Sequence-Only Based Prediction of β-Turn Location and Type Using Collocation of Amino Acid Pairs

Kevin Campbell and Lukasz Kurgan*

Department of Electrical and Computed Engineering, University of Alberta, Canada

Abstract: Development of accurate β-turn (beta-turn) type prediction methods would contribute towards the prediction of the tertiary protein structure and would provide useful insights/inputs for the fold recognition and drug design. Only one existing sequence-only method is available for the prediction of beta-turn types (for type I and II) for the entire protein chains, while the proposed method allows for prediction of type I, II, IV, VII, and non-specific (NS) beta-turns, filling in the gap. The proposed predictor, which is based solely on protein sequence, is shown to provide similar performance to other sequence-only methods for prediction of beta-turns and beta-turn types. The main advantage of the proposed method is simplicity and interpretability of the underlying model. We developed novel sequence-based features that allow identifying beta-turns types and differentiating them from non-beta-turns. The features, which are based on tetrapeptides (entire beta-turns) rather than a window centered over the predicted residues as in the case of recent competing methods, provide a more biologically sound model. They include 12 features based on collocation of amino acid pairs, focusing on amino acids (Gly, Asp, and Asn) that are known to be predisposed to form beta-turns. At the same time, our model also includes features that are geared towards exclusion of non-beta-turns, which are based on amino acids known to be strongly detrimental to formation of beta-turns (Met, Ile, Leu, and Val).

Keywords: Secondary protein structure, Beta-turns, Beta-turn types, Prediction, Collocation of amino acid pairs, Support vector machine.

INTRODUCTION

The secondary structure of a protein consists of helices, beta-strands and coils, where the coil region comprises tight turns, bulges and random coil structures [1]. Tight turns are believed to be important structural elements in regards to protein folding and molecular recognition processes between proteins, which has lead to interest in mimicking beta-turns for medicine synthesis [2, 3]. Tight turns are classified as δ-turns, γ-turns, β-turns, α-turns and π-turns, where a β-turn (beta-turn) is a four-residue reversal in the protein chain that is not in an α-helix. While characterization and prediction of α-turns attracted some research attention [4-7], our research focuses on beta-turns. We observe that beta-turns are the most common turn type, and make up, on average one quarter of all residues in proteins [8]. Formation of beta-turns is also a vital stage during the process of protein folding [3]. Therefore, development of accurate beta-turn prediction methods would be a valuable step towards the overall prediction of the three-dimensional structure of a protein from its amino acid sequence and could provide insights and inputs for the fold recognition and drug design.

The beta-turns can be classified into nine different types based on the φ and ψ angles of the two central residues [9]. As a result, prediction of the location of beta-turn types, in contrast to a binary prediction that would identify location of beta-turns, would provide additional, structural, information that concerns the φ and ψ angles. A commonly used benchmark dataset of 426 non-homologous protein chains [10], which have been used to rank and test several methods for prediction of beta-turn types [11, 12], reveals that some of these types are infrequent and thus they are commonly combined together [11]. To this end, we focus on prediction of beta-turn types I, II, IV and VIII, while the remaining types I', II', VIa1, VIa2 and VIb, which only make up 304, 165, 44, 17 and 70 respectively out of the total 7153 beta-turns in the aforementioned dataset have been combined into one set referred to as non-specific (NS), which is consistent with [11]. The challenging aspect of the beta-turn type prediction is that these turns are not isolated in a chain. Quite the opposite, in fact, Hutchinson and Thornton (1994) report that 58% of beta-turns overlap with another beta-turn, i.e., they share one or more residues with another beta-turn [9].

There exist a number of recent works that address prediction of beta-turn types, which can be divided into two categories, statistical methods and machine learning based methods. Statistical methods utilize probabilities computed using information regarding the preference of individual amino acid types at each position of the beta-turn to form a turn. The most promising of which is COUDES [13], which is based on propensities of individual residues augmented with the information coming from multiple sequence alignment. The position-specific score matrix (PSSM), which is calculated with PSI-BLAST [14], was used to weigh propensities for a given residue, so that all the residues present in the multiple alignment at this position are taken into account. Secondary structure information predicted by PSIPRED

*Address correspondence to this author at the Department of Electrical and Computed Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4; E-mail: lkurgan@ece.ualberta.ca

1875-0362/08

[15], SSPRO2 [16], and PROF [17] and the flanking residues around the beta-turn tetrapeptide were utilized to improve the prediction accuracy. The COUDES method uses a window of size 12 with prediction being made on the four central (with respect to the window) residues.

The machine learning methods include BETATURNS [11] and BTPRED [12], which are based on artificial neural networks (ANN), and a hybrid multinomial logistic regression and ANN model [18]. BTPRED encodes the sequence using a large window of 11 residues centered over the predicted residue together with secondary structure predicted with PHDsec [19] to perform predictions. The window is used to incorporate the effects of neighboring residues on the formation of beta-turns. BETATURNS is an improved neural network design, in which two networks are used. The first network uses the sequence together with the PSSM as the input, and its output is fed along with the PSIPRED predicted secondary structure of the central residue into the second network that produces the predictions. BETATURNS employs a window of 9 residues where prediction is made on the central residue. Finally, the multinomial logistic regression model uses a two-stage hybrid approach and considers only beta-turns. The latter method is not used for the prediction of the location of beta-turns, but it allows differentiating different types of beta-turns based on the underlying tetrapeptides while it does not consider non-beta-turn sequence segments, i.e., it predicts a beta-turn type for a given tetrapeptide that corresponds to a beta-turn.

In comparison with the methods that predict beta-turn types for entire protein sequences outlined above, which include [11-13] and which use significant amount of auxiliary information such as PSSM and predicted secondary structure, it is clear that a method based solely on the protein sequence would be simpler to design and execute. However, this may lead to reduced quality as only limited information (sequence) relative to the competition would be used. The motivation for our sequence-based design comes from work of Chou and colleagues who found that support vector machine (SVM) classifier can be used to express the relation between different beta-turns types or non-beta-turns and the underlying tetrapeptides [20]. They observed that the accuracies of self-consistency (prediction on the training set) test for beta-turn types I, I', II, II', VI and VIII and non-beta-turns are over 97%. This was a follow up on their previous study in which they verified that the relation between the tetrapeptides and the beta-turns types can be expressed using a probabilistic approach [21]. The authors applied their sequence-coupled model [5, 21, 22] to perform prediction for a selected set of tetrapeptides, and applied this model to predict beta-turn types for the rubredoxin protein [1]. This is similar to work done in [18], except that in this case the non-beta-turns were considered in building the predictive model. We also note that only two sequence-only based method (method that uses as the input only the protein sequence and no sequence derived information such as PSSM or predicted secondary structure) are available for prediction of beta-turn types [1, 23]. The method in [23] addresses prediction of only type I and type II beta-turns, while the method in [1] predicts beta-turn types I, I', II, II', VI, and VIII, which are different than targets addressed by newer prediction methods [11]. This provides additional motivation for the development of the proposed method. The consideration and employment of the window is a fundamental difference in our approach relative to those listed above. BTPRED, BETATURNS, and COUDES methods predict the beta-turn type of individual residues (using a sliding window centered over the predicted residue), whereas our method predicts entire tetrapeptides as either a given beta-turn type or a non-beta-turn. Unlike the other methods, this results in features that are more biologically relevant and that are better for describing full beta-turns vs. non- beta-turns as opposed to simply identifying residues that are apt to be in beta-turns.

Our intention is to develop a method that gives similar performance to the aforementioned methods but with a main goal of creating a simple predictive model that allows derivation of sequence based factors which facilitate differentiation between different beta-turn types and non-beta-turns.

