

Developing a new SVM classifier for the extended ES protein structure prediction

Piotr Fabian

Silesian Technical University
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: pfabian@polsl.pl

Katarzyna Stapor

Silesian Technical University
ul. Akademicka 16, 44-100 Gliwice, Poland
Email: kstapor@polsl.pl

Abstract—This article presents a new SVM classifier for the prediction of the extended early-stage (ES) protein structures. The classifier is based on physicochemical features and position-specific scoring matrix (PSSM). Experiments have shown that prediction results for specific classes are significantly better than those already obtained.

I. INTRODUCTION

THE INTEREST of biologists is to determine the shape of amino acid chains that make up proteins. Accurate measurement methods involving the observation and measurement of real chains are troublesome, so the shape is often predicted basing on the amino acid sequence itself. In addition to methods that simulate the behavior of atoms and larger particles in accordance with known physical laws, methods based on machine learning are used. The shape of an unknown protein is predicted on the basis of the assumption of a similar effect of the amino acid sequence on the shape of the different proteins.

The shape of the chain may be predicted with less details than exact coordinates of atoms in the three-dimensional space. The concept of secondary structure concerns the local shape of the chain, classified as a helix, strand and similar cases, defined by biologists. Predicting the secondary structure consists in assigning individual segments of the chain to one of several classes describing the local shape of the chain. The number of classes may vary in different algorithms. For example, the DSSP algorithm defines eight different classes. Methods operating on the so-called structural code define seven classes. A review of methods used to predict the shape of proteins can be found in the literature. As mentioned in [1], identical sequences of pentapeptides exist with completely different tertiary structures in proteins. On the other hand, different amino acid sequences can have approximately the same three-dimensional structure. However, the patterns of sequence conservation can be used for protein structure prediction. The secondary structure local shape is commonly used for “ab initio” methods as a common starting conformation for precise protein structure prediction. A large number of experiments and theoretical evidence suggests that local structure is frequently encoded in short segments of protein sequence. A definite relation between the amino acid sequences of a region folded into a supersecondary structure has been found. It was also found that they are independent of the remaining

sequence of the molecule. Early studies of local sequence-structure relationships and secondary structure prediction were based on either simple physical principles or statistics. Nearest neighbor methods use a database of proteins with known three-dimensional structures to predict the conformational states of test proteins. Some methods are based on nonlinear algorithms known as neural nets or Hidden Markov Models. In addition to studies of sequence-to-structure relationships focused on determining the propensity of amino acids for predefined local structures, others involve determining patterns of sequence-to-structure correlations. The evolutionary information contained in multiple sequence alignments has been widely used for secondary structure prediction. Prediction of the percentage composition of α -helix, β -strand and irregular structure based on the percentage of amino acid composition, without regard to sequence, permits proteins to be assigned to groups, as all α , all β , and mixed α/β . Structure representation is simplified in many models. Side chains are limited to one representative virtual atom; virtual $C\alpha$ - $C\alpha$ bonds are often introduced to decrease the number of atoms present in the peptide bond. The search for structure representation in other than the ϕ , ψ angles conformational space has been continuing. Other models are based on limitation of the conformational space. One of them divided the Ramachandran map into four low-energy basins. In another study, all sterically allowed conformations for short polyalanine chains were enumerated using discrete bins called mesostates. The need to limit the conformational space was also asserted.

The model introduced in [1] is based on limitation of the conformational space to the particular part of the Ramachandran map. This part is represented by an elliptical path which traverses areas corresponding to well defined secondary structural motifs on the Ramachandran plot. The structures created according to this limited conformational subspace are assumed to represent early-stage structural forms of protein folding in silico. In contrast to commonly used base of final native structures of proteins, the early-stage folding conformation of the polypeptide chain is the criterion for structure classification.

