Note: Accepted in Journal of Theoretical Biology for publication in 2015, DOI:10.1016/ j.jtbi.2015.10.015.

A Balanced Secondary Structure Predictor

Md Nasrul Islam[1], *Sumaiya Iqbal*[1], *Ataur R Katebi*[2] *and Md Tamjidul Hoque*[1,*]
[1]*Computer Science, University of New Orleans, Louisiana, 70148, USA*
[2]*National Cancer Institute, National Institute of Health (NIH), USA.*

Abstract: Secondary structure (SS) refers to the local spatial organization of a polypeptide backbone atoms of a protein. Accurate prediction of SS can provide crucial features to form the next higher level of 3D structure of a protein accurately. SS has three different major components, helix (H), beta (E) and coil (C). Most of the SS predictors express imbalanced accuracies by claiming higher prediction performances in predicting H and C, and on the contrary having low accuracy in E predictions. E component being in low count, a predictor may show very good overall performance by over-predicting H and C and under predicting E, which can make such predictors biologically inapplicable. In this work we are motivated to develop a balanced SS predictor with higher accuracies in predicting all three SS components. Our approach uses three different support vector machines for binary classification of the major classes and then form optimized multiclass predictor using genetic algorithm (GA). The trained three binary SVMs are E versus non-E (i.e., E/¬E), C/¬ C and H/ ¬ H. This GA based optimized and combined three class predictor, called cSVM, is further combined with SPINE X to form the proposed final balanced predictor, called MetaSSPred. This novel paradigm assists us in optimizing the precision and recall. We prepared two independent test datasets (CB471 and N295) to compare the performance of our predictors with SPINE X. MetaSSPred significantly increases beta accuracy (Q_E) for both the datasets. Q_E score of MetaSSPred on CB471 and N295 were 71.7% and 74.4% respectively. These scores are 20.9% and 19.0% improvement over the Q_E scores given by SPINE X alone on CB471 and N295 datasets respectively. Standard deviations of the accuracies across three SS classes of MetaSSPred on CB471 and N295 datasets were 4.2% and 2.3% respectively. On the other hand, for SPINE X, these values are 12.9% and 10.9% respectively. These findings suggest that the proposed MetaSSPred is a well-balanced SS predictor compared to the state-of-the-art SPINE X predictor.

1. INTRODUCTION

Proteins are the most versatile macromolecules in living organisms and play significant roles in the biological processes [1]. Functions and three dimensional structures, developed through folding of proteins have significant evolutionary relationship [2-5]. Due to the importance, the prediction of three dimensional structure of protein is a widely researched area of bioinformatics. Despite numerous experimental and analytical endeavor, protein 3D structure prediction is still an unresolved problem. Accurate prediction of protein 3D structure significantly relies on the precise secondary structure prediction (SSP) [6-8], since the secondary structure (SS) defines the local spatial organization of a polypeptide's backbone atoms [9]. To analyze the massive amount of protein sequences generated by genome project highly efficient theoretical methods for predicting SS is of immense importance as experimental methods are highly time consuming, costly and in some cases inefficient [6, 10]. The SSP problem is defined as a biological sequence analysis problem, and the solution is to assign a right label of structure on every residue or amino acid of a protein sequence [11]. In other words, it is a classification problem and the major classes are helix (H), sheet (E) and turn (C) or coil. Scientists have been attempting to solve SSP problem by applying Bayesian statistics and a wide variety of theoretical models, machine learning approaches such as neural networks (NN), hidden Markov model (HMM), support vector machine (SVM), and so on. One of the earliest methods used for SSP problem was GOR method, which was established by Robson and colleagues through a series of papers in 70's [12-15]. GOR method is the first real method that applied computer program for SSP problem by coding the relation between secondary structure and amino acids utilizing information theory and Bayesian statistics and its accuracy was 64.4% [16]. NN for SSP was first employed in 1988 by Qian and Sejnowski [17]. Their overall accuracy on a test set, that was non-homologous to the training set, was 64.3%. Rost and Sander established new standard of SSP method introducing PHD method in 1993 [18]. They applied a set of feed forward neural networks trained by back-propagation algorithm [19] with non-

* Corresponding author
Email address: thoque@uno.edu

redundant data set of 130 protein chains. Most important aspect of their method was that they used multiple sequence alignment (MSA) as evolutionary information instead of single sequences. Accuracy of PHD method was 70.8%. If larger data set is used along with PSI-BLAST [20], accuracy of PHD method may increase to 75.0% [21]. Chandonia and Karplus [22], developed a two level neural network and trained their model with 681 protein sequences. Their model yielded around 75.0% accuracy. One of the most successful NN based SSP method is PSIPRED [23]. Instead of doing MSA, this used intermediate PSI-BLAST profiles as a direct input to its SSP model. This prediction method has three stages: generating sequence profile, predicting initial SS, and finally filtering the predicted structure. PSIPRED achieved an overall accuracy between 76.5 to 78.3%. PSIPRED and PHD shared similar network topology. The improvement in PSIPRED over PHD may be attributed to the better alignment fed to the NN because of the filtering strategy applied by PSIPRED to exclude unrelated proteins and also to the increase in the size of database [24].

Asai K *et al* [25] implemented HMM based protein secondary structure prediction model for the first time. They trained four HMMs for predicting helix, sheet, turn and others separately. They used 120 sequence from Brookhaven PDB for training and testing purpose. They achieved a Q_3 score (see table 3 for the definition) of 54.7% which was not very good. Further, using HMM, noticeable improvement of secondary structure prediction accuracy was done by Bystroff, Thorsson and Baker [26]. They proposed a novel HMM, HMMSTR based on I-sites library of sequence-structure motifs. I-sites are short sequence-structure motifs that show strong correlation with local three dimensional structural elements of proteins. Their secondary structure prediction accuracy (Q_3 score) was 74.3% using homologous sequence information. Martin J *et al* introduced a new type of HMM that does not use prior knowledge [27]. They applied genetic algorithm (GA) for DNA sequence analysis [28, 29] in order to automatically select an HMM topology. The Q_3 score of the model was 68.8% for single sequence and 75.5% when multiple sequence alignment was used.

Another machine learning approach scientists applied to solve SSP problem is the SVM, which generates hyperplane that partitions the given data into two classes with maximum accuracy based on a kernel function. Specially, when the data is not linearly separable, SVM can cast the data into a higher dimensional space where the data becomes classifiable. SVM was first applied to SSP in 2001 by Hua and Sun [30]. They employed six binary classifiers: H versus not H, E versus not E, C versus not C, H versus not E, E versus not C and C versus not H using radial basis kernel function based SVM. Their method's accuracy (Q_3) was 73.5%. Guo *et al.* [31] developed a dual-layer SVM and used PSSM generated from PSI-BLAST to predict secondary structure. They reported Q_3 as 75.2% on CB513 data set prepared by Cuff and Burton [32]. Another SVM based approach [33], used patterns of consecutive amino acids of any length that are frequently found in a protein database. The main theme of this approach was to find all such patterns within a defined protein language occurring with a frequency greater than a user-defined minimum frequency or support. They generated three different feature sets based on frequent pattern to test their model on 150 targets from the EVA [34] contest. Their accuracy ranges from 75.34% to 77.0%.

A recent work on SSP is SPINE X by Faraggi *et al.* [6]. This method is a multi-step NN algorithm that combined the SSP with the prediction of the real value residue solvent accessibility (RSA) and backbone torsion angles in an iterative fashion. It has total six steps. In the first, fourth and last steps, they predicted secondary structure and in the second step they predicted RSA and in the third and fifth steps, they predicted backbone torsion angles. The theme is to boost up any subsequent prediction steps, utilizing previous steps' predicted information as input features. They tested their model on multiple data sets and their overall accuracy ranges from 81.3% to 82.0% while DSSP [35] assignment was used for the secondary component assignments.

