Journal of Bioinformatics and Computational Biology
Vol. 9, No. 1 (2011) 67–89
© Imperial College Press
DOI: 10.1142/S0219720011005355

Imperial College Press
www.icpress.co.uk

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

IMPROVED SEQUENCE-BASED PREDICTION OF STRAND RESIDUES

KANAKA DURGA KEDARISETTI*, MARCIN J. MIZIANTY†,
SCOTT DICK‡ and LUKASZ KURGAN§

Department of Electrical and Computer Engineering
University of Alberta, Edmonton, AB, Canada
**kanaka@ece.ualberta.ca*
†mizianty@ualberta.ca
‡dick@ece.ualberta.ca
§lkurgan@ece.ualberta.ca

Received 5 September 2010
Revised 19 November 2010
Accepted 19 November 2010

Accurate identification of strand residues aids prediction and analysis of numerous structural and functional aspects of proteins. We propose a sequence-based predictor, BETArPRED, which improves prediction of strand residues and β-strand segments. BETArPRED uses a novel design that accepts strand residues predicted by SSpro and predicts the remaining positions utilizing a logistic regression classifier with nine custom-designed features. These are derived from the primary sequence, the secondary structure (SS) predicted by SSpro, PSIPRED and SPINE, and residue depth as predicted by RDpred. Our features utilize certain local (window-based) patterns in the predicted SS and combine information about the predicted SS and residue depth. BETArPRED is evaluated on 432 sequences that share low identity with the training chains, and on the CASP8 dataset. We compare BETArPRED with seven modern SS predictors, and the top-performing automated structure predictor in CASP8, the ZHANG-server. BETArPRED provides statistically significant improvements over each of the SS predictors; it improves prediction of strand residues and β-strands, and it finds β-strands that were missed by the other methods. When compared with the ZHANG-server, we improve predictions of strand segments and predict more actual strand residues, while the other predictor achieves higher rate of correct strand residue predictions when under-predicting them.

Keywords: Strand residues; beta-strands; secondary structure; prediction.

1. Introduction

Protein secondary structure (SS), which includes helix, strand and coil states, concerns spatially local structures formed through hydrogen bonds between backbone atoms. The last three decades have seen intense research in the sequence-based prediction of the SS.[1] In the last 15 years, state-of-the-art three-state predictive accuracy (Q_3) improved from about 70%[2] to over 80%.[3] Recent SS predictors employ

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

a variety of machine learning-based models such as neural networks, support vector machines, and regression. They can be categorized into standalone methods and ensembles that combine multiple SS predictors. A majority of the standalone predictors are based on different types of neural networks, including PHD,[4] PSIPRED,[5,6] SABLE,[7] SSpro,[8] YASPIN,[9] PORTER,[10] and SPINE.[11] Their Q_3 is relatively high and ranges between 73% and 78% on the benchmark EVA5 dataset.[9,12] The ensemble predictors include CoDe,[13] PROTEUS,[3,14] and CDM,[15] and they achieve Q_3 of up to 89.9% on their test datasets.[3]

The above methods attempt to solve the general three-state prediction problem; however, recent research shows that predicting specific SS types, such as coil types including β- and γ-turns,[16,17] also produces high-quality results. Empirical analysis of two SS predictors, YASPIN and PORTER, reveals that their Q_E values (accuracy of strand predictions) are lower than Q_H (accuracy of helix predictions) by 7 to 16 percentage points.[9,10] Jones *et al.* show that binary classification of strand versus non-strand residues (either coils or helices) is characterized by lower improvement over a baseline than the binary classification of helices or coils.[18] Furthermore, fragments of protein sequence that fold into strands are characterized by numerous patterns with respect to the occurrence of certain amino acid types, which were investigated in numerous studies over the last 30 years,[19–23] and which could be exploited to build effective predictors. At the same time, accurate identification of strand residues aids numerous applications including predictions of β-sheets[24–26] and tertiary structure,[27] elucidation of protein folding pathways,[28] protein design,[29] characterization of super-secondary structures and protein folding patterns,[30] and in investigations of certain mechanisms causing neurodegenerative diseases.[31] There is thus a clear need for methods that accurately predict strand residues.

Virtually all modern SS predictors, including PSIPRED, SSpro, PORTER, and PROTEUS, exploit local information in the sequence using a windowing approach to compute their predictions. Their designs imply independence between positions in the window, i.e. the predictions are based on neighboring AAs but do not exploit relations between them. While this is acceptable when considering the AA sequence, windowing the predicted SS sequence (e.g. in the second stage of PRIPRED), loses vital information. This recently prompted development of a method that post-processes predicted SS,[32] and it inspires the development of our feature set. We also note that certain residue characteristics, such as burying depth, that can be predicted relatively accurately from the sequence[33,34] and have not been considered by existing SS predictors, could provide valuable predictive input. More specifically, recent analysis shows that helices are about three times more abundant on the protein surface when compared with strands, while their abundance in the protein core is comparable, and twice as high compared to coils.[33]

We propose the BETArPRED predictor, which tackles the binary classification problem of predicting strand residues from protein sequences. Our approach is motivated by recent works demonstrating that ensemble-based SS predictors

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

outperform standalone solutions.[3,14,35] Similarly, combining multiple predictors results in improvements in related predictive efforts, including prediction of protein fold types,[36,37] structural classes,[38] quaternary structure type,[39] transmembrane helices,[40] and disorder,[41–43] to name a few. Encouraged by the success of the ensemble-based predictors and the fact that correlations between neighboring SSs are stronger than that between neighboring residues,[44,45] we use SS predicted by SSpro, PSIPRED and SPINE as our inputs. BETArPRED also uses residue depth predictions computed with RDpred[34] and sequence-derived information to generate a small set of nine features that are fed into a logistic regression classifier. The features utilize local (window-based) patterns in the predicted SS to exploit relations between adjacent residues. They further combine information about the predicted SS and residue depth, and consider global (sequence-wide) information concerning chain length.

2. Methods

2.1. *Datasets*

We extracted a large dataset of low-similarity protein chains, which were deposited into PDB[46] between January 2007 and December 2008, to design and test the proposed method. We selected recent depositions to remove potential bias with respect to templates in the base SS prediction methods utilized in our solution. We further filtered these proteins to consider only chains with high-quality structures; those determined using X-ray crystallography with resolution <2.5 Å and R-value <0.25. As in Cheng and Baldi[24] and Lippi and Frasconi[25] we retained the sequences that have at least 50 residues and contain at least 10% strand residues. Next, using CD-hit[47] we reduced the sequence similarity within the dataset by selecting a subset of chains that has pairwise sequence identity <40%. We additionally removed any sequence that has >25% similarity to the sequences deposited in PDB before January 2007 using pairwise identity computed by BLAST. The final dataset consists of 861 protein sequences. For each sequence we annotated strand residues using DSSP.[48] The dataset was randomly divided into two subsets, the TRAINING and the TEST sets. The TRAINING dataset contains 429 protein sequences (103,390 residues and 25,697 strand residues), which are used to design and train the predictive model using five-fold cross validation. We chose this type of the test which randomizes the selection of the five folds, instead of the jackknife cross validation which leads to a unique non-randomized result[49,50] and which was recently used in related works,[51–66] to reduce the computational time. The TEST dataset contains 432 sequences (106,405 residues and 25,648 strand residues) and is used to determine the out-of-sample prediction quality of BETArPRED. We further evaluate BETArPRED on targets from the most recent CASP8 competition,[67] excluding three targets which we could not process using DSSP and another seven for which the predictions of the top-performing tertiary structure predictor in CASP8[68] were missing or could not be processed by the DSSP. The CASP8 dataset thus includes

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

111 sequences (22,875 residues and 5,358 strand residues). The datasets are available at http://biomine.ece.ualberta.ca/BETArPred/BrP.htm.