This article presents two methods for predicting structural codes and results for selected classes of the structural code. The methods assume the possibility of determining the local protein structure only on the basis of a known sequence of

amino acids, without implementing any physical or chemical relationships between the particles other than precomputed features for specific amino acids. The sequence of amino acids is described by strings of symbols (letters) representing 20 amino acids, while the structural code with a sequence of symbols denoting individual classes of the local shape. For structural code, there are seven classes. Thus, the task of predicting a secondary structure can be defined as a search for a function mapping a set of words over a 20-character alphabet into a set of words over a 7-character alphabet. There is a set of learning data, containing proteins with known shape. Methods for predicting the secondary structure do not usually produce accurate results and are therefore evaluated by quality measures that specify the fraction of correctly-enrolled classes for experiments involving previously examined proteins. For the secondary structure prediction, modern methods achieve accuracy of about 80%. Achieving high accuracy (over 90%) is hampered by the ambiguous classification of the local shape, especially at the ends of the chain fragments belonging to one class. For the structural code, the most commonly method uses contingency tables described later.

II. METHODS AND ALGORITHMS

A. The structural code

Predicting the three-dimensional shape of proteins may be implemented as a simulation of atoms and particles, where all known physical forces are involved and influence the dynamics of the whole system. This approach is called “ab initio”. The final shape of the protein is a result of minimizing the energy of the whole system. However, the number of variables to take into account is enormous and the time complexity of algorithms implementing this idea makes it difficult to use this approach for longer chains of amino acids. The “ab initio” method needs a starting point - an initial conformation of the chain. Good starting point results from predicted secondary structure or structural codes. In our experiments we have tried to predict the structural code, which is described in [1], [2]. The local shape of amino acid chain results from the values of the ϕ and ψ dihedral angles. Observations of angles occurring in chains are presented in the so-called Ramachandran graph. Observations show that in that two-dimensional space (ϕ , ψ) clusters are formed describing possible pairs of angular values. A method used to classify angle pairs (ϕ , ψ) into one of those clusters defines an ellipse in the plane (ϕ , ψ), divides it into seven segments, and determines which segment is the closest to the sample. The structural code does not directly map into classes defined for secondary structures. For certain codes there is a rough mapping: the code C corresponds to an α -helix, E and F represent β -sheets while other codes correspond to a loop.

B. The SVM method applied to the structural code

The SVM (support vector machine) ([3]), method is widely used in machine learning and protein shape prediction [4]. This method allows to classify vectors in a multidimensional

feature space. However, the method was mainly applied to the secondary structure of proteins, not to the structural code.

As stated in the paper [4], SVM has shown promising results on several biological pattern classification problems. This method became a standard tool in bioinformatics. SVMs have been successfully applied to the recognition of protein translation-initiation sites in DNA sequences and functional annotation of genes from expression profiles.

C. Feature extraction

For predicting the shape of proteins, we have tried to used different features, mapped to numbers. Our experiments involve some physicochemical features and features based on statistics.

Physicochemical features have been already used to predict the protein secondary structure, as described in [6]. Following features have been used: hydrophobic values (F1), net charge (F2), side chain mass (F3), probabilities of conformation for the three secondary structures H, E and C (F4, F5, F6). The values have been defined for each of 20 amino acids and are presented in table I.

1) *Hydrophobic values*: For protein folding, polar residues prefer to stay outside of protein to prevent non-polar (hydrophobic) residues from exposing to polar solvent, like water. Therefore, hydrophobic residues appearing periodically can be used to predict protein secondary structure. In general, the residues in α -helix structure are made up of two segments: hydrophobic and hydrophilic. However, β -sheet structure is usually influenced by the environment, so this phenomenon is not obvious. In other words, hydrophobic residue affects the stability of secondary structure. The hydrophobic values of amino acids can also be obtained from Amino Acid index database (or AAindex, [5]). Higher positive values mean, that the residue is more hydrophobic.

