, and the count of Cys (with a value equal to or less than 7.654) was used to create the first tree branches (bacterial proteases) and the frequency of Gly - Gly was used to generate the next subgroup (< 0.021 for bacterial proteases and > 0.021 for halolysin proteins). In the plant subgroup, the frequency of Lys (0.113) was used to create fungal and plant proteases. The same results occurred when a feature selection filtering was applied.

A tree with a depth of 2 was generated when the CHAID model was applied to the data with and without feature selection. If the frequency of Lie-Ala was < 0.005, the mode was F; if it was > 0.005 and the frequency of Lie-Ala was equal to or less than 0.006, the mode was T. The same trees with the same features and values were generated when exhaustive CHAID models were applied on datasets with and without the feature selection. When H/B/F/P/T groups were compared and the CHAID model was applied to the data with and without the feature selection, a tree with a depth of 3 was generated. If the count of hydrophobic residues was < 180, the mode was bacterial proteases; if it was > 417, the mode was plant proteases. If the count of hydrophobic residues range from 180 to 196 and the frequency of hydrogen was equal to or less than < 0.492, the mode was bacterial proteases; otherwise it was the termitase. When the counts of hydrophobic residues were > 196 and < 225, it formed the next branch, and three other branches were created when the same feature was between 225 and 268, 268 and 341, 341 and 386, and 386 and 417 (figure 3). The same trees with the same features and values were generated when exhaustive CHAID models were applied on datasets with and without the feature selection. The best percentage of correctness, performance evaluation, and mean correctness in the decision tree models were observed in the C5.0 model, followed by the CR&T, CHAID, and finally the Quest models (table 2).

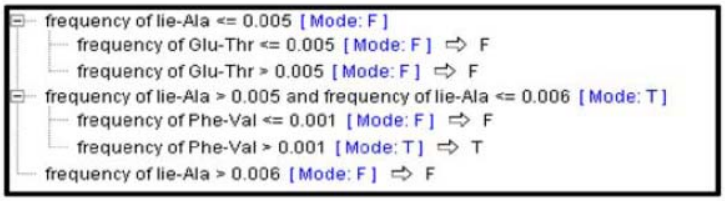


**Figure 3.** A decision tree generated by the CHAID modeling method without feature selection filtering, comparing halolysin proteins with the others (T/F groups).

**Table 2**. Percentage of correctness, wrongness, performance evaluation (T & F), and mean correct and incorrect in various decision tree models, in datasets without feature selection (a) and with feature selection (b), comparing halolysin proteins with others (T/F groups).

| **(a)** | | | | |
|---|---|---|---|---|
| | **% Correct** | **% Wrong** | **Performance evaluation (T)** | **Performance evaluation (F)** |
| **C5.0** | 100 | 0 | 3.173 | 0.043 |
| **C5.0 with 10-fold validation** | 100 | 0 | 3.173 | 0.043 |
| **CR&T** | 98.14 | 1.6 | - | - |
| **QUEST** | 99.53 | 0.47 | - | - |
| **CHAID** | 99.53 | 0.47 | - | - |
| **Exhaustive CHAID** | 99.53 | 0.47 | - | - |
| **(b)** | | | | |
| **C5.0** | 99.53 | 0.47 | 3.173 | 0.038 |
| **C5.0 with 10-fold validation** | 99.53 | 0.47 | 3.173 | 0.038 |
| **CR&T** | 98.14 | 1.84 | 2.933 | 0.033 |
| **QUEST** | 99.53 | 0.47 | 3.173 | 0.038 |
| **CHAID** | 100 | 0 | 3.173 | 0.043 |
| **Exhaustive CHAID** | 100 | 0 | 3.173 | 0.043 |

