Combining PSSM and physicochemical feature for protein structure prediction with support vector machine

Article *in* Journal of Physics Conference Series · April 2017
DOI: 10.1088/1742-6596/835/1/012006

4 authors, including:

Toto Haryanto
University of Indonesia
17 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Lailan Sahrina Hasibuan
Bogor Agricultural University
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Muhammad Asyhar Agmalaro
Bogor Agricultural University
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Combining PSSM and physicochemical feature for protein structure prediction with support vector machine

2017 J. Phys.: Conf. Ser. 835 012006

(http://iopscience.iop.org/1742-6596/835/1/012006)

IOP Conf. Series: Journal of Physics: Conf. Series **835** (2017) 012006 doi:10.1088/1742-6596/835/1/012006

Combining PSSM and physicochemical feature for protein structure prediction with support vector machine

I Kurniawan, T Haryanto*, L S Hasibuan, M A Agmalaro

Department of Computer Science, Bogor Agricultural University, Indonesia

E-mail: totoharyanto@apps.ipb.ac.id

Abstract. Protein is one of the giant biomolecules that act as the main component of the organism. Protein is formed from building blocks namely amino acids. Hierarchically, the structure of protein is divided into four levels: primary, secondary, tertiary, and quaternary structure. Protein secondary structure is formed by amino acid sequences that would form three-dimensional structures and have information about the tertiary structure and function of proteins. This study used 277,389 protein residue data from enzyme categories. Position-specific scoring matrix (PSSM) profile and physicochemical are used for features. This study developed support vector machine models to predict the protein secondary structure by recognizing patterns of amino acid sequences. The Q3 results showed that the best scores obtained are 93.16% from the dataset that has 260 features with the radial kernel. Combining PSSM and physicochemical feature additions can be used for prediction.

1. Introduction

Protein is one of the giant biomolecules that act as the main component of the organism. Protein is formed by amino acid sequence linked to each other by peptide bonds through the carbon chain so that to form complex structures. The process of protein formation process involves translation and transcription. The transcription process is writing the genetic code of deoxyribonucleic acid (DNA) into messenger ribonucleic acid (mRNA) to encode each nitrogenous bases of DNA into RNA in the form of nitrogenous bases. The translation process is translating each nitrogenous bases in mRNA into the amino acid sequence [1]

Hierarchically, proteins are divided into four levels, primary structure, secondary structure, tertiary structure, and tertiary structure. The primary structure is the sequence of amino acids that form a polypeptide chain. Secondary structure is a series of amino acids that form a three-dimensional structure alpha helix (H), beta-sheet (B), or coil (C) which is the result of amino acid sequences that bind to the peptide bond [1]. Tertiary structure is a combination of the secondary structure after folding process. The role of the protein can be known if it formed the tertiary structure in 3D. However, the tertiary structure can be determined if the structures previously known.

Conventionally, the structure of a protein can be detected by X-ray crystallography [2] and Nuclear Magnetic Resonance (NMR) [3]. This technique is used to find the primary structure and new structure of proteins and contribute to the validation of protein structures. But both these techniques take time and cost are relatively expensive. Thus, the computing-based approach is widely used to

* Corresponding author

predict the secondary structure of proteins. Protein secondary structure prediction is performed to find the 3D structure of proteins based on the primary structure of proteins. There are two methods of protein secondary structure prediction, the method of comparative modeling and modeling of de novo or ab initio. Comparative protein modeling to predict protein structure based on the structure of other proteins that are known, while the method of ab initio or de novo to determine the structure of proteins from primary sequence without comparing with other protein structure [4].

The basis of the protein secondary structure classification system is by recognizing pattern in the protein amino acid sequences. Many related types of research have been conducted to predict the secondary structure of this protein in order to get the best accuracy from previous studies. Huang and Chen [5] predict the secondary structure of proteins by using Support Vector Machine with Position Specific Scoring Matrix (PSSM) and four feature extraction consisting of conformation parameters, net charges, hydrophobic, and side chain mass. This research resulted in the accuracy of the value of Q3 score of 75.92%. Position based structure was proposed for secondary structure prediction with accuracy in ranging 87.6% up to 95.6% [6].