It is important to note that a major drawback of most of these computational methods is that their accuracy in predicting E is comparatively low. For example, SPINE X reported 86.6%, 75.3% and 81.5% accuracy on H, E and C segments on their 2,640 dataset when DSSP assignment was used. One of the top SVM based SS predictors proposed by Wang *et al.* [36] reported average accuracy on H, E and C as 78.2%, 67.2% and 82.5% respectively on CB513 dataset [32]. Because E plays a fundamental role in protein structure, function, and evolution [37], to be able to enhance the specificity of the prediction of E will be of higher biological significance towards practical applications. Here we report a novel approach to build an SS predictor that can yield better accuracy in E segment without compromising the overall accuracy.

As demonstrated by a series of recent publications ([38], [39], [40], [41], [42], [43]), to establish a really useful sequence-based statistical predictor for a biological system, we aligned the outline of our paper accordingly towards the steps of Chou's 5-step rule [44] as: *(a)* construct or select a valid benchmark dataset to train and test the predictor, described in section 2.1; *(b)* formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, described in section 2.2; *(c)* introduce or develop a powerful algorithm (or engine) to operate the

prediction, described in section 2.3; *(d)* properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor, described in sections 3; and *(e)* establish a user-friendly web-server for the predictor that is accessible to the public, if not available immediately, then at least the stand alone code is provided[1].

2. MATERIALS AND METHODS

2.1 Datasets

We have collected one training dataset and two independent test datasets. How the training and the test datasets were collected, have been explained in subsections 2.1.1 and 2.1.2.

2.1.1 Training dataset

We collected protein sequences from PDB [45] with the specifications: sequence length ≥ 50, resolution ≤ 2.5 Å , sequence identity ≤ 30%, and refinement *R* factor 0 to 0.25. We obtained 6,521 protein sequences from PDB by applying these search criteria. To eliminate bias arising from repeated instances of some similar sequences, we ran BLASTclust [46] to filter out similar ones with a cut-off of 25% similarity from each cluster, keeping just one per cluster which resulted in 2,150 sequences in this step. Then we ran DSSP [35] to collect secondary structure assignment for each residue using the collected dataset. We discarded the sequences for which DSSP failed to fully assign secondary structure. We also discarded sequences where we found that the length of a sequence given by DSSP assignment and fasta sequence length provided by PDB differs. We refined this data set further to discard the protein sequences that contain unknown amino acids labelled as "X" and the sequences which contain amino acids of unknown coordinates. Finally, we discarded 2 more incomplete sequences. Our final training data set consists of 552 sequences with no more than 25% identity among themselves. From now on we will call these dataset as T552. T552 composed of 149,093 residues with 18.2%, 51.6% and 30.2% of E, C and H residues, respectively.

2.1.2 Test datasets

We have two different test datasets. First, we have collected CB513 dataset [32] for testing purpose. Then we ran BLASTclust at 25% identity cut-off on T552 and CB513 to ensure that these datasets are independent of our training dataset. Here we extracted 475 sequences from CB513 at 25% identity cut-off with respect to T552. After further refinement based on failure to generate features such as angle fluctuations or ASA for some sequences, we had 471 sequences as our first test set, named CB471. To confirm the robustness of our predictors, we collected another comparatively new independent dataset with 25% identity cut-off criteria from PDB generated in 2014. This dataset consists of 295 sequences and we named this test dataset as N295.

2.2 Features

We collected a comprehensive and independent set of residue level features which may sufficiently capture sequence information, evolutionary information as well as structural information of the amino acids in the protein sequences. A list of used features is presented in Table 1.

Table 1: A list of features extracted and used from a given protein sequence.

Category	Feature count
Amino acid (AA)	1
Physiochemical properties (PP)	7
Position specific scoring matrix (PSSM)	20
Monogram (MG)	1
Bigram (BG)	20
Disorder probability (DP)	1
Accessible surfae area (ASA)	1
Torsion angles (φ, ψ) fluctuation (AF)	2
Terminal indicator (TI)	1

Amino acid (AA) can be any integers from 1 to 20, each uniquely represents one of the 20 different amino acids. Each amino acid has 7 unique properties, collectively named as physiochemical properties (PP). They

[1] http://cs.uno.edu/~tamjid/Software/BalancedSSP/BalancedSSP.tar.gz

are steric parameter, polarizability, hydrophobicity, isoelectric point, helix probability and sheet probability [47]. PSSM were generated by running PSI-BLAST [20]. Monogram (MG) and bigram (BG) were calculated from PSSM as described in [48] to infer structural information from sequence level evolutionary information. We predicted disorder probability (DP) using DisPredict [49]. It gives the probability of an amino acid residue being disordered, i.e., having no well−defined three dimensional structure. Accessible surface area (ASA) is the surface area of a biomolecule that is accessible to the solvent in which the molecule is dissolved. Conformational dynamics of proteins which is crucial for their diverse functionalities, is strongly correlated with the ASA of each of the residue of a protein [50, 51]. ASA is directly related to the protein-protein interactions [52, 53] and is also an important factor in beta pair formation [37]. Therefore, we used a very recently developed high accuracy ASA predictor, REGAd^{3}p [54], to predict ASA in our work. REGAd^{3}p uses regularized exact regression with 3rd degree polynomial kernel and also applied GA to optimize the weights computed by regularized regression. Angle fluctuations (AFs) represent the flexibility of protein backbone as derived from the ensembles of NMR structure. These are two very important features, since these two torsion angles, φ and ψ, are sufficient for a nearly complete description of the backbone of a protein structure [55]. First five and last five residues in any sequence are considered as terminal residues. For the first and last one we assigned -1 and +1, respectively, as the terminal information and then gradually increased or decreased the value by 0.2 as we moved forward from the starting terminal or move back-ward from the end terminal. For example, the second and the penultimate residue gets -0.8 and 0.8, respectively as the terminal information value. All other intermediate residues, except those ten residues, are assigned a terminal value of 0.

We used these features in a variety of combinations in our search to come up with an optimal feature set. Some features were common in all models used, whereas some were excluded in some models to gauge the impact of the excluded features in our prediction. Different sets of features we used in our model search are listed in Table 2.

Table 2: Compositions of the explored feature sets.

Name of the feature set	Features	Details of the features (feature count)
f_{29}	29	AA(1), PP (7), PSSM(20), and TI(1)
f_{31}	31	AA(1), PP (7), PSSM(20), DP(1), ASA(1) and TI(1)
f_{33}	33	AA(1), PP (7), PSSM(20), DP(1), ASA(1), AF(2) and TI(1)
f_{51}	51	AA(1), PP (7), PSSM(20), MG(1), BG(20), DP(1) and TI(1)

We compared the performance of our model using different feature sets shown in Table 2. Finally we found that f_{33} is the best feature set in terms of accuracy across three SS classes. Therefore, we trained our final predictor using f_{33} feature set. We also used sliding window information to incorporate neighboring information for each residue. After experimenting with different window sizes, starting from 7 through 25, we found that 15 is the optimal window size for this computation. Finally, selecting 15 residue based window, the feature count per residues becomes (33×15) or, 495.

2.3 Methods

We have used binary SVMs coupled with genetic algorithm in our investigations. We have trained three binary SVMs – E/ ¬ E, C/ ¬ C and H/ ¬ H using libsvm [56]. Although SVM could have been used directly for three class classification, we rather choose to use three binary SVM classifiers, so that we can attain a balanced accuracy in all three classes. Another reason for using binary SVMs is that SVM was built for binary classification problem and performance may degrade if used for multi class classification problem. These three SVMs provide us the probability for each residue that belongs to beta, coil and helix structure. Combining the results from these three predictors' outcomes into the balanced three class classification solution is a crucial challenge. We combine these three binary predictors using GA to form final three class prediction with optimal weights of the individual class. A brief discussion of our implemented GA is presented in later in this section. GA finds separate real value parameter for each class as an additive factor for each class probability given by three binary SVMs. For example if the probabilities that a particular residue belongs to either E, C or H classes are p_1, p_2 and p_3 respectively, GA finds three real values v_1, v_2 and v_3 and the revised class probabilities become (p_1+v_1), (p_2+v_2) and (p_3+v_3) for E, C or H class

respectively. Finally the class, for which this revised probability is highest, is accepted as the predicted class of that particular residue. We refer to our combined SVM predictor as cSVM. Our final predictor is a meta predictor, named as MetaSSPred, combines the outputs of cSVM and SPINE X [6]. Algorithm for combining the prediction of SPINE X and cSVM to build MetaSSPred is shown in Figure 1.