The GRI node analysis created 100 rules with 215 valid transactions with minimum and maximum support of 3.26% and 8.37%, respectively. Maximum confidence reached 100% and minimum confidence decreased to 50.0%. When the feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence changed to 0.47%, 9.3%, 100%, and 50.0%. The highest confidence (100%) and support (4.19%) occurred when the frequency of oxygen was > 0.12 and count of hydrogen was < 3652 or N-terminal was Met or both together. When the feature selection filtering was applied the highest confidence and support were 100% and 3.72 when the frequency of Gly - Gly was > 0.022 and the frequency of Ala-Lys < 0.0001 or the frequency of Gly was higher than 0.008. When the halolysin protein was compared with other individual protein classes (H/B/F/P/T groups), a GRI node analysis created 100 rules with 215 valid transactions with minimum and maximum support of 14.88% and 17.21%, respectively. Maximum confidence reached 100% and minimum confidence decreased to 97.22%. When the feature selection was used, minimum support, maximum support, maximum confidence, and minimum confidence changed to 14.88%, 17.67%, 100%, and 97.74%. The highest confidence (100%) and support (16.28%) in both methods (with/without feature selection filtering) occurred when the count of Lys was lower than 28.5, the frequency of Gly-Pro was greater than 0.002, and the frequency of Asp-Leu was less than 0.006 (table 3).

Statistical analyses showed significant differences ($p < 0.01$) in positively and negatively charged amino acids between halolysins and other proteins. Halolysin proteins had higher average of negatively charged amino acids comparing to other proteins. Asp and Glu, two negatively charged amino acids with average of 0.091 and 0.053, showed higher average comparing to other amino acids in halolysin proteins. More than 20% of amino acids in halolysin proteins were negatively charged comparing with just 9% in other proteins; resulting in at least two times more negatively charged amino acids presence in halophilic proteins. The Ratio of negatively charged amino acids to positively charged amino acids in halolysin and other proteins were 3 and 1.3 times.

A significant difference ($p < 0.01$) was found in 21 features of primary protein structure in halolysins and plant proteases. Positively charged amino acids (such as Lys, Arg and His) showed higher frequencies in plant proteases. A highly significant difference ($p < 0.0001$) was found in hydrophobic amino acids (Val, Pro, Phe, Ile, Leu and Met) of plant proteases and halolysin proteins resulting in the same significant differences of hydrophobic compounds in those proteins. Cys and Met, as N-terminal amino acids, were found to be more frequent in plant proteases than halolysins and other proteases studied in this paper forming more di-sulphid bonds in plant proteases. In halolysins, about 50% of Cys were in the N-terminal position while just 20% of the N-terminal amino acid in plant proteases was Cys.

A significant difference ($p < 0.05$) was found in aliphatic index in plant proteases and halolysin proteins, which could be due to aliphatic amino acids (Ile, Val, Pro, Met and Leu). More beta-strand was found in plant proteases which could be due to higher number of Lys, His and Cys. The frequency of Pro in plant proteases was higher than its frequency in halolysin proteins ($14.11 \pm 9.54$ and $13.78 \pm 1.09$, (mean ± SD), respectively). Some dipeptid bonds (such as Met-Met, Met-Cys and Cys-Cys) were more frequent in plant proteases and they could contribute in more beta-strand formation.

## Discussion

Salt dependence and salt tolerance microorganisms are newly discovered microorganisms, classified as new taxa with new names within the microbial taxonomy. Some use the term for all organisms that require some level of salt for growth, including concentrations around 35 g/l as found in seawater. Halobacterium species are obligatory halophilic microorganisms that have been adapted to optimal growth under conditions of extremely high salinity. They contain a correspondingly high concentration of salts internally and exhibit a variety of unusual and unique molecular characteristics. Since their discovery, extreme halophiles have been studied extensively by chemists, biochemists, microbiologists, and molecular biologists to define both molecular diversity and universal features of life. A notable list of early research milestones on halophiles includes the discovery of a cell envelope composed of an S-layer glycoprotein, Archaea l either lipids and purple membrane, and metabolic and biosynthetic processes operating at saturating salinities. These early discoveries established the value of investigations directed at extremophiles and set the stage for pioneering phylogenetic studies leading to the three-domain view of life and