2.2. *Evaluation measures*

The performance of BETArPRED is assessed using measures that quantify prediction quality at the residue and the segment (β-strand) levels. The residue-level measures include:

$$\text{Acc (accuracy)} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}),$$

$$\text{Q}_{\text{e\_obs}}\ \text{(sensitivity)} = (\text{TP})/(\text{TP} + \text{FN}),$$

$$\text{Q}_{\text{e\_pred}} = (\text{TP})/(\text{FP} + \text{TP}),$$

where TP (true positives) is the number of correctly predicted strand residues, TN (true negatives) is the number of correctly predicted non-strand residues, FP (false positives) is the number of non-strand residues incorrectly predicted as strand residues, and FN (false negatives) is the number of strand residues incorrectly predicted as non-strand residues. As in Lin *et al.*[9] and McGuffin and Jones,[69] we also compute four quality measures that quantify different types of prediction errors, see Fig. 1. Over-prediction error, O_e, is defined as the number of FP residues where the entire segment of the predicted strand residues (β-strand) does not overlap with the actual strand residues. Similarly, under-prediction error, U_e, quantifies the number of FN residues where none of the residues in the entire actual β-strand is correctly predicted. The length error, L_e, represents the total number of FN residues and FP residues where some of the predicted strand residues overlap with an actual β-strand and where the incorrect predictions form a segment that extends to a terminus of the actual β-strand. The inner-segment error, W_e, is defined as the number of FN residues which are inside an actual β-strand, i.e. the segment of these incorrect predictions does not extend to a terminus of the actual β-strand.

The segment-level measures include segment overlap (SOV_e) scores[70] for β-strands, and average strand segments coverage:

$$ASSC = \frac{\sum_{i=1}^{N} \frac{S_{io}}{S_{ia}}}{N},$$

where S_{io} is the number of predicted strand residues that overlap with an actual β-strand S_i, S_{ia} is the number of residues in the actual β-strand S_i, and N is the total number of β-strands in the dataset.

```
-----EEEE-----EEEE------EEEEEE----EEEE-   actual (native) structure
-EE---EEEE------EEEEE--EE--EE----------   prediction
-OO--LEEEL----LLEELLL--LEWWEEL----UUUU-   prediction errors
```

Fig. 1. Illustration of four types of prediction errors. The top line gives the true positions of strand residues (E) and non-strand residues (-), the middle line shows a prediction, and the bottom line annotates the errors using bold font where the over-prediction, under-prediction, length, and inner-segment errors, are denoted with O, U, L, and W, respectively.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

We also compute the $Q_{\text{h\_obs}}$, $Q_{\text{h\_pred}}$, $Q_{\text{c\_obs}}$, and $Q_{\text{c\_pred}}$ which evaluate the prediction of helix and coil residues, Q_3 to quantify the overall 3-state secondary structure predictions, as well as SOV_3 (for the 3-state secondary prediction), SOV_h, and SOV_c measures. The measures are consistent with the measures applied in the EVA platform.[71]

2.3. *Secondary structure predictions*

We considered the key SS predictors listed by Rost[72] which include PORTER, PSIPRED, SSpro, SABLE, and YASPIN, as well as two recent predictors, SPINE and PROTEUS2. PSIPRED is widely applied in prediction of various structural properties such as solvent accessibility,[73] fold,[37] structural class,[74] outer membrane beta barrel protein types,[75] folding rate,[76] and β- and γ-turns,[16,17] to name a few. PROTEUS2 is a recent ensemble method that was selected due to its reported favorable performance when compared with eight competing SS predictors.[3] YASPIN was reported to provide high quality predictions of strand residues.[9] PORTER (the standalone version provided at http://distill.ucd.ie/porter/) and SSpro 4.0 were selected due to their strong performance on the EVA server.[71] We computed Acc, $Q_{\text{e\_obs}}$, $Q_{\text{e\_pred}}$, SOV_3, SOV_h and SOV_e values on the TRAINING set for each of the seven predictors, see Table 1. We select the three methods with highest accuracy (SSpro, PSIPRED, and SPINE) for BETArPRED. These methods also have high SOV_e, SOV_h, SOV_3 and $Q_{\text{e\_pred}}$ values, while their $Q_{\text{e\_obs}}$ is also relatively large. SABLE and PORTER have low $Q_{\text{e\_obs}}$, while PROTEUS and YASPIN over-predict strand residues, leading to low $Q_{\text{e\_pred}}$. The selected predictors have $\text{SOV}_\text{e} < \text{SOV}_\text{h}$, again showing that helix residues are better predicted than strands.

2.4. *Overall design*

Our preliminary experiments indicated that a simple ensemble of SS predictions yields a model that tends to mimic the strongest base method, and provides only marginal improvements. Instead, we propose a novel type of ensemble in which we accept the strand residue predictions of the strongest base method and (re)predict the remaining residues. Table 1 shows that SSpro has the highest $Q_{\text{e\_pred}}$ values.

Table 1. The quality of the SS predictions on the TRAINING dataset for the seven considered SS predictors. The methods are sorted by Acc.

SS predictor	Acc	$Q_{\text{e\_obs}}$	$Q_{\text{e\_pred}}$	SOV_e	SOV_h	SOV_3
SSpro	89.02	70.49	82.64	74.76	80.78	77.33
PSIPRED	88.71	73.76	79.24	75.49	80.25	77.51
SPINE	88.68	72.01	80.27	75.51	80.02	77.23
SABLE	88.31	68.60	81.29	74.14	78.46	76.48
PROTEUS	87.95	82.42	72.65	78.89	79.24	77.51
PORTER	87.03	66.82	77.74	71.37	78.34	76.08
YASPIN	85.57	72.67	70.18	73.00	76.21	73.41

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

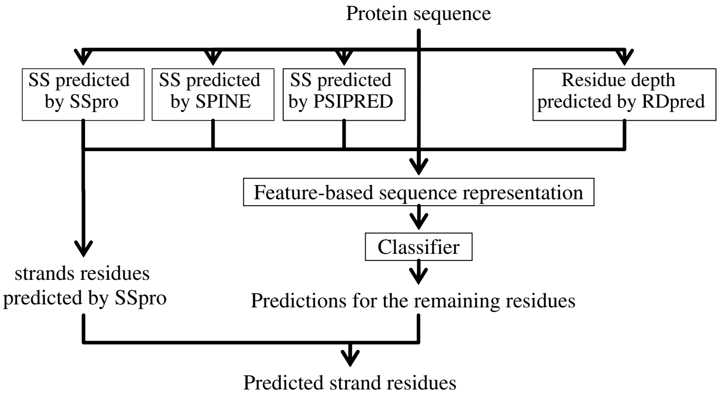

Protein sequence
SS predicted by SSpro
SS predicted by SPINE
SS predicted by PSIPRED
Residue depth predicted by RDpred
Feature-based sequence representation
Classifier
strands residues predicted by SSpro
Predictions for the remaining residues
Predicted strand residues

Fig. 2. The overall design of the proposed prediction method.

Only about 17% of its strand residue predictions are incorrect. However, 29% of the actual strand residues are missed by SSpro, and our ensemble is designed to find them. The overall design of the proposed method is shown in Fig. 2. The input protein sequence is fed into SSpro, SPINE and PSIPRED to obtain predicted SS. The strand residues predicted by SSpro are passed to the final prediction. The predicted SS, residue depth predicted with RDpred method,[34] and the sequence itself are used to compute a feature vector for the remaining residues. These features combine both local (window-based) and global (sequence-based) information from these sources. The feature vector is passed to a classifier, and the predicted strand residues are merged with the predictions from SSpro. This design is empirically compared against a typical ensemble that predicts all residues to demonstrate the benefits of the proposed design in Sec. 2.6.

2.5. *Features*

We employ features generated at three levels: the predicted residue itself (raw values), from a local window centered over the predicted residues (aggregated local information), and the entire protein sequence (aggregated global information). Window- and sequence-level features are used because strand residues form β-strand segments (formation of β-strands involves local interactions) and since β-strands form β-sheets that involve strand segments that may be dispersed over the entire sequence. The features are obtained from three sources: the sequence, the predicted SS, and the residue depth predicted with RDpred. RDpred predicts three types of residue depths: two distance-based depths based on the MSMS[77] and DPX[78] methods, and a volume-based depth based on the SADIC algorithm.[79] We employ all three depth predictions as they are complementary. Specifically, the absolute correlations between these depth predictions range between 0.63 and 0.77.[34] In total we extract 214 features. The first two letters in the prefixes of the feature names indicate the information level (r, w, and p denote residue, window,

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

and sequence-level features, respectively) and the type of the aggregation (a and s correspond to no aggregation (values for individual amino acids are used), and to aggregation of the predicted SS (SS segments or SS states), respectively).

