Understanding the Performance of Protein Stability Predictors

TEAM MEMBERS

1) Vikrant Kaushal: vkaushal@umail.iu.edu

2) Kedar Gundlupet: kgundlup@umail.iu.edu

3) Srikanth Srinivas Holavanahalli: sriksrin@umail.iu.edu

Under guidance of:

Jose Lugo-Martinez: jlugomar@indiana.edu

OBJECTIVE

Understanding the effects of single point mutations on protein stability is an important problem in bioinformatics. Toward this goal, we shall perform an in-depth analysis on a representative set of established sequence- and structure-based approaches for the problem of protein stability prediction. In particular, we are interested in replicating and exploring the conclusions from the research article “Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines” by J. Cheng et al.[2] (2006) which states that sequence information alone can be used to accurately predict protein stability for single-site mutations, even in the presence of additional protein structural information. We believe that a detailed analysis on the features and performance evaluation scenario will shed light on the unexpected results that structural information has no significant benefits on predicting protein stability. We expect that the various evaluation settings will help us identify the reasons underlying Cheng et al.’s results. Additionally, we are interested in identifying potential biases on the original data set.

Our motivation for this project is two-fold: (1) work on a relevant and challenging problem, and (2) apply the concepts and methods learned in the data mining course on a real-world data set. Furthermore, new insights on the problem of protein stability prediction should improve our understanding of the interplay between genetic variants and disease.

BACKGROUND

Mutation is a process where the DNA sequence is altered, and this change can occur in a number of ways. In this project, we are only interested in missense variants which occur when a DNA base of a gene is changed such that it leads to a single amino acid substitution in the gene sequence. These single amino acid mutations can have varying effects on the protein structure or functionality.

In this project, we dealt with prediction of protein stability changes resulting from single amino acid mutation. There are four approaches to predict the protein stability: (1) physical potential approach, (2) statistical potential approach, (3) empirical potential approach, and (4) machine learning approach. Some of these approaches require 3D structural information of the protein to make accurate predictions of protein stability. However, previous work has shown that it is possible to predict the protein stability using sequence information without the knowledge of protein's structural information. For instance, E. Capriotti et al. [1] demonstrated the prediction of protein stability based on sequence information using artificial neural networks. Similarly, J. Cheng et al. [2] predicted protein stability by using sequence composition with the help of Support Vector Machines (SVM). Therefore, it is feasible to predict protein's stability in the absence of a protein 3D structure. Moreover, J. Cheng et al. [2] showed that predicting protein stability based on sequence composition alone performs as good as methods which rely on the tertiary structural information. In this work, we would like to systematically characterize the performance of a set of sequence-based and structure-based protein stability predictors.

METHODS

a) **Dataset**

We used S1615 dataset compiled by Capriotti et al. [1]. Below are the important attributes of the data.

- *PDB ID*: Protein Data Bank ID is unique ID given to each unique protein structure.
- *Chain ID*: A chain of amino acids joined together to form a protein.
- *Protein Sequence*: Connected sequence of amino acids in protein.
- *Position:* Position in sequence where single mutation has occurred.
- *Original Residue*: Amino acid which has been mutated
- *Substituted Residue*: Amino acid which substituted original residue.
- *P^H Value*: P^H value of the protein
- *Temperature*: Temperature of protein
- *Energy Change*: Positive indicates that mutation has stabilized he protein. Negative value indicates that mutation has destabilized the protein.

We removed duplicate records in S1615 dataset. To identify and remove duplicates, we considered following attributes: *protein ID*, *chain ID*, *sequence*, *position*, *original and substituted residue*, *temperature* and *energy change sign*. As a result, we identified 119 redundant records. After removing them, we got 1496 unique records.

In the case of the structure dataset, we collected the amino acid residues in the structural 3D neighborhood of the position of interest (i.e. atomic distance between neighboring residues was less than 4Å).

Data Set Analysis

We analyzed the data to determine the number of positive and negative classes for per protein ID and per chain. The results are tabulated in the following tables.

Protein ID + Chain ID	**No. Of Positive Records**	**No. Of Negative Records**
1BVCA	12	47
1IOBA	1	6
1VQBA	13	79
3SSIA	14	35
1HFYA	4	0
1BTAA	1	0
2LZMA	121	289
1ROPA	12	9
5CROO	10	1
1RN1B	1	0
1IGVA	3	0
4LYZA	17	26
1LZ1A	11	46
1MBGA	2	1
1CSPA	4	3
1TUPA	0	5
2ABDA	4	23
1C2RA	1	2
1G6NA	1	1
3MBPA	1	5
2CI2I	12	102
1ARRA	3	0
1PGAA	1	4

1RN1A	11	51
1SUPA	6	0
1POHA	4	1
1SARA	0	3
2RN2A	83	37
1RTBA	2	13
1STNA	7	37
1ONCA	0	9
1A23A	1	0
1CYOA	1	9
2AKYA	0	4
1AARA	5	4
1BPIA	2	47
2TRXA	0	2
1ANKA	0	4
1LRPA	7	5
1DDRA	2	11
1C9OA	5	10
1BNIA	16	164

Table1: Number of positive and negative records for each unique protein ID + chain ID combination

Observation: All the records of protein chain such as 1HFYA, 1BTAA, 1ONCA and so on belong to one of the two classes. Due to this, while classification, the records of same protein chain will be classified to the class which has non-zero number of records. For example, the protein chain 1ONCA will be always be classified as negative class.

Also, protein chain such as 1BNIA, 2CI2I, 2LZMA, among others have a major chunk of records. Hence, if any of these protein chains comes as part of test data, this will reduce the training records, which affects the accuracy of the model.

Protein ID	No. Of Positive Records	No. Of Negative Records
1IOB	1	6
1IGV	3	0

1MBG	2	1
1C9O	5	10
2ABD	4	23
1HFY	4	0
1VQB	13	79
1LZ1	11	46
1BTA	1	0
1ANK	0	4
3MBP	1	5
1CYO	1	9
2LZM	121	289
1STN	7	37
1ONC	0	9
1BNI	16	164
1A23	1	0
1CSP	4	3
2RN2	83	37
1RTB	2	13
1C2R	1	2
1RN1	12	51
1G6N	1	1
1DDR	2	11
1LRP	7	5
5CRO	10	1
1POH	4	1
2TRX	0	2
1BPI	2	47
4LYZ	17	26
1BVC	12	47
3SSI	14	35
1SUP	6	0
1TUP	0	5
1PGA	1	4

1ROP	12	9
1AAR	5	4
2AKY	0	4
2CI2	12	102
1ARR	3	0
1SAR	0	3

Table 2: **Number of positive and negative records for each unique protein ID combination**

Observation: All the records of protein IDs such as 1SUP, 1TUP, 1HFY etc belong one of the two classes. Due to this, while classification, the records of same protein ID will be classified to the class which has non-zero number of records. For example, the protein chain 1ONCA will be always be classified as negative class.