**Table 3:** The association rules found in the data by the generalized rule induction (GRI) method, comparing halolysin proteins with the others (T/F groups)

| Antecedent | Confidence % |
|---|---|
| Freq. of Gly - Gly > 0.022 and Freq. of Ala -Lys < 0.000 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. of Gly > 0.122 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. of Phe < 0.018 and Freq. of Gly > 0.122 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. of Glu > 0.048 and Ile < 18.500 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. of Asp > 0.083 and Ile < 18.500 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Phe < 9.500 and Freq. of Gly > 0.122 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. Positively Charged < 0.048 and Ile < 18.500 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. Negatively Charged > 0.134 and Ile < 18.500 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Positively Charged residues < 25.500 and Freq. of Gly > 0.122 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Isoelectric point < 4.480 and Ile < 18.500 | 100.0 |
| Freq. of Gly > 0.122 and Isoelectric point < 4.480 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Phe < 9.500 and Freq. sulphur < 0.002 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. Positively Charged < 0.048 and Freq. Positively Charged > 0.046 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. Negatively Charged > 0.134 and Isoelectric point > 4.385 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Positively Charged residues < 25.500 and Freq. sulphur < 0.002 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Freq. sulphur < 0.002 and Isoelectric point < 4.480 | 100.0 |
| Freq. of Glu-Leu > 0.008 and Isoelectric point < 4.480 and Isoelectric point > 4.385 | 100.0 |
| Freq. Negatively Charged > 0.198 | 100.0 |
| Isoelectric point < 4.170 and Isoelectric point > 4.040 | 100.0 |
| Freq. of Gly - Gly > 0.022 and Isoelectric point < 4.480 | 88.89 |
| Freq. of Glu-Leu > 0.008 and Ile < 18.500 | 88.89 |
| Freq. of Glu-Leu > 0.008 and Phe < 9.500 and Isoelectric point > 4.385 | 87.5 |
| Freq. of Glu-Leu > 0.008 and Freq. Positively Charged < 0.048 and Isoelectric point > 4.385 | 87.5 |
| Freq. of Glu-Leu > 0.008 and Positively Charged residues < 25.500 and Isoelectric point > 4.385 | 87.5 |
| Freq. of Glu-Leu > 0.008 and Isoelectric point < 4.480 | 72.73 |
| Freq. of Gly - Gly > 0.022 | 66.67 |
| Isoelectric point < 4.480 | 60.0 |
| Freq. of Gly > 0.122 | 57.14 |
| Freq. Negatively Charged > 0.134 and Isoelectric point < 4.480 | 55.56 |
| Freq. of Gly - Gly > 0.016 | 55.0 |
| Freq. of Ala -Lys < 0.000 and Isoelectric point < 5.265 | 55.0 |
| Freq. of Asp > 0.078 and Positively Charged residues < 32.500 | 55.0 |
| Lysine < 9.500 and Isoelectric point < 4.480 | 52.94 |
| Freq. of Asp-Gln > 0.004 and Isoelectric point < 5.025 | 52.63 |
| Freq. of lie-Arg < 0.000 and Isoelectric point < 5.065 | 50.0 |
| Freq. of Asp-Gln > 0.004 and Freq. of Ile < 0.042 | 50.0 |
| Freq. of Asp-Gln > 0.004 and Freq. of Gly > 0.106 and Freq. Positively Charged < 0.062 | 50.0 |
| Freq. of Asp-Gln > 0.004 and Freq. of Glu > 0.042 and Isoelectric point < 5.025 | 50.0 |
| Freq. of Asp-Gln > 0.004 and Phe < 9.500 and Isoelectric point < 5.445 | 50.0 |
| Freq. of Asp-Gln > 0.004 and Positively Charged residues < 32.500 and Freq. of Asp > 0.058 | 50.0 |
| Freq. of Ala -Lys < 0.000 and Freq. Negatively Charged > 0.114 | 50.0 |
| Freq. of Ile < 0.038 and Isoelectric point < 5.555 | 50.0 |
| Freq. of Phe < 0.018 and Isoelectric point < 5.610 | 50.0 |
| Freq. of Asp > 0.078 and Phe < 10.500 | 50.0 |
| Ile < 19.500 and Isoelectric point < 5.405 | 50.0 |
| Phe < 9.500 and Isoelectric point < 5.265 | 50.0 |
| Freq. of Ala -Asp > 0.010 | 50.0 |
| Freq. of Ala - Cys > 0.002 and Isoelectric point < 4.760 | 50.0 |
| Freq. of Lysine < 0.018 and Isoelectric point < 5.155 | 50.0 |
| Freq. of Asp > 0.078 and Isoelectric point > 4.385 | 50.0 |
| Lysine < 9.500 and Freq. of Gly > 0.114 | 50.0 |
| Lysine < 9.500 and Freq. of Phe < 0.018 and Positively Charged residues < 28.500 | 50.0 |
| Lysine < 9.500 and Lysine > 5.500 and Positively Charged residues < 25.500 | 50.0 |
| Lysine < 9.500 and Phe < 9.500 and Freq. Positively Charged < 0.052 | 50.0 |
| Lysine < 9.500 and Positively Charged residues < 25.500 and Isoelectric point < 4.760 | 50.0 |
| Freq. Positively Charged < 0.048 and Isoelectric point < 4.480 | 50.0 |
| Freq. Negatively Charged > 0.134 and Freq. Positively Charged < 0.050 | 50.0 |
| Positively Charged residues < 25.500 and Isoelectric point < 4.760 | 50.0 |
| Isoelectric point < 4.480 and Isoelectric point > 4.285 | 50.0 |
| Isoelectric point < 4.170 | 50.0 |