Support Vector Network had been introduced [7] and known as Support Vector Machine (SVM). This method was first presented at the Annual Workshop on Computational Learning Theory. The basic concept of SVM is a harmonious combination of computational theories that already exist such as margin hyperplane and kernel. SVM perform a technique to discover the function of separator that separates the two sets of data from two different classes. This method is a machine learning method that works on the principle of Structural Risk Minimization (SRM) with the goal of finding the best hyperplane that separates two classes in the input space. Basically, SVM working with the principle of linear classifier, then developed to work on nonlinear case by using kernel concept in high-dimensional workspace.

Neural Network (NN) and SVM with their variance are prominent technique for secondary structure prediction. Knowledge based and NN as one approach for secondary structure prediction [8], variance of NN, SOM and SOGR also proposed in [9] and JNet algorithm also proposed [10]. Mix modal of SVM proposed in [11] with PSI-BLAST for input feature. Even, improved SVM and Neural Network was introduced in [12] for protein secondary structure pereddiction.

According to the problems described in the background, problem formulation of this research is how to predict the secondary structure of proteins using SVM method and analyze the effect of using PSSM profile and physicochemical features on SVM-based model accuracy. Feature extraction on protein structure prediction is one of the main interest research focus. This study purposed to combine PSSM and physicochemical as feature extraction with 13 sliding window refers to [5] on protein secondary structure prediction. This research is expected to produce a good model to predict the secondary structure of proteins before predict the 3D structure of proteins in order to identify further structure and function of the protein. The assignemnt of secondary structure label derived from Dictionary of Secondary Structure of Proteins in short DSSP with enzim commision data from Protein Data Bank (PDB).

2. Methodology

This chapter is divided into eight sections includes literature review, data collection secondary structure of proteins, PSSM feature extraction, physicochemical feature extraction, SVM models, testing and evaluation. Flowchart of the research methods can be illustrated in Figure 1.

2.1. Literature review

Things to do at this stage is to study the subjects related to protein secondary structure prediction and classification method of SVM. From the study, modeling with SVM classification can be implemented and used to predict the secondary structure of proteins by the addition of physicochemical features to improve the accuracy of the classification.