Also, protein ID such as 2LZM, 2CI2, 1BNI contain a majority of records. Hence, if any of these protein IDs comes as part of test data, this will reduce the training records, which affects the accuracy of the model.

METHODOLOGY

Sequence Features

For each of the 1496 unique records, by using mutation position, we took a window size of 7. That is, we considered 3 amino acids to left and right of the mutation position in the sequence. As we used SVM light[4] which expects the features to be present in the increasing order, we encoded each amino acid in the window from left to right (except for mutated amino acid) by adding the respective amino acids ordinal value to increasing multiples of 26. We obtained 6 features. The value of each of this features will be one. Then, we encoded original residue and substitute residue by using same approach and giving value –1 for original residue and +1 for substitute residue. Thus, we obtained 8 features. The class variable value was given as +1 for stabilize and –1 for destabilized mutation.

If the window size is less than 7 as there is a possibility that the mutation position can be in beginning or at the end of sequence, then we are considering features obtained from smaller window.

While fetching neighbors of the position where mutation occurred we were expecting only 20 different amino acids but by analyzing carefully we figured out that we were getting more than 20 amino acids. We analyzed that extra amino acid ‘X’ is present in 13 different dataset examples. As it is not relevant to data so we removed it.

Structure Features

The structures dataset has 1496 records. The neighbors count, original residue and substitute residue were encoded using same ordinal and multiples of 26 approach used for sequence features.

Sequence + Structure Features

We mapped 1496 records in structure records to 1496 records in sequence records. We then concatenated their features to form sequence + structure features.

Clustering

In order to cluster the records in sequence dataset based on per protein and per chain, we used CD-HIT[3] tool. As CD-HIT[3] expects the input in FATSA format, we converted the 1496 sequence records into per protein and per chain FATSA format files. The per protein FASTA file contains protein ID followed by its sequence data and per chain FATSA file has protein ID and chain ID followed by sequence data.

We used per protein and per chain FASTA files to cluster the files on similarity values of 40% to 100% with 10% intervals to observe the number of clusters per each cluster similarity value. The below charts visualize the results.

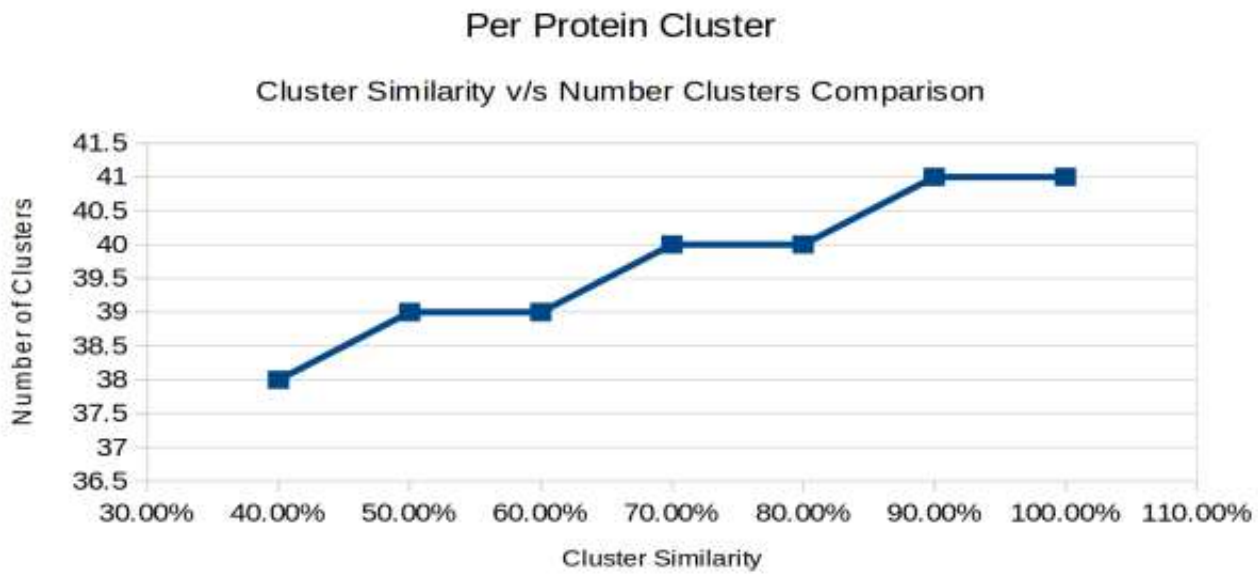

Fig 1: Per protein comparison of cluster similarity v/s no. of clusters

In clustering per protein, the number of clusters is directly proportional to number of clusters. The cluster size varies for similarity values of 40%, 50, 70% and 90%