classification of Halobacterium as a member of the Archaea l domain. It has been shown that some proteins and enzymes are responsible for living organism's tolerance against hypersaline conditions; therefore defining features contribute to this valuable characteristics of proteins paves roads toward engineering new strains of plants growing in harsh salty conditions. To date, some studies have looked at phylogeny, taxonomy and nomenclature of halophilic strains and various models have been employed to determine the most important features that contribute to these organisms' ability to stand hypersalinity media. In this study, we applied different modeling techniques to study more than 70 features of some halophilic proteins and compared them with similar proteases and termitase (found after multiple alignments) in an attempt to understand their ability to withstand salty conditions. We used different screening, clustering, and decision tree modeling on two datasets: one with and the other without feature selection filtering.

The phylogenic tree (figure 1) showed that halophilic organisms can be placed in a separate phylum between eukaryotes and bacteria, *Archea*, which is in line with previous studies (Pruess et al., 2003; Li et al., 2008; Wimmer et al., 2008). Although the results of feature selection modeling showed that 83 features (from 252) had a value greater than 0.95, the frequency of Gly - Gly ranked as the most important feature (table 1), and it was used in some decision tree models to create the main subgroups and branches. The number of peer group (one group) did not change when feature selection filtering was applied but anomaly index cutoff decreased from 2.42 (without feature selection) to 1.92 (with feature selection) showing the positive effects of feature selection filtering on removing outliers. Although the number of records in the clusters changed between the models with and without feature selection, the number of clusters generated by K-Means modeling did not. In the TwoStep model, the number of clusters decreased from three (without feature selection) to just two (with feature selection) groups.

The depth of trees generated by the various decision tree models varied from 1 (in the C&RT model with T/F comparison, with/without the feature selection dataset) to 5 (in the C5.0 model with 10-fold cross-validation on H/B/F/P/T groups) branches. The best cross-validation results were obtained in the C5.0 model when H/B/F/P/T groups compared. The protein features were used by various decision tree models to create trees varied from the count of sulphur (in the C5.0, C5.0 with 10-fold cross-validation and C&RT model on H/B/F/P/T groups) to the frequency of Gly - Gly (in the C&RT T/F and Quest models) and the count of Cys, Leu-Ala and hydrophobic residues in Quest, T/F CHAID and H/BFPT CHAID. In most GRI association rules (100 rules), the frequency of Gly - Gly was used as an antecedent to support the rules. Although previous studies have shown the importance of acidic amino acids (Glu and Asp) residues (Lanyi, 1969; Lanyi, 1974) and Gly (Lai, Hong et al. 2000; Robert, Le Marrec et al. 2000) in halophilic proteins, in this study, for the first time, we looked not only at individual amino acid composition, but also the importance of dipeptid amino acid composition in salt stability of these proteins and found Gly - Gly as the most important feature contributes to halotolerant capacity of these proteins. Performance evaluations in the decision tree models tested were found to be the same in all models. No significant differences in the percent of correctness, performance evaluation, and mean correctness of various decision tree models were found when feature selected datasets were used, but when feature selection datasets were used the number of peer-groups in clustering models reduced significantly.