1. Generate secondary structure probabilities and classes by cSVM.
2. IF cSVM's output class is E THEN
   3a.ACCEPT E as the output of MetaSSPred
   ELSE
   3b. ACCEPT SPINE X's output as the output of MetaSSpred
   ENDIF.

Figure 1: Algorithm for combining cSVM and SPINE X.

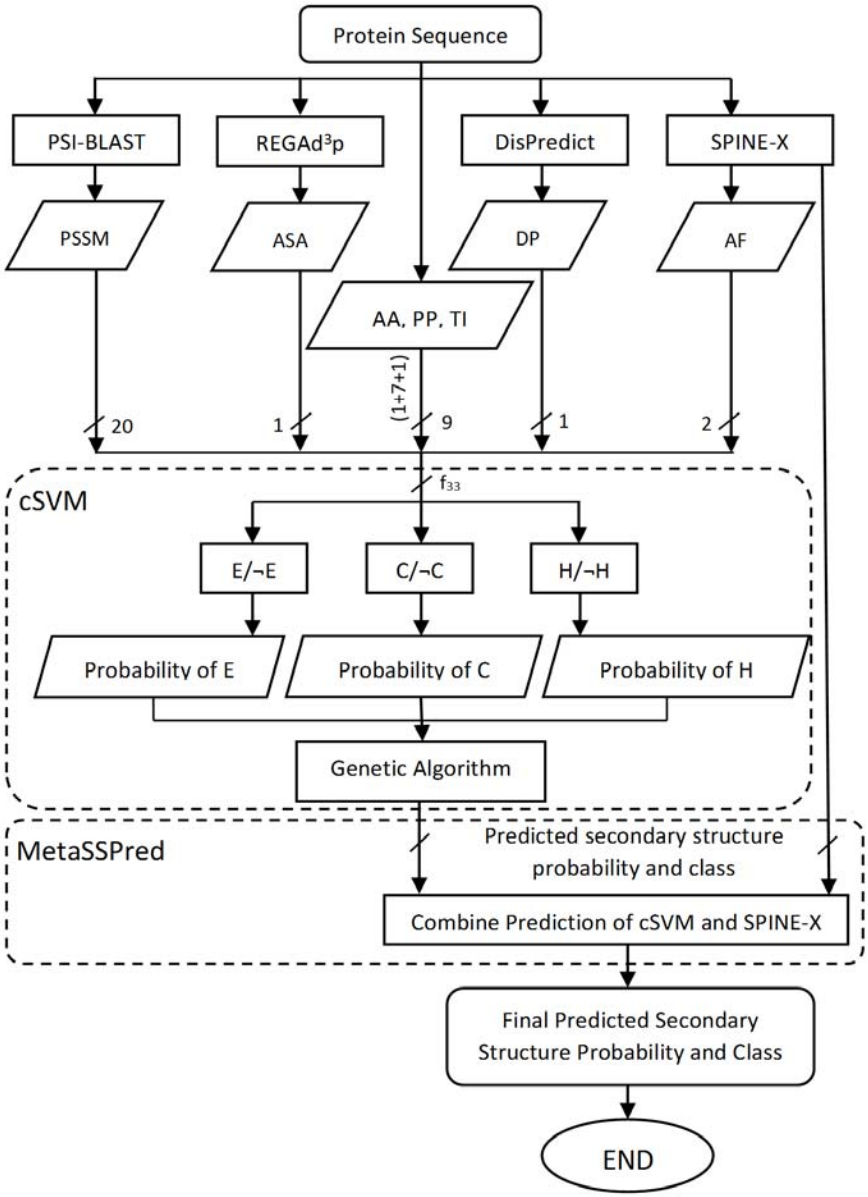

Figure 2: Flow diagram of MetaSSPred predictor. Abbreviated feature names used, can be found in Table 1.

GA is an evolutionary learning based heuristic algorithm which maintains a population of individuals for each iteration. Each individual in the population, commonly known as chromosome, represents a potential solution to the problem to be solved. In our GA, for the first iteration the population was generated randomly where each chromosome is a 36 bit long binary number, which essentially contains the three additive factors, each 12 bit long. In the next iteration, a new generation of population is created from the previous generation after some evolutionary transformation.

The evolutionary transformation contains three mechanisms – elite preservation, crossover and mutation. Candidates for elite, crossover or mutation were randomly selected based on fitness based probability of each chromosome. We used Q_3 as fitness for each chromosome. The higher the Q_3, the more likely a chromosome will be selected for crossover or mutation. Elites are the top 10% chromosomes in terms of fitness in any population. We used 80% cross over rate and 10% mutation rate. We retained the population size of 2,000 chromosomes for each population and we ran 50 generations. A comprehensive flow diagram of our method is shown in Figure 2.

2.4 Performance Evaluation

To compare and evaluate the performance of each predictor, we used 4 performance parameters: accuracy, precision, recall and over prediction rate. To measure these metrics, we calculated of true positive (TP), false positive (FP), true negative (TN) and false negative (FN). TP is the number of instances that are predicted as positive and are actually positive. FP is the number of instances that are predicted as positive and are actually negative. TN is the number of instances that are labelled as negative but are actually negative. FN is the number of instances that are labelled as negative and are actually positive. Calculations of these measures along with their evaluation focus are presented in Table 3.

Table 3*:* Evaluation criteria of the classifer.

Measure	Formula	Evaluation Focus
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall effectiveness of a classifier
Precision	$\frac{TP}{TP + FP}$	Agreement on class of the data labels with the positive labels given by the classifier
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Effectiveness of a classifier in identifying positive labels
Over Predicton	$\frac{\#Predicted\ class}{\#Actual\ class}$	Measures whether higher accuracy for particular class is due to over prediction or not
Overall Precision	$\frac{\sum_{c=1}^{n} Precision_c}{n}$	Average agreement on n different classes of the data labels with the positive labels given by the classifier
Overall Recall	$\frac{\sum_{c=1}^{n} Recall_c}{n}$	Average effectiveness of a classifier in identifying positive labels for *n* classes
Q_3	$\frac{Total\ number\ of\ correctely\ predicted\ (tri)\ states\ (C,E,H)\ of\ the\ resudues}{Total\ number\ of\ residues}$	
Q_C, Q_E, Q_H	*Fraction of correctly predicted coils (C), sheets (E), helices (H), respectively, out of total residues.*	

3. RESULTS AND DISCUSSION

In this section, we have presented the result of SSP obtained from our predictors and from SPINE X on CB471 and N295 test datasets and compared those results. In order to assess the performance of each

predictors, we have calculated Q_3 of each predictors as well as accuracy for beta, coil and helix class (Q_E, Q_C and Q_H respectively) along with precision and recall. We also have calculated the over prediction rate for each class to investigate whether any higher measure is due to over prediction.

3.1 Performance on CB471 Test Dataset

To facilitate comparison, in Figure 3, accuracies as well as over prediction rates of each predictors for each class are presented using bar chart.