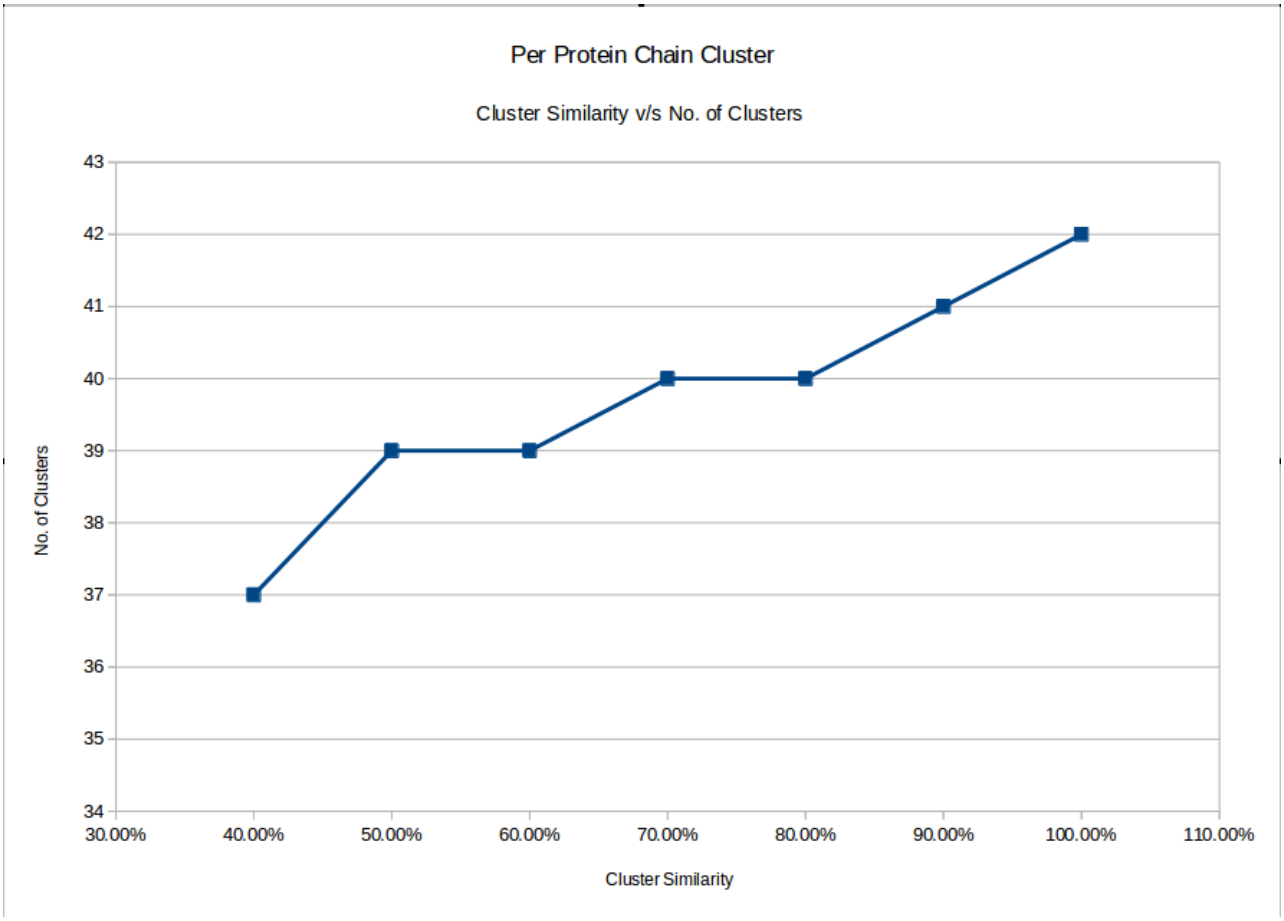

Fig 2: Per protein chain comparison of cluster similarity v/s no. of clusters

In clustering per protein chain, the number of clusters is directly proportional to number of clusters. The cluster size varies for similarity values of 40%, 50, 70%, 90% and 100%.

To perform the experiment as same setting as done by Capriotti et al., we determined the evaluation measures such as Precision, F1- measure, Recall, Accuracy and Correlation Coefficient, we performed 10-fold cross validation using SVM light[4] on sequence features, structure features, sequence + structure features and cluster results. The evaluation measures mentioned above were determined by calculating True Positive Rate(TPR), True Negative Rate(TNR), False Positive Rate(FPR) and False Negative Rate(FNR) by using output of SVM light[4] classify command. In order to perform this evaluation, we appended the class labels' value in SVM light's[4] input file using '#' symbol. The SVM light[4] classification output values will be of range [-∞, +∞]. We used log sigmoid function to convert it to posterior probability of range [0,1]. If the log sigmoid value is greater than 0.5 then it means SVM light[4] classified it as positive class else as negative class.

However, to make sure that all the proteins belonging to same protein ID, protein chain or cluster are either in training or test data, we created auxiliary file for sequence, structure and cluster data. We repeated the cross validation 5 times and took their average value to get closer to accurate values.

RESULTS

Structure

The below mentioned table is evaluated by just considering structure only of proteins for both per protein and per chain unique id.

	Per Protein	Per Chain
Precision	45.2621398281%	44.4061131718%
Recall	35.9%	36.15%
F1-Measure	40.0305484274%	39.8073754625%
Accuracy	71.1215580927%	71.403626595%
Correlation Coefficient	0.2160197868	70.6917394224%

ROC Curves

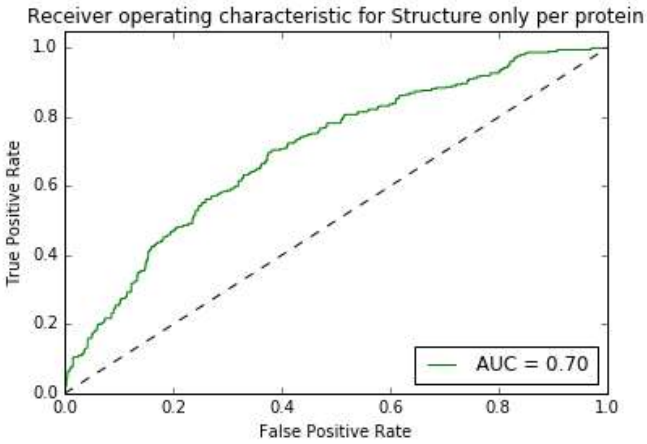

Fig 3: ROC Curve for Structure only Per Protein ID

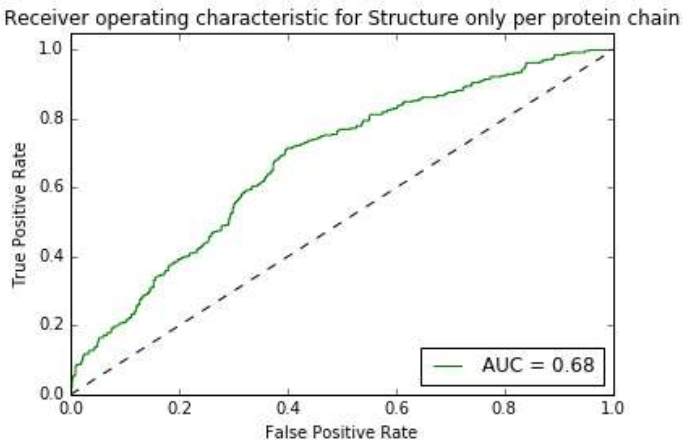

Fig 4: ROC Curve for Structure only Per Protein Chain

Sequence + Structure Results

The below mentioned table is evaluated by just considering sequence + structure of proteins for both per protein and per chain unique id.