2.5.1. *Residue-level features*

For each residue in a given sequence, the following nine features are computed:

- *r_a_from_N* and *r_a_from_C* quantify linear distance from the N- and C-termini of a sequence, respectively (two features).
- $r\_a\_ss_i$ is the SS predicted using the ith method, where $i = \{$PSIPRED, SSPro, SPINE$\}$ (three features).
- *r_a_score* is the PSIPRED reliability score (one feature). The other two SS predictors do not provide reliability scores.
- $r\_a\_depth_j$ is the depth predicted by RDpred using the jth definition, where $j = \{MSMS, DPX, SADIC\}$ (three features).

2.5.2. *Window-level features*

One hundred and seventy-four features are computed for each residue in a given protein sequence using a local widow. The maximal window size that we use is nine (four residues on each side of the predicted residue). This size was selected since our previous work suggests that formation of strands appears to be affected by residues within three positions in the sequence.[80] We extended the resulting seven residues-wide window to include one more position, assuming that feature selection (which is described in the next subsection) will remove features that are irrelevant. Among the 174 features, 63 are generated using the predicted residue depths (in some cases combined with the predicted SS):

- $w\_a\_depth_j\_frag_s$ is the average predicted depth according to the jth depth predictor in windows of size $s = \{3, 5, 9\}$ centered on the predicted residue ($3 \times 3 = 9$ features).
- $w\_s\_m_i\_avgdepth\_state_k\_depth_j$ is the average depth predicted for residues in kth $= \{h, e, c\}$ SS state in the window of size nine using the SS predictions of the ith method and the jth depth definition ($3 \times 3 \times 3 = 27$ features). A value of -1 is used when a given SS state is not predicted in the window.
- $w\_s\_m_i\_avgdepth\_seg_l\_depth_j$ is the average predicted depth for the SS segment that includes the predicted residue, where $l = \{h, e, c\}$ is the SS type of the segment extracted from the SS predicted by the ith method and where the jth depth definition is used. The values for the remaining two SS types, l, for a given i and j are set to -1. ($3 \times 3 \times 3 = 27$ features).

Another 87 features quantify composition of the predicted SS:

- $w\_s\_m_i\_state_k$ is the count of the residues in the kth SS state in a window of size nine using the SS from the ith method ($3 \times 3 = 9$ features).

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

- $w\_s\_m_i\_state_k\_norm\_len$ is the count of the residues in the kth SS state in a window of size nine, normalized by the window size, and using the SS predicted by the ith method ($3 \times 3 = 9$ features).
- $w\_s\_m_i\_dipep_m$ is the count of the mth SS dipeptide in a window of size nine using the SS predicted by the ith method, where $m = \{hh, ee, cc, hc, ec, ch, ce\}$ ($7 \times 3 = 21$ features). We do not consider the dipeptides where strand residues are next to helix residues, since they very rarely occur naturally, if at all.
- $w\_s\_m_i\_tripep_n\_central\_res$ is the binary feature that denotes whether the predicted residue is in the nth SS tripeptide conformation (centered on the predicted residue) using the SS predicted by the ith method, where $n = \{hhh, hcc, cch, hhc, chh, hch, eee, ecc, cce, cec, eec, cee, ece, ccc, ech, hce\}$ ($16 \times 3 = 48$ features). We do not consider the tripeptides where strand residues are next to helix residues.

The following nine features utilize the reliability scores for the SS predicted by PSIPRED:

- $w\_s\_m_{PSIPRED}\_avg\_rel\_score\_state_k$ is the average reliability score in a window of size nine for the kth SS state (three features).
- $w\_s\_m_{PSIPRED}\_max\_rel\_score\_state_k$ is the maximal reliability score in a window of size nine for the kth SS state (three features).
- $w\_s\_m_{PSIPRED}\_min\_rel\_score\_state_k$ is the minimal reliability score in a window of size nine for the kth SS state (three features).

The next nine features quantify the number and size of predicted SS segments:

- $w\_s\_m_i\_max\_seg\_len$ and $w\_s\_m_i\_min\_seg\_len$ are the maximal and minimal length of SS segments in a window of size nine using the SS predicted by the ith method ($3 + 3 = 6$ features).
- $w\_s\_m_i\_seg\_number$ is the number of the SS segments in a window of size nine using the SS predicted by the ith method (three features).

The final six features quantify the position of the predicted residue with respect to the predicted SS segment that includes this residue:

- $w\_s\_m_i\_max\_interface\_distance$ and $w\_s\_m_i\_min\_interface\_distance$ are the maximal and minimal distances between the position of the predicted residue and the two termini of the SS structure segment that includes this residues using the SS predicted by the ith method ($3 + 3 = 6$ features).

2.5.3. *Sequence-level features*

A total of 31 features are computed by exploring the entire protein sequence:

- $p\_a\_chain\_len$ is the length of the protein sequence (one feature).
- $p\_s\_m_i\_segs$ is the number of the SS segments predicted by the ith method (three features).

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

- $p\_s\_m_i\_seg_l\_norm\_len$ and $p\_s\_m_i\_seg_l\_norm\_total$ are the counts of the SS segments of lth type using the SS predicted by the ith method, normalized by the chain length and by the total number of SS segments in the chain, respectively ($3 \times 3 + 3 \times 3 = 18$ features).
- $p\_s\_m_i\_Eseg\_+/-1$, $p\_s\_m_i\_Eseg\_+/-2$ and $p\_s\_m_i\_Eseg\_+/-3$ are the counts of β-strands that are of length of up to $+/-1$, $+/-2$, and $+/-3$ residues, respectively, when compared with the length of the β-strand that includes the predicted residue using the SS predicted by the ith method. These features are set to -1 when the predicted residue in not in a β-strand ($3 + 3 + 3 = 9$ features).

The $p\_s\_m_i\_seg_l$ features bias the ensemble when predicting sequences that are rich in certain types of SS conformations. The nine $p\_s\_m_i\_Eseg$ features exploit the fact that β-sheets include up to several β-strands with similar segment lengths.

2.6. *Feature and classifier selection*

Empirical feature selection will identify a subset of features that are effective in predicting strand residues. At the same time, we require a classifier with favorable predictive quality. These tasks were performed using the WEKA[81] and LIBLINEAR[82] software packages. We considered two feature selection strategies: a filter-based method in which feature sets are evaluated by their "association" with the prediction outcomes, and a wrapper-based method in which feature sets are assessed based on prediction quality using a given classification method.[83]

We applied two filter-based methods, consistency-based (CONS)[84] and correlation-based (CFS).[85] The CONS method uses a ratio between the number of inconsistent versus total number of data samples (residues) when the input data are projected onto a given subset of features. Samples are considered inconsistent if they have the same feature values and different predictions. The CFS method defines a ratio between a correlation-based estimate of the predictive value of a given feature set and its estimated redundancy. These two methods were shown to reduce the dimensionality of the feature vector while maintaining or improving prediction quality.[84,85] For efficiency, we used best-first search with forward feature selection to search through the space of the feature sets. We used these two selection methods on the TRAINING dataset using five-fold cross validation and we combined the features selected in each fold together. We also took the union and intersection of these two feature sets (denoted UNION and INTER, respectively).

The wrapper-based selection was performed simultaneously with classifier selection. We consider three classifiers: logistic regression (LOG),[86] a normalized Gaussian radial basis function (RBF) network,[87] and a linear-kernel based Support Vector Machine (SVM).[82] The RBF network requires setting the number of clusters, k, and we use two variants with $k = 1$ and $k = 2$, referred to as RBF(1) and RBF(2), respectively. We also parameterized the value of complexity constant C for the SVM for each of the feature sets using five-fold cross validation. As with the

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

filter-based selection, we used best-first search to generate feature subsets that were inputted into the four classifiers, LOG, RBF(1), RBF(2), and SVM. Each of the feature sets was evaluated on the TRAINING dataset using five-fold cross validation. We evaluate the classifiers using three indices: Accuracy (ACC), average of Q_{e_pred} and Q_{e_obs} (AVG), and SOV_e. Consequently, we have three feature sets for each classifier.