MATERIALS AND METHODOLOGY

Datasets

Three nonredundant datasets were used through the course of this study. The first, which was used for feature selection, was prepared in [18] to design method that differentiates different types of beta-turns (excluding non-beta-turns) and was based on 565 non-homologous protein chains (and will be hereby referred to as 565). The chains were selected using the PAPIA system [24], contain no chain breaks, have structure determined by X-ray crystallography at 2.0A resolution or better, and no two chains have more than 25% sequence identity. The PROMOTIF program was used to assign the beta-turns in protein chains [25]. The original dataset includes only the tetrapeptides that correspond to all beta-turns in the 565 non-homologous proteins, i.e. it does not include the entire protein chains. We augmented the original beta-turn tetrapeptides with randomly chosen set of tetrapeptides that correspond to non-beta-turns assuming that the number of the selected non-beta-turns should approximately equal the number of the most frequent beta-turn type. More specifically, the 565 dataset includes 4115, 1442, 4128, 1100, and 1028 beta-turns of type I, II, IV, VIII, and NS and 4448 non-beta-turns.

The second dataset, used for testing and comparing the prediction method, was comprised of 426 protein chains and 95,289 residues and was prepared in [26]. This dataset (hereby referred to as 426) has been widely used to validate and compare beta-turn prediction methods [10, 11, 13, 26-28] and includes chains that are non-redundant at 25% and that have been resolved with X-ray crystallography at 2.0A resolution or better. Again, the PROMOTIF program [25] was used to assign the beta-turns in protein chains using the classification scheme proposed by Hutichinson and Thornton (1994) [9]. Every chain in this dataset includes at least one beta-turn. In order to assess the accuracy of the proposed model and to remain consistent with recent beta-turn prediction literature [10, 11, 13, 26-28], sevenfold cross-validation was employed on dataset 426. The dataset was divided into 6 folds of 61 sequences and 1 fold of 60 sequences. Six of the

folds were then used to train the model, while the seventh was used to test it, and the process was repeated seven times.

The third dataset, which is used for model parameterization, involves 183 sequences from the 426 dataset. These sequences constitute three of the seven folds of the 426 dataset. Additionally, these 183 sequences were randomly down sampled to 20% of the original residues. This dataset will be referred to as 183.

Quality Indices

To assess the accuracy of the prediction method, as well as for comparison purposes, the standard quality indices of beta-turn prediction literature were employed [10-13, 27, 28].

The percentage of correct predictions for each beta-turn type is defined as follows:

$$Q_{total} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (true positive) is the number of residues observed and predicted as a given beta-turn type, TN (true negative) is the number of residues observed and predicted as not the given beta-turn type, FP (false positive) is the number of residues not observed but predicted as a given beta-turn type, and FN (false negative) is the number of residues observed but not predicted as a given beta-turn type.

When describing accuracy, Q_{total} tends to overestimate predictive performance due to the high number of true negatives, which underemphasises the false negatives and false positives [12, 13, 28]. Therefore, it is better to use the Matthews Correlation Coefficient (MCC) [29] which takes underprediction and overprediction into account:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

Underprediction can be evaluated using Q_{obs}, which is the fraction of observed given beta-turn types predicted correctly:

$$Q_{obs} = \frac{TP}{TP + FN}$$

Finally, overprediction is evaluated using Q_{pred}, which is the fraction of correctly predicted given beta-turn types:

$$Q_{pred} = \frac{TP}{TP + FP}$$

While these quality indices are used consistently, they will be applied in two different ways, first comparing predicted beta-turn type to actual beta-turn type on a residue by residue basis denoted as follows, $Q^{res}_{pred/obs/total}$ and second comparing predicted beta-turn type to actual beta-turn type on a turn by turn basis, denoted as $Q^{turn}_{pred/obs/total}$. In the latter case, the unit of the prediction is a tetrapeptide.

Model Overview

Fig. (**1**) compares the proposed prediction system and the existing methods [11-13]. The competing methods use a window centered on the predicted residue as the input information that is processed with PSI-BLAST and a secondary structure prediction method. These inputs are converted into features and next fed into a classifier that predicts a given beta-turn type / non-beta-turn for a single residue. In our design, the processing unit is the tetrapeptide, i.e. four adjacent amino acids, that forms a given beta-turn type or a non-beta-turn. Thus, the sequences in each dataset were broken down into four residue fragments *via* a sliding window. Next, these segments are represented using a feature set that consists of three vectors, which is tagged by turn type if the start of the window was also the start of a turn. The resulting vector is passed to the classifier and the prediction is applied to all four residues in the window. As each residue is predicted four times as part of four separate possible turns, in the case of overlap between turn and non-turn, a turn prediction overrides a non-turn prediction. We believe that this design results in features that are more biologically relevant as they describe full beta-turns (and non-beta-turns) as op-

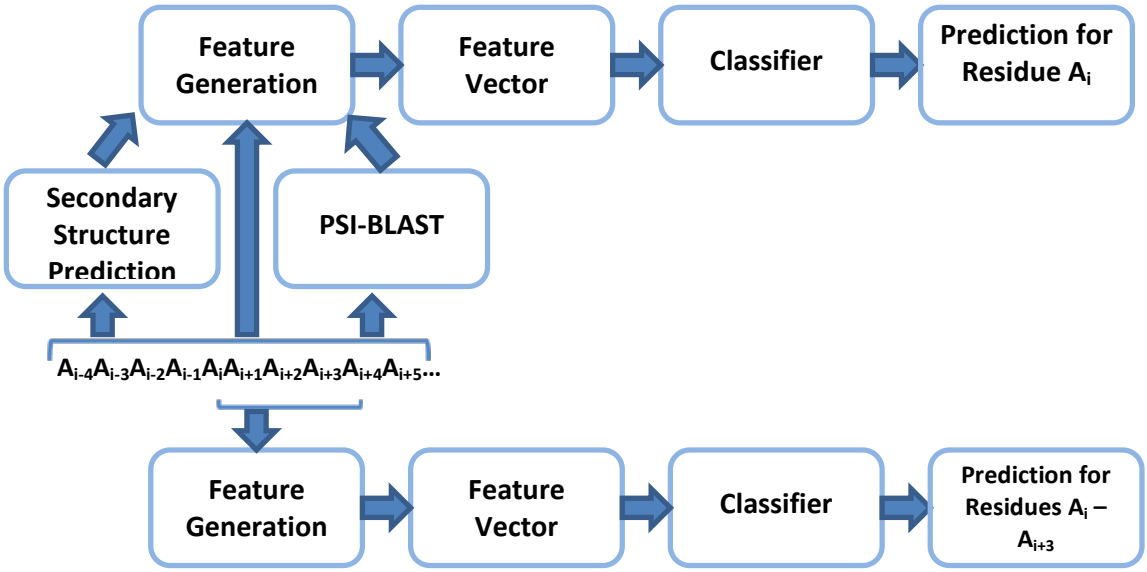

Fig. (1). Comparison of existing and the proposed beta-turn prediction methods. The upper portion of the figure represents design of the classical methods, while the bottom portion shows the proposed design.

posed to describing information concerning a window centered over the predicted residue.

Feature Generation

Composition Vector

The composition vector is a simple sequence representation which is widely used in prediction of various aspects of protein structure [30-37]. The vector is composed of the twenty amino acids, alphabetically ordered, and stores the number of occurrences of the amino acid in the sequence window (in our case the tetrapeptide). With 20 amino acids, this results in 20 corresponding features.