	Per Protein	**Per Chain**
Precision	43.3600046147%	42.1100648592%
Recall	35.05%	33.3%
F1-Measure	38.7431578634%	37.1675045924%
Accuracy	70.2619207522%	69.805238415%
Correlation Coefficient	0.196087125293	0.178827590699

ROC Curves

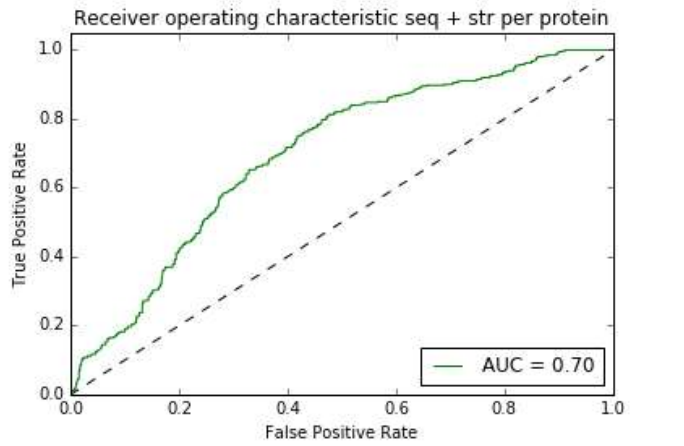

Fig 5: ROC Curve for Structure + Sequence Per Protein ID

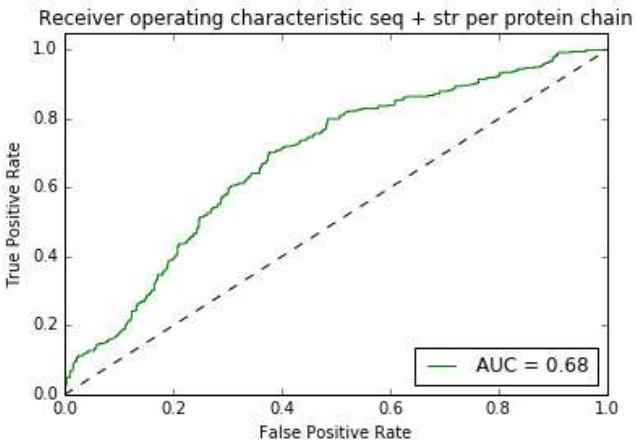

Fig 6: ROC Curve for Structure + Sequence Per Protein

Sequence only:

The below mentioned table is evaluated by just considering sequence of proteins for both per protein and per chain unique id.

	Per Protein	**Per Chain**
Precision	39.92381012978%	38.92258606822%
Recall	21.64588528678%	21.59600997506%
F1-Measure	28.04608502156%	27.75648381236%
Accuracy	70.21390374332%	69.8796791444%
Correlation Coefficient	0.1215693618887	0.114113438146

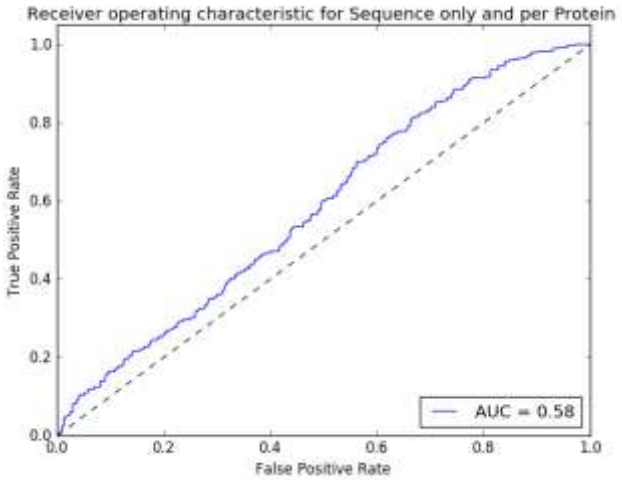

Fig 7: ROC curve for Sequence only, per Protein

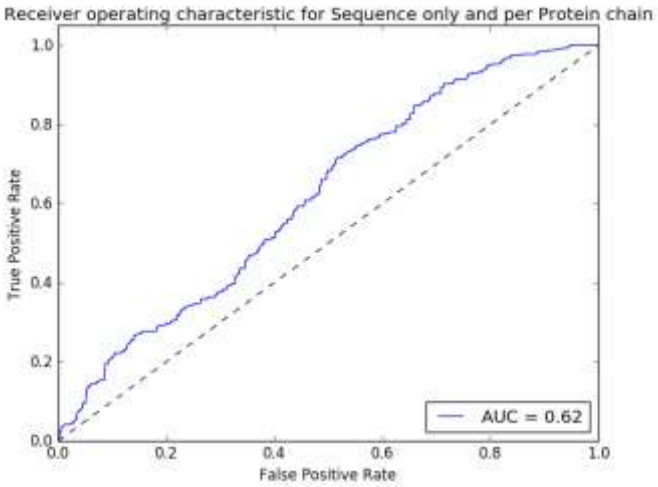

Fig 8: ROC curve for Sequence only, per Protein Chain

Clustering Results

Sequence Only Protein ID Results

The below table tabulates the results for sequence clustering per protein ID for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	**50% Similarity**	**70% Similarity**	**90% Similarity**
Precision	41.8632299361%	40.5007079048%	41.2343808559%	39.851866303%
Recall	21.2%	22.3%	22.5%	21.0%
F1 Measure	28.1360047162%	28.7152417388%	29.4338578382%	27.436190134%
Accuracy	70.9798657718%	70.3087248322%	71.0738255034%	70.268456375%
Correlation Coefficient	0.134791285966	0.127894589142	0.145062431633	0.118888880695