Charged amino acids prevent charged ions from attaching to proteins and they have a significant role in stabilizing protein against salty conditions, and keep water molecules around these components. Sequence comparisons showed that, in general, the halophilic proteins contain an excess of negatively charged amino acids over positively charged amino acids, and the number of negatively charged amino acid residues is higher than that in their non-halophilic homologs (Kushner and Onishi, 1966; Rao and Argos, 1981; Tokunaga et al., 2008). The additional negative charges are located mostly on the protein surface, presumably helping to stabilize the protein molecule by competing with the salt for hydration (Lanyi, 1974). It has also been proposed that hydrophobic interactions play an important role in the ability of these proteins to cope with the salt stress in a hypersaline environment (Mevarech et al., 2000; Kastritis et al., 2007; Memmi et al., 2008). It has been shown that negatively charged amino acids such as Asp and Glu may contribute to protein ability to resist salty conditions; as shown in a higher percentage of negatively charged amino acid residues (18.5%) in halophilic strains than its non-halophilic counterparts (Pieper et al., 1998). Our finding were in line with the previous studies showing higher average of negatively charged amino acids in halolysin proteins with highly significant difference ($p < 0.001$) comparing to other proteins. It has been shown the cumulative

amount of Lys and Arg amino acids and even the content of Val were remarkably high in salt stability Archaea (Ferrer et al., 1996). Higher hydrophobic amino acids found in plant proteins could be related to their function as inside proteins tending to aggregate as a sphere surrounded by water to increase their stability inside the cells and this may clarify more positively charged amino acid such as Lys, Arg and His found in plant proteases, although it have been mentioned that this feature may also contribute to salt stability in some organisms (White and Jacobs, 1990; Srimathi et al., 2007; Valery et al., 2008). The results showed that Met was the sole N-terminal amino acid in halolysin proteins whereas other amino acids such as Cys, Thr, Tyr, Try and Leu were also found at this position of other proteases and termitase. In similar studies, it have been shown the N-terminal sequence of halophilic species play important role in their resistance to salty conditions (Baker et al., 1992; Wakai et al., 1995; Ferrer et al., 1996; Ihara et al., 1997; Porciero et al., 2005). A significant difference ($p<0.05$) in aliphatic index was found between plant proteases and halolysin proteins which could be due to the presence of more aliphatic amino acids such as Ile, Val, Pro, Met and Leu in plant proteases and this difference or higher number of dipeptid bonds may be responsible for more beta-strands in plant proteases (Hose et al., 2001; Lahav et al., 2002; Mishra and Jha, 2009).

We analyzed the performance of different screening, clustering, and decision tree algorithms for discriminating halophilic and non-halophilic proteins. Our results showed that the amino acid composition can be used to discriminate between protein groups. We found that most of the mentioned algorithms can be used to discriminate between halophilic and non-halophilic proteins with accuracy in the range of 98–100 %. Our analysis detected no significant difference in performance between different methods used in this paper. Interestingly, all decision tree models had a similar accuracy (higher than 98 %), and no differences were observed between analysis with and without feature selection. The best performance and correctness results were obtained with C5.0 and CHAID algorithms. Thus, we suggest that these decision tree models can be used as an effective tool to discriminate halophilic and non-halophilic proteins.

