# Protein Content Prediction Based on Principal Component Analysis and Support Vector Machine Regression

Mandana Rahbari, Dmitri Truhachev and Lukasz Kurgan

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4
{mandana,dmitrytr,lkurgan}@ece.ualberta.ca

## Abstract

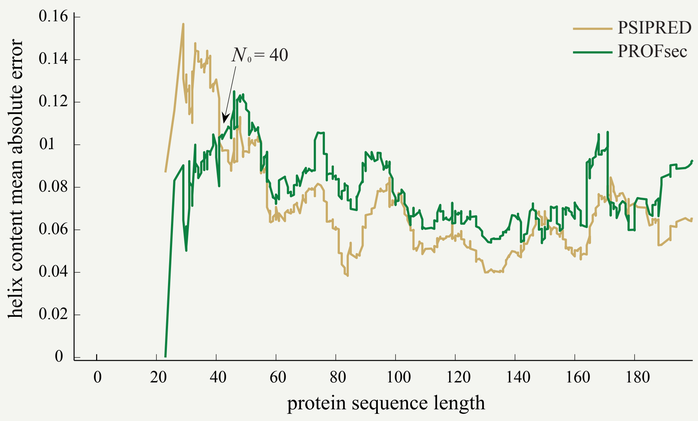
*Protein content prediction is an important intermediate problem in understanding higher-level protein conformations. We propose a novel machine learning algorithm for protein secondary structure content prediction. Two groups of features, structural and physicochemical, are constructed and an iterative feature selection process based on Principal Component Analysis (PCA) of the data is performed. The resulting learning model based on Support Vector Machine Regression outperforms state-of-the-art prediction methods in terms of content prediction error.*

## Motivation

The protein content was recently applied to prediction of structural classes, folding rates and transition, enzyme proteins and types, and analysis of protein interactions. Establishing a method which is capable of supplying more accurate content estimates is an important intermediate problem in setting up accurate methods for prediction of higherlevel (2D and 3D) structures.

## Method

In this work a new protein secondary structure content prediction method (LAMICA) based on machine learning is introduced. One training set (EVA977), and two test sets (EVA149 and EVA150) are employed. The proposed learning models are trained on features extracted from two complementary secondary structure prediction methods, PSIPRED and PROFsec, and amino acid physicochemical properties of the protein sequences. Although PSIPRED has lower overall absolute average error of content, PROFsec showed better performance for short protein sequences.




Protein 1o75

3D shape


Amino acid Sequence:
ETSYGYATLSYADYWAGELGQSRDVLLADR...
2D (secondary) structure:
CEEEEEEEEEHHHHHHHHHCCCCHHHCCCC...

## Feature Generation

In this section 552 features are computed. The features are generated using the predicted secondary structure by PSIPRED and PROFsec (structural features), and amino acid physicichmical properties (sequence based features).

**Structural features**: We compute a number of structural features using secondary structure prediction result of PROFsec for short sequences and PSIPRED for long sequences. For each protein sequence, we define $n_s^j$ to be the number of segments of length $j$ having predicted structure S. By $T_s$, we denote the total number of segments in a protein sequence with predicted secondary structure S.
Normalized segment counts are calculated by the following equations:

$$N_E^K = \frac{\sum_{j=k}^{20} n_E^j}{T_H + T_E} \qquad N_C^K = \frac{\sum_{j=k}^{20} n_C^j}{T_H + T_E + T_C} \quad \text{for } k = 3, 4, ..., 20. \quad \text{and} \quad N_H^K = \frac{\sum_{j=k}^{20} n_H^j}{T_H + T_E} \quad \text{for } k = 2, 3, ..., 20.$$

Composition moment vector and normalized maximum and average segment lengths for a protein of length $N$ are given by:

$$CMV_S^k = \frac{\sum_{j=1}^{n_S} n_{S,j}^k}{\prod_{i=0}^{k}(N-i)} \quad \text{for } k = 0, 1, ..., 5, \qquad \overline{M}_S = \frac{M_S}{N} \quad \text{and} \quad \overline{m}_S = \frac{m_S}{N}$$

Where $M_s$ denotes the length of the longest segment having structure S in the protein sequence and $m_s$ indicates the average length of segments with structure S.