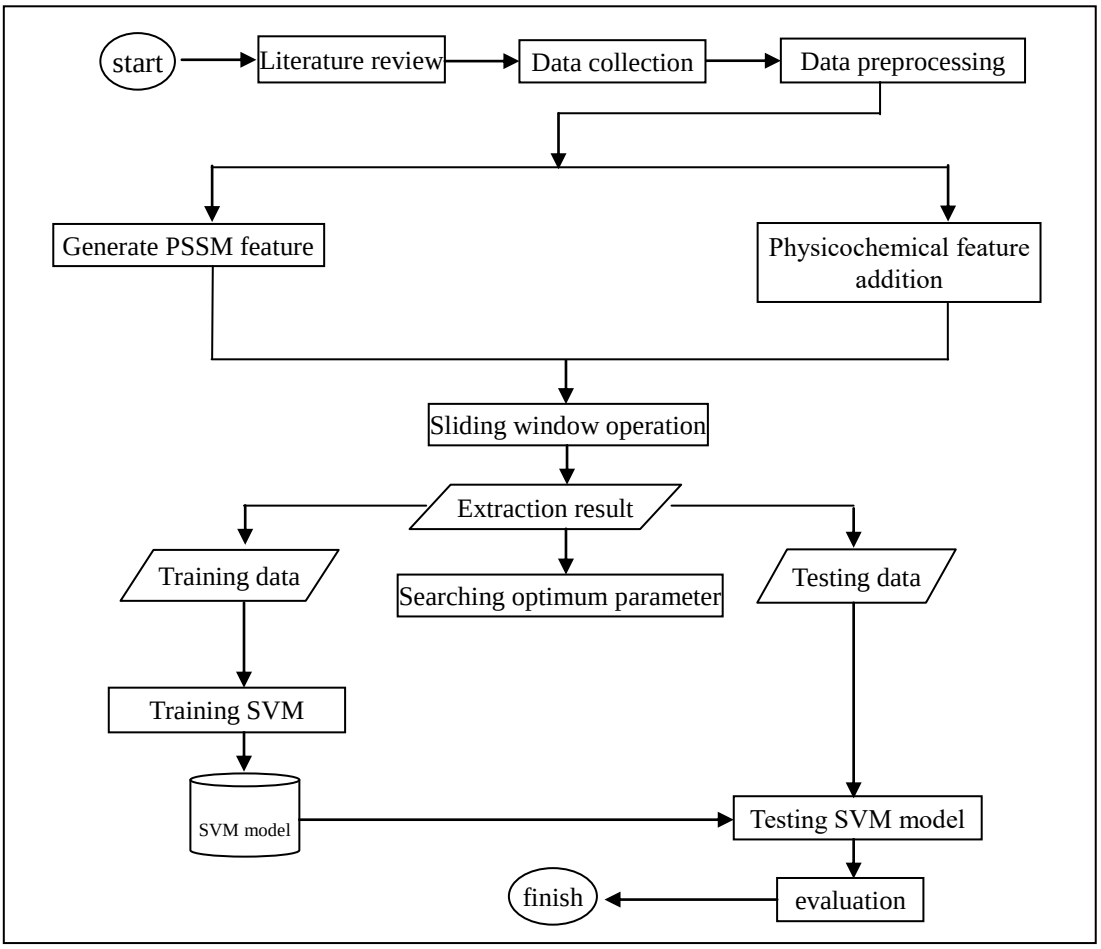


Figure 1. Flowchart of research methods

2.2. Data Collection and preprocessing

The data retrieval process begins with collecting a secondary protein sequence data through the website of PDB in http://www.rcsb.org/ to get PDB id. The id indicating different structure We choose enzim sequence based on enzim commission information. Furthermore, the process generates a file DSSP (Dictionary Protein Secondary Structure). DSSP is database of protein secondary structure assignment. Tools used to generate file DSSP is XSSP in http://www.cmbi.ru.nl/xssp website. In this study, the existing class segment in the DSSP is reduced to three classes, namely alpha-helix (H), beta-sheet (B) and the coil (C) [13]. Detail of secondary structure assignment for data grouping can be shown in Table 1.

Table 1. Grouping of protein secondary data

DSSP 8-Class	3-Class
α-Helix (H) ,3/10 helix (G)	Helix (H)
β-Helix (E), β-Bridge (B)	Strand(E)
Π-Helix (I), Turn (T), Bend (S), Coil (C)	Coil (C)

On preprocessing task, we just use two column in DSSP file, amino acid (aa) and structure. Amino acid describe the residue of sequence protein and structure inform us as secondary structure assignment.

2.3. Feature Extraction of PSSM

In conducting protein secondary structure prediction, the features required to help define the secondary structure. The data from 20 amino acids used as a determinant of input pattern classification used. The data file that is processed in DSSP will be formed into PSSM profile. Forming PSSM profile is done by entering DSSP amino acids data files with tools on the web http://www.cbs.dtu.dk/biotools/Seq2Logo. To extract the features of PSSM profile, Sliding window used for pattern recognition process. The use of sliding window will acquire more information from protein residues with the optimal number of windows 13[5].

2.4. Physicochemical feature addition

Physicochemical features of each amino acid were added to PSSM as input for the classification process. Physicochemical features will work as identifier in this study that includes conformation parameters, net charge, hydrophobic side chain mass[5]. The value of conformation parameters indicates the chances each amino acid residue of the secondary structure of H, E, and C. Net charge feature obtained by the index table of amino acids index. And hydrophobic side chain mass is used as a feature in the protein secondary structure prediction because it is associated with the folding process.