Positional Vectors

The positional vectors are similar to the compositional vector in that they are a simple sequence representation composed of the twenty amino acids, alphabetically ordered, and identify the presence/absence of a given amino acid in a given position in the tetrapeptide. As the window size considered includes 4 amino acids, this results in 80 corresponding features.

Collocation Vector

Finally, a relatively new representation based on the frequency of collocation amino acid pairs [38-41] in the sequence window was applied. Our motivation is that the composition and positional vectors are insufficient to represent the sequence and the interactions between local amino acid pairs. As interactions between short-range amino acid pairs, not just dipeptides have the potential to impact beta-turn formation [9, 10], the representation considers collocated pairs of amino acids, which are separated by p amino acids. Collocated pairs for p = 0, 1 and 2 are considered, where p = 0 pairs reduce to dipeptides and p = 1 and 2 can be understood as dipeptides with gaps. For each value of p, there are 400 corresponding features that store the number of occurrences of the collocated pairs. We emphasize that this feature set was not yet utilized for prediction of beta-turns types.

As a result, we consider a feature set which includes a total of 400 * 3 + 80 + 20 = 1300 features.

SVM Classifier

We employed a support vector machine (SVM) classifier [42] which was previously applied to beta-turn prediction [43, 44] and was shown to provide promising results in identifying beta-turn types [20]. Given a training set of data point pairs (x_i, c_i), i = 1, 2, … n, where x_i denotes the feature vector, c_i ={-1, 1} denotes binary class label, n is the number of training data points, finding the optimal SVM is achieved by solving:

$$\min \|w\|^2 + C\sum_i \xi_i$$

such that $c_i(wz_i - b) \geq 1 - \xi_i$ and $1 \leq i \leq n$

where w is a vector perpendicular to $wx–b$=0 hyperplane that separates the two classes, C is a user defined complexity constant, ξ_i are slack variables that measure the degree of misclassification of x_i for a given hyperplane, b is an offset that defines the size of a margin that separates the two classes, and $z=\Phi(x)$ where $k(x,x')= \Phi(x)\Phi(x')$ is a user defined kernel function.

The SVM classifier was trained using Platt's sequential minimal optimization algorithm [45] that was further optimized by Keerthi and colleagues [46]. The prediction that includes multiple types of beta-turns and non-beta-turns is solved using pairwise binary classification, namely, a separate classifier is build for each pair of classes (beta-turns types and non-beta-turns). We used RBF kernel and performed parameterization (selection of the value of the complexity constant C and RBF kernel width γ) based on 3-fold cross validation on the dataset 183. The final classifier uses C = 3 and the RBF kernel

$k(x_i, x_i') = e^{-\gamma\|x-x\|^2}$ where $\gamma = 0.3$

The classification algorithm and feature selection algorithms used to develop and compare the proposed method were implemented in Weka [47].

Feature Selection

As the proposed representation includes a relatively large number of features, three feature selection methods were employed in tandem to reduce the dimensionality and potentially improve the prediction: an Information Gain based method (IG) [48]; a Chi-Squared method (CHI) [49]; and the Relief algorithm (REL) [50]. We used three different methods in order to reduce the bias introduced by each of the methods. In these algorithms, each feature was ranked based on its merit (etc., information gain in IG, the value of the chi-squared statistic in CHI and the weights in REL), and next they were sorted by their average rank across the three algorithms. The measurement of the merit for the three algorithms is defined below.

Information gain (IG) measures the decrease in entropy when a given feature is used to group values of another (class) feature. The entropy of a feature X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i))$$

where $\{x_i\}$ is a set of values of X and $P(x_i)$ is the prior probability of x_i. The conditional entropy of X, given another feature Y (in our case the beta-turn type or non-beta-turn) is defined as

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j))$$

where $P(x_i|y_j)$ is the posterior probability of X given the value y_j of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called information gain

$$IG(X|Y) = H(X) - H(X|Y)$$

According to this measure, Y has stronger correlation with X than with Z if $IG(X|Y)>IG(Z|Y)$.

Chi-Squared statistic (CHI) is a statistical test that measures divergence from the expected distribution assuming that the occurrence of a given feature is independent of the class

value. Given that X is a feature with $m = 6$ possible outcomes $x_1, x_2, \ldots, x_m$, which correspond to the type I, II, IV, VIII, and NS beta-turn as well as non-beta-turn, with probability of each outcome $P(X=x_i) = p_i$. The Pearson-chi-squared statistic is defined as:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

where n_i is the number of instances which will result the outcome x_i. A feature that gives higher value of χ receives lower rank.

Relief algorithm (REL) is based on the feature weighting approach, which estimates the features according their performance in distinguishing similar instances. REL searches the two nearest neighbors for each instance: one from the same class (nearest hit) and another from any other class (nearest miss). The algorithm to calculate the weights as follows

(1) Initialization: given $D = \{(\boldsymbol{X}_n, y_n)\}$ $(n = 1,2,\ldots,N)$ where $\boldsymbol{X}_n$ is the feature set, y_n is class label, and N is the number of instances, set $w_i = 0$, $1 \le i \le I$ where I is the number of features and T is the number of iterations.

(2) For { $t = 1:T$

Randomly select an instance x from D;

Find the nearest hit NH(x) and miss NM(x) of x;

For { $i = 1:I$

calculate $w_i = w_i + |x^{(i)} - \mathrm{NM}^{(i)}(x)| - |x^{(i)} - \mathrm{NH}^{(i)}(x)|$ } }

Using the average rank generated by the three methods outlined above, an SVM model using an RBF with default parameters C = 1.0 and $\gamma = 0.1$ was used with 9-fold cross-validation on dataset 565 in order to select a subset of the ranked features. We note that the same cross validation was used in [18]. We selected the top ranked features in increments of 10 and computed

$$Accuracy^{turn} = \frac{\sum_{i \in \{type\,I, II, Iv, VIII, NS, non-\beta-turn\}} TP_i}{total}$$

where i denotes a given prediction outcome (a given beta-turn type and non-beta-turns), TP_i denotes the number of true positive predictions for i^{th} outcome, and *total* denotes the total number of the tetrapeptides in the dataset. The $Accuracy^{turn}$, which quantifies aggregated (over all prediction outcomes) quality of prediction, was computed over the 9 cross-validated folds, see Fig. (**2**).

We observe that the 50 highest average ranked features give high $Accuracy^{turn}$ values relative to the number of used features. In considering the trade off between $Accuracy^{turn}$ and the number of features selected, we attempted to minimize the feature count in an endeavor to be able to better explain them and to obtain less complex classification model. The selection of 50 features allows for a large improvement in $Accuracy^{turn}$, i.e., 1.4%, when compared with using 40 features, while the subsequent improvements (when using more features) are relatively small when compared with the additional number of employed features. Although the highest $Accuracy^{turn}$ = 46.2% was obtained with 250 features, this is only 1.3% higher than $Accuracy^{turn}$ = 44.9% obtained with 50 features that corresponds to a reduction of 200 features.