ROC Curves

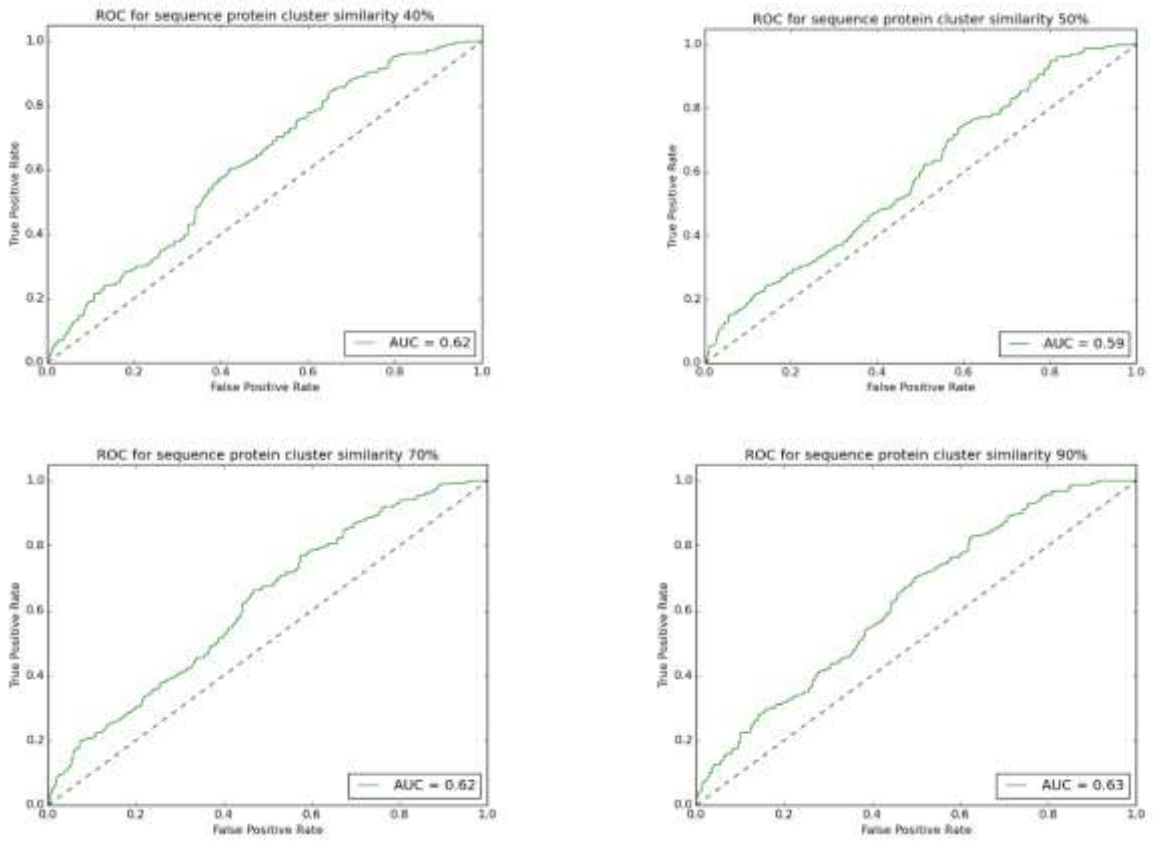

Fig 9: ROC Curves for sequence only per protein for cluster similarity values 40%, 50%, 70% and 90%

Structure Only Protein ID Results

The below table tabulates the results for structure clustering per protein ID for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	50% Similarity	70% Similarity	90% Similarity
Precision	39.9622909742%	42.3547528639%	39.0272485633%	39.754483428%
Recall	35.5610972569%	37.2568578554%	38.0548628429%	37.206982543%
F1 Measure	37.6172264098%	39.6333764891%	38.440145816%	38.3849716434%
Accuracy	68.322147651%	69.4630872483%	67.1812080537%	67.838926174%
Correlation Coefficient	0.165510655445	0.193865771667	0.161574215213	0.167204240095

ROC Curves

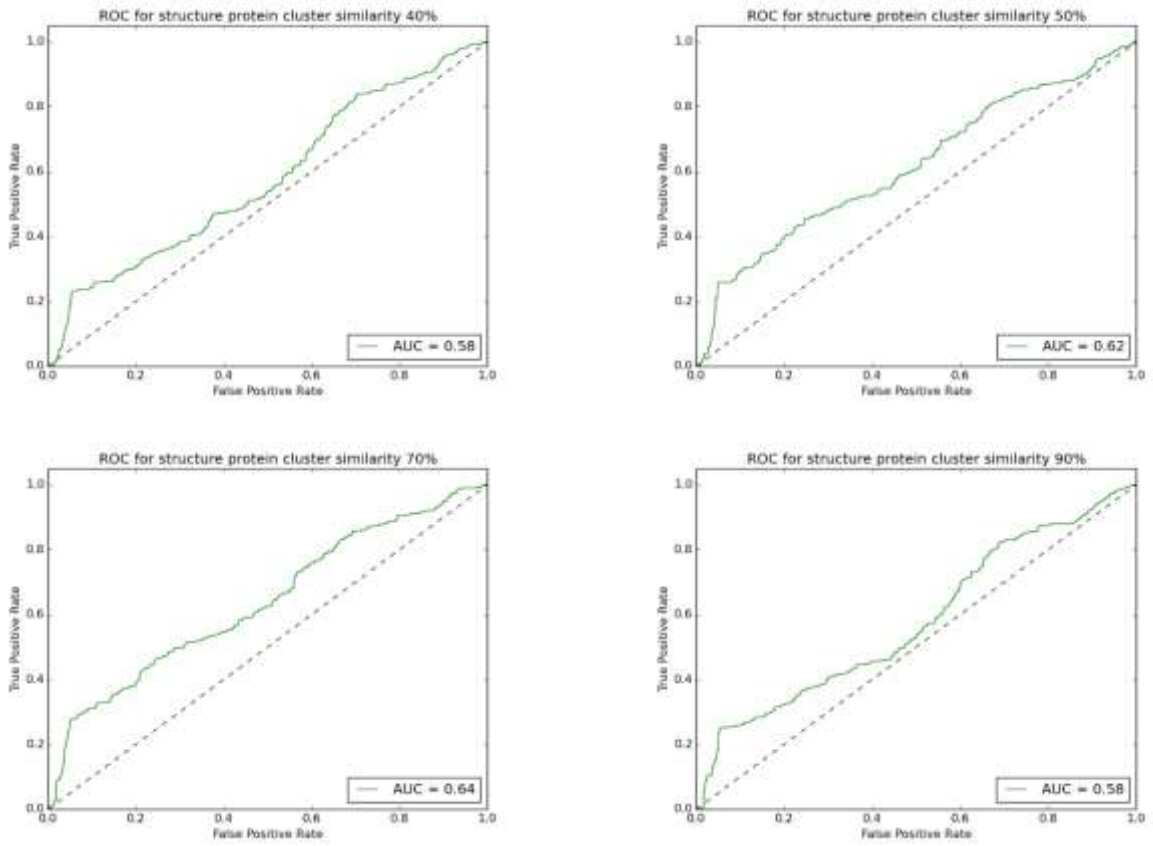

Fig 10: ROC Curves for structure only per protein for cluster similarity values 40%, 50%, 70% and 90%

Sequence + Structure Protein ID Results

The below table tabulates the results for sequence + structure clustering per protein ID for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	**50% Similarity**	**70% Similarity**	**90% Similarity**
Precision	39.7911095138%	41.2343808559%	43.6921938403%	40.7990903237%
Recall	42.0448877805%	43.7406483791%	45.5860349127%	43.391521197%
F1-Measure	40.858104875%	42.4277664028%	44.6160659348%	42.0538814369%
Accuracy	67.1946308725%	67.9865771812%	69.5436241611%	67.8255033557%
Correlation Coefficient	0.18225143934	0.203229944775	0.236388305838	0.198324993002