2.5. Sliding window operation

Combination between PSSM and physicochemical are main focus on this research. The combination of two features then operated via sliding window operation to generate number of feature as input of Support Vector Machine. Output of this process as data model known Model I, model II and Model III that will be explained in next section.

2.6. Searching Optimal SVM parameter and training

SVM as classifier fore predicting secondary structure must be tuned to obtain optimal parameter. Empirically, we use kernel on SVM from linear, radial and polynomial kernal with 2, 3 and 4 degree. Before training the model, one optimal kernel must be obtained first to ensure that error in training process can be minimized. The output of this process is optimal parameter for the kernel for training.

2.7. Classification method Support Vector Machine (SVM)

In this study, dataset was built to three models in order to compare the accuracy of the results between the three models. The first model (Model I) was built with data that have been done PSSM feature extraction without the addition of physicochemical features do the sliding windows. The second model (Model II) was built with data that have been done and done PSSM feature extraction process of sliding windows. Then added six physicochemical features. The third model (Model III) was built with feature extraction data has been done PSSM then added six physicochemical features. Then do the process of sliding windows.

The difference between the three models is the number of features in each model into a dimension of a vector space SVM. Model I without addition of physical-chemical features while Model II and III are additional features with different physical chemistry physical chemistry feature placement position in data processing. The number of feature describe in greater detail below. Model I has 260 features obtained from 20 (amino acid) x 13 (sliding window). Model II with 266 features obtained from 260 + 6 (physicochemical). Last but not least, Model III position sliding window operation after physicochemical were added. Therefore, we obtain (20+6) x 13 = 338 features.

Before the classification process, the distribution of the data in the dataset is divided into two parts, the training data part as much as 75% of the dataset and test data part as much as 25% of the dataset. Distribution of the proportion of secondary structure classes can be seen in Figure 2.

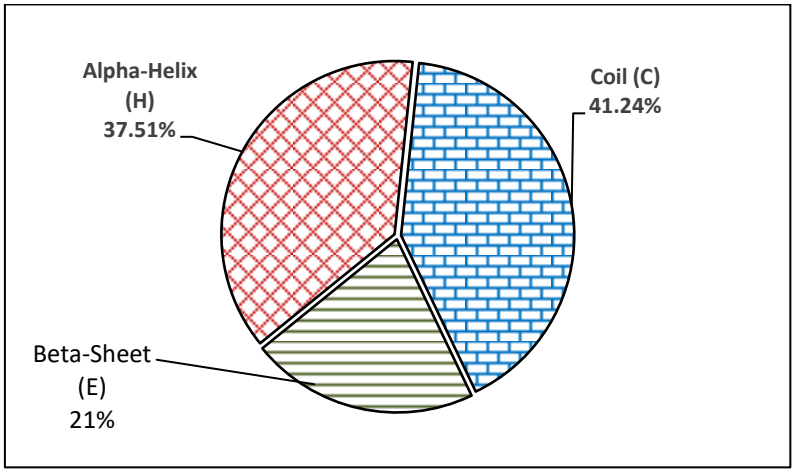


Figure 2. Data distribution on dataset

2.8. Testing and Evaluation

Testing is performed to test the model generated by Support Vector Machine. Q3 accuracy assessment will use a score that contains information of data classes and predicted results. The value equation of accuracy using the equation 1 where i is the class type of protein secondary structure, alphahelix (H), beta-sheet (B), and coil (C). Equation 2 is the Q3 score equation which is the average accuracy value of the entire class. In Q3 similarity score, N_H is accuracy for alphahelix (H), N_B is accuracy to beta-sheet (B), N_C is accuracy to coil (C), and N_{total} is the number of classes on the test data.