SVM Parameterization

The most relevant 50 features, as determined by average ranking of three different feature selection methods and reduced with a SVM model on dataset 565, were then used with dataset 183 to parameterize the SVM classifier. This reduced dataset was used since parameterization is computationally expensive. We apply 3-fold cross-validation as the 183 dataset corresponds to three out of seven folds of the dataset 426. Classifier parameterization was done in greedy fashion using three phases:

1. The extent of down sampling of the non-beta-turns was estimated to force the model to predict beta-turn types. This step is necessary due to highly skewed nature of the dataset. More specifically, the dataset 426 includes 72064 non-beta-turn residues which corresponds to 75.6% of all residues. In contrast, the fraction of residues that constitute type I, II, IV, VIII, and NS beta-turns equals 9.5%, 3.8%, 10%, 2.8%, and 2%, respectively. Note that the beta-turns of different

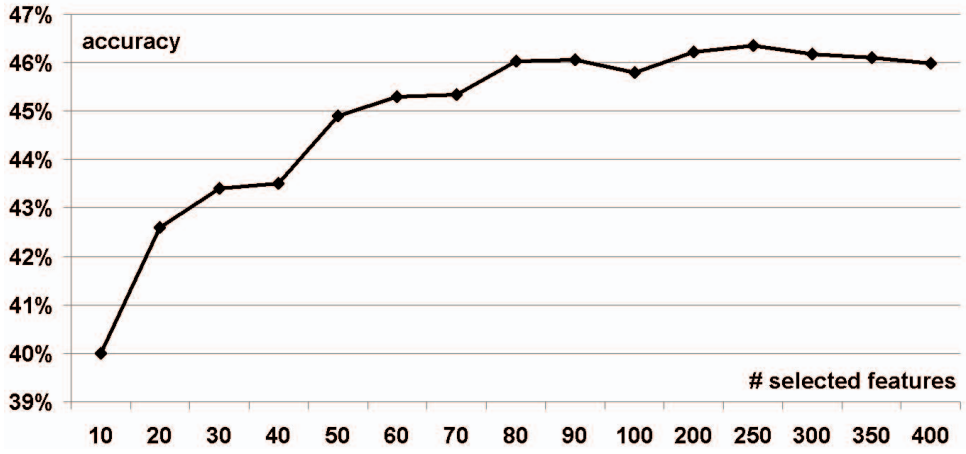

Fig. (2). The values of $Accuracy^{turn}$ (y-axis) against the number of the selected top features (x-axis) for 9-fold cross validation on dataset 565.

types may overlap and thus the total number does not equal 100%. The down sampling was performed at random and was applied to the training set, while the original (no sampling) test set was used for the prediction.

2. Parameter C was optimized using the down sampled training sets. In this case, the default value of $\gamma = 0.1$ was assumed and we varied values of C to optimize the prediction performance.

3. Finally, Parameter γ was optimized using the down sampled training sets and the optimized value of C.

Fig. (**3**) shows results associated with different degrees of downsampling performed on dataset 183. Varying downsampling rates results in a trade-off between Q_{obs}^{turn} and Q_{pred}^{turn}. Although there is no truly optimal configuration (there is no common optimum for both quality indices), two different downsamplings of non-beta-turns were selected. First, 3% of the non-beta-turns were kept (3%NT) as it resulted in the number of non-beta-turns being approximately the same size as the largest beta-turn type, type IV. Additionally, 8% of the non-beta-turn were randomly selected (8%NT), as this resulted in the closest number of predicted beta-turns when compared with the actual count of beta-turns in the 183 dataset. Fig. (**3A**) shows the values of Q_{obs}^{turn} and Q_{pred}^{turn} weighted by the beta-turn type counts when considering all beta-turn types and non-beta-turns, while Fig. (**3B**) shows the same but when non-beta-turns are excluded. The Figure shows that 3%NT results in low Q_{pred}^{turn} / high Q_{obs}^{turn} for the beta-turns and high Q_{pred}^{turn} / low Q_{obs}^{turn} when considering both beta-turns and non-beta-turns. This means that the method overpredicts beta-turns at the expense of underpredicting non-beta-turns. At the same time, 8%NT is associated with the best trade-off between the prediction of beta-turns and non-beta-turns, i.e., the corresponding Q_{pred}^{turn} and Q_{obs}^{turn} lines cross at this point. Therefore, in the case of our application, the 8%NT downsampling is considered optimal.

In the second and thirst parameterization steps, the C was optimized for both 3%NT and 8%NT and found to be 3.0 for both cases. Then, γ was optimized and found to be 0.15 for 3%NT and 0.3 for 8%NT.

RESULTS

Analysis of the Proposed Prediction Model

Of the 50 features selected, 32 were collocated pair features, 15 were positional vector features, and 3 were composition vector features. Of the 32 collocated pair features, 30 included Gly (G) at one of the two positions, see Fig. (**4**).

According to Hutchinson and Thornton [9], Gly (G) has the highest potential of any residue to form a beta-turn. Gly is also characterized by the highest potential to form beta-turn when it occupies positions 3 (position i+2 in the tetrapeptide with starting position i) and 4 (i+3). Additionally, Gly at position 3 of a beta-turn type II is experimentally observed to occur at least four times as often as any other amino acid [51]. Hence, Gly is present more often as the second residue in the collocated pairs when compared with its occurrence as the first residue in the pair. This is particularly transparent when considering that Gly is the first residue in the collocated pair with p = 0 (dipeptide), while larger gaps sizes are observed when it constitutes the second residue in the pair. Among the residues that are involved in 3 or more collocated pairs, Asp (D) and Asn (N) are characterized by the highest potential to form beta-turns for positions 1 and 3, Pro (P) by the highest potential for position 2, and Gly (G) by the highest potential for positions 3 and 4, as presented in [9, 26].

The novel features considered in this work include 12 collocated pairs with p > 1, with ten pairs that have p = 1 and two with p = 2. The two pairs with p = 2 include DxxG and NxxG, and we note that both of them are based on amino acids that are known to be predisposed to form beta-turns [9, 26]. The only collocated pair that does not include Gly is DxN, which incorporates the same amino acids as the above two pairs for p = 2. We emphasize that Gly (G), Asp (D), and Asn (N) are the three amino acids that have the highest

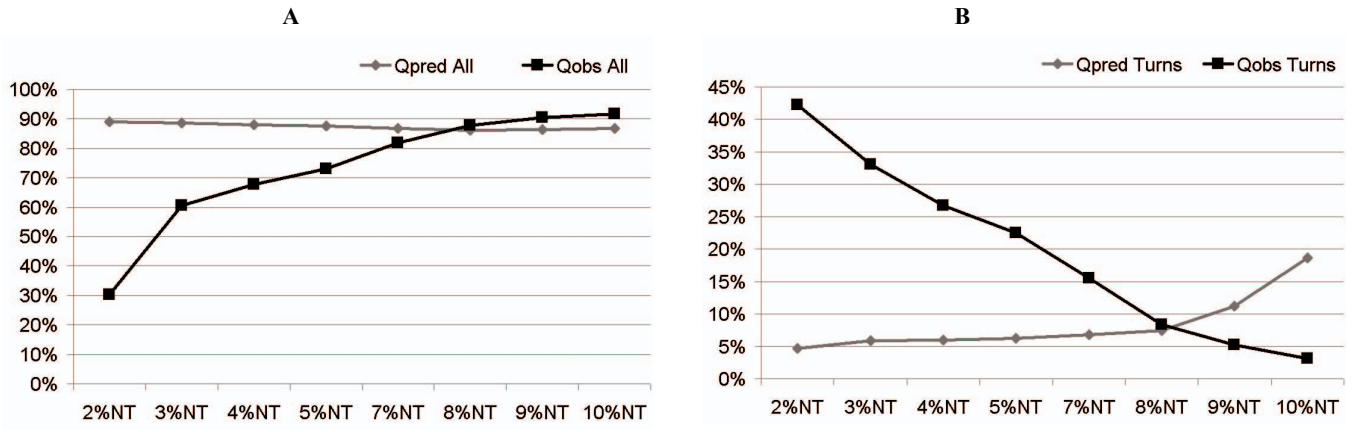

Fig. (3). The Q_{obs}^{turn} and Q_{pred}^{turn} values (y-axis) for different downsampling amounts (x-axis) obtained based on 3-fold cross validation on dataset 183 using SVM with default parameters C = 1.0 and γ = 0.1. Panel **A** shows results when considering all beta-turn types and non-beta-turns. Panel **B** shows results when considering only all beta-turn types.

potential to form a beta-turn and that Asp, Asn, and Gly also have the highest positional potentials, as discussed above [9,26].