ROC Curves

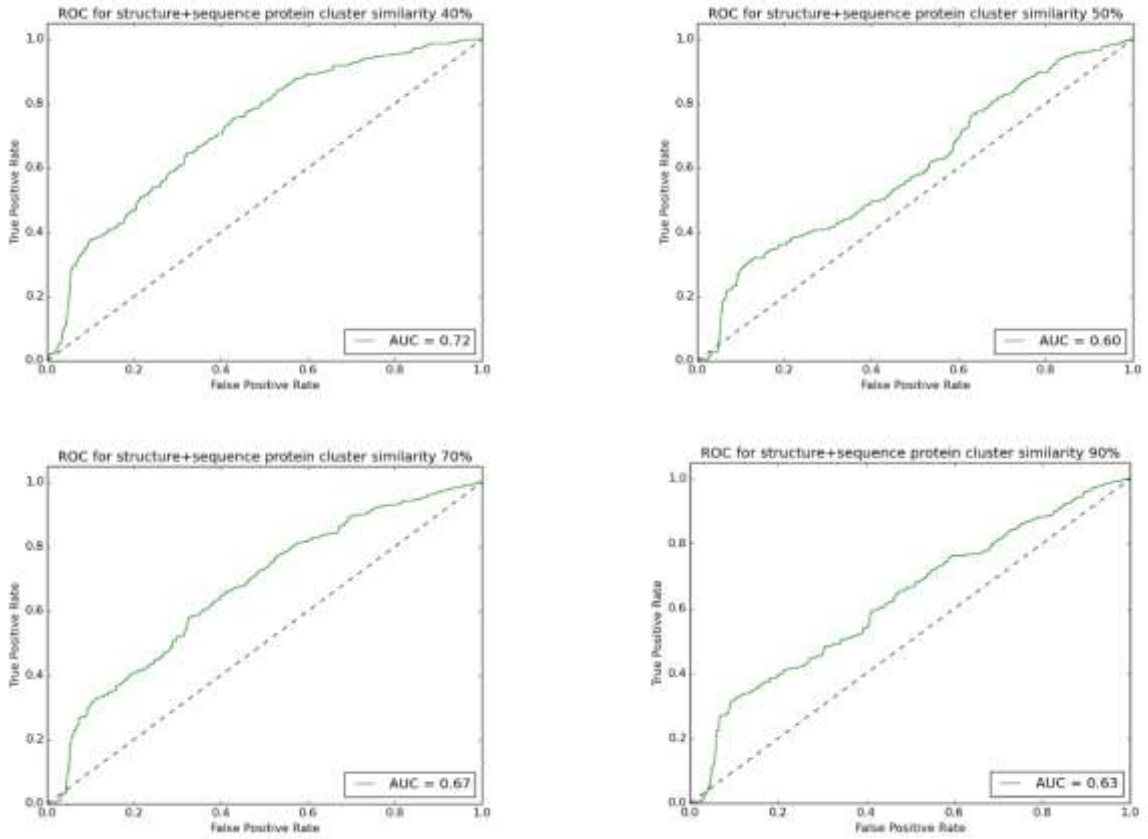

Fig 11: ROC Curves for structure + sequence only per protein for cluster similarity values 40%, 50%, 70% and 90%

Sequence Only Protein Chain Results

The below table tabulates the results for sequence clustering per protein chain for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	**50% Similarity**	**70% Similarity**	**90% Similarity**
Precision	41.8665540745%	44.2930976053%	40.4311517439%	40.186513396%
Recall	22.95%	23.45%	21.55%	21.4%
F1 Measure	29.6257063764%	30.6533612233%	28.0870400287%	27.913106965%
Accuracy	70.711409396%	71.5033557047%	70.4026845638%	70.348993288%
Correlation Coefficient	0.140269830215	0.159942613474	0.125099660809	0.122934150258

ROC Curves

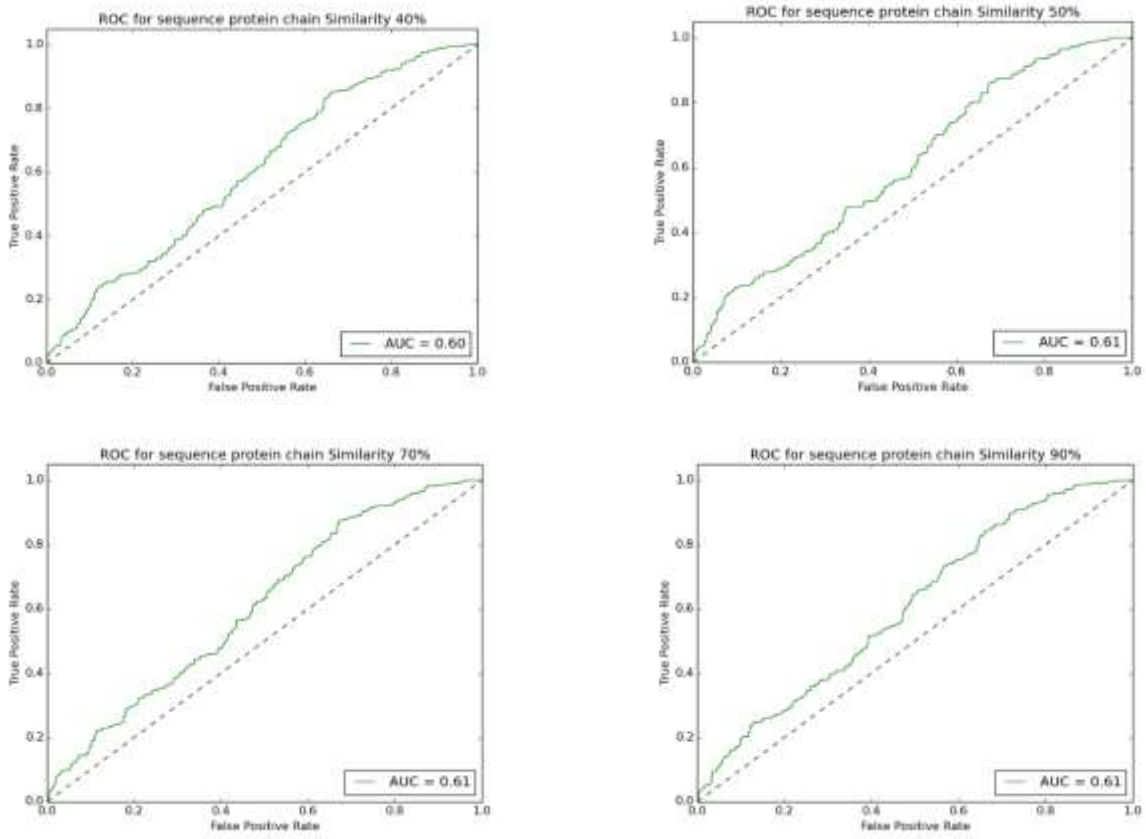

Fig 12: ROC Curves for sequence only per protein chain for cluster similarity values 40%, 50%, 70% and 90%

Structure Only Protein Chain Results

The below table tabulates the results for structure clustering per protein chain for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	**50% Similarity**	**70% Similarity**	**90% Similarity**
Precision	39.4900313061%	40.651987182%	40.152568375%	43.099203606%
Recall	37.0074812968%	37.4563591022%	37.4064837905%	37.805486284%
F1 Measure	38.1009982994%	38.9764175962%	38.6726315932%	40.271744850%
Accuracy	67.5570469799%	68.4697986577%	68.0402684564%	69.852348993%
Correlation Coefficient	0.162574095673	0.178061592142	0.171674838625	0.20309966319