In Figure 3, we see that Q_E of MetaSSPred is significantly higher than Q_E of the other two predictors. More specifically, Q_E of MetaSSPred is a 20.9% improvement over SPINE X. Further, MetaSSPred does not heavily under or over predict beta compared to other two methods. Therefore, MetaSSPred is certainly a better predictor for the beta class. Lower Q_E of SPINE X here may be attributed to the very high under prediction rate (24.6%). In case of Q_C, SPINE X gives the highest score (81.4%), cSVM is just after SPINE X with a Q_C of 80.6% and MetaSSPred's Q_C is 76.0% only. If we look at the over prediction rates of this class, we find that both cSVM and SPINE X highly over predict coil class (13.4% and 14.5% respectively). On the other hand, MetSSPred's over prediction rate is very low, only 1.9%. This is a strong reason why the Q_Cs of cSVM and SPINE X are higher than that of MetaSSPred. SPINE X comes up with the highest Q_H (81.90%) and MetaSSPred closely follows SPINE X with 80.10%. cSVM is the worst performer in this case. The reason for low Q_H of cSVM may be attributed to the fact that it under predicts helix by 10.3%. In Q_3 measure, MetaSSPred and SPINE X are almost equal with a 0.1% gap in favor of SPINE X. cSVM is also not far behind. Overall, comparatively MetaSSPred appears as a very balanced predictor yielding good accuracies for all three classes separately as well as in Q_3. To be specific about measuring the performances, we have computed precision and recall next.

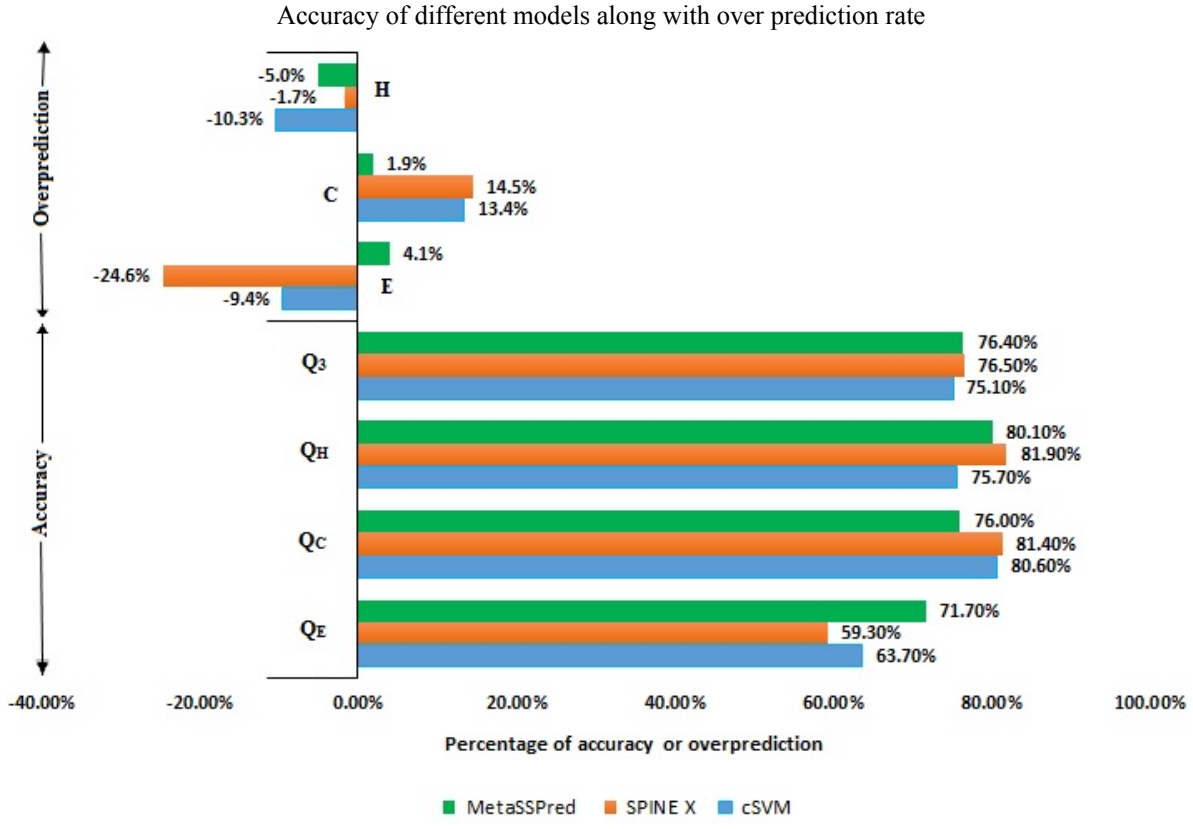

Figure 3: Comparison of accuracy along with over prediction rate on CB471 dataset.

Precision and recall measures for CB471 datset are presented in Figure 4, where we see that SPINE X has the highest precision and lowest recall value for beta class. Gap between the precision and recall score

of SPINE X for beta class is 7%. This suggests that overall SPINE X gives lower false positives, however it gives a very high false negatives in case of beta prediction. Therefore, for any application, where the cost of failure to detect beta residue is high, SPINE X may not be suitable. MetaSSPred provides relatively high balanced precision and recall value for beta prediction with a gap of only 2.8% between recall and precision. The proposed MetaSSPred provides the highest recall value for beta class among the three predictors discussed. Therefore, applications where detection of beta is very important, MetaSSPred would perform well. cSVM has a recall value 7.4% higher than that of SPINE X, however, cSVM achieves 10.7% lower precision compared to SPINE X.

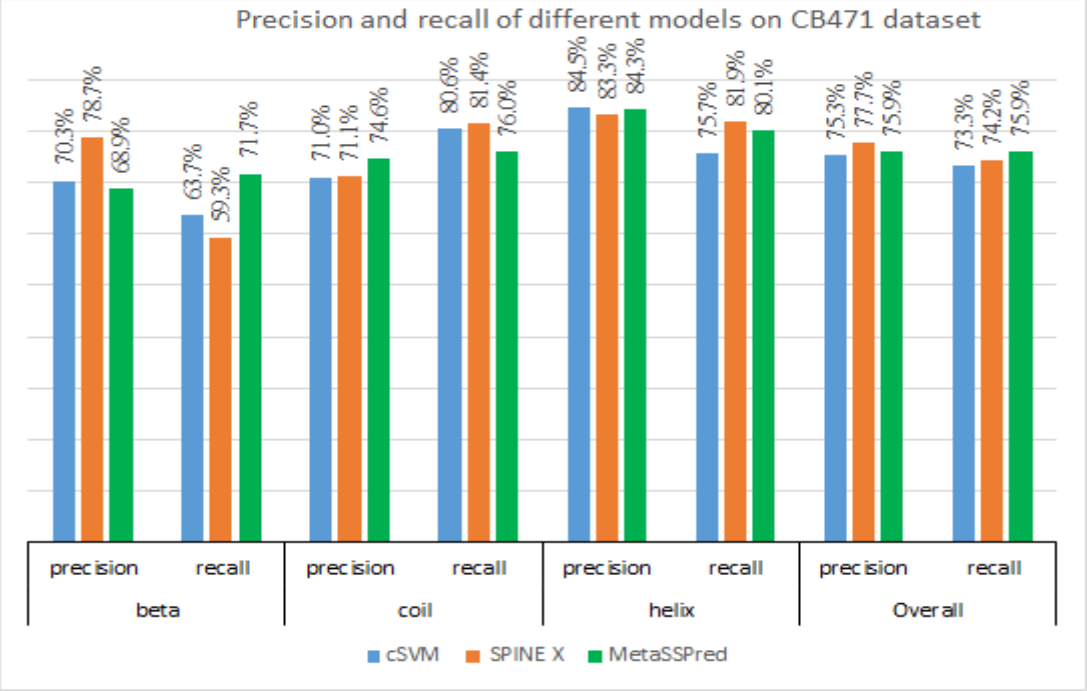

Figure 4: Precision and recall on CB471 dataset obtained for the three predictors: cSVM, SPINE X and MetaSSPred.