2) *Net charges*: There are five amino acids with charges: R, D, E, H and K. Because residues with similar electric charges repel each other and interrupt the hydrogen bond of the main chain, they are disadvantageous for α -helix formation. Besides, succeeding residues of β -sheet cannot be with similar charges. This information helps to predict the secondary structure. The net charge of amino acids can be taken from the Amino Acid index database (or AAindex). The value 1 represents positive charge, the value -1 represents a negative charge.

3) *Side chain mass*: Although the basic structure is the same for 20 amino acids, the size of the side chain group still influences protein folding. First, the side chain R group is distributed in the outside of the main chain of α -helix structure, but the continuous large R groups can make α -helix structure unstable, thereby disabling amino acids from forming α -helix structure. Next, the R group with ring structure like proline (P) is not easy to form α -helix structure. Proline is composed of 5 atoms in a ring, which is difficult to reverse and is also not easy to generate a hydrogen bond. Finally, we observe that the R group of β -sheet structure is smaller than those of other structures, in general.

TABLE I
FEATURE VALUES FOR INDIVIDUAL AMINO ACIDS

AA	F1	F2	F3	F4	F5	F6
A	1.8	0	15.0347	0.49	0.16	0.35
R	-4.5	1	100.1431	0.42	0.19	0.39
N	-3.5	0	58.0597	0.27	0.13	0.60
D	-3.5	-1	59.0445	0.31	0.11	0.58
C	2.5	0	47.0947	0.26	0.29	0.45
E	-3.5	-1	73.0713	0.49	0.15	0.36
Q	-3.5	0	72.0865	0.46	0.16	0.38
G	-0.4	0	1.0079	0.16	0.14	0.70
H	-3.2	1	81.0969	0.30	0.22	0.48
I	4.5	0	57.1151	0.35	0.37	0.28
L	3.8	0	57.1151	0.45	0.24	0.31
K	-3.9	1	72.1297	0.40	0.17	0.43
M	1.9	0	75.1483	0.44	0.23	0.33
F	2.8	0	91.1323	0.35	0.30	0.35
P	-1.6	0	41.0725	0.18	0.09	0.74
S	-0.8	0	31.0341	0.28	0.19	0.54
T	-0.7	0	45.0609	0.25	0.27	0.48
W	-0.9	0	130.1689	0.37	0.29	0.35
Y	-1.3	0	107.1317	0.34	0.30	0.36
V	4.2	0	43.0883	0.30	0.41	0.29

4) *Conformation parameters*: Conformation parameters are the probabilities of creating particular types of the secondary structure by a given amino acid. In general, protein secondary structure is divided into three types: α -helix (H), β -sheet (E) and coil (C), so that there are three values for each amino acid. In the feature extraction, all the conformation parameters are calculated from a data set. The conformation parameters for each amino acid S_{ij} are defined as follows: $S_{ij} = \frac{a_{ij}}{a_i}$, where $i = 1, \dots, 20$, $j = 1, 2, 3$. In that formula, i indicates one of 20 amino acids, j indicates the 3 types of secondary structure: H, E and C. Here, a_i is the amount of the i -th amino acid in a data set whereas a_{ij} is the amount of the i -th amino acids with the j -th secondary structure. The conformation parameters for each amino acid in a data set are shown in table I as F4, F5 and F6. The reason of using conformation parameters as features is that the folding of each residue has some correlation with forming a specific structure.

5) *PSSM profiles*: The position-specific scoring matrix (PSSM) is a commonly used representation of motifs in biological sequences. The matrix is defined for a given set of proteins and specifies the probability of finding a given amino acid at a given position. There are 20 amino acids, so there are 20 values from the PSSM matrix for each position. When generated for a sliding window of the length 15, we have additionally $20 \cdot 15 = 300$ features.