Next, we used the same four classifiers to compare predictive quality of all selected feature sets (four selected using filter-based methods and three from using the wrapper-based method). Each of the 28 experiments (4 classifiers × 7 sets) used the five-fold cross validation on the TRAINING dataset. Additionally, we repeated the same procedure with a standard ensemble, in contrast to the proposed design that accepts strand residues predicted by SSpro and predicts the remaining residues. The complete results are given in Table A.1 in the Supplementary Materials. Table 2 compares the two best models, with the highest accuracy and the highest SOV_e, for both ensemble configurations against the three base SS predictors. Table A.1 shows that two solutions attained highest accuracy for the proposed design and we chose the solution with higher Q_{e_pred}. Table 2 reveals that the best results, in terms of both high accuracy and SOV_e, are obtained by the LOG classifier and wrapper-based feature selection evaluated using accuracy (the last row in Table 2). This feature set includes only nine features (details are shown in Sec. 3.2). We also compared the two best models for the proposed and the alternative designs on the TEST and CASP8 datasets. The results, which are summarized in Table A.2 in the Supplementary Materials, confirm that the chosen ensemble provides favorable predictive quality as measured by accuracy, SOV_e and the best trade-off between Q_{e_obs} and Q_{e_pred}. Thus, the proposed BETArPRED method uses the strand residues predicted by SSpro and predicts the remaining residues utilizing the LOG classifier and the nine features.

Table 2. Results of five-fold cross-validation on the TRAINING dataset for the two best performing feature sets, according to accuracy and SOV_e, using the proposed design (by taking strand residues predicted by SSpro and predicting the remaining positions), the alternative design that predicts all residues, and for the PSIPRED, SSpro and SPINE. The proposed/alternative design rows encode the classifiers (SVM, RBF(1), and LOG) and feature selections (SOV_e, AVG, and ACC) used.

Predictor		Acc	SOV_e	Q_{e_obs}	Q_{e_pred}
SSpro		89.02	74.76	70.49	82.64
PSPRED		88.71	75.49	73.76	79.24
SPINE		88.68	75.51	72.01	80.27
Alternative design	SOV_e + SVM	55.39	79.34	82.73	33.69
	AVG + RBF(1)	89.54	77.31	75.62	80.92
Proposed design	SOV_e + SVM	55.59	80.27	84.64	34.06
	ACC + LOG	89.51	78.19	76.63	80.15

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

3. Results

3.1. *Comparison with related methods*

Our predictions are assessed using residue (Acc, Q_{e_obs}, Q_{e_pred}, O_e, U_e, L_e, and W_e) and β-strand segment quality indices (ASSC and SOV_e). We compare BETArPRED with the seven SS predictors on the TEST and CASP8 datasets. For the CASP8 dataset we also include the best automated 3D structure predictor from the CASP8 competition,[68] the ZHANG-server, with the predicted structure processed using DSSP to obtain the positions of beta residues. We include results on the entire CASP8 dataset and also on its two subsets that include sequences with at least one strand residue and 10% of strand residues, respectively. This is because most of the quality indices (Q_{e_obs}, U_e, L_e, W_e, ASSC, and SOV_e) could not be measured for chains with no strand residues and they may provide statistically unreliable estimates when the number of strand residues is low. In particular, for chains with no strand residues they would default to zero and cannot quantify how many strand residues are incorrectly predicted. The complete results are given in Table A.3 in the Supplementary Materials. Table 3 summarizes results on the TEST set (chains with at least 10% strand residues) and the CASP8 set (again, chains with at least 10% strand residues) and gives statistical significance of improvements on both datasets. We compared results for individual proteins between BETArPRED, each of the seven SS predictors, and the ZHANG-server. When a given quality measure for both predictors is normally distributed (per the Shapiro–Wilk test of normality with $p < 0.05$), we applied the paired t-test and otherwise we used the Wilcoxon rank sum test. Table A.4 in the Supplementary Materials provides these results for different versions of the CASP8 dataset.

BETArPRED achieves the highest SOV_e and accuracy on the TEST dataset. The ASSC, SOV_e, Q_{e_obs}, and U_e of BETArPRED are statistically significantly better at 0.05 when compared with six out of the seven SS predictors. When compared with the remaining PROTEUS which over-predicts strand residues, BETArPRED significantly improves Q_{e_pred}, accuracy, L_e and W_e. The results on the CASP8 confirm these findings. We note statistically significant improvements in SOV_e, ASSC, Q_{e_obs}, and U_e. Overall, the results indicate that BETArPRED better predicts individual strand residues (highest accuracy among SS predictors on both CASP8 and TEST sets) as well as β-strands (highest SOV_e, except for YASPIN on the CASP8 set). Importantly, the low values of the U_e, which are significantly lower than most of the other predictors including SSpro, demonstrate that our method finds β-strands that were missed by other methods. When compared with the ZHANG-server, our method significantly improves prediction of strand segments (as measured by ASSC and SOV_e) and is inferior in the context of prediction of strand residues. We note that ZHANG-server under-predicts strand residues and these predictions have high quality, while BETArPRED finds significantly more actual strand residues (higher Q_{e_obs}).

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

Table 3. Summary of results of the BETArPRED and the seven representative SS predictors on the TEST dataset and the CASP8 dataset with chains that include at least 10% of β-residues. Results on the CASP8 dataset also include the top-performing automated 3D predictor, ZHANG-server. The second and the last sets of rows report results of the statistical significance tests which compare the proposed BETArPRED against the seven SS predictors (and ZHANG-server on the CASP8). The "– – –"/"– –"/"–" means that BETArPRED is worse with $p < 0.02/0.05/0.1$, the "+++"/"++"/"+" means that BETArPRED is better with $p < 0.02/0.05/0.1$, and "=" denotes that the BETArPRED and the other methods are not significantly different.

Dataset/test type	Predictor	ASSC	SOV_e	Q_{e_obs}	Q_{e_pred}	Acc	O_e	U_e	L_e	W_e
TEST	BETArPRED	76.20	79.46	76.60	78.95	89.41	1.46	2.22	6.89	0.02
	SSpro	70.58	75.72	71.08	82.03	89.25	1.16	2.86	6.69	0.03
	PSIPRED	72.13	75.53	72.97	77.96	88.49	1.29	3.13	7.07	0.03
	SPINE	70.00	74.87	70.71	79.33	88.48	1.31	2.99	7.17	0.06
	SABLE	67.01	73.37	67.43	79.45	87.92	1.24	3.42	7.41	0.02
	PROTEUS	80.23	77.68	79.99	71.59	87.50	1.35	2.16	8.64	0.09
	PORTER	65.01	70.90	66.17	76.97	87.05	1.47	2.57	7.53	0.02
	YASPIN	73.00	73.16	72.81	68.32	85.28	2.71	3.45	8.55	0.01
Significance of the differences when compared to BETArPRED on the TEST set	SSpro	+++	+++	+++	– – –	=	– – –	+++	=	=
	PSIPRED	+++	+++	+++	=	+	– –	+++	=	=
	SPINE	+++	+++	+++	=	+++	=	+++	=	+++
	SABLE	+++	+++	+++	=	+++	=	+++	+	=
	PROTEUS	– – –	=	– – –	+++	+++	=	=	+++	+++
	PORTER	+++	+++	+++	=	+++	– –	++	+	=
	YASPIN	+++	+++	+++	+++	+++	+++	+++	+++	– – –
CASP8 99 chains with at least 10% strand residues	BETArPRED	76.15	80.15	76.23	80.22	89.05	1.43	2.39	7.19	0.04
	ZHANG-server	67.95	72.77	68.35	90.47	90.06	0.46	3.01	6.43	0.04
	SSpro	70.60	76.63	70.97	83.30	88.99	0.96	2.99	7.01	0.05
	PSIPRED	72.47	75.30	73.37	77.60	87.84	1.34	3.27	7.54	0.02
	SPINE	71.76	76.53	71.86	80.59	88.43	1.28	3.06	7.15	0.07
	SABLE	67.63	75.00	67.95	80.43	87.64	1.32	3.52	7.50	0.01
	PROTEUS	75.35	74.98	75.95	72.70	86.62	1.50	3.17	8.69	0.02
	PORTER	63.51	69.70	64.19	75.26	85.52	1.48	3.73	9.25	0.02
	YASPIN	79.16	80.33	79.38	73.84	87.60	1.77	2.74	7.88	0.01
Significance of the differences when compared to BETArPRED on the CASP8 set with 99 chains	ZHANG-server	++	+	++	– – –	– –	– – –	=	– –	=
	SSpro	++	+	+++	–	=	=	=	=	=
	PSIPRED	=	++	=	=	=	=	+++	=	=
	SPINE	+	+	+	=	=	=	++	=	=
	SABLE	+++	+++	+++	=	=	=	+++	=	=
	PROTEUS	+++	++	=	+++	+++	=	+++	=	=
	PORTER	+++	+++	+++	+	+++	=	+++	=	=
	YASPIN	=	=	=	+++	+	=	=	=	–