Second residue in pair

First residue in pair	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A						0														
C																				
D						0,2						1								
E						0														
F						1														
G	0		0	0	0	0,1			0			0		0	0	0	0	0		
H																				
I						1														
K						0,1														
L						1														
M																				
N						0,2														
P			0			0,1														
Q						1														
R																				
S																				
T						1														
V						0,1														
W																				
Y																				

Fig. (4). Selected collocation vector features. The rows show the first amino acid and columns show the second amino acid in the pair. Values in cells show the corresponding *p* value of the selected collocated pair, while empty cells show features that were not selected.

We also observe that many of the selected pairs formed with Gly involve hydrophilic residues. More specifically, total of 13 pairs involve Asn (N), Asp (D), Glu (E), Lys (K), and Gln (Q). In contrast, only 6 pairs are formed with hydrophobic residues that include Ile (I), Leu (L), Phe (F), and Val (V). This is again consistent with [12], where authors show that beta-turns tend to be found at the solvent-exposed surface, which explains the prevalence of hydrophilic residues.

In Fig. (**5**), the features selected from the composition and position vectors are summarized. Most notably, Asp (D), Gly (G), Pro (P), and Asn (N) make up 14 of the 18 features. These four amino acids also have the highest overall potential to form beta-turns according to Hutchinson and Thornton [9] and Guruprasad and Rajkumar [26]. This selection is also motivated by the fact that amino acids with short and polar side chains, e.g. Ser (S), Asp (D), and Asn (N), are preferred at the position 3 of a type I beta-turn [51].

According to [23], type I beta-turns favor Asp (D), Asn (N), Ser (S) and Cys (c) at position 1, Asp (D), Ser (S), Thr (T) and Pro (P) at position 2, Asp (D), Ser (S), Asn (N) and Arg (R) at position 3, and Gly (G), Trp (W) and Met (M) at position 4. At the same time, type II beta-turns prefer Pro (P) at position 2, Gly (G) and Asn (N) at position 3, and Gln (G) and Arg (R) at position 4. These preferences have been explored statistically and explained by specific side-chain interactions observed within the X-ray structures [23]. They also motivate selection of 9 out of the 15 positional vector features:

- selection of Asp (D) and Asn (N) at position 1 is explained by their abundance in type I beta-turns,
- selection of Pro (P) at position 2 is motivated by its abundance in type I and II beta-turns,
- selection of Asp (D) and Asn (N) at position 3 is associated with their abundance in type I beta-turns, and selection of Gly (G) and Asn (N) by their abundance in type II beta-turns,
- selection of Gly (G) at position 4 is explained by its abundance in type I and II beta-turns.

According to [9, 26], the residues with the lowest potential to form beta-turns include Met (M), Ile (I), Leu (L), and Val (V). These residues are shown to be strongly detrimental to formation of beta-turns when in position 3 (*i*+4), which is consistent with their selection shown in Fig. (**5**). This shows that our method uses not only features that allows identify particular beta-turn types but also those that can identify non-beta-turns.

Table **1** estimates contribution of each of the three feature sets, i.e., composition vector, positional vectors, and collocation vector, on the prediction of beta-turn types and compares these predictions against results when using the complete set of 50 features. The best results are obtained with the use of the positional features. The second best set is based on the collocation features, while composition vector features contribute very little. We observe that very few predictions are made when using only the composition vector features, i.e., only about 250 residues were predicted as type I tu**r**n and 36 as type IV (with corresponding Q_{pred}^{res} equal 26% and 14%, respectively), and no residues were predicted to assume the remaining turn types. This poor result is expected due to very low number of features, i.e., 3, in this set. We note that high Q_{total}^{res} values are due to large number of true negative predictions. On the other hand, both the collocation vector and the positional vector features strongly contribute to the prediction. Although predictions with positional vector features have the highest MCC values, i.e., overall they are

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
CV			1			1							1							
P1			1									1								
P2						1						1	1							
P3			1			1		1		1		1	1				1	1		
P4						1							1							

Fig. (5). Selected composition and positional vector features. Rows show the type of the feature, i.e., CV stands for composition vector and P*i* denotes positional vector for i^{th} position in the tetrapeptide, while columns correspond to amino acids.

Table 1. Comparison of contribution of individual feature sets (composition vector, positional vectors, and collocation vector) on prediction of beta-turn types. The table reports MCC, Q^{res}_{pred}, Q^{res}_{obs}, and Q^{res}_{total} for 7-fold cross validation on dataset 426. All results are based on 8% sampling of non-beta-turns (8%NT) and use optimized SVM with $C = 3$ and $\gamma = 0.3$

Beta-turn type	MCC			
	Composition vector	**Positional vectors**	**Collocation vector**	**All features**
I	0.03	0.18	0.14	0.18
II	0.00[1]	0.23	0.23	0.23
VIII	0.00[1]	0.00	0.00[1]	-0.01
NS	0.00[1]	0.12	0.06	0.15
IV	0.00	0.11	0.06	0.12
Beta-turn type	Q^{res}_{total}			
	Composition vector	**Positional vectors**	**Collocation vector**	**All features**
I	90.4	78.1	86.7	78.3
II	96.2	79.6	94.0	86.1
VIII	97.2	97.2	97.2	97.0
NS	98.0	95.6	97.9	95.0
IV	90.0	81.5	87.2	80.4
Beta-turn type	Q^{res}_{obs}			
	Composition vector	**Positional vectors**	**Collocation vector**	**All features**
I	0.7	42.7	18.2	42.6
II	0.0[1]	69.3	26.8	53.5
VIII	0.0[1]	0.0	0.0[1]	0.0
NS	0.0[1]	18.1	2.5	25.3
IV	0.1	24.9	8.4	27.9
Beta-turn type	Q^{res}_{pred}			
	Composition vector	**Positional vectors**	**Collocation vector**	**All features**
I	26.4	19.8	23.7	20.0
II	0.0[1]	12.1	24.5	14.5
VIII	0.0[1]	0.0	0.0[1]	0.0
NS	0.0[1]	11.4	18.3	12.2
IV	13.9	18.4	18.8	18.3

[1]the corresponding 0's are due to lack of predictions of the corresponding beta-turn type, while the remaining 0's are due to lack or very low number of true positive predictions.

better than the predictions with the other sets of features, the collocation features are characterized by higher Q^{res}_{pred}, which indicates that they generate fewer false positives when compared with the number of true positives, i.e., they are more selective than the positional features. This is again expected since the collocation features are based on information about two positions in the beta-turn tetrapeptide, while positional vector features are associated with only one position.

We observe that for type VIII beta-turn TP = 0 in case of all individual feature sets, and when using the complete feature set. This shows that the proposed method is not capable of predicting this type of beta-turns. We hypothesize that this is due to lack of features that would allow differentiating these turns from other beta-turn types. As observed in [13], type VIII turns are characterized by high conformational heterogeneity (they cannot be stabilized by backbone hydrogen bond), and thus the underlying conformational preferences of amino acid are much harder to capture when compared with other beta-turn types. This is even more difficult when considering that beta-turns of type VIII have to be differentiated from non-beta-turns. We note that competing prediction methods are characterized by similarly poor predictions for this beta-turn type [11-13], see Table **2**.