To use the SVM method, we need a feature vector for each position of the amino acid chain. The feature vector should include information about the context in which a given amino acid occurs. To get a clear result for a given position in the chain, we choose a window of 15 elements

and describe feature for the element in the middle of it (at the 8th position). Therefore, we construct a feature vector using a sliding window of 15 elements. We slide it through the amino acid chain and for each position, we retrieve six features from the table I. In this approach, we get a vector of 90 features. The window size (15) was chosen arbitrarily after experimenting with shorter and longer windows. With additional PSSM features, we get a vector of 390 features for each position of the chain. Initial values at the ends of the chain, where the windows contains positions outside of the chain, are impossible to compute. So we have decided to cut the analyzed part of all chains by 7 positions from both sides, obtaining full coverage of data and complete feature vectors.

D. Contingency tables

The structural code for amino acid chains may be predicted using statistical methods, as described in [1]. The idea of contingency tables described in this article assumes, that the sequence of amino acids determines or at least influences the local shape of the protein chain. To reduce the complexity of computations it was assumed, that a sequence of only four amino acids (so called tetrapeptide) influences the secondary code within this sequence. Unfortunately the tetrapeptide does not strictly determine the shape, because there are cases, where identical tetrapeptides lead to different shapes. The contingency table collects information about tetrapeptides-shape relation in a given set of training data (over 1.5 million of tetrapeptides in the cited paper [1]). There are 7 structural codes and 20 different amino acids, so the table is a matrix of the size $7^4 \times 20^4 = 2401 \times 160000$ elements. Based on the training set, statistics are generated to describe how many times a given 4-element structural code occurred for a given tetrapeptide (so we have 2401×160000 counters). Then, probability values are computed and stored in the array to predict structural codes. After collecting data, regularities may be observed in the contingency table. Results of structural code prediction using contingency tables are presented e.g. in [7] and summarized in table III (in the first row).

III. TRAINING AND TEST DATA

To test the performance of the SVM classifier, we have taken a set of proteins called CB513 (<http://comp.chem.nottingham.ac.uk/disspred/datasets/CB513>). Training and testing sets in such experiments should contain carefully selected proteins to avoid distortion in results of experiments. If the training set contained proteins similar to proteins selected for testing, results of prediction would be distorted, possibly improved. The presence of training samples similar to testing samples makes the classification task easier for the classifier and thus is avoided in experiments. The CB513 is a set of selected proteins, where no pair of proteins shares more than 25% sequence identity over a length of more than 80 residues. All proteins are available in the PDB protein database (<http://www.rcsb.org/pdb/home/home.do>) with precise three-dimensional shape.

TABLE II
RESULTS OF STRUCTURAL CODE PREDICTION WITH SVM ON CB513, 9-FOLD EXPERIMENT

S. code	s1	s2	s3	s4	s5	s6	s7	s8	s9	Total
A	17.39%	18.29%	12.90%	17.86%	18.37%	15.38%	19.35%	22.86%	12.90%	17.28%
B	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C	62.37%	64.19%	61.13%	54.75%	65.68%	64.02%	66.62%	60.30%	62.95%	62.45%
D	14.77%	18.51%	19.78%	13.07%	17.33%	14.97%	17.65%	17.64%	15.24%	16.55%
E	87.73%	83.33%	85.94%	88.57%	90.27%	88.07%	86.30%	87.87%	86.59%	87.19%
F	0.97%	1.14%	0.86%	0.94%	0.51%	0.73%	0.71%	0.56%	0.68%	0.79%
G	15.31%	19.45%	18.29%	17.87%	12.39%	16.20%	20.22%	19.72%	20.16%	17.74%
Total	51.25%	51.17%	49.96%	47.57%	53.65%	51.68%	55.06%	51.22%	49.74%	51.26%

TABLE III
COMPARISON OF STRUCTURAL CODE PREDICTION WITH CONTINGENCY TABLES AND SVM

S. code	A	B	C	D	E	F	G
Accuracy cont. table	2.26%	1.95%	82.32%	5.95%	38.05%	17.56%	10.95%
Accuracy SVM	17.28%	0.00%	62.45%	16.55%	87.19%	0.79%	17.74%