3.2. *Comparison of 3-state secondary structure predictions*

The outputs of BETArPRED are combined with the predictions from SSpro, which obtained the highest accuracy on the training dataset (see Table 1), to generate the three-state secondary structure predictions. More specifically, we predict strands for all residues predicted by BETArPRED as strands and we use predictions from SSpro for the non-strand residues predicted by BETArPRED. Since by design BETArPRED predicts all strand residues predicted by SSpro as strands, the

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

SSpro predictions for the non-strand residues predicted by BETArPRED are either coil or helix residues. Our objective is to evaluate the impact of the improved strand residue and segment predictions provided by BETArPRED on the prediction of the other two secondary structure states. We compare these three-state predictions with the corresponding predictions produced by the considered secondary structure predictors on the TEST and the CASP8 sets, see Table 4. The results show that the improved prediction of the strand residues provided by BETArPRED does not have a detrimental effect on the prediction of helix and strand residues. The overall three-state predictive quality measured using Q_3 and SOV_3 for the three-state secondary structure generated using predictions from BETArPRED is the highest for both datasets. A direct comparison between the three-state predictions generated by SSpro and the SSpro predictions augmented using the BETArPRED outputs demonstrates that the latter increases both Q_3 and SOV_3 values on the TEST and CASP8 datasets. We observe a small decrease in the SOV_h, similar SOV_c, and substantially improved SOV_e values when comparing the SSpro and the BETArPRED-based predictions.

3.3. *Analysis of the selected features*

Table 5 lists the features used by BETArPRED. They utilize all considered input predictions at all three information levels. The features use the residue-level SS predicted by PSIPRED and SPINE, local information extracted from the SS predicted by PSIPRED, SPINE and SSpro, a combination of the local predicted SS and residue depth quantified using both volume and distance based definitions, and sequence-level information concerning the chain length. Figure 3 visualizes values of two pairs of these features. Both plots show how a given combination of a predicted depth-based feature with a feature that utilizes predicted SS is helpful in annotation of the strand versus non-strand residues. Note that "predicted SS" that defines features on the x-axis comes from a different predictor than for the y-axis. When the predicted SS of the residue is a strand (x-axis in the bottom panel) or when this residue is located inside a strand segment (x-axis in the top panel), the values of the average depth of the predicted helical conformations in the vicinity of this residue (y-axes) provide evidence on its proper classification. If there are no predicted helices (-1 on the y-axis), then it is most likely a strand conformation (the marker is green). The higher the average depth of the predicted helices (shown on the y-axis), the smaller the likelihood that our prediction should be a strand (the marker is more red). This agrees with the underlying biology, as it is more likely that the predicted helical conformation is correct if its depth is higher.[33]

3.4. *Case study*

We selected the galactose mutarotase related enzyme Q5FKD7 (PDBid 3DCD) among the CASP8 targets to demonstrate predictions of our method. This chain contains about 45% strand residues with several short and longer segments. Figure 4

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

Table 4. Summary of results for the three-state secondary structure predictions generated by combining predictions of BETArPRED with SSpro and the seven representative SS predictors on the TEST dataset and the CASP8 dataset with chains that include at least 10% of β-residues. We predict strands for all residues predicted by BETArPRED as strands and we use predictions from SSpro for the non-strand residues predicted by BETArPRED to obtain the three-state secondary structure predictions. Results on the CASP8 dataset also include the top-performing automated 3D predictor, ZHANG-server.

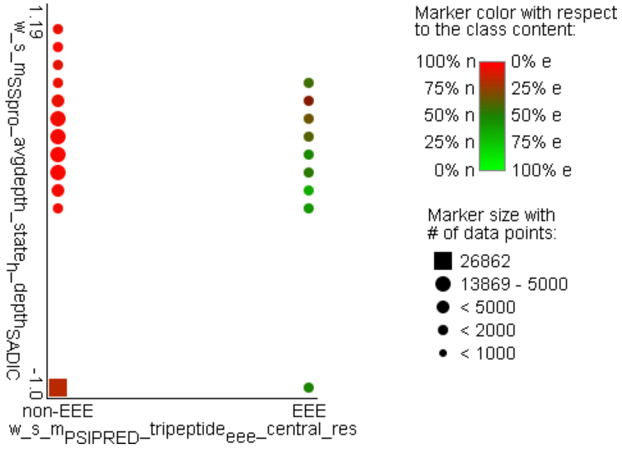
Dataset	Predictor	Q_3	SOV_3	SOV_h	SOV_e	SOV_c	Q_{h_obs}	Q_{h_pred}	Q_{e_obs}	Q_{e_pred}	Q_{c_obs}	Q_{c_pred}
TEST	BETArPRED	80.45	77.56	81.27	79.46	73.70	84.91	85.30	76.60	78.95	78.85	77.17
	SSpro	80.12	77.07	81.76	75.72	73.73	85.25	84.56	71.08	82.03	81.04	75.56
	PSIPRED	78.73	76.68	80.25	75.53	72.16	85.47	82.00	72.97	77.96	76.28	76.19
	SPINE	78.82	76.32	80.59	74.87	72.10	84.56	82.89	70.71	79.34	78.65	75.09
	SABLE	77.87	75.90	79.01	73.37	72.40	81.60	84.70	67.43	79.45	80.84	72.04
	PROTEUS	78.48	77.08	78.57	77.68	71.27	84.21	86.40	79.99	71.59	72.58	76.22
	PORTER	76.81	75.50	79.22	70.91	71.91	81.90	82.75	66.17	76.97	78.72	72.04
	YASPIN	75.20	73.56	77.03	73.16	68.56	80.94	81.00	72.81	68.32	71.61	74.48
CASP8 99 chains with at least 10% strand residues	BETArPRED	80.15	78.83	80.16	80.15	73.57	84.61	84.98	76.23	80.22	78.96	76.20
	ZHANG-server	78.14	73.36	78.51	72.77	68.96	85.90	77.42	68.35	90.47	78.19	73.43
	SSpro	80.10	78.16	81.92	76.63	73.82	85.92	84.23	70.97	83.30	80.99	75.20
	PSIPRED	77.93	77.06	78.88	75.30	71.70	83.81	81.82	73.37	81.82	75.85	74.83
	SPINE	78.71	76.77	78.88	76.53	72.21	83.20	83.16	71.86	80.59	79.29	74.19
	SABLE	77.54	75.94	77.65	75.00	71.60	79.85	85.46	67.95	80.43	81.75	70.74
	PROTEUS	76.38	74.13	73.08	74.98	67.94	81.22	83.27	75.95	72.70	72.57	73.09
	PORTER	74.43	73.29	75.00	69.70	68.80	80.44	80.95	64.19	75.26	75.85	69.00
	YASPIN	78.26	77.39	80.75	80.33	69.99	84.96	82.83	79.38	73.84	71.84	77.25

Table 5. Features used by the BETArPRED.