## Acknowledgements

The authors greatly appreciate and acknowledge the support of Bioinformatics Research Groups, Green Research Center, Qom University, and the School of Agriculture at Shiraz University for supporting the project.

## References

1- Baker P. J., Britton K. L., Engel P. C., Farrants G. W., Lilley K. S., Rice D. W. and Stillman T. J. (1992) Subunit assembly and active site locatin in the structure of glutamate dehydrogenase. Proteins Structure Function and Genetic 12: 75-86.

2- Bijanzadeh E., Emam Y. and Ebrahimie E. (2010) Determining the most important features contributing to wheat grain yield using supervised feature selection model. Australian Journal of crop science 4: 402-407.

3- Binbuga B., Boroujerdi A. F. and Young J. K. (2007) Structure in an extreme environment: NMR at high salt. Protein Sci 16: 1783-1787.

4- Conover R. K. and Doolittle W. F. (1990) Characterization of a gene involved in histidine biosynthesis in Halobacterium (Haloferax) volcanii: isolation and rapid mapping by transformation of an auxotroph with cosmid DNA. J Bacteriol 172: 3244-3249.

5- Cui H. L., Yang Y., Dilbr T., Zhou P. J. and Liu S. J. (2006) Biodiversity of halophilic archaea isolated from two salt lakes in Xin-Jiang region of China. Wei Sheng Wu Xue Bao 46: 171-176.

6- De Castro R. E., Ruiz D. M., Gimenez M. I., Silveyra M. X., Paggi R. A. and Maupin-Furlow J. A. (2008) Gene cloning and heterologous synthesis of a haloalkaliphilic extracellular protease of Natrialba magadii (Nep). Extremophiles 12: 677-687.

7- Ebrahimi M. and Ebrahimie E. (2010) Sequence-based prediction of enzyme thermostability through bioinformatics algorithms. Current Bioinformatics 5: 195-203.

8- Ebrahimi M., Ebrahimie E. and Ebrahimi M. (2009) Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree alogorithms. EXCLI Journal 8: 218-233.

9- Elevi Bardavid R. and Oren A. (2008) Sensitivity of Haloquadratum and Salinibacter to antibiotics and other inhibitors: implications for the assessment of the contribution of Archaea and Bacteria to heterotrophic activities in hypersaline environments. FEMS Microbiol Ecol 63: 309-315.

10- Feng D. and Yang S. (2008) Current status on proteomics of extremophilic microorganisms--a review. Wei Sheng Wu Xue Bao 48: 1675-1680.

11- Ferrer J., Perez-Pomares F. and Bonete M. J. (1996) NADP-glutamate dehydrogenase from the halophilic archaeon *Haloferax mediterranei*: enzyme purification, N-terminal sequence and stability. FEMS Microbilogy Letter 141: 59-63.

12- Fine A., Irihimovitch V., Dahan I., Konrad Z. and Eichler J. (2006) Cloning, expression, and purification of functional Sec11a and Sec11b, type I signal peptidases of the archaeon Haloferax volcanii. J