In coil prediction, MetaSSPred provides balanced precision and recall score with a gap of 1.4% only, and its precision is the highest among all predictors. SPINE X and cSVM provides 7.1% and 6.0% higher recall score than that of MetaSSPred. However, MetaSSPred provided 4.9% and 5.1% improvement in precision score than those of SPINE X and cSVM, respectively. The reason behind this phenomenon is that both cSVM and SPINE X highly over predict coil class which are 13.4% and 14.5% higher prediction rate, respectively.

In helix prediction, precisions of each class are very close. Although cSVM provides the highest precision for helix, it gives the lowest recall. Therefore, if failure to detect helix is very costly, we should prefer other predictors to cSVM. SPINE X and MetaSSPred are close competitors for helix prediction. Both of them have good and balanced precision and recall score in this case.

3.2 Performance on N295 Test Dataset

We will start our analysis of the performances of our predictors on N295 dataset with accuracy measures. Accuracies and over prediction rates on N295 dataset are shown in Figure 5.

Both SPINE X and cSVM give comparatively lower Q_E score and both of them highly under predict beta (22.5% and 15.0% under prediction respectively). On the other hand MetaSSPred gives the highest Q_E almost without over or under prediction of beta class. Here improvement in Q_E achieved by MetaSSPred

over SPINE X is 19.0%. Therefore, MetaSSPred is clearly the best predictor for beta class for this N295 dataset.

In coil prediction, cSVM and SPINE X give almost similar accuracy with a gap of 0.5% in favor of cSVM. On the other hand, Q_C of MetaSSPred is the lowest among all. Q_C of SPINE X is 6.1% higher than that of MetaSSPred. If we look at the over prediction rates, we see that both cSVM and SPINE X highly over predicts coil, by 17.7% and 19.2% respectively. MetaSSPred also over predicts coil, however by only 5.1%. Therefore, the higher Q_C of cSVM and SPINE X than that of MetaSSPred could be because of the over predictions by the first two methods mainly.

The highest Q_H is obtained from SPINE X, and MetaSSPred is closely following SPINE X by a gap of 1.6%. Q_H score of cSVM is significantly lower than those of two others. If we look at the over prediction rate, we see that cSVM under predicts helix by 12.2%. This may be a strong reason why cSVM gives so poor Q_H. Other two methods, SPINE X and MetaSSPred also under predict helix. Interestingly, slightly lower Q_H of MetaSSPred results from its around 3% of higher under prediction rate compared to that of SPINE X. Overall, MetaSSPred provides the highest Q_3 score for N295 dataset.

Accuracy of different models along with over prediction rate

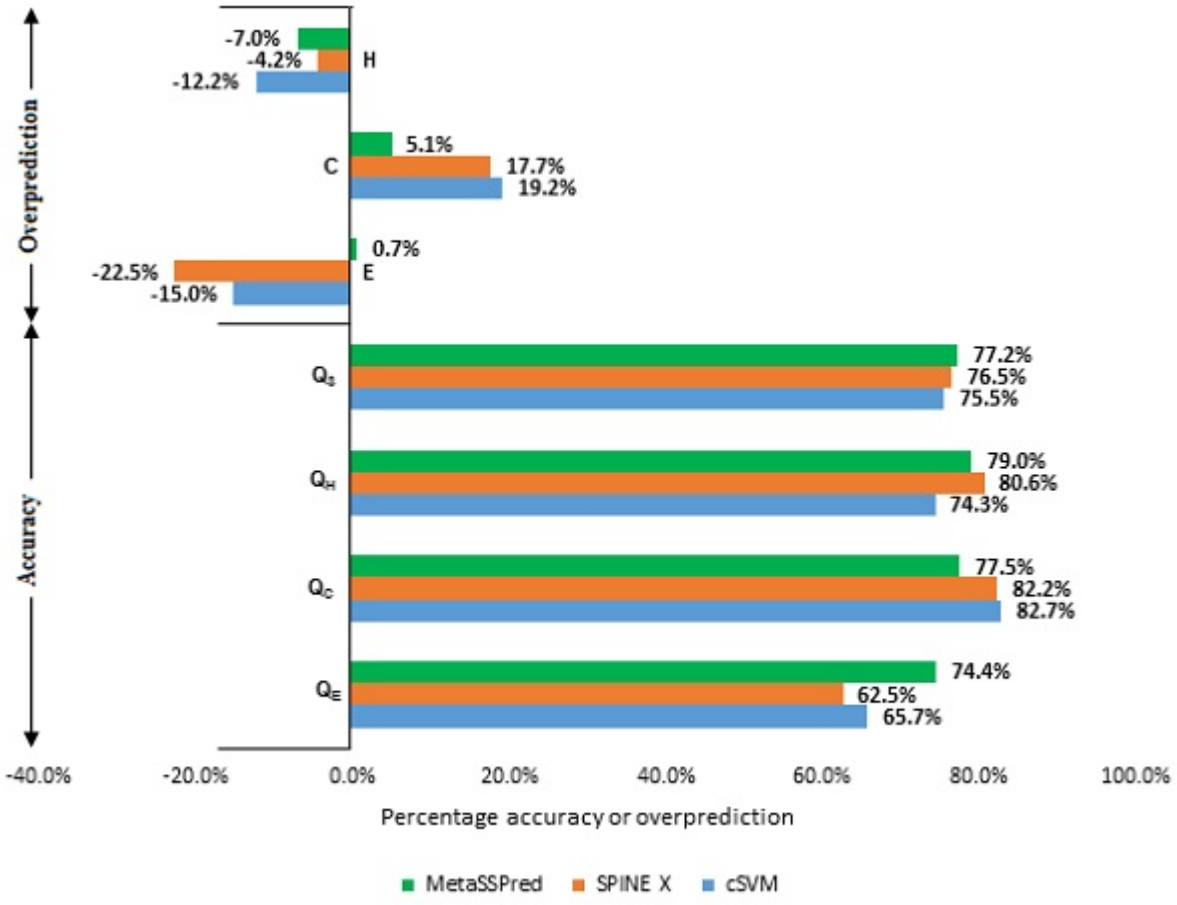

Figure 5: Comparison of accuracy along with over prediction rate on N295 dataset for the three predictors: MetaSSPred, SPINE X and cSVM.

Highest Q_H is obtained from SPINE X, and MetaSSPred is closely following SPINE X by a gap of 1.6%. Q_H score of cSVM is significantly lower than those of two others. If we look at the over prediction rate, we see that cSVM under predicts helix by 12.2%. This may be a strong reason why cSVM gives so poor Q_H. Other two methods, SPINE X and MetaSSPred also under predict helix. Interestingly, slightly lower Q_H of

MetaSSPred results from its around 3% of higher under prediction rate compared to that of SPINE X. Overall, MetaSSPred provides the highest Q_3 score for N295 dataset.

To further gauge the performances of our predictors on N295 dataset, we can focus on the precision and recall scores. Precision and recall scores of all three predictors on N295 dataset are presented in Figure 6. We see in Figure 6 that SPINE X gives the highest precision and lowest recall in beta prediction. The reason behind such imbalance prediction is that SPINE X highly under predicts (by 22.5%) beta residues. Second highest precision is given by cSVM and again it also has a recall score comparatively much lower than that of MetaSSPred. Main reason is again under prediction. cSVM under predicts beta residues by 15%. On the other hand, MetaSSPred gives a very balanced precision and recall and it has the highest recall score for beta prediction. Over prediction rate of MetaSSPred is only 0.7% for beta residues. Therefore, false positive rate of MetaSSPred is also very low.

