## Poster I-98

**Sequence based prediction of distortion in alpha-helix using Support Vector Machine**


**Authors:**
Ashish V Tendulkar (*KReSIT, IIT Bombay*)
Pramod P. Wangikar (*Dept of Chemical Engg, IIT Bombay*)

**Short Abstract:** The structure conservation in alpha-helix subclasses reveals sequence and context dependent factors causing variations. We capture these factors in form of sequence features and train support vector machine(SVM) for sequence based discrimination between regular and distorted helix. The SVM performs with 80.97% precision and 88.05% recall.

**Long Abstract:**

alpha-helices are the most regular structures in the proteins. However, they are distorted for a variety of reasons, such as occurrence of proline beyond third position causing a kink in the helix. In our earlier work, we had classified the helices into five important subclasses: regular, kinked, curved, n-cap and c-cap based on their structural properties. All the subclasses except the regular alpha-helix represent distortion of one form or the other. The structure conservation in the subclasses reveal the sequence and context dependent factors causing the distortions. The sequence-structure relationship in the subclasses can be used to predict structural variations in alpha-helix purely based on its sequence.

In our work, we train an support vector machine(SVM) to discriminate a given amino acid sequence between regular and distorted alpha-helix. The SVM is trained with amino acid sequences corresponding to regular alpha-helix as the positive examples and the sequences corresponding to all the other subclasses as the negative examples. We have total 28205 alpha-helix sequences, out of which 17500 sequences correspond to regular alpha-helix, while the rest correspond to the distorted alpha-helix. The sequences are represented with the features containing overall and position specific amino acid propensities in its various regions. The features are constructed by splitting the sequence into its subparts: first four amino acids corresponds to the N-terminus region, the last four amino acids corresponds to the C-terminus region, and the middle four amino acids corresponds to the middle region. We also construct the overall and position specific amino acid propensities corresponding to a sequence of four amino acids prior and following the helix sequence. Thus, a sequence is represented with 440 features. The SVM is trained with 70% of the total available sequences and 30% of the sequences are set aside for testing. The rigorous testing of the classifier is carried out with 10 fold cross-validation.

The weights are computed for different features based on the support vectors obtained after the training. The features corresponding to highest positive and negative weights are the most discriminating features. We found that the topmost positive weight is assigned to the overall propensity of Leu followed by the overall propensities of Ile, Ala and Met. The strong positive weight is also assigned to the Gly at the last position in helix. Moreover, positive weights are also assigned to overall amino acid propensities of Val, Arg, Gln, Trp, Lys and Cys. The overall amino a acid propensities of Phe and Tyr are assigned small positive

weights. It implies that regular alpha helix strongly prefers amino acids like Leu, Ile, Ala, Met, Val, Arg, Lys, Gln, Trp, Tyr and Cys. On the other hand, the strongest negative weight is assigned to overall propensity of Pro, followed by Asp and Gly. The position specific propensity of proline beyond the first turn of alpha-helix also receives high negative weights. It implies that Pro beyond the first turn in alpha-helix is instrumental in various structural perturbations in the helix. Although amino acids like Ser and Asp are not good helix formers, they are strongly preferred at the N-terminus of the alpha-helix. Moreover, we found that the regular alpha-helix strongly prefers acidic or neutral polar amino acids in its N-terminus, while basic polar amino acids in its C-terminus. It is also observed that the occurrences of amino acids such as Gly, Ser, and Lys in the middle as well as occurrences of hydrophobic amino acids at either termini of alpha-helix causes distortion in it. Based on 10-fold cross validation, we found that the SVM has average precision of 80.97% with standard deviation of 0.8 and average recall of 88.05% with standard deviation of 1.2. The average F1 score is 84.51%.

The correlation between structural variation in alpha-helices and their sequences is manifested by the performance of SVM based on sequence features. The splitting of the sequence into N-terminus, middle as well as C-terminus region along with prior and following amino acids helps in correctly capturing the context dependent factors in helix formation. The results presented here are useful for computational design of helices as well as for prediction of structural perturbations in alpha-helix purely based on its sequence.