ROC Curves

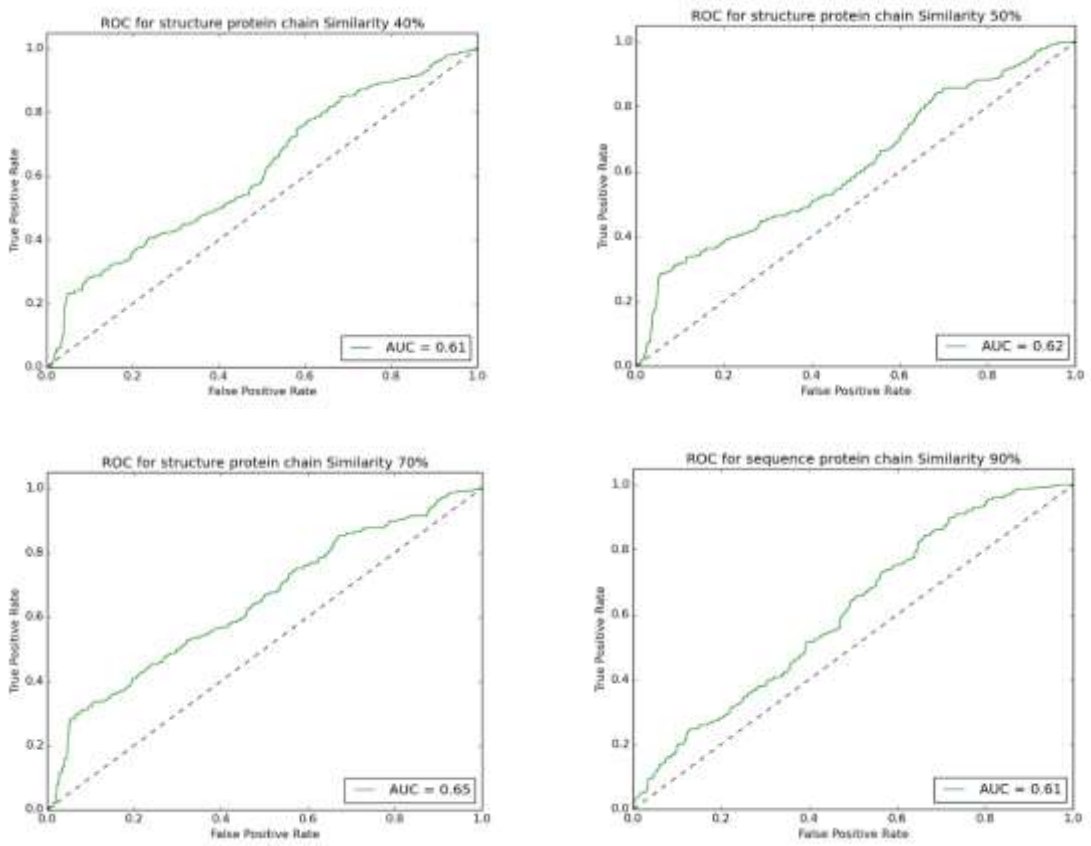

Fig 13: ROC Curves for structure only per protein chain for cluster similarity values 40%, 50%, 70% and 90%

Sequence + Structure Protein Chain Results

The below table tabulates the results for sequence + structure clustering per protein chain for cluster similarity values of 40%, 50%, 70% and 90%.

	40% Similarity	**50% Similarity**	**70% Similarity**	**90% Similarity**
Precision	41.7032628356%	42.2151449817%	39.3637467842%	42.126629953%
Recall	43.7905236908%	43.9401496259%	42.5935162095%	43.940149625%
F1 Measure	42.7129161667%	43.0578957186%	40.9004608985%	42.998829548%
Accuracy	68.3624161074%	68.7248322148%	66.8456375839%	68.604026845%
Correlation Coefficient	0.208979372513	0.215224331093	0.179517282873	0.213717809814

ROC Curves

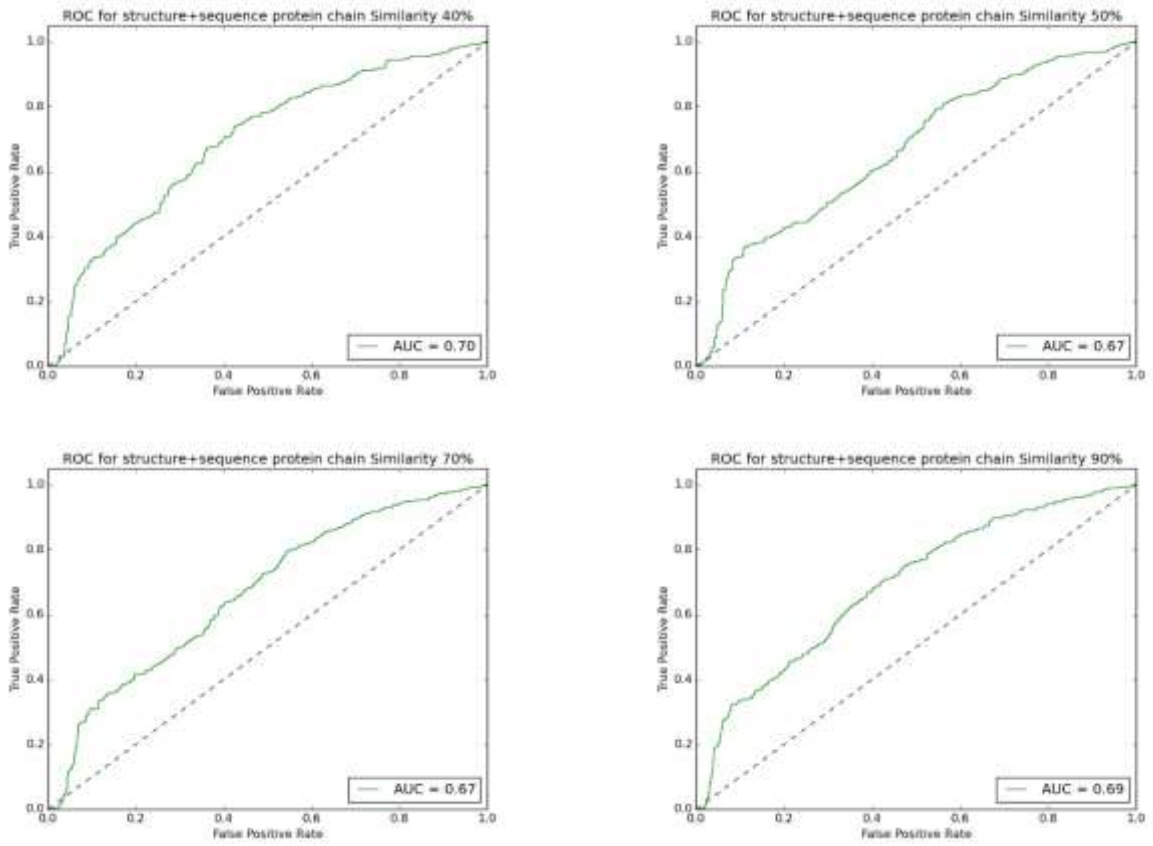

Fig 14: ROC Curves for structure+sequence per protein chain for cluster similarity values 40%, 50%, 70% and 90%