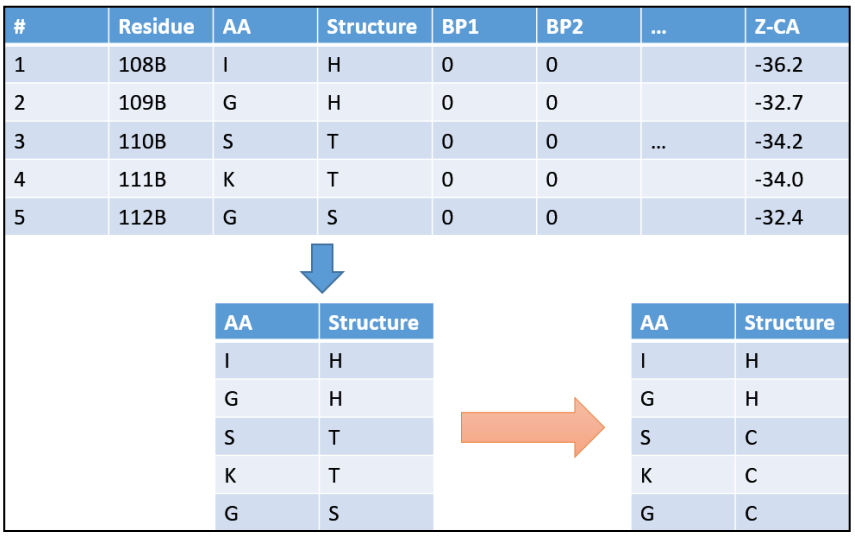
$$Accuracy = \frac{\sum_{i=1}^{n} data \quad of \quad true \quad i \quad class}{\sum_{i=1}^{n} number \quad of \quad class \quad i} \times 100\% \tag{1}$$

$$Q3 \quad Score = \frac{N_H + N_B + N_C}{N_{total}} \tag{2}$$

3. Result and Discussion

3.1. Data Collection and preprocessing

At this stage, the protein sequence data retrieved from the file extension .dssp. Inside that file, there are 19 fields of data. This study only takes two columns, AA column (Amino Acid) and Structure. Secondary structures grade that exists on Structure column will be reduced from 8 class into 3 classes. Illustration of retrieval data column can be seen in Figure 3 and the reduction classes are conducted based on Table 1.



#	Residue	AA	Structure	BP1	BP2	...	Z-CA
1	108B	I	H	0	0		-36.2
2	109B	G	H	0	0		-32.7
3	110B	S	T	0	0	...	-34.2
4	111B	K	T	0	0		-34.0
5	112B	G	S	0	0		-32.4