Comparison with Competing Prediction Methods

Quality of Beta-Turn Type Prediction

Using the parameterized SVM classifiers, the training sets of the 426 dataset were downsampled to 3%NT and 8%NT, and 7-fold cross-validated tests were run on the complete test sets. Although our predictions were run at the level of tetrapeptides, the predictions were aggregated and each residue was tested with respect to prediction of a given beta-turn type and non-beta-turn in order to allow compari-

son with competing predictors. The resulting quality indices, which were computed for prediction of each beta-turn type, are summarized in Table **2**. The proposed method is based on the 8%NT sampling, while the results for the 3%NT sampling are provided for comparative purposes.

Considering the MCC, we observe that all methods are characterized by relatively poor performance on type VIII beta-turns. The proposed method has comparable performance when predicting NS turns against BETATURNS, and when predicting type IV beta-turns when compared with BTPRED and COUDES. However, the proposed method is outperformed when predicting all other beta-turn types. The most similar quality of predictions is provided by BTPRED. It is important to bear in mind the limited, and easily explicable information, that is used in the proposed model.

Considering the Q_{total}^{res}, the results are high and vary little between different prediction methods due to the high number of true negatives. Therefore, Q_{obs}^{res} and Q_{pred}^{res} are examined. The proposed method exhibits Q_{obs}^{res} over 40% for type I beta-turns, over 50% for type II beta-turns, and over 25% for NS and IV types of beta-turns. The results for type I and II beta-turns are relatively consistent with the competing methods, however, the proposed method surpasses COUDES and BTPRED for type IV beta-turns. Q_{pred}^{res} is relatively compara-

Table 2. Comparison of beta-turn type prediction quality expressed with MCC, Q_{pred}^{res}, Q_{obs}^{res}, and Q_{total}^{res} for 7-fold cross validation test on dataset 426[1]. The quality indices concerning the proposed method are shown in 8%NT and 3%NT columns

Beta-turn type	MCC				
	8%NT	**3%NT**	**COUDES**	**BTPRED**	**BETATURNS**
I	0.18	0.15	0.31	0.22	0.29
II	0.23	0.23	0.30	0.25	0.29
VIII	-0.01	0.01	0.07	0.06	0.02
NS	0.15	0.17	---	---	0.17
IV	0.12	0.12	0.11	0.03	0.23
Beta-turn type	Q_{total}^{res}				
	8%NT	**3%NT**	**COUDES**	**BTPRED**	**BETATURNS**
I	78.3	55.4	84.5	91.2	74.5
II	86.1	81.4	91.0	95.5	93.5
VIII	97.0	96.7	90.7	97.5	96.5
NS	95.0	94.1	---	---	98.1
IV	80.4	57.1	84.9	96.8	67.9
Beta-turn type	Q_{obs}^{res}				
	8%NT	**3%NT**	**COUDES**	**BTPRED**	**BETATURNS**
I	42.6	72.8	50.0	46.6	74.1
II	53.5	65.8	52.8	58.4	52.8
VIII	0.0	0.8	18.7	18.0	2.8
NS	25.3	31.8	---	---	13.3
IV	27.9	63.5	17.7	2.2	72.0
Beta-turn type	Q_{pred}^{res}				
	8%NT	**3%NT**	**COUDES**	**BTPRED**	**BETATURNS**
I	20.0	14.2	30.8	13.9	22.1
II	14.5	12.7	22.2	12.2	25.5
VIII	0.0	3.9	6.9	3.3	7.2
NS	12.2	12.0	---	---	23.7
IV	18.3	13.9	20.7	9.3	18.6

[1]Results for BTPRED taken from [12] and results for BETAETATURNS were taken from [11]. We note that BTPRED used a different test set. We also note that COUDES [13] predicts type I' and type II' turns whereas our method and BETATURNS combines them into NS type, and BTPRED ignores the entire NS category (and thus the corresponding cell in the table are left empty). All quality indices are based on per residue comparison.

Table 3. Comparison of beta-turn vs. non-beta-turn prediction accuracy with competing prediction methods. The prediction quality is expressed with MCC, Q^{res}_{total}, Q^{res}_{pred}, and Q^{res}_{obs} for 7-fold cross validation test on dataset 426. The quality indices concerning the proposed method are shown in 8%NT and 3%NT rows

Prediction Method		Q^{res}_{total}	Q^{res}_{pred}	Q^{res}_{obs}	MCC
Sequence-only based methods	proposed method (8%NT)	64.5	36.7	63.0	0.24
	proposed method (3%NT)	44.8	29.4	90.6	0.20
	Chou-Fasman [10, 52]	65.2	37.6	63.5	0.26
	Thornton [10, 23]	68.0	38.6	52.4	0.23
	GORBTURN [10, 53]	70.5	39.3	37.3	0.19
	1-4 & 2-3 correlation [10, 54]	59.1	32.4	61.9	0.17
	Sequence coupled [10, 22]	53.3	32.4	72.8	0.17
Methods that utilize PSSM and/or predicted secondary structure	SVM (multiple alignment) [43]	77.3	53.1	67.0	0.45
	BTSVM [44]	78.7	56.0	62.0	0.45
	BETATPRED2 (multiple alignment) [11, 27]	75.5	49.8	72.3	0.43
	COUDES ($\psi_{threshold}$= 0 for PSSM) [13]	74.8	48.8	69.9	0.42
	COUDES ($\psi_{threshold}$= -100 for PSSM) [13]	75.5	49.8	66.6	0.41
	SVM (single sequence) [43]	74.8	49.1	67.9	0.41
	BETATPRED2 (single sequence) [11, 27]	74.3	48.4	71.2	0.41
	KNN [28]	75.0	46.5	66.7	0.40
	BTPRED [10, 12]	74.4	48.3	57.3	0.35

ble across the methods, excluding type VIII beta-turns, with the proposed method ranging between 12 and 20% of the predictions being the actual beta-turns, where COUDES ranges from 20 – 30%, BTPRED 9 – 14% and BETATURNS 18 – 26%. We observe that the proposed method outperformed BTPRED with respect to this quality index.

Quality of Beta-Turn vs. Non-Beta-Turn Predictions

We also present comparison of results when combining prediction of all beta-turn types into the prediction of generic beta-turns. In this case, any predicted beta-turn type is considered as a generic beta-turn and the proposed method is compared against several related methods [13, 22, 23, 27, 43, 44, 52-54] based on 7-fold cross validation on dataset 426, see Table **3**. We note that several other beta-turn prediction methods, which are not included in our comparison, were also developed [55, 56].

We observe that the proposed method matches the prediction quality (measured using MCC) of the most accurate sequence-only based methods, and as expected has quality lower than the quality of methods that utilize multiple alignments and predicted secondary structure. We note that the Chou-Fasman method [52] which is a sequence-only based method characterized by similar MCC as the proposed method, is based on a set of probabilities assigned to each residue, conformational parameters, and positional frequencies determined by computing the relative frequency of a given secondary structure type as well as the fraction of residues appearing in that type of secondary structure. This means that the design of this method was based on data with known secondary structure, while the proposed method was based purely on knowledge of beta-turns and non-beta-turn, and without of knowledge of any other secondary structures. Most importantly, we note that only two sequence-only based methods, i.e., Thornton's method [23] and Chou's method [1], are available. In contrast to Thornton's method that addressed prediction of only type I and II beta-turns, the proposed method addressed prediction of five turn types, and provides quality comparable with that of the Thornton's method. Direct comparison with Chou's method, which predicts six beta-turn types (types I, I', II, II', VI, and VIII), is relatively difficult since these beta-turn types differ from the types predicted by the proposed method.

