Protein Family Identification using Markov Chain as Feature Extraction and Probabilistic Neural Net....

Conference Paper · September 2015
DOI: 10.13140/RG.2.1.4620.2968

3 authors, including:

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

2015 International Conference on Information, Communication Technology and System

Protein Family Identification using Markov Chain as Feature Extraction and Probabilistic Neural Network (PNN) as Classifier

Toto Haryanto
Department of Computer Science,
Bogor Agricultural University
Jalan Meranti Wing 20 Level 5 IPB Dramaga, Indonesia
email : totoharyanto@apps.ipb.ac.id

Rizky Kurniawan, Sony Muhammad
Department of Computer Science,
Bogor Agricultural University
Jalan Meranti Wing 20 Level 5 IPB Dramaga, Indonesia

***Abstract*—Proteins are organic molecules formed from 20 amino acid combination with various function for living things such as transportation system, catalyst of chemical reaction for metabolism and food reserves. This research aims to classify proteins family based on sequences of amino acid as primary structure. There are 300 amino acid fragment obtained from ftp://ftp.sanger.ac.uk/pub/databases/Pfam. Subset of proteins family database with three classes, including 1-cysPrx_C, 4HBT and ABC_Tran were obtained. In this research, we use first and second order of Markov chain for extracting features. Moreover, we use Probabilistic Neural Network (PNN) as classifier compared to joint probability technique with markov assumptions. We evaluate the results by comparing the sensitivity and the specificity of both classification techniques. The evaluation results show that overall PNN has slightly better performance than joint probability technique for classifying protein family**.

Keywords- joint probability, markov chain, probabilistic neural network, protein family

I. INTRODUCTION

Protein is an essential element for living thing. Structurally, protein made up of primary structure, secondary structure and tertiary structure. Sequence of amino acid as primary structure can be categorized into several types [1]. Classification protein family has important role such as to improve protein identification when it is difficult to be characterized, to help maintaining database of protein family, to retrieve biological information with enormous data effectively and to represent gene expression from protein family for analyzing phylogenetic [1]. Protein has several databases with specific purposes, such as hierarchical family database (PIR-PSD and ProtoMap), protein domain database (PFam and ProDom), motif sequence database (PROSITE and PRINT), structural database (SCOP and CATH) and integrated family database (iProClass and InterPro). Classification of protein is also significant to information retrieval such as structure, activity and annotation, and also its metabolism system [2].

Classification of Protein family and protein annotation were conducted using rule base method [2]. Previous studies are based on Sparse Markov Transducer (SMT) [3]. The study is focuses on two techniques of SMT. SMT prediction model and SMT classifier model, was implemented. Classification using SMT prediction model obtained the highest accuracy up to 100% for FGF, MCP signal and MHC_I classes.

However, the lowest accuracy was obtained from UPAR_LT6 class. Moreover, for SMT classifier model, the highest accuracy was obtained from TIM, S12, MCP signal MHC_I, FGF, Cys_knot, ATP-synt-ab and 7tm_3 classes with the value of 100%.

Markov chain is one of the model probabilistic which is often used in bioinformatics research area. Markov chain can be conducted to analyze DNA sequences, RNA, and sequence of amino acid [4]. Amino acid sequences formed in two process, DNA transcription and RNA translation. Amino acid can be yielded by RNA codon or triplet. Sequences of amino acid will represent a protein. In Markov chain, new sequences can be identified using joint probability with Markov assumption [5].

Alternatively, Probabilistic Neural Network (PNN) as classifier also has an important role in bioinformatics research. In proteomic area, PNN is used as an approach for classifying protein superfamily with 497 features, not only from amino acid information [6]. PNN also is used in genomic research area for DNA analysis sequences pattern [7].

This research aimed to classify protein family which is particularly obtained from sequences of amino acid as primary structure. We use joint probability with Markov assumption and PNN as classification techniques. The evaluation was conducted by comparing the performance of both classification techniques.

II. METHODOLOGY

This research divided into two big schemas, feature extraction and classification. Terminology of protein family also will be explained in this paper.

A. Protein Family

Proteins are classified into the same class, principally they have evolutionally relationships. Proteins with similar class have same ancestor and generally are similar in sequences,

three dimensional structure and functionality. Sometimes it is difficult to evaluate the similarity significance among protein families. However, proteins with different ancestor have different sequences of amino acid. This is the reason why many techniques were developed to classify protein family based on sequence of amino acid. Recently, there are more than 60.000 classes of protein family [8]. Database of protein family can be accessed at Uniform Resource Locator (URL) of Sanger Genome Institute research center at ftp://ftp.sanger.ac.uk/pub/databases/Pfam.

B. Markov chain method

Markov chain method was introduced by Russian Mathematician, A.A Markov in early of 20 century. Using Markov process, the stochastic phenomena in the real world can be modeled. Problem that underlie in Markov chain is how to determine state transition appropriately so that the stochastic process fulfills the Markov property. These mean that the information about state is available enough for predicting next stochastic behavior [5].

A Markov chain can be called as a discrete time Markov chain if a space of Markov process is a finite set or countable. If the value of state in a given period depends only on one previous state, this is called as the first order Markov chain. Equation (1) describes the first order Markov chain.

$$P\{X_{n+1} = j | X_n = i\} \quad (1)$$

If the value of state in a given state period depends on m previous state, then this is called as the m-order Markov chain refers to (2)

$$P\{X_{n+1} = j | X_{(n+1)-m} = i_1, X_{(n+1)-m+1} = i_2, \dots, X_n = i_n\} \quad (2)$$

Set of states in this research was represented as sequence of amino acid. Probability X_{n+1} at state j given Xn at the state i was called as the first order probability refers to (3)

$$P_{ij}^{n.n+1} = P\{X_{n+1} = j | X_n = i\} \quad (3)$$

Probability of transition was state represented as transition matrix known as transition matrix probability P. Pij defined as $P\{X_{n+1} = j | X_n = i\}$ refers to (4)

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{14} \\ p_{21} & p_{23} & \cdots & p_{11} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \quad (4)$$

Every class in this research has matrix as model. For identifying new sequence of amino acid, joint probability with Markov assumption was conducted refers to (5)

$$P(X_1, \dots X_n) = \prod_{i=1}^{n} P(X_i | X_{i-1}) \quad (5)$$

Equation (5) resulted