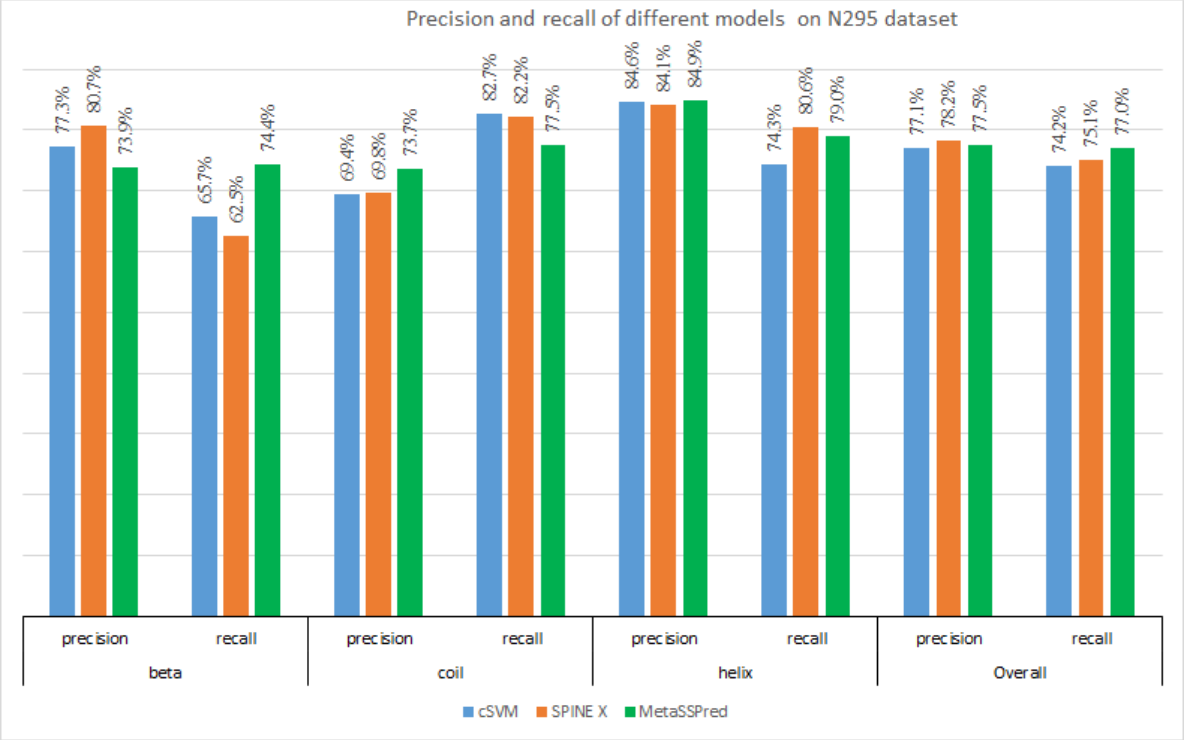

Figure 6: Precision and recall on N295 dataset obtained for the three different predictors: cSVM, SPINE X and MetaSSPred.

In case of coil prediction, precision scores of SPINE X and cSVM are comparatively lower than that of MetaSSPred. On the other hand, recall score of MetaSSPred is lower than that of both cSVM and SPINE X in coil prediction. The reason is again that cSVM and SPINE X over predicts coils by 19.2% and 17.7% respectively. MetaSSPred also over predicts coil, however by only 5.1%. Therefore, false positive rates of SPINE X and cSVM are higher than that of MetaSSPred.

MetaSSPred provides the highest precision for helix prediction. Other predictors are also very close. Highest recall is given by SPINE X and MetaSSPred closely follows it. If we look at the over prediction rate, we see that all three methods under predict helix, however under prediction rate is lowest for SPINE X. Considering precision, recall and over prediction rate, it seems that SPINE X is the best method for helix prediction and then comes MetaSSPred.

4.3 Overall Ranking of the Predictors

In this section, we summarize the performance of each predictor by ranking them with respect to Q_3, Q_E, Q_C, Q_H, overall and class wise precision, recall and absolute over/under prediction rate for each test dataset.

Higher absolute over/under prediction rate yields lower ranking, while higher scores for all other measures result in higher ranking. Absolute value of over prediction is taken assuming that both over and under prediction are equally bad. Best scorer in any measure gets 3 points, second best gets 2 and the third gets 1. The ranks of testing CB471 and N295 are presented in Table 4 and Table 5 respectively.

Table 4: Rank of all predictors across different performance measure on CB471 test data set.

	Rank (higher value ranks better**)**		
Measure	**cSVM**	**SPINE X**	**MetaSSPred**
Q_3	1	3	2
Q_E	2	1	3
Q_C	2	3	1
Q_H	1	3	2
Precision (E)	2	3	1
Precision (C)	1	2	3
Precision (H)	3	1	2
Overall precision	1	3	2
Recall (E)	2	1	3
Recall (C)	2	3	1
Recall (H)	1	3	2
Overall recall	1	2	3
Absolute over prediction (E)	2	1	3
Absolute over prediction (C)	2	1	3
Absolute over prediction (H)	1	3	2
Total	**24**	**33**	**33**

Table 5: Rank of all predictors across different performance measure on N295 test data set.

	Rank (higher value ranks better**)**		
Measure	**cSVM**	**SPINE X**	**MetaSSPred**
Q_3	1	2	3
Q_E	2	1	3
Q_C	3	2	1
Q_H	1	3	2
Precision (E)	2	3	1
Precision (C)	1	2	3
Precision (H)	2	1	3
Overall precision	1	3	2
Recall (E)	2	1	3
Recall (C)	3	2	1
Recall (H)	1	3	2
Overall recall	1	2	3
Absolute over prediction (E)	2	1	3
Absolute over prediction (C)	1	2	3
Absolute over prediction (H)	1	3	2
Total	**24**	**31**	**35**

We see in Table 3 that SPINE X and MetaSSPred performed equally on CB471 test dataset. On the other hand, performance ranking of cSVM is comparatively much lower. However in many parameters, cSVM is better than SPINE X. For example, recall of E or precision of H is better in cSVM compared to those of SPINE X. Therefore, our test result on CB471 justifies the development of meta predictor. We see in Table 5 that performance of MetaSSPred is the best on N295 test dataset, whereas SPINE X ranked second. Therefore, we would like to conclude that MetaSSPred is a more generalized predictor with balanced accuracy across all the secondary structure classes.

4.4 Case Study: HIV-1 protease

In this case study, we have used two different practical cases to highlight the contribution of MetaSSPred, particularly related to E segments in the sequence. We choose a 99 residue long monomer of HIV-1 protease [57], PDB ID: 1EBY, which has high number of beta components (Figure 7). Two such monomers form the HIV-1 protease [58] (Figure 7).

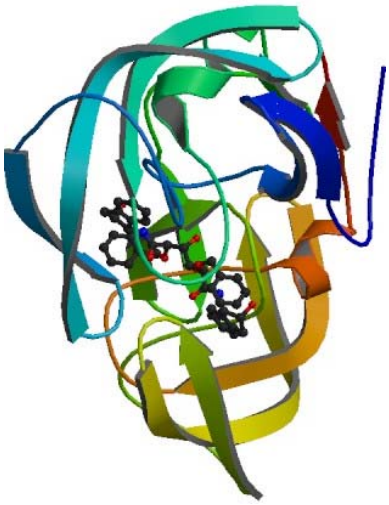

Figure 7: HIV-1 protease, a 99 residue long monomer, PDB ID: 1EBY, having high number of beta components.

Practically, the HIV-1 protease is an important case, since it is a very important target for the treatment of AIDS. The active site of HIV-1 protease is located at the interface of two monomers. The two beta hairpin structures, located in the active site, bind the HIV inhibitor molecule by undergoing structural changes. This beta hairpin structure is very advantageous for inhibitor design as it offers a good number of tight interactions between the enzyme and the inhibitor [57]. Therefore, identifying such beta structures in proteins can assist in drug design for complex disease cases.