Bacteriol 188: 1911-1919.
13- Gonzalez-Hernandez J. C. and Pena A. (2002) Adaptation strategies of halophilic microorganisms and Debaryomyces hansenii (halophilic yeast). Rev Latinoam Microbiol 44: 137-156.
14- Hose E., Clarkson D. T., Steudle E., Schreiber L. and Hartung W. (2001) The exodermis: a variable apoplastic barrier. J Exp Bot 52: 2245-2264.
15- Ihara K., Watanabe S., Sugimura K., Katagiri I. and Mukohata Y. (1997) Identification of proteolipid from an extremely halophilic archaeon Halobacterium salinarum as an N,N'-dicyclohexyl-carbodiimide binding subunit of ATP synthase. Arch Biochem Biophys 341: 267-272.
16- Ingoldsby L. M., Geoghegan K. F., Hayden B. M. and Engel P. C. (2005) The discovery of four distinct glutamate dehydrogenase genes in a strain of Halobacterium salinarum. Gene 349: 237-244.
17- Joo W. A. and Kim C. W. (2005) Proteomics of Halophilic archaea. J Chromatogr B Analyt Technol Biomed Life Sci 815: 237-250.
18- Kamekura M. (1998) Diversity of extremely halophilic bacteria. Extremophiles 2: 289-295.
19- Kamekura M. and Seno Y. (1993) Partial sequence of the gene for a serine protease from a halophilic archaeum Haloferax mediterranei R4, and nucleotide sequences of 16S rRNA encoding genes from several halophilic archaea. Experientia 49: 503-513.
20- Kamekura M., Seno Y. and Dyall-Smith M. (1996) Halolysin R4, a serine proteinase from the halophilic archaeon Haloferax mediterranei; gene cloning, expression and structural studies. Biochim Biophys Acta 1294: 159-167.
21- Kamekura M., Seno Y., Holmes M. L. and Dyall-Smith M. L. (1992) Molecular cloning and sequencing of the gene for a halophilic alkaline serine protease (halolysin) from an unidentified halophilic archaea strain (172P1) and expression of the gene in Haloferax volcanii. J Bacteriol 174: 736-742.
22- Kastritis P. L., Papandreou N. C. and Hamodrakas S. J. (2007) Haloadaptation: insights from comparative modeling studies of halophilic archaeal DHFRs. Int J Biol Macromol 41: 447-453.
23- Kristjansson H., Sadler M. H. and Hochstein L. I. (1986) Halobacterial adenosine triphosphatases and the adenosine triphosphatase from Halobacterium saccharovorum. FEMS Microbiol Rev 39: 151-157.
24- Kushner D. J. and Onishi H. (1966) Contribution of protein and lipid components to the salt response of envelopes of an extremely halophilic bacterium. J Bacteriol 91: 653-660.
25- Lahav R., Fareleira P., Nejidat A. and Abeliovich A. (2002) The identification and characterization of osmotolerant yeast isolates from chemical wastewater evaporation ponds. Microb Ecol 43: 388-396.
26- Lai M. C., Hong T. Y. and Gunsalus R. P. (2000) Glycine betaine transport in the obligate halophilic archaeon Methanohalophilus portucalensis. J Bacteriol 182: 5020-5024.
27- Lanyi J. K. (1969) Studies of the electron transport chain of extremely halophilic bacteria. II. Salt dependence of reduced diphosphopyridine nucleotide oxidase. J Biol Chem 244: 2864-2869.
28- Lanyi J. K. (1974) Salt-dependent properties of proteins from extremely halophilic bacteria. Bacteriol Rev 38: 272-290.
29- Li T., Wang P. and Wang P. (2008) [Bacterial and archaeal diversity in surface sediment from the south slope of the South China Sea]. Wei Sheng Wu Xue Bao 48: 323-329.
30- Madern D., Pfister C. and Zaccai G. (1995) Mutation at a single acidic amino acid enhances the halophilic behaviour of malate dehydrogenase from Haloarcula marismortui in physiological salts. Eur J Biochem 230: 1088-1095.
31- Memmi S., Kyndt J., Meyer T., Devreese B., Cusanovich M. and Van Beeumen J. (2008) Photoactive yellow protein from the halophilic bacterium Salinibacter ruber. Biochemistry 47: 2014-2024.
32- Mevarech M., Frolow F. and Gloss L. M. (2000) Halophilic enzymes: proteins with a grain of salt. Biophys Chem 86: 155-164.
33- Mishra A. and Jha B. (2009) Isolation and characterization of extracellular polymeric substances from micro-algae Dunaliellasalina under salt stress. Bioresour Technol 100: 3382-3386.
34- Mukohata Y., Ihara K., Tamura T. and Sugiyama Y. (1999) Halobacterial rhodopsins. J Biochem 125: 649-657.
35- Oren A. (1994) Enzyme diversity in halophilic archaea. Microbiologia 10: 217-228.
36- Pesenti P. T., Sikaroodi M., Gillevet P. M., Sanchez-Porro C., Ventosa A. and Litchfield C. D. (2008) Halorubrum californiense sp. nov., an extreme archaeal halophile isolated from a crystallizer pond at a solar salt plant in California, USA. Int J Syst Evol Microbiol 58: 2710-2715.
37- Pieper U., Kapadia G., Mevarech M. and Herzberg O. (1998) Structural features of halophilicity derived from the crystal structure of dihydrofolate reductase from the Dead Sea halophilic archaeon, Haloferax volcanii. Structure 6: 75-88.
38- Porciero S., Receveur-Brechot V., Mori K., Franzetti B. and Roussel A. (2005) Expression, purification, crystallization and preliminary crystallographic analysis of a deblocking aminopeptidase from Pyrococcus horikoshii. Acta Crystallogr Sect F Struct Biol Cryst Commun 61: 239-242.
39- Pruess M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E., Mittard V., Mulder N., Phan I., Servant F. and Apweiler R. (2003) The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res 31: 414-417.
40- Rao J. K. and Argos P. (1981) Structural stability of halophilic proteins. Biochemistry 20: 6536-6543.
41- Robert H., Le Marrec C., Blanco C. and Jebbar M. (2000) Glycine betaine, carnitine, and choline enhance salinity tolerance and prevent the accumulation of sodium to a level inhibiting growth of Tetragenococcus halophila. Appl Environ Microbiol