CONCLUSION

In this project, we provided a simple but effective in depth analysis of prediction of protein stability due to single point mutation. The analysis was carried out by considering train or test examples based on similarity criteria. Similarity criteria were protein, protein plus chain and cluster based.

The striking observation obtained from our project is that accuracies obtained by Cheng et al. are over estimated. The over estimation is because their incorrect division of the data into train and test examples. For e.g. 10 examples of a protein 1IOB are positive, if 1 positive example is in train set, other 9 test examples would be classified as positive(so 100% accuracy in 1 fold) , this would lead to over estimation.

The claim that sequence data is enough and structural data is not required for stability prediction is true. We have obtained consistent accuracies when dataset was analyzed with sequence information, structural information and sequence + structural information. So we confirm their claims.

Due to lack of time, we could not perform classification using neural network. As a future work, we would like to perform classification using neural network, logistic regression and random forests. Also, we would want to perform linear regression to find exact energy change values and verify results obtained by Cheng et al. [2]

We have also analyzed the skewed nature of the positive and negative examples. We were getting precision and recall as 0 % in some folds of the 10 fold cross-validation. So when we dig deep we analyzed that when a test set is a mixture of protein with large number of negative examples and another protein with only positive examples, all the positive examples are incorrectly classified as negative. This might be happening due to skewed nature of the test data.

INDIVIDUAL TASKS

Roles and Contributions

The project is an extension of work done by J Cheng et al. We believed and proved that the accuracy was over-estimated as they didn't take care of appropriately diving the training and testing data sets while cross validation. New technique that we have used is while cross-validation we divided the data set on the basis of below mentioned parameters on each occasion in addition to 'per protein':

1. Per chain
2. Per cluster

We have used auxiliary dataset to perform cross-validation efficiently which authors haven't done. It helped us to keep all proteins belonging to each protein id either in test or in train set while doing cross-validation. We have also used logsig function to map our predictions from [-∞,∞] to [0,1].

Here is the work done by me:

3. I parsed the dataset and extracted relevant data from raw data for sequence. I extracted all the sequential neighbors (window size=7), pdb code, ph and temperature.
4. While parsing data I noticed that some proteins contains 21 amino acids instead of 20 as per the paper. Then after careful analysis I found 'X' amino acid (In 13 protein examples) which was the cause of discrepancy so I removed it.
5. I removed the duplicates from extracted data using pdb code, wild type, position of mutation, mutation, ph, temperature and sign of energy change.
6. After removing duplicates I created feature file using encoding, wild type, mutation and sequential neighbors for Sequence.

7. Created hashmap for ‘per protein’ and ‘per protein chain’ that can be used for creating Auxiliary file for sequence only.
8. Created n-cross-Validation that divide train and test data on the basis of hashmap and auxiliary file eventually to be used by SVM light for classification.
9. Learnt and used SVM light for sequence data with specified parameters of penalty ratio, gamma and cost factor for sequence setup.
10. Learnt and used sklearn on how to draw ROC curve, find AUC in python and then incorporated it in the code for sequence setup.
11. Evaluated various results such as Accuracy, F1- measure, Precision and Recall explicitly for sequence setup for ‘per protein’ and ‘per protein+chain’.
12. Plotted ROC for setups mentioned in point 9.
13. In some folds of 10-cross validation, we were getting precision and recall as zero so I analyzed it very thoroughly. I found that it is happening because of skewed nature of data as we provided analysis in conclusion.

Here is the work done by Srikanth Srinivas Holavanahalli:

Structure

1. He parsed the structure data given for initial analysis
2. He analyzed how many positive and negative examples are present per protein and protein + chain that helped us to find the skewed nature of the data.
3. He created hashmaps and auxiliary file for structure and cluster setup.
4. He again created feature file using wildtype, mutation and neighbors count of structure setup.
5. He cross validated the dataset and trained SVM light by maintaining same parameters as required for structure.
6. Learnt how to use SVM light and used it for structure setup.
7. Evaluated various results such as Accuracy, F1- measure, Precision and Recall explicitly for structure setup for ‘per protein’ and ‘per protein+chain’.
8. Plotted ROC for setups mentioned in point 7.

Structure + Sequence

9. He divided data into 2 files containing positives and negatives from sequence raw data and mapped this these files to structure data
10. He parsed the Mupro data to eliminate the duplicates and prepare 1496 data set for sequence data
11. Mapped the examples between structure and sequence and created a feature file by extracting same 182 features of sequence and append count of the neighbors as features.
12. Repeated all the points for Structure + Sequence setup as above mentioned for structure

Work done by Kedar Gundlupet:

1. Using CD hit to divide data into clusters.
2. He converted the mupro data into FATSA format for per protein and per protein chain. Learnt how to use CD-HIT and used it to cluster the per protein and per protein chain FATSA format files for similarity values from 40% to 100% with 10% intervals. Created the auxilary file for per protein and per chain sequence dataset.
3. Performed 10-fold cross validation 5 times for cluster similarity values of 40%, 50%, 70% and 90% on following feature files received from Srikanth and me:
    a. Sequence only per protein and per chain
    b. Structure+sequence per protein and per chain
    c. Structure only per protein and per chain
4. He plotted 24 ROC curves for the above mentioned setups.
5. Evaluated various results such as Accuracy, F1- measure, Precision and Recall explicitly for sequence, structure and clusters.
6. Learnt and helped us to use sigmod function to convert the output of SVM light into posterior probability of range [0, 1]

REFERENCES

[1] Capriotti E, Fariselli P, Casadio R. A neural network based approach to predict the protein stability based on the single-site mutations in protein structures. 2004.

[2] Cheng J, Randall A, Pierre B. Prediction of protein stability changes for single site mutations using support vector machines. 2006.

[3] CD-HIT: http://weizhongli-lab.org/cd-hit/

[4] SVM light by Thorsten Joachims. http://svmlight.joachims.org/