**Sequence based features:** This set of features are computed based on 53 amino acid proprties such as average coil tendency, average medium contact, molecular weight, and hydrophobicity. We denote the value of the property $k$ of the $j$th amino acid in the protein sequence by $p_j^{(k)}$. For each property $k$, we define the autocorrelation $\rho_j^{(k)}$ (with $n$ shift) and standard deviation $\sigma^{(k)}$ as:

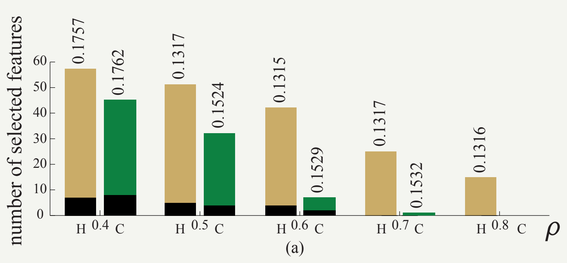
$$\rho_n^{(k)} = \frac{\sum_{j=1}^{N-n} p_j^{(k)} p_{j+n}^{(k)}}{N-n} \quad \text{for } n = 1, 2, ..., 6. \qquad \overline{p}^{(k)} = \frac{\sum_{j=1}^{N} p_j^{(k)}}{N} \quad \text{and} \quad \sigma^{(k)} = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(p_j^{(k)} - \overline{p}^{(k)}\right)^2}$$

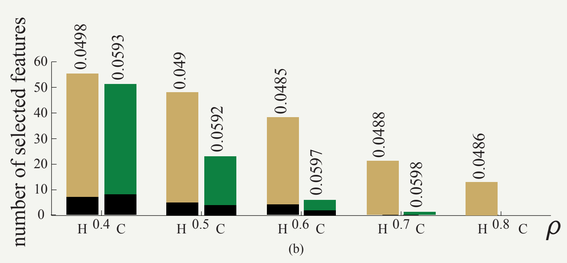
The composition vector and composition moment vectors of first and second degree are defined by:

$$CMV_i^k = \frac{\sum_{j=1}^{n_i} n_{ij}^k}{\prod_{d=0}^{k}(N-d)} \quad \text{for } k = 0, 1, 2.$$

## Feature Selection

Feature selection is done in two steps. In the first step, Pearson correlation coefficient between each feature and the content value in the training set (EVA977) is computed. We define a feature to be significant if the absolute correlation coefficient between the feature and secondary structure content is higher than a selected threshold $\overline{\rho}$. The threshold $\overline{\rho} = 0.6$ gives the lowest error for the prediction of the helix content for short sequences (figure a) and for long sequences (figure b), while, for the coil content prediction, the optimal value of the correlation threshold was found to be $\overline{\rho} = 0.5$ for both sets.

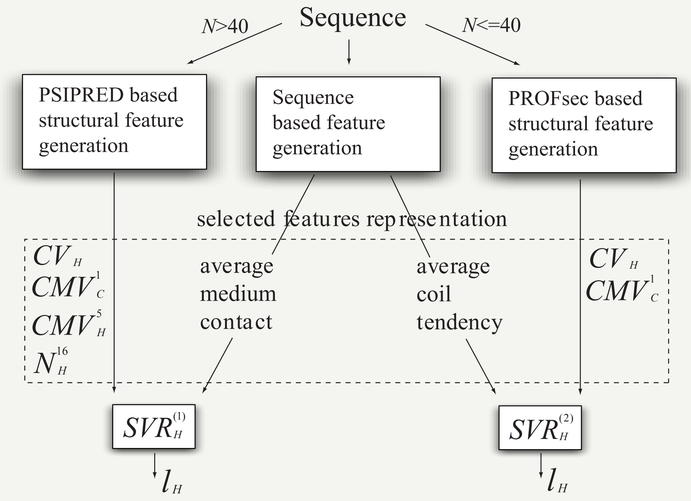



In the second stage of the feature selection, we further reduce the number of features using an iterative selection process. At every iteration of the process we perform principal component analysis (PCA) with 95% threshold for the explained variance. After transforming the data back to the original space we rank the features and eliminate the weakest feature.