66: 509-517.
42- Soppa J. (2006) From genomes to function: haloarchaea as model organisms. Microbiology 152: 585-590.
43- Srimathi S., Jayaraman G., Feller G., Danielsson B. and Narayanan P. R. (2007) Intrinsic halotolerance of the psychrophilic alpha-amylase from Pseudoalteromonas haloplanktis. Extremophiles 11: 505-515.
44- Strahl H. and Greie J. C. (2008) The extremely halophilic archaeon Halobacterium salinarum R1 responds to potassium limitation by expression of the K+-transporting KdpFABC P-type ATPase and by a decrease in intracellular K+. Extremophiles 12: 741-752.
45- Sumper M. (1987) Halobacterial glycoprotein biosynthesis. Biochim Biophys Acta 906: 69-79.
46- Tokunaga H., Arakawa T. and Tokunaga M. (2008) Engineering of halophilic enzymes: two acidic amino acid residues at the carboxy-terminal region confer halophilic characteristics to Halomonas and Pseudomonas nucleoside diphosphate kinases. Protein Sci 17: 1603-1610.
47- Valery C., Pouget E., Pandit A., Verbavatz J. M., Bordes L., Boisde I., Cherif-Cheikh R., Artzner F. and Paternostre M. (2008) Molecular origin of the self-assembly of lanreotide into nanotubes: a mutational approach. Biophys J 94: 1782-1795.
48- Wakai H., Takada K., Nakamura S. and Horikoshi K. (1995) Structure and heterologous expression of the gene encoding the cell surface glycoprotein from Haloarcula japonica strain TR-1. Nucleic Acids Symp Ser: 101-102.
49- White S. H. and Jacobs R. E. (1990) Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution. Biophys J 57: 911-921.
50- Wimmer F., Oberwinkler T., Bisle B., Tittor J. and Oesterhelt D. (2008) Identification of the arginine/ornithine antiporter ArcD from Halobacterium salinarum. FEBS Lett 582: 3771-3775.
51- Yang Y., Cui H. L., Zhou P. J. and Liu S. J. (2006) Halobacterium jilantaiense sp. nov., a halophilic archaeon isolated from a saline lake in Inner Mongolia, China. Int J Syst Evol Microbiol 56: 2353-2355.
52- Zaccai G., Wachtel E. and Eisenberg H. (1986) Solution structure of halophilic malate dehydrogenase from small-angle neutron and X-ray scattering and ultracentrifugation. J Mol Biol 190: 97-106.
53- Zhu D., Cui S. and Nagata S. (2008) Isolation and characterization of salt-sensitive mutants of the moderately halophilic bacterium Salinivibrio costicola subsp. yaniae. Biosci Biotechnol Biochem 72: 1977-1982.