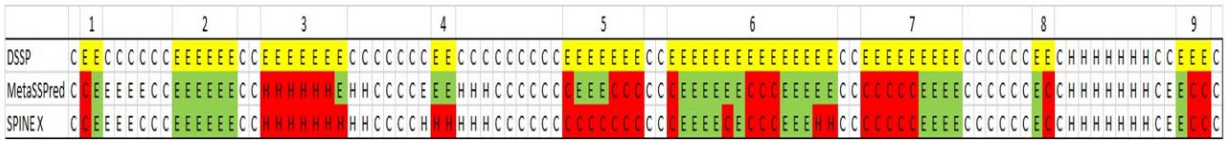

Figure 8: Overlapped regions of predicted E residues by MetaSSPred and SPINE X for the protein with PDB ID: IEBY. Different E segments of the protein as per DSSP assignment are highlighted (in yellow). Green color indicates correctly overlapped predicted E segments and red color indicates the regions where the prediction does not overlap with actual E segment.

We predicted the secondary structure of 1EBY using both SPINE X and our MetaSSPred. As indicated in Figure 8, we have shown different segments of E (highlighted yellow) of the protein. Here, we see that the protein has 9 different E segments according to DSSP assignment. SPINE X could not identify any E structure in segment 3, 4 and 5. MetaSSPred is partially correct in identifying E segments. Total number of

overlapped E residue for MetaSSPred is 30 and 21 for SPINE X out of 53 E residue in the DSSP assignment. Precision of MetaSSPred and SPINE X for the E segments are quite close, 83.3% and 84.0% respectively. On the other hand, recall score of MetaSSPred and SPINE X are 56.6% and 39.6% respectively, differs heavily. Q_3 of MetaSSPred and SPINE X were 65.7% and 57.6% respectively for this protein. Therefore, MetaSSPred also gives around 14% improvement in Q_3 over SPINE X for this important case.

4. CONCLUSIONS

In this section, the outcomes of our investigation is briefly summarized and some future directions for further improvement are also suggested.

Our basic model was a combined version of the three binary class SVMs, which were optimally combined into a multiclass classifier (cSVM) using GA. We compared the performance measures of our classifier with those of SPINE X, which is the state-of-the-art secondary structure predictor in terms of reported accuracy Q_3 score of SPINE X was 76.5% on both CB471 and N295 datasets in our investigation. For our cSVM, we obtained 75.5% and 74.2% Q_3 score based on CB471 and N295 datasets respectively. However, we observed that Q_E scores of SPINE X were comparatively lower for both CB471 and N295 datasets and which were found to 59.3% and 62.5% respectively to be exact. On the other hand, our cSVM provided better Q_E scores than SPINE X on both datasets and the scores were were 63.7% and 65.7% respectively. We also have observed that SPINE X highly under predicted helix by 24.6% and 22.5% for dataset CB471 and dataset N295 respectively. Q_C score of cSVM and SPINE X for both datasets were close. For example, Q_C of cSVM were 80.6% and 82.7% and those of SPINE X were 81.4% and 82.2% for dataset CB471 and dataset N295 respectively. On CB471, SPINE X gave higher accuracy in coil than cSVM. However, the difference was only 0.8%. On the other hand, cSVM gave higher coil accuracy on N295 dataset and the gap was 0.5% only. In helix prediction, SPINE X performed better than cSVM on both the test datasets. Gaps in Q_H of SPINE X and cSVM were 6.2% and 1% based on dataset CB471 and N295 respectively.

In a nutshell, SPINE X was better for helix prediction. On the other hand cSVM was better for beta prediction. In coil prediction, both are almost equally accurate. Therefore, we took this opportunity to combine cSVM and SPINE X to achieve better accuracy in all three classes and developed our meta predictor, MetaSSPred, combining the result from cSVM and SPINE X. The outcome is found to be very promising and well balanced.

MetaSSPred significantly increases Q_E for both datasets. Q_E score of MetaSSPred on CB471 and N295 were 71.7% and 74.4% respectively. Importantly, these are 20.9% and 19.0% improvement over the Q_E scores given by SPINE X on CB471 and N295 datasets respectively. The increased accuracy in E prediction by MetaSSpred is also supported by the important HIV-1 protease case (section 4.4). Improvements of Q_E scores by MetaSSPred over those of cSVM were also significant, which were 12.6% and 13.3% on CB471 and N295 datasets respectively. However, the improvement in Q_E brought some cost for coil prediction mainly. For example, Q_C scores of MetaSSPred were 5.4% and 4.7% lower in absolute value than those of SPINE X on CB471 and N295 datasets respectively. Average drop of Q_H score in MetaSSPred over SPINE X was 1.7% on both datasets. Overall accuracy of MetaSSPred decreased by 0.1% in absolute value on CB471 dataset, however increased by 0.7% in absolute value on N295 dataset compared to those of SPINE X.

Further, MetaSSPred decreased the volatility of accuracies across three secondary structure classes. For example, standard deviations of accuracies across three classes were 12.9% and 10.9% on CB471 and N295 test sets respectively for SPINE X. On the other hand, for the same data sets standard deviations of the three class accuracies were 4.2% and 2.3% respectively for MetaSSPred. Precision and recall gap volatility also decreased in MetaSSPred. For example, standard deviations of the gaps between respective precision and recall scores across three secondary structure class are 10.0%, 15.0% and 3.7% for cSVM, SPINE X and MetaSSPred respectively on CB471 dataset. The same volatility reduction in the gaps between precision and recall was observed for N295 dataset. Standard deviations of such gaps for cSVM, SPINE X and MetaSSPred on N295 dataset were 14.0%, 15.3% and 4.9% respectively. Therefore, we conclude that MetaSSPred is a more balanced secondary structure predictor.

We have observed that though our MetaSSPred provides more balanced SSP accuracies across three secondary structure classes, overall accuracies of MetaSSPred on different datasets were not significantly different from those of SPINE X. To improve further the overall accuracy of SSP without compromising the balance achieved, various measures can be taken. For example, boosting may be used while training the SVMs. It may be a good idea to use consensus SS assignment instead of DSSP assignment, since different assigning methods assign SS based on different factors, overall error in assignment might be lower if consensus assignment is used for our proposed MetaPredictor method. Some notable SS assignment methods are KAKSI [59], STRIDE [60], P-SEA [61], etc. We also recommend further investigation of the efficacy of the bigram and monograms as features along with the boosting.

ACKNOWLEDGEMENTS

MNI, SI and MTH gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19.

References:

1. Berg, J.M., J.L. Tymoczko, and L. Stryer, *Biochemistry*. 5th ed. 2002, New York: W. H. Freeman & Company.
2. Kinch, L.N. and N.V. Grishin, *Evolution of protein structures and functions.* Current Opinion in Structural Biology, 2002. **12**: p. 400-408.
3. Gu, J. and P.E. Bourne, *Structural Bioinformatics*. 2nd ed. 2009.
4. Hvidsten, T.R., et al., *A Comprehensive Analysis of the Structure-Function Relationship in Proteins Based on Local Structure Similarity.* PLoS ONE, 2009. **4**(7): p. e6266.
5. Xia, F., et al., *FPGA accelerator for protein secondary structure prediction based on the GOR algorithm.* BMC Bioinformatics, 2011. **12**(Suppl 1): p. S5.
6. Faraggi, E., et al., *SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles.* Journal of Computational Chemistry, 2012. **33**(3): p. 259-267.
7. Simossis, V.A. and J. Heringa, *Integrating Protein Secondary Structure Prediction and Multiple Sequence Alignment.* Current Protein & Peptide Science, 2004. **5**: p. 1-15.
8. Pirovano, W. and J. Heringa, *Protein secondary structure prediction.* Current Protein & Peptide Science, 2004. **5**(4): p. 249-266.
9. Voet, D. and J.G. Voet, *Biochemistry*. 4th ed. 2010: John Wiley & Sons, Inc.
10. Chopra, G., C.M. Summa, and M. Levitt, *Solvent dramatically affects protein structure refinement.* PNAS, 2008. **105**(51): p. 20239 –20244.
11. Eddy, S.R., *What is a hidden Markov model?* Nature Biotechnology 2004. **22**(10): p. 1315-1316.
12. Robson, B. and R.H. Pain, *Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Possible Implications for the Mechanism of Formation of Helical Regions.* Journal of Molecular Biology, 1971. **58**: p. 237-256.
13. Robson, B., *Analysis of the Code Relating Sequence to Conformation in Globular Proteins: Theory and Application of Expected Information.* Biochemical Journal, 1974. **141**: p. 853-867.
14. Robson, B. and R.H. Pain, *Analysis of the Code Relating Sequence to Conformation in Globular Proteins The distribution of residue pairs in turns and kinks in the backbone chain.* Biochemical Journal, 1974. **141**: p. 899-904.
15. Garnier, J., D.J. Osguthorpe, and B. Robson, *Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins.* Journal of Molecular Biology, 1978. **20**(1): p. 97-120.
16. Garnier, J., J.-F. Gibrat, and B. Robson, *GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence* Methods in Enzymology, 1996. **266**: p. 540-553.
17. Qian, N. and T.J. Sejnowski, *Predicting the Secondary Structure of Globular Proteins Using Neural Network Models* Journal of Molecular Biology, 1988. **202**: p. 865-884.
18. Rost, B. and C. Sander, *Prediction of Protein Secondary Structure at Better than 70% Accuracy.* Journal of Molecular Biology, 1993. **232**: p. 584-599.