Table **3** shows that the proposed method finds 63% of all beta-turns, and that 36% of the predicted beta-turns are the actual beta-turns. This indicates that the selected features that were applied in the proposed method are in fact associated with beta-turns. We note that the Q^{res}_{obs} values of the proposed method are similar to values obtained by the competing methods that utilize the PSI-BLAST and/or predicted secondary structure, while we suffer lower values of Q^{res}_{pred}. The latter indicates that inclusion of the predicted secondary structure and evolutionary information allows for more selective predictions, i.e., removal of some false positive predictions.

Fig. (**6**) shows ROC curve (TP rate = TP / (TP + FN) vs. FP rate = FP / (FP + TN)) for the beta-turn predictions performed with the proposed method. The Figure shows that our results are substantially better than a random prediction, i.e., the ROC curve is above the diagonal line. For example, the results show that our method obtains 26.8% TP rate for 10% FP rate. We note that we could not draw ROC curve for beta-turn type predictions since different turn types overlap

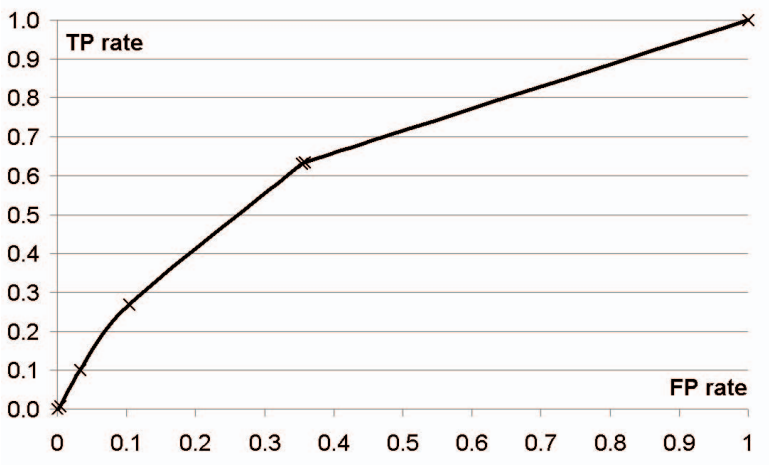

Fig. (6). ROC curve for the β-turn predictions of the proposed method for 7-fold cross validation on dataset 426; *x*-axis shows FP rate = FP / (FP + TN) and *y*-axis shows TP rate = TP / (TP + FN).

in the sequence, i.e., the same residue is often classified into several beta-turn types.

DISCUSSION AND CONCLUSION

The proposed method succeeds in providing similar performance to other methods that utilize the same information (only the sequence). The results for beta-turn type prediction have proven similar in certain regards to other machine learning methods that use additional information and the results for beta-turn vs. non-beta-turn predictions are consistent with the sequence-only based methods. At the same time, we observe that only two sequence-only methods are available for the prediction of beta-turn types from the entire protein chains, i.e., Thornton's method [23] that predicts only type I and type II beta-turns and Chou's method [1] that predicts beta-turn types I, I', II, II', VI, and VIII. In contrast, the proposed method predicts types I, II, IV, VII, and non-specific (NS) beta-turns, which are consistent with the targets of modern prediction methods [11]. We also observe that inclusion of additional information such as predicted secondary structure and PSI-BLAST profiles provides reduction of false positive predictions.

The main advantage of the proposed method is simplicity and interpretability of the underlying model. It uses only on the input protein sequence and it does not rely on additional methods. The main contribution of this work is the development of novel sequence-based features that allow identifying different beta-turns and differentiating them from non-beta-turns. The computed features, which are based on tetrapeptides (entire beta-turns) rather than a window centered over the predicted residues, provide a more biologically sound model. They include 12 novel features that are based on collocation of amino acid pairs with a single or double gap (which denotes inclusion of any amino acid) between them. The selected features further reaffirm the biological relevance of the model, focusing on amino acids that are known to be predisposed to form beta-turns. Virtually all collocated pairs used by the proposed method include Gly (G) that has the highest potential of any residue to form a beta-turn [9, 26]. The two motifs based on the double gap include DxxG and NxxG, and the only motif that does not include Gly is DxN. The above three amino acids, i.e., Gly (G), Asp (D), and Asn (N), are the top three that have the highest potential to form beta-turns, and additionally Asp and Asn are characterized the highest positional potential at positions 1 and 3, while Gly has the highest potential at positions 3 and 4 [9, 26]. At the same time, our model also includes several features that are geared towards exclusion of non-beta-turns. According to [9, 26], the residues with the lowest potential to form beta-turns include Met (M), Ile (I), Leu (L), and Val (V), and they are shown to be strongly detrimental to formation of beta-turns being in position 3. To this end, the proposed model includes features that encode occurrence of Ile, Leu, and Val at the position 3.

The datasets used to develop and test the proposed method can be freely accessed at http://biomine.ece. ualberta.ca/BTcollocation/BTcollocation.htm

ACKNOWLEDGEMENTS

This work was supported in part by NSERC Canada.

REFERENCES

[1] K.C. Chou, "Prediction of tight turns and their types in proteins", *Anal. Biochem*., vol. 286, pp. 1-16, November 2000.

[2] K.S. Kee and S.D. Jois, "Design of β-turn based therapeutic agents", *Curr. Pharm. Des*., vol. 9, pp. 1209-1224, 2003.

[3] K. Takano, Y. Yamagata, K. Yutani, "Role of amino acid residues at turns in the conformational stability and folding of human lysozyme", *Biochemistry*, vol. 39, pp. 8655-8665, June 2000.

[4] V. Pavone, G. Gaeta, A. Lombardi, F. Nastri, and O. Maglio, "Discovering protein secondary structures: classification and description of isolated alpha-turns", *Biopolymers*, vol. 38, pp. 705-721, June 1996.

[5] K.C. Chou, "Prediction and classification of alpha-turn types", *Biopolymers*, vol. 42, pp. 837-853, December 1997.

[6] Y.D. Cai and K.C. Chou, "Artificial neural network for predicting alpha-turn types", *Anal. Biochem*., vol. 268, pp. 407-409, March 1999.

[7] Y.D. Cai, K.Y. Feng, Y.X. Li, and K.C. Chou, "Support vector machine for predicting alpha-turn types", *Peptides*, vol. 24, pp. 629-630, April 2003.

[8] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers*, vol. 22, pp. 2577-2637, December 1983.

[9] E.G. Hutchinson and J.M. Thornton, "A revised set of potentials for β-turn formation in proteins", *Protein Sci.*, vol. 3, pp. 2207-2216, December 1994.

[10] H. Kaur and G.P. Raghava, "An evaluation of β-turn prediction methods", *Bioinformatics*, vol. 18, pp. 1508-1514, November 2002.

[11] K.S. Kaur and G.P. Raghava, "A neural network method for prediction of β-turn types in proteins using evolutionary information", *Bioinformatics*, vol. 16, pp. 2751-2758, November 2004.

[12] A.J. Shepard, D. Gorse, and J.M. Thornton, "Prediction of location and type of β-turns in proteins using neural networks", *Protein Sci.*, vol. 8, pp. 1045-1055, May 1999.

[13] P.F. Fuchs and A.J. Alix, "High accuracy prediction of β-turns and their types using propensities and multiple alignments", *Proteins*, vol. 59, pp. 828-839, June 2005.

[14] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSIBLAST, A new generation of protein database search programs", *Nucleic Acid Res*., vol. 25, pp. 3899-3402, September 1997.

[15] D.T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices", *J. Mol. Biol.*, vol. 292, pp. 195-202, September 1999.

[16] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles", *Proteins*, vol. 47(2), pp. 228-35, May 2002.