IV. EXPERIMENTS AND EVALUATION OF THE RESULTS

For the secondary structure prediction, the traditional measure of the prediction quality is called Q_3 , which is defined as the number of correctly predicted residues divided by the length of the chain. However, it was shown, that the evaluation should be more specific. For seven codes, a slightly modified version of Q_3 called Q_7 has been used. Q_7 was described in [8]: $Q_7 = \frac{N_{r7}}{N} \cdot 100$, where N expresses the total number of amino acids in the polypeptide under consideration, N_{r7} expresses the number of correctly predicted amino acids representing the structural form r .

Experiments have been implemented in the R language with the Machine Learning package (*mlr*). A set of precomputed PSSM matrices was used. The Radial Basis Function kernel (RBF, Gaussian kernel) was used in the SVM classifier - *classif.ksvm* from the *mlr* package, which implements multi-class classification. The RBF kernel for two samples y and y' representing feature vectors, is defined as:

$$K(y, y') = \exp\left(-\frac{\|y - y'\|^2}{2\sigma^2}\right).$$

The term $\|y - y'\|^2$ is a squared Euclidean distance between feature vectors y and y' . Values of the RBF kernel are in the range from 0 (for very distant samples, in the limit) to 1 (for equal samples). It is interpreted as a measure of similarity.

Results of experiments on the CB513 set are shown in the table III. As this table shows, the results on some structural

codes differ significantly for two tested methods. Especially, the code A, D and E are predicted better by the SVM method. Code E is usually found at the end of β -twists, which may lead to the conclusion, that SVM is better at borders of motifs. Reasons of differences on other codes need further research.

REFERENCES

- [1] M. Brylinski, L. Konieczny, P. Czerwonko, W. Jurkowski and I. Roterman, "Early-Stage Folding in Proteins (In Silico) Sequence-to-Structure Relation," *Journal of Biomedicine and Biotechnology*, vol. 2 (2005), pp. 65–79, <http://dx.doi.org/10.1155/JBB.2005.65>.
- [2] B. Kalinowska, P. Alejster, K. Sałapa, Z. Baster and I. Roterman, "Hypothetical in silico model of the early-stage intermediate in protein folding," *Journal of Molecular Modeling*, vol. 19, 20 13, pp. 4259–4269, <https://dx.doi.org/10.1007%2Fs00894-013-1909-6>.
- [3] Bishop C., "Pattern Recognition and Machine Learning," Springer-Verlag, New York, 2006.
- [4] J. J. Ward, J. L. McGuffin, B. F. Buxton and D. T. Jones, "Secondary structure prediction with support vector machines," *Bioinformatics* vol. 19, 2003, <https://doi.org/10.1093/bioinformatics/btg223>.
- [5] S. Kawashima, M. Kanehisa, "AAindex: Amino Acid index database," *Nucleic Acids Research* 28(1):374, 2000, <http://dx.doi.org/10.1093/nar/28.1.374>.
- [6] Yin-Fu Huang and Shu-Ying Chen, "Extracting Physicochemical Features to Predict Protein Secondary Structure," *The Scientific World Journal*, vol. 20 13, <http://dx.doi.org/10.1155/2013/347106>.
- [7] B. Kalinowska, P. Fabian, K. Stapor and I. Roterman, "Statistical dictionaries for hypothetical in silico model of the early-stage intermediate in protein folding," *Journal of Computer-Aided Molecular Design*, vol. 29, 2015, <https://dx.doi.org/10.1007%2Fs10822-015-9839-2>.
- [8] M. Bryliński, L. Konieczny and I. Roterman, "SPI - Structure Predictability Index for Protein Sequences," *In Silico Biology*, vol. 5, no. 3, pp. 227–237, 2005.