Feature name	Description
r_s_ss$_{\text{PSIPRED}}$ r_s_ss$_{\text{SPINE}}$	Residue-level predicted SS by PSIPRED and SPINE
w_s_m$_{\text{PSIPRED}}$_tripep$_{\text{eee}}$_central_res w_s_m$_{\text{PSIPRED}}$_tripep$_{\text{ece}}$_central_res w_s_m$_{\text{SPINE}}$_tripep$_{\text{ece}}$_central_res w_s_m$_{\text{SSpro}}$_tripep$_{\text{cch}}$_central_res	Local predicted by PSIPRED, SPINE, and SSpro SS of tripeptides, including EEE, ECE, and CCH combinations, centered on the predicted residue
w_s_m$_{\text{PSIPRED}}$_avgdepth_seg$_{\text{h}}$_depth$_{\text{MSMS}}$ w_s_m$_{\text{SSpro}}$_avgdepth_state$_{\text{h}}$_depth$_{\text{SADIC}}$	Local average predicted depth of the predicted helix residues and helical segments predicted by PSIPRED and SSpro
p_a_chain_length	Sequence-level chain length

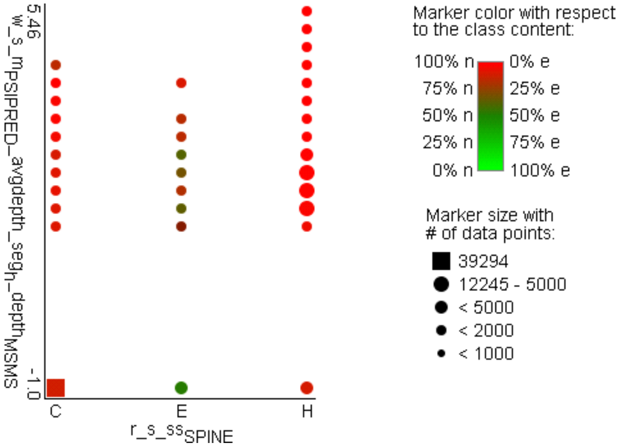
J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

shows side-by-side the actual DSSP-derived SS, the results from BETArPRED, and the predictions from the ZHANG-server and SSpro. The results reveal that the proposed predictor finds three β-strands in the middle of the sequence that were missed by SSpro, adding a total of 16 strand residues to the SSpro predictions, out of which 12 are correct and 4 are incorrect. BETArPRED correctly finds additional β-strands as a trade-off for a few over-predicted strand residues located at the termini of the correctly predicted β-strands. The ZHANG-server under-predicts



(a)

Fig. 3. Scatter plots of two pairs of features used by the BETArPRED. Size of the markers denotes number of residues and color denotes their membership (green for strand residues and red for non-strand residues). (a) The y-axis quantifies the average predicted depth of helical residues predicted by SSpro in a window of size 9 centered on the predicted residue. Value of -1 is used when there are no predicted helices in the window. The x-axis shows whether the predicted residue is in the EEE segment, as predicted by PSIPRED. (b) The y-axis quantifies the average predicted depth for the helix segment predicted by PSIPRED that includes the predicted residues. Value of -1 is used when the predicted residue in not in a helix segment. The x-axis shows the SS predicted by SPINE for the predicted residue.



(b)

Fig. 3. (*Continued*)

```
DSSP:     -EEEEE---EEEEEE-E--EEEEEEE-----E-E-------------EEE---------EEEE--EEEE----E-----E-EEEEEE--
SSpro:    --EEEE---EEEEEE----EEEEEE-----EEEEE------------EEEE--------EEEE--EEEE-----------EEEEE----
BrP:      --EEEE---EEEEEE----EEEEEEE----EEEEE------------EEEE--------EEEE--EEEE----------EEEEEE----
ZHANG:    -EEEEE----EEEEE----EEEEEE-----E----------------EE-----------EEE--EEE-------------EEE-----

DSSP:     EEEEEEE------------EEEEEEEEEE--EEEEEEEEEE-----EEE-EEE--EEE-----------EEEEEE----EEE--EE--E
SSpro:    EEEEEEE-----------EEEEEEEEEE---EEEEEEEEEE---------------EE-----------EEE-----------------
BrP:      EEEEEEEE---------EEEEEEEEEEE---EEEEEEEEEE------EEE-----EEE-----------EEEE-----EEEE-------
ZHANG:    --EEEE-------------EEEEEEEEEE--EEEEEEEEEE-----EE---EE--E--------------EEEE----E----------

DSSP:     E-----EEE-----EE----------EEEE-----EEEEEEE-----EEEEEEE---EEEEE--------EEEEEEEE----E-----E
SSpro:    -------------------------EEEEE------EEEEE------EEEEE-----EEEEE--------EEEEE--------------
BrP:      --------------EE---------EEEEE------EEEEEE-----EEEEEE----EEEEE--------EEEEE--------------
ZHANG:    --------------EE--------EE-----------EEEE------EEEEEE-----EE------------E----------------

DSSP:     -------EEE----EEEEEEEEEEE--------
SSpro:    -------EEE----EEEEEEEEEEE----EE--
BrP:      -------EEE----EEEEEEEEEEE----EE--
ZHANG:    --------EE----EEEEEEEEEE---------
```

Fig. 4. Comparison of the SSpro, BETArPRED (BrP), and ZHANG-server (ZHANG) predictions with the actual DSSP-derived SS structure for the galactose mutarotase related enzyme Q5FKD7 (PDBid 3DCD). The DSSP, SSpro, BrP and ZHANG are shown in four consecutive rows where "–" and "E" denote non-strand and strand residues, respectively. The sequences are split into multiple rows. DSSP is annotated such that bold indicates strand residues missed by SSpro and BrP, and underlined bold shows β-residue segments found by BrP and missed by SSpro. BrP is annotated such that bold/underlined bold indicate mistakes/improvements when compared with SSpro.

the strand residues; only 88 residues were correctly identified, while BETArPRED correctly predicts 115 out of a total of 140 strand residues. The $\mathrm{SOV_e}$ and accuracy of BETArPRED are 86.6 and 87.0, respectively, while for SSpro and ZHANG-server they are 79.8 and 84.0, and 74.5 and 80.7, respectively. The $\mathrm{U_e}$ of BETArPRED is 3.7 which is lower by 3.7 and 3.3 when compared with SSpro and ZHANG-server, indicating that our method finds a few extra β-strands. At the same time, this comes as a trade-off for the $\mathrm{Q_{e\_pred}}$ of BETArPRED that is lower by 2 and 5.8 when compared with SSpro and ZHANG-server.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

4. Conclusions

BETArPRED is shown to improve predictions of strand residues and strand segments when compared to a wide range of modern SS predictors. It could thus be useful in prediction of higher level structures such as β-sheets.[24–26] Its predictions are also competitive when compared with the best-performing tertiary structure predictor. Since BETArPRED performs well for low identity chains, its outputs could be useful in the context of the development of improved sequence profile-profile alignments.[27] The improvements stem from the novel design, which uses features that aggregate and combine information coming from three SS predictors and the residue depth predictor. The dataset and the prediction model are freely available at http://biomine.ece.ualberta.ca/BETArPred/BrP.htm. Since development of user-friendly and publicly accessible web-servers increases the practical value of predictors,[88] we plan to provide a web-server for the method presented in this paper.

Although the BETArPRED provides high quality predictions, there is still room for further improvement. One potential approach could be to exploit strand-strand interactions. This could be done with the help of scoring profiles that reflect inter-strand amino acid pairing preferences; these were recently proposed and successfully used to predict relative orientation of a pair of strand segments.[89] These profiles could be utilized to score the predicted strand residues with respect to their potential match with strand residues on another predicted strand segment. Such an approach would reflect the long range interactions between strand segments that are not covered by the current local window-based predictors. Another useful source of information that could be used to improve the strand predictions is related to position-specific propensities of amino acid types in strand segments. Recent work shows that these propensities are highly position-specific and that they follow a characteristic periodic pattern in inner positions with respect to the cap residues at both termini of the strand segments.[23] Finally, a simple extension of the current method could be to use flexible windows as proposed by Chou and colleagues,[90–92] instead of the fixed-size windows, to extract local information.