[17] M. Ouali and R.D. King, "Cascaded multiple classifiers for secondary structure prediction", *Protein Sci*., vol. 9, pp. 1162–1176, June 2000.

[18] M. Asgary, S. Jahandideh, P. Abdolmaleki, and A. Kazemnejad, "Analysis and prediction of β-turn types using multinomial logistic regression and artificial neural network", *Bioinformatics*, vol. 23(23), pp. 3125-3130, December 2007.

[19] B. Rost, "PHD, predicting one-dimensional protein structure by profile-based neural networks", *Methods Enzymol*., vol. 266, pp. 525-539, 1996.

[20] Y.D. Cai, X.J. Liu, X.B. Xu, and K.C. Chou, "Support vector machines for the classification and prediction of beta-turn types", *J. Pept. Sci*., vol. 8(7), pp. 297-301, July 2002.

[21] K.C. Chou and J.R. Blinn, "Classification and prediction of beta-turn types", *J. Protein Chem*., vol. 16(6), pp. 575-595, August 1997.

[22] K.C. Chou, "Prediction of beta-turns in proteins", *J. Peptide Res*., vol. 49, pp. 120-144, February 1997.

[23] C.M. Wilmot and J.M. Thornton, "Analysis and prediction of the different types of beta-turn in proteins", *J. Mol. Biol*., vol. 203(1), pp. 221-32, September 1988.

[24] T. Noguchi, T.H. Matsuda, and T. Akiyama, "PDB-REPRDB, a database of representative protein chains from the Protein Data Bank (PDB)", *Nucleic Acids Res*., vol. 29(1), pp. 219-220, January 2001.

[25] E.G. Hutchinson and J.M. Thornton, "PROMOTIF, A program to identify and analyze structural motifs in proteins", *Protein Sci*., vol. 5, pp. 212-220, February 1996.

[26] K. Guruprasad and S. Rajkumar, "Beta- and gamma-turns in proteins revisited, a new set of amino acid turn-type dependent positional preferences and potentials", *J. Biosci.*, vol. 25, pp. 143-156, June 2000.

[27] H. Kaur and G.P. Raghava, "Prediction of β-turns in proteins from multiple alignment using neural networks", *Protein Sci.*, vol. 12, pp. 627-634, March 2003.

[28] S. Kim, "Protein β-turn prediction using nearest-neighbor method", *Bioinformatics*, vol. 20, pp. 40-44, January 2004.

[29] B.W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme", *Biochim. Biophys. Acta*, vol. 405, pp. 442-451, October 1975.

[30] D.W. Elrod and K.C. Chou, "A study on the correlation of G-protein-coupled receptor types with amino acid composition", *Protein Eng*., vol. 15(9), pp. 713-715, September 2002.

[31] Y.D. Cai, X.J. Liu, and K.C. Chou, "Artificial neural network model for predicting protein subcellular location", *J. Comput. Chem*., vol. 26(2), pp. 179-82, January 2002.

[32] Z. Yuan, "Better prediction of protein contact number using a support vector regression analysis of amino acid sequence", *BMC Bioinformatics*, vol. 6, p. 248, October 2005.

[33] K.D. Kedarisetti, L. Kurgan, and S. Dick, "Classifier ensembles for protein structural class prediction with varying homology", *Biochem. Biophys. Res. Commun*., vol. 348, pp. 981-988, September 2006.

[34] L. Kurgan and L. Homaeian, "Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy", *Pattern Recognit.*, vol. 39(12), pp. 2323-43, December 2006.

[35] L. Homaeian, L. Kurgan, K.J. Cios, J. Ruan, and K. Chen, "Prediction of Protein Secondary Structure Content for the Twilight Zone Sequences", *Proteins*, vol. 69(3), pp. 486-498, November 2007.

[36] Y. Shi, J. Zhou, D. Arndt, D.S. Wishart, and G. Lin, "Protein contact order prediction from primary sequences", *BMC Bioinformatics*, vol. 9(1), p. 255, May 2008.

[37] J.T. Huang and J.P. Cheng, "Differentiation between two-state and multi-state folding proteins based on sequence", *Proteins*, vol. 72(1), pp. 44-9, July 2008.

[38] K. Chen, L. Kurgan, and M. Rahbari, "Prediction of protein crystallization using collocation of amino acid pairs", *Biochem. Biophys. Res. Commun.*, vol. 355, pp. 764-769, April 2007.

[39] K. Chen, L. Kurgan, and J. Ruan, "Prediction of flexible/rigid regions in proteins from sequences using collocated amino acid pairs", *BMC Struct. Biol.*, vol. 7, p. 25, April 2007.

[40] K. Chen, L. Kurgan, and J. Ruan, "Prediction of protein structural class using novel evolutionary collocation based sequence representation", *J. Comput. Chem*., vol. 29(10), pp. 1596-1604, July 2008.

[41] Y.Z. Chen, Y.R. Tang, Z.Y. Sheng, and Z. Zhang, "Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs", *BMC Bioinformatics*, vol. 9, p. 101, February 2008.

[42] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1999.

[43] Q. Zhang, S. Yoon, and W.J. Welch, "Improved method for predicting β-turn using support vector machine", *Bioinformatics*, vol. 21, pp. 2370-2374, May 2005.

[44] T.H. Pham, K. Satou, and T.B. Ho, "Prediction and analysis of β-turns in proteins by support vector machine", *Genome Inform*., vol. 14, pp. 196-205, December 2003.

[45] J. Platt, "Fast training of support vector machines using sequential minimal optimization", in *Advances in kernel methods - support vector learning*, B. Schoelkopf, C. Burges, and A. Smola, Eds., MIT Press, 1998, pp. 185-208.

[46] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murphy, "Improvements to platt's smo algorithm for SVM classifier design", *Neural Comput.*, vol. 13, pp. 637-49, March 2001.

[47] I. Witten and E. Frank, *Data Mining, Practical machine learning tools and techniques*, Morgan Kaufmann, 2005.

[48] L. Yu and H. Liu, "Feature selection for high-dimensional data, a fast correlation-based filter solution", in Tenth International Conference on Machine Learning, August 2003, pp. 856-863.

[49] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *J. Mach. Learn. Res*., vol. 3, pp.1289-305, March 2003.

[50] K. Kira and L.A. Rendell, "A practical approach to feature selection", in Ninth International Workshop on Machine Learning, July 1992, pp. 249-256.

[51] A. Perczel, I. Jákli, M.A. McAllister, and I.G. Csizmadia, "Relative stability of major types of beta-turns as a function of amino acid composition, a study based on Ab initio energetic and natural abundance data", *Chemistry*, vol. 9(11), pp. 2551-2566, June 2003.

[52] P.Y. Chou and G.D. Fasman, "Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins" *Biochemistry*, vol. 13, pp. 211-222, January 1974.

[53] C.M. Wilmot and J.M. Thornton, "β-turns and their distortions, a proposed new nomenclature", *Protein Eng*., vol. 3, pp. 479-493, May 1990.

[54] C.T. Zhang and K.C. Chou, "Prediction of beta-turns in proteins by 1-4 & 2-3 correlation model", *Biopolymers*, vol. 41, pp. 673-702, May 1997.

[55] Y.D. Cai, H. Yu, and K.C. Chou, "Prediction of beta-turns", *J. Protein Chem.*, vol. 17, pp. 363-376, May 1998.

[56] Y.D. Cai, Y.X. Li, and K.C. Chou, "Classification and prediction of beta-turn types by neural networks", *Adv. Eng. Software*, vol. 30, pp. 347-352, May 1999.

Received: June 20, 2008 Revised: July 11, 2008 Accepted: July 14, 2008