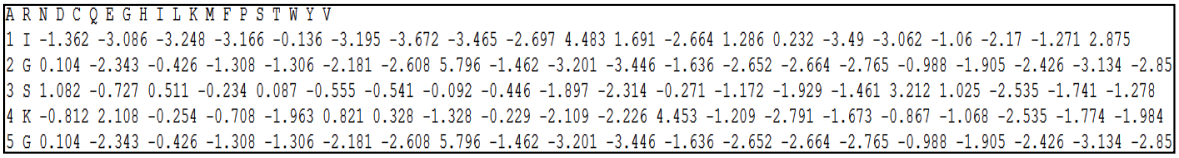
AA	Structure
I	H
G	H
S	T
K	T
G	S

AA	Structure
I	H
G	H
S	C
K	C
G	C

Figure 3. Data collection illustration

3.2. Feature Extraction of PSSM

After getting the protein sequence data collection, the sequence included as input in http://www.cbs.dtu.dk/biotools/Seq2Logo/ to do the generation characteristic of amino acids process. This process is a search for value changes in amino acid residues of the other amino acid residues. Interesting patterns contained in the value of the change is an amino acid value becomes highest when met with the amino acid itself. Once the protein sequence is inserted, then the resulting PSSM matrix measuring the number of inputs × 20 (amino acids). PSSM matrix form can be seen in Figure 4.



```
A R N D C Q E G H I L K M F P S T W Y V
1 I -1.362 -3.086 -3.248 -3.166 -0.136 -3.195 -3.672 -3.465 -2.697 4.483 1.691 -2.664 1.286 0.232 -3.49 -3.062 -1.06 -2.17 -1.271 2.875
2 G 0.104 -2.343 -0.426 -1.308 -1.306 -2.181 -2.608 5.796 -1.462 -3.201 -3.446 -1.636 -2.652 -2.664 -2.765 -0.988 -1.905 -2.426 -3.134 -2.85
3 S 1.082 -0.727 0.511 -0.234 0.087 -0.555 -0.541 -0.092 -0.446 -1.897 -2.314 -0.271 -1.172 -1.929 -1.461 3.212 1.025 -2.535 -1.741 -1.278
4 K -0.812 2.108 -0.254 -0.708 -1.963 0.821 0.328 -1.328 -0.229 -2.109 -2.226 4.453 -1.209 -2.791 -1.673 -0.867 -1.068 -2.535 -1.774 -1.984
5 G 0.104 -2.343 -0.426 -1.308 -1.306 -2.181 -2.608 5.796 -1.462 -3.201 -3.446 -1.636 -2.652 -2.664 -2.765 -0.988 -1.905 -2.426 -3.134 -2.85
```

Figure 4. Matrix PSSM Illustration

3.3. Physicochemical feature addition

At this stage of the calculation of the parameter value parameter conformation, net charge, hydrophobic, and side chain mass. Conformation parameter is the composition of the amino acid sequence of a particular class of secondary structures that exist on the dataset. Net charges are physical-chemical parameters which only five amino acid arginine (R), aspartic acid (D), glutamine acid (E), histidine (H), and lysine (K). This relates to the formations that will be formed by the electric charge on the five amino acids. Values electrical charge provided are +1 and -1 selainnya is worth 0. Hydrophobic properties that affect the stability of the secondary structure of proteins. The more positive the value of these properties, the hydrophobic nature of the stronger. Side chain mass is a value that is calculated based on the mass of the chemical bond, an amino acid that affects protein structure folding process. In this study, the parameter value parameter conformation, net charge, hydrophobic, and side chain mass.

3.4. Sliding window operation

This method is a process of feature extraction from PSSM matrix with secondary structure proteins class column written derived from tables of data retrieval. This process uses sliding windows which amounted to 13 refer to [5]. Windows that exist in the sliding window will take the data or protein

sequence as many as 13 rows with row-7 as a point of interest and secondary structure are assigned as class label to this point of interest. The result of this process becomes the input data or the dataset used for the purposes of training and testing process. There are three data model are generated from this process. Data with 260, 266 and 338 respectively. Number of features are described on subsection 2.7

3.5. SVM Optimum Parameters

Empirically, combination of kernel and parameter are examined at this stage. We spent eight hours from our research for getting optimal parameter. C and γ are parameter will be tuned. The parameter C and γ used in the classification of SVM with radial and polynomial kernel. The data used were 9000 data from the dataset and tuning process is carried out with a grid search method. The function used is tune.svm() from e1071 R package. The parameters C and γ are determined by a combination of 5 × 4 matrix. The combination matrix formed from $[2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}]$ and $[2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}]$. Optimum parameters obtained by the $C = 2^{1}$ and $\gamma = 2^{-6}$.

3.6. Support Vector Machine Classification

At this stage performed the optimum parameter search, SVM model training, and testing of SVM models. Searches conducted to find optimum parameters for training with radial and polynomial kernel. Package used in this study is a package e1071 with svm(), predict(), and tune.svm() function. The details of the process are as follows:

3.7. Training Model SVM

At this stage, SVM model training conducted with training data which has been divided from the dataset with the optimum parameters of the search parameters. Training is done on every model with linear kernel, kernel radial and polynomial kernel of degree 2, 3, and 4. The resulting model as many as 15 pieces. Training SVM models using svm() function from e1071 R package.

3.8. Testing Model SVM

In this stage, testing conducted to models that have been trained with a dataset and a different kernel. Tests conducted by the predict() from e1071 R package. The test results in the chart of Q3 Score values can be illustrated in Figure 5.