## Algorithm

For both helix content ($l_H$) and coil content ($l_C$) prediction, two SVR based models were computed, one for large and the other for small proteins. The diagram demonstrates the architecture of the proposed method for helix content. The strand content ($l_E$) is calculated as $l_E = 1 - l_C - l_H$. All four SVRs, employed for helix and coil content prediction, use polynomial kernels. For helix content prediction, the complexity parameter $C$ is $C = 1$ for $SVR_H^{(1)}$ and $C = 25$ for $SVR_H^{(2)}$. For coil content prediction, the complexity parameter is $C = 21$ for (long sequences) $SVR_C^{(1)}$ and $C = 42$ for $SVR_C^{(2)}$ (short sequences).

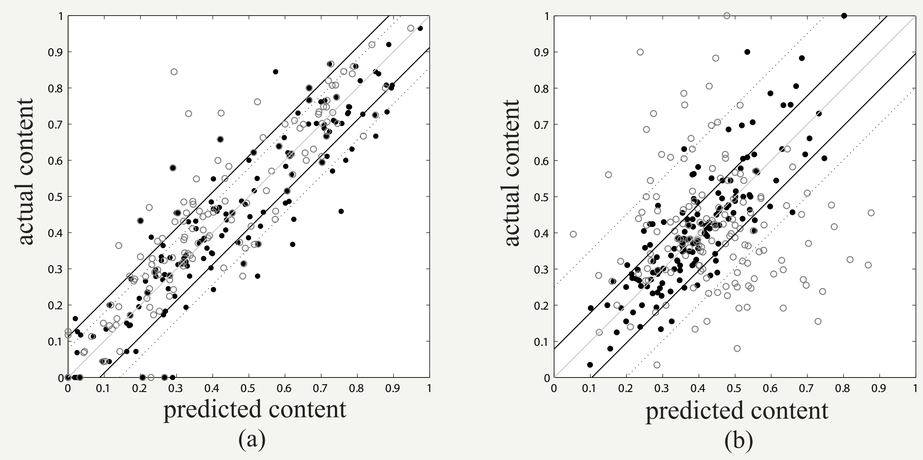



## Results

Comparison of the average absolute content prediction error for helix, coil and strand prediction between LAMICA and nine competing methods is shown in the table below. Experiments indicate that the proposed method outperforms state-of-the-art prediction methods in terms of content prediction error. For helix content prediction, our method shows 5-11% improvement over the second best competing method. For coil and strand content prediction, the achieved improvements are 0.5-4% and 4-10% respectively.