References

1. Rost B, Protein secondary structure prediction continues to rise, *J Struct Biol* **134**:204–218, 2001.
2. Rost B, Sander C, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Proc Natl Acad Sci USA* **90**:7558–7562, 1993.
3. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS, Improving the accuracy of protein secondary structure prediction using structural alignment, *BMC Bioinformatics* **7**:301, 2006.
4. Rost B, PHD: Predicting one-dimensional protein structure by profile based neural networks, *Meth in Enzymol* **266**:525–539, 1996.
5. Jones DT, Protein secondary structure prediction based on position-specific scoring matrices, *J Mol Biol* **292**:195–202, 1999.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

6. McGuffin LJ, Bryson K, Jones DT, The PSIPRED protein structure prediction server, *Bioinformatics* **16**:404–405, 2000.
7. Adamczak R, Porollo A, Meller J, Combining prediction of secondary structure and solvent accessibility in proteins, *Proteins* **59**:467–475, 2005.
8. Pollastri G, Przybylski D, Rost B, Baldi P, Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles, *Proteins* **47**:228–235, 2002.
9. Lin K, Simossis V, Taylor W, Heringa J, A simple and fast secondary structure prediction method using hidden neural networks, *Bioinformatics* **21**:152–159, 2005.
10. Pollastri G, McLysaght A, Porter: A new, accurate server for protein secondary structure prediction, *Bioinformatics* **21**:1719–1720, 2005.
11. Ofer D, Zhou Y, Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training, *Proteins* **66**:838–845, 2007.
12. Rost B, Eyrich VA, EVA: Large-scale analysis of secondary structure prediction, *Proteins* **5**:192–199, 2001.
13. Selbig J, Mevissen T, Lengauer T, Decision tree-based formation of consensus protein secondary structure prediction, *Bioinformatics* **12**:1039–1046, 1999.
14. Montgomerie S, Cruz JA, Shrivastava S, Arndt D, Berjanskii M, Wishart DS, PROTEUS2: A web server for comprehensive protein structure prediction and structure-based annotation, *Nucleic Acids Res* **36**:W202–W209, 2008.
15. Cheng H, Sen TZ, Jernigan RL, Kloczkowski A, Consensus data mining (CDM) protein secondary structure prediction server: Combining GOR V and fragment database mining, *Bioinformatics* **19**:2628–2630, 2007.
16. Zheng C, Kurgan L, Prediction of ß-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments, *BMC Bioinformatics* **9**:430, 2008.
17. Hu X, Li Q, Using support vector machine to predict beta- and gamma-turns in proteins, *J Comput Chem* **29**(12):1867–1875, 2008.
18. Ward JJ, McGuffin LJ, Buxton BF, Jones DT, Secondary structure prediction with support vector machines, *Bioinformatics* **19**(13):1650–1655, 2003.
19. Chou KC, Pottle M, Nemethy G, Ueda Y, Scheraga HA, Structure of beta-sheets: Origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets, *J Mol Biol* **162**:89–112, 1982.
20. Chou KC, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA, Interactions between two beta-sheets: Energetics of beta/beta packing in proteins, *J Mol Biol* **188**:641–649, 1986.
21. Chou KC, Carlacci L, Energetic approach to the folding of alpha/beta barrels, *Proteins* **9**:280–295, 1991.
22. Mandel-Gutfreund Y, Gregoret LM, On the significance of alternating patterns of polar and non-polar residues in beta-strands, *J Mol Biol* **323**(3):453–461, 2002.
23. Bhattacharjee N, Biswas P, Position-specific propensities of amino acids in the β-strand, *BMC Struct Biol* **10**:29, 2010.
24. Cheng J, Baldi P, Three–stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms, *Bioinformatics* **21**:75–84, 2005.
25. Lippi M, Frasconi P, Prediction of protein beta-residue contacts by markov logic networks with grounding specific weights, *Bioinformatics* **25**(18):2326–2333, 2009.
26. Max N, Hu C, Kreylos O, Crivelli S, BuildBeta — A system for automatically constructing beta sheets, *Proteins* **78**(3):559–574, 2010.
27. Wu ST, Zhang Y, MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information, *Proteins* **72**(2):547–556, 2008.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

28. Mandel-Gutfreund Y, Zaremba SM, Gregoret LM, Contributions of residue pairing to beta-sheet formation: Conservation and co-variation of amino acid residue pairs on antiparallel beta-strands, *J Mol Biol* **305**:1145–1159, 2001.
29. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D, Design of a novel globular protein fold with atomic-level accuracy, *Science* **302**:1364–1368, 2003.
30. Kamat A, Lesk A, Contact patterns between helices and strands of sheet define protein folding patterns, *Proteins* **66**:869–876, 2007.
31. Stöhr J, Weinmann N, Wille H, Kaimann K, Nagel-Steger L, Birkmann E, Panza G, Prusiner SB, Eigen M, Riesner D, Mechanisms of prion protein assembly into amyloid, *Proc Natl Acad Sci USA* **105**(7):2409–2414, 2008.
32. Madera M, Calmus R, Thiltgen G, Karplus K, Gough J, Improving protein secondary structure prediction using a simple k-mer model, *Bioinformatics* **26**(5):596–602, 2010.
33. Yuan Z, Wang ZX, Quantifying the relationship of protein burying depth and sequence, *Proteins* **70**:509–516, 2008.
34. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L, Sequence based residue depth prediction using evolutionary information and predicted secondary structure, *BMC Bioinformatics* **9**:388, 2008.
35. Albrecht M, Tosatto SC, Lengauer T, Valle G, Simple consensus procedures are effective and sufficient in secondary structure prediction, *Protein Eng* **16**(7):459–462, 2003.
36. Shen HB, Chou KC, Ensemble classifier for protein fold pattern recognition, *Bioinformatics* **22**(14):1717–1722, 2006.
37. Chen K, Kurgan L, PFRES: Protein fold classification by using evolutionary information and predicted secondary structure, *Bioinformatics* **23**(21):2843–2850, 2007.
38. Kedarisetti K, Kurgan L, Dick S, Classifier ensembles for protein structural class prediction with varying homology, *Biochem Biophys Res Commun* **348**(3):981–988, 2006.
39. Shen HB, Chou KC, QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information, *J Proteome Res* **8**(3):1577–1584, 2009.
40. Shen HB, Chou JJ, MemBrain: Improving the accuracy of predicting transmembrane helices, *PLoS One* **3**(6):e2399, 2008.
41. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B, Improved disorder prediction by combination of orthogonal approaches, *PLoS One* **4**:e4433, 2009.
42. Mizianty M, Stach W, Chen K, Kedarisetti KD, Miri Disfani F, Kurgan L, Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, *Bioinformatics* **26**(18):i489-i496, 2010.
43. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN, PONDR-FIT: A meta-predictor of intrinsically disordered amino acids, *Biochim Biophys Acta* **1804**(4):996–1010, 2010.
44. Crooks GE, Brenner SE, Protein secondary structure: Entropy, correlations and prediction, *Bioinformatics* **20**:1603–1611, 2004.
45. Liu Y, Carbonell J, Klein-Seetharaman J, Gopalakrishnan V, Comparison of probabilistic combination methods for protein secondary structure prediction, *Bioinformatics* **20**(17):3099–3107, 2004.
46. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE, The protein data bank, *Nucleic Acids Res* **28**:235–242, 2000.
47. Li W, Jaroszewski L, Godzik A, Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics* **18**:77–82, 2002.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

48. Kabsch W, Sander C, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**(12):2577–2637, 1983.
49. Chou KC, Shen HB, Recent progresses in protein subcellular location prediction, *Anal Biochem* **370**:1–16, 2007.
50. Chou KC, Shen HB, Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms, *Natural Science* **2**:1090–1103, 2010.
51. Vilar S, Gonzalez-Diaz H, Santana L, Uriarte E, A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer, *J Theor Biol* **261**:449–458, 2009.
52. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML, Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach, *J Theor Biol* **259**:366–372, 2009.
53. Chen C, Chen L, Zou X, Cai P, Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine, *Protein Pept Lett* **16**:27–31, 2009.
54. Ding H, Luo L, Lin H, Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition, *Protein Pept Lett* **16**:351–355, 2009.
55. Li S, Li H, Li M, Shyr Y, Xie L, Li Y, Improved prediction of lysine acetylation by support vector machines, *Protein Pept Lett* **16**:977–983, 2009.
56. Lin ZH, Wang HL, Zhu B, Wang YQ, Lin Y, Wu YZ, Estimation of affinity of HLA-A * 0201 restricted CTL epitope based on the SCORE function, *Protein Pept Lett* **16**:561–569, 2009.
57. Lu J, Niu B, Liu L, Lu WC, Cai YD, Prediction of small molecules' metabolic pathways based on functional group composition, *Protein Pept Lett* **16**:969–976, 2009.
58. Nanni L, Lumini A, A further step toward an optimal ensemble of classifiers for peptide classification, a case study: HIV protease, *Protein Pept Lett* **16**:163–167, 2009.
59. Jiang Y, Iglinski P, Kurgan L, Prediction of protein folding rates from primary sequences using hybrid sequence representation, *J Comp Chem* **30**(5):772–783, 2009.
60. Wang T, Xia T, Hu XM, Geometry preserving projections algorithm for predicting membrane protein types, *J Theor Biol* **262**:208–213, 2010.
61. Joshi RR, Sekharan S, Characteristic peptides of protein secondary structural motifs, *Protein Pept Lett* **17**:1198–1206, 2010.
62. Liu T, Zheng X, Wang C, Wang J, Prediction of subcellular location of apoptosis proteins using pseudo amino acid composition: An approach from auto covariance transformation, *Protein Pept Lett* **17**:1263–1269, 2010.
63. Mohabatkar H, Prediction of cyclin proteins using Chou's pseudo amino acid composition, *Protein Pept Lett* **17**:1207–1214, 2010.
64. Yang XY, Shi XH, Meng X, Li XL, Lin K, Qian ZL, Feng KY, Kong XY, Cai YD, Classification of transcription factors using protein primary structure, *Protein Pept Lett* **17**:899–908, 2010.
65. Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L, Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility, *Proteins* **78**(9):2114–2130, 2010.
66. Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, Suganthan PN, Martinetz T, Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition, *Protein Pept Lett* **17**:1473–1479, 2010.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

67. Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A, Critical assessment of methods of protein structure prediction-Round VIII, *Proteins* **77**(Suppl 9):1–4, 2009.
68. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A, Evaluation of template-based models in CASP8 with standard measures, *Proteins* **77**(Suppl 9):18–28, 2009.
69. McGuffin LJ, Jones DT, Benchmarking secondary structure prediction for fold recognition, *Proteins* **52**:166–175, 2003.
70. Zemla A, Venclovas C, Fidelis K, Rost B, A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment, *Proteins* **34**(2):220–223, 1999.
71. Koh IY, Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Eswar N, Graña O, Pazos F, Valencia A, Sali A, Rost B, EVA: Evaluation of protein structure prediction servers, *Nucleic Acids Res* **31**:3311–3315, 2003.
72. Rost B, Prediction of protein structure in 1D: Secondary structure, membrane regions, and solvent accessibility, in *Structural Bioinformatics*, 2nd edition, Gu J, Bourne PE (eds.), pp. 679–714, 2009.
73. Garg A, Kaur H, Raghava GP, Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* **61**(2):318–324, 2005.
74. Mizianty M, Kurgan L, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, *BMC Bioinformatics* **10**:414, 2009.
75. Mizianty M, Kurgan L, Improved identification of outer membrane beta barrel proteins using primary sequence, predicted secondary structure and evolutionary information, *Proteins* **79**(1):294–303, 2011.
76. Ivankov DN, Finkelstein AV, Prediction of protein folding rates from the amino-acid sequence-predicted secondary structure, *Proc Natl Acad Sci USA* **101**:8942–8944, 2004.
77. Sanner MF, Olson AJ, Spehner JC, Reduced surface: An efficient way to compute molecular surfaces, *Biopolymers* **38**:305–320, 1996.
78. Pintar A, Carugo O, Pongor S, DPX, for the analysis of the protein core, *Bioinformatics* **19**:313–314, 2003.
79. Varrazzo D, Bernini A, Spiga O, Ciutti A, Chiellini S, Venditti V, Bracci L, Niccolai N, Three-dimensional computation of atom depth in complex molecular structures, *Bioinformatics* **21**(12):2856–2860, 2005.
80. Chen K, Kurgan L, Ruan J, Optimization of the sliding window size for protein structure prediction, *Proc 2006 IEEE CIBCB Symposium*, pp. 366–372, 2006.
81. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, The WEKA Data Mining software: An update, *SIGKDD Explorations* **11**(1):10–18, 2009.
82. Fan, R-E, Chang, K-W, Hsieh, C-J, Wang, X-R, Lin, C-J, LIBLINEAR: A library for large linear classification, *J Mach Learn Res* **9**:1871–1874, 2008.
83. Hall M, Smith L, Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper, *Proc FLAIRS*, pp. 235–239, 1999.
84. Liu H, Setiono R, A probabilistic approach to feature selection — A filter solution, *Proc ICML*, pp. 319–327, 1996.
85. Hall M, Correlation-based feature selection for discrete and numeric class machine learning, *Proc ICML*, pp. 359–366, 2000.
86. Cessie S, Houwelingen J, Ridge estimators in logistic regression, *Applied Statistics* **41**(1):191–201, 1992.

87. Bugmann G, Normalized Gaussian radial basis function networks, *Neurocomputing* **20**:97–110, 1998.
88. Chou KC, Shen HB, Recent advances in developing web-servers for predicting protein attributes, *Natural Science* **2**:63–92, 2009.
89. Zhang N, Duan G, Gao S, Ruan J, Zhang T, Prediction of the parallel/antiparallel orientation of beta-strands using amino acid pairing preferences and support vector machines, *J Theor Biol* **263**(3):360–368, 2010.
90. Chou KC, Using subsite coupling to predict signal peptides, *Protein Eng* **14**:75–79, 2001.
91. Chou KC, Prediction of protein signal sequences, *Curr Prot Pept Sci* **3**:615–622, 2002.
92. Chou KC, Shen HB, Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides, *Biochem Biophys Res Commun* **357**:633–640, 2007.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.

Kanaka Durga Kedarisetti received her M.Sc. degree from the Department of Electrical and Computer Engineering at the University of Alberta in 2005 and she is currently working towards Ph.D. degree in the same department. During her M.Sc. and Ph.D. programs Mrs. Kedarisetti received several awards, the most prestigious being the Izaak Walton Killam Memorial Scholarship, the Canada Graduate Scholarships from NSERC, and the Alberta Ingenuity/iCORE Graduate Student Scholarship. Her research interests include development and application of computational methods in structural bioinformatics, with focus on analysis of sequence and structure of proteins.

Marcin Mizianty is currently a research assistant and a Ph.D. student at the University of Alberta. He received his M.Sc. degree in Applied Computer Science from the AGH University of Science and Technology, Krakow, Poland in 2008. His work focuses on structural bioinformatics, rational drug-design, knowledge discovery, and machine learning. His website is located at http://www.ece.ualberta.ca/∼mizianty/.

Scott Dick received his B.Sc. degree in 1997, his M.Sc. degree in 1999, and his Ph.D. in 2002, all from the University of South Florida. His Ph.D. dissertation received the USF Outstanding Dissertation Prize in 2003. From 2002 to 2008 he was an Assistant Professor of Electrical and Computer Engineering at the University of Alberta in Edmonton. Since 2008, he has been an Associate Professor in the same department and has published over 40 scientific articles in journals and conferences. Dr. Dick is a member of the IEEE Computational Intelligence Society's Fuzzy Systems Technical Committee. He is a member of the ACM, IEEE, ASEE, and an associate member of Sigma Xi.

Lukasz Kurgan received his M.Sc. degree (with honors) in Automation and Robotics from the AGH University of Science and Technology, Krakow, Poland in 1999 and his Ph.D. degree in Computer Science from the University of Colorado at Boulder in 2003. He joined the Department of Electrical and Computer Engineering at the University of Alberta in 2003, where he was promoted to the rank of an Associate Professor in 2007. His research interests include development and application of modern data mining methods in structural bioinformatics, with focus on analysis of sequence, structure, and function of biologically interesting macromolecules. He has published close to 100 peer-reviewed articles. Dr. Kurgan is an associate editor of the *BMC Bioinformatics*, *Neurocomputing*, *Open Proteomics Journal*, *Journal of Biomedical Science and Engineering*, *Open Bioinformatics Journal*, and *Protein and Peptide Letters* journals, and serves(ed) on program committees of numerous conferences and workshops related to bioinformatics and data mining. The web site of his lab is located at http://biomine.ece.ualberta.ca/.

J. Bioinform. Comput. Biol. 2011.09:67-89. Downloaded from www.worldscientific.com
by VIRGINIA COMMONWEALTH UNIVERSITY LIBRARIES on 02/03/17. For personal use only.