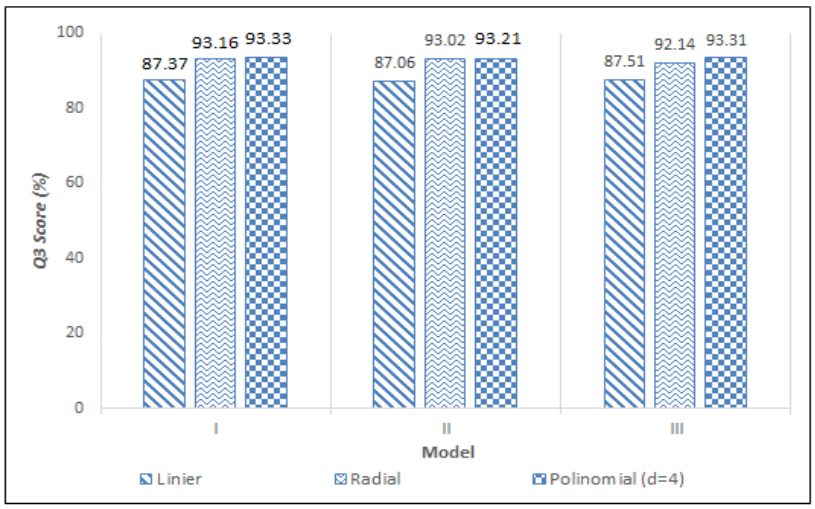


Figure 5. Q3 Score diagram

The figure 5 indicate that the highest score value Q3 owned by model I with polynomial kernel of degree 4. Model I have 260 PSSM profile features without physical chemical features. Values obtained is 93.33%. Lowest score Q3 Value owned by Model II with linear kernel. Model II has 266

features (260 features six feature PSSM profile and physicochemical). Values obtained for this model is 87.06%. The highest Q3 value score obtained by the model dataset is processed by the polynomial kernel of degree 4. From the figure 5 indicate that the addition of physicochemical after and before sliding window not have significance accuracy on prediction. This is because the values of physicochemical are static and relatively too little compared to PSSM features so does not give more information as a feature.

4. Conclusion

Secondary structure can be performed using SVM with PSSM and physicochemical as feature. Furthermore, evaluation of models produced by calculating the value of Q3 Score. Additional features of physical chemistry can be implemented, however does not have a significant influence on the value of accuracy. The best results obtained by the model derived from the dataset Model I by using kernel polynomial of degree four.

References

[1] Polanski A and Kimmel M 2007 *Bioinformatics* (New York: Springer)
[2] Bray A, Johnson H, Raff L and Walter R 2009 *Essential cell biology* Ed 3th (London: Garland Science)
[3] Zhuravleva A and Korzhnev D M 2017 *Prog. Nucl. Magn. Reson. Spectrosc.* **100** 52–77
[4] Martin J, Letellier G, Marin A and De Brevern A G 2005 *BMC Structural Biology* **17** 1–17
[5] Huang Y and Chen S 2013 *Sci. World J.* **2013** 1–8
[6] Wang J, Wang C, Cao J, Liu X, Yao Y and Dai Q 2015 *Gene* **554** 241–8
[7] Vapnik V N and Cortes C 1995 *Machine Learning* **20** 273–97
[8] Patel M S and Mazumdar H S 2014 *J. Theor. Biol.* **361** 182–9
[9] Atar E, Ersoy O, and Ozyilmaz L 2005 *2005 ICSC Congress* (Istanbul: Computational Intelligence Methods and Applications)art. no. 1662358
[10] Drozdetskiy A, Cole C, Procter J and Barton G J 2015 *Nucleic Acids Res* **43** 389–94
[11] Yang B, Wu Q, Ying Z and Sui H 2011 *Knowledge-Based Systems* **24** 304–13
[12] Johal A K 2014 *International Journal Of Engineering And Computer Science* **3** 3593–7
[13] Aydin Z, Altunbasak Y and Borodovsky M 2006 *BMC Bioinformatics* **7** 178