| target | dataset | Seq. size | PSIPRED | PROFsec | PHD | PHDPSI | SSPRO | PG | Zhang98 | Zhang01 | PSSC-core | LAMICA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Helix | EVA149 | N ≤ 40 | 0.2080 | 0.1192 | 0.1370 | 0.1447 | 0.2176 | - | 0.1698 | 0.2969 | 0.1814 | 0.1192 |
| | | N > 40 | 0.0471 | 0.0560 | 0.0671 | 0.0603 | 0.0580 | - | 0.1131 | 0.0928 | 0.1023 | 0.0467 |
| | | All | 0.0589 | 0.0607 | 0.0723 | 0.0666 | 0.0698 | - | 0.1173 | 0.0949 | 0.1081 | 0.0521 |
| | EVA150 | N ≤ 40 | 0.1559 | 0.0931 | 0.1450 | - | - | 0.1037 | 0.2167 | 0.1698 | 0.1421 | 0.0933 |
| | | N > 40 | 0.0695 | 0.0783 | 0.0766 | - | - | 0.0703 | 0.1198 | 0.0929 | 0.1152 | 0.0680 |
| | | All | 0.0851 | 0.0810 | 0.0889 | - | - | 0.0763 | 0.1372 | 0.0944 | 0.1201 | 0.0725 |
| Coil | EVA149 | N ≤ 40 | 0.1954 | 0.1475 | 0.1434 | 0.1411 | 0.1694 | - | 0.1369 | 0.1035 | 0.1250 | 0.1452 |
| | | N > 40 | 0.0641 | 0.0577 | 0.0740 | 0.0739 | 0.0729 | - | 0.1094 | 0.0960 | 0.1004 | 0.0577 |
| | | All | 0.0738 | 0.0644 | 0.0791 | 0.0788 | 0.0800 | - | 0.1114 | 0.0961 | 0.1022 | 0.0641 |
| | EVA150 | N ≤ 40 | 0.1353 | 0.0786 | 0.1142 | - | - | 0.0982 | 0.1245 | 0.1686 | 0.0976 | 0.0832 |
| | | N > 40 | 0.0761 | 0.0814 | 0.0824 | - | - | 0.0842 | 0.1213 | 0.1112 | 0.0644 | 0.0642 |
| | | All | 0.0867 | 0.0809 | 0.0881 | - | - | 0.0867 | 0.1219 | 0.1123 | 0.0704 | 0.0676 |
| Strand | EVA149 | N ≤ 40 | 0.1231 | 0.1429 | 0.1578 | 0.1679 | 0.1568 | - | 0.1873 | 0.1934 | 0.1396 | 0.1382 |
| | | N > 40 | 0.0489 | 0.0488 | 0.0674 | 0.0588 | 0.0597 | - | 0.1025 | 0.1054 | 0.0976 | 0.0471 |
| | | All | 0.0544 | 0.0558 | 0.0741 | 0.0669 | 0.0668 | - | 0.1088 | 0.1063 | 0.1007 | 0.0483 |
| | EVA150 | N ≤ 40 | 0.0643 | 0.068 | 0.1163 | - | - | 0.0931 | 0.2086 | 0.0195 | 0.1555 | 0.0806 |
| | | N > 40 | 0.0508 | 0.0637 | 0.0632 | - | - | 0.0617 | 0.1092 | 0.0880 | 0.0939 | 0.0488 |
| | | All | 0.0532 | 0.0645 | 0.0728 | - | - | 0.0674 | 0.1271 | 0.0867 | 0.1050 | 0.0508 |

Comparison of the predicted content for LAMICA and PROFsec




Hollow circles correspond to the PROFsec predictions and the black dots correspond to the proposed prediction method on EVA150 for helix content (figure a) and coil conent (figure b) prdiction.

## Conclusion

A novel machine learning method called LAMICA for prediction of protein secondary structure content is proposed. Two sets of learning features are generated, the structural features and sequence based physicochemical features. To reduce the prediction error, two separate SVR based learning models, one for short and one for long sequences, are constructed. Experimental results obtained using two independent test sets demonstrate that the prediction error of LAMICA is smaller than the error of current prediction techniques reported in the literature, including content predictions performed directly from sequence and predictions computed from predicted secondary structure.

## References

[1] Birzele,F., Kramer,S. (2006) A new representation for protein secondary structure prediction based on frequent patterns, Bioinformatics, 22, 2628-2634.
[2] Homaeian,L., Kurgan,L.A., Ruan,J., Cios,K.J., Chen,K. (2007) Prediction of protein secondary structure content for the twilight zone sequences, Proteins, 69, 486-498.
[3] Jones,D.T. (1999) Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices, J. Mol. Biol., 292, 195-202.
[4] Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks, Methods Enzymol, 266, 525-539.
[5] Rost,B., Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy, J Mol Biol, 2, 584-599.
[6] Zhang,C.T., Zhang,Z., He,Z. (1998) Prediction of the secondary structure contents of globular proteins based on three structural classes, J Protein Chemistry, 17, 261-272.
[7] Zhang,Z., Sun,Z., Zhang,C.T. (2001) A New Approach to Predict the Helix/Strand Content of Globular Proteins, J. theor. Biol., 208, 65-78.