19. Rumelhart, D.E., G.E. Hinton, and R.J. Williams, *Learning internal representations by error propagation*. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, ed. D.E. Rumelhart and J.L. McClelland. Vol. 1. 1986, Cambridge MA: MIT Press.
20. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Research, 1997. **25**(17): p. 3389–3402.
21. Przybylski, D. and B. Rost, *AlignmentsGrow,SecondaryStructurePrediction Improves.* PROTEINS: Structure, Function, and Genetics, 2002. **46**: p. 197-205.
22. Chandonia, J.-M. and M. Karplus, *New methods for accurate prediction of protein secondary structure.* . Proteins: Structure, Function, and Genetics, 1999. **35**(3): p. 293-306.
23. Jones, D.T., *Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices.* Jounal of Molecular Biology 1999. **292**: p. 195-202.
24. Bruni, R., *Mathematical Approaches to Polymer Sequence Analysis and Related Problems*. 2011, New York: Springer.
25. Kiyoshi Asai, S. Hayamizu, and K.i. Handa, *Prediction of Protein Secondary Structure by the Hidden Markov Model.* Computer Applications in the Biosciences, 1993. **9**(2): p. 141-146.
26. Bystroff, C., V. Thorsson, and D. Baker, *HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins.* Journal of Molecular Biology, 2000. **301**: p. 173-190.
27. Martin, J., J.-F. Gibrat, and F. Rodolphe, *Analysis of an optimal hidden Markov model for secondary structure prediction.* BMC Structural Biology, 2006. **6**(25).
28. Yada, T., et al., *DNA Sequence Analysis using Hidden Markov Model and Genetic Algorithm.* Genome Informatics, 1994. **5**(178-179).
29. Won, K.-J., A. Prügel-Bennett, and A. Krogh, *Training HMM Structure with Genetic Algorithm for Biological Sequence Analysis.* Bioinformatics, 2004. **20**(18): p. 3613-3619.
30. Hua, S. and Z. Sun, *A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach.* Journal of Molecular Biology, 2001. **308**: p. 397-407.
31. Guo, J., et al., *A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles.* Proteins: Structure, Function and Bioinformatics, 2004. **54**: p. 738-743.
32. Cuff, J.A. and G.J. Barton, *Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction.* PROTEINS: Structure, Function, and Genetics, 1999. **34**: p. 508-519.
33. Birzele, F. and S. Kramer, *A new representation for protein secondary structure prediction based on frequent patterns.* Bioinformatics, 2006. **22**(21): p. 2628-2634.
34. Eyrich, V.E., et al., *EVA: continuous automatic evaluation of protein structure prediction servers.* Bioinformatics, 2001. **17**(12): p. 1242-1243.
35. Kabsch, W. and C. Sander, *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, 1983. **22**: p. 2577-2637.
36. Wang, L.-H., et al., *Predicting Protein Secondary Structure by a Support Vector Machine Based on a New Coding Scheme.* Genome Informatics 2004. **15**(2): p. 181-190.
37. Cheng, J. and P. Baldi, *Three-stage prediction of proteinβ-sheets by neural networks, alignments and graph algorithms.* Bioinformatics, 2005. **21**(1): p. i75–i84.
38. Chen, W., et al., *iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition.* Nucleic Acids Res. , 2013. **41**(6): p. e68.
39. Lin, H., et al., *iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition.* Nucleic Acids Res. , 2014. **42**(21): p. 12961 - 72.
40. Ding, H., et al., *iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels.* BioMed Research International, 2014. **2014**(2014).
41. Xu, Y., et al., *iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition.* PLoS One, 2014. **9**(8): p. e105018.
42. Liu, Z., et al., *iDNA-Methyl: identifying DNA methylation sites via pseudo trinucleotide composition.* Anal Biochem, 2015. **474**: p. 69 - 77.
43. Jia, J., et al., *iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC.* J Theor Biol, 2015. **377**: p. 47 - 56.
44. Chou, K.C., *Some remarks on protein attribute prediction and pseudo amino acid composition.* J Theor Biol, 2011. **273**(1): p. 236 - 47.

45. PDB. *Protein Data Bank*. 2014 [cited 2014 September 27]; Available from: http://www.rcsb.org/pdb/secondary.do?p=v2/secondary/search.jsp#AdvancedSearch.
46. NCBI, *Using BLASTClust to Make Non-redundant Sequence Sets*, in *NCBI* 2014.
47. Meiler, J., et al., *Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks.* Journal of Molecular Modeling, 2001. **7**(9): p. 360-369.
48. Sharmaa, A., et al., *A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition.* Journal of Theoretical Biology, 2013. **320**: p. 41-46.
49. Iqbal, S. and M.T. Hoque, *DisPredict: A Predictor of Disordered Protein from Sequence using RBF Kernel.* Tech. Report TR-2014/1, 2014.
50. Marsh, J.A., *Buried and Accessible Surface Area Control Intrinsic Protein Flexibility.* Journal of Molecular Biology, 2013. **425**(17): p. 3250 - 3263.
51. Zhang, H., et al., *On the relation between residue flexibility and local solvent accessibility in proteins.* Proteins, 2009. **76**(3): p. 617 - 36.
52. Lee, B. and F. Richards, *The interpretation of protein structures: estimation of static accessibility.* J Mol Biol, 1971. **55**(3): p. 379-400.
53. Connoly, M., *Solvent accessibility surfaces of protein and nucleic acids.* Science, 1983. **221**: p. 709 - 713.
54. Iqbal, S., A. Mishra, and M.T. Hoque, *Improved Prediction of Accessible Surface Area Results in Efficient Energy Function Application.* Tech. Report TR-2015/1, 2015.
55. Zhang, T., E. Faraggi, and Y. Zhou, *Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction.* Proteins, 2010. **78**(16): p. 3353–3362.
56. Chang, C.-C. and Chih-Jen, *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011. **2**(3): p. 27:1--27:27.
57. Andersson, H.O., et al., *Optimization of P1–P3 groups in symmetric and asymmetric HIV-1 protease inhibitors.* European Journal of Biochemistry, 2003. **270**: p. 1476-1758.
58. Navia, M.A., et al., *Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1.* Nature, 1989. **337**: p. 615–620.
59. Martin, J., et al., *Protein secondary structure assignment revisited: a detailed analysis of different assignment methods.* BMC Structural Biology, 2005. **5**(17).
60. Heinig, M. and D. Frishman, *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.* Nucleic Acids Research, 2004. **1**(32): p. W500–W502.
61. G.Labesse, et al., *P-SEA:anewefficientassignmentof secondarystructurefromCOLtraceofproteins.* CABIOS, 1997. **13**(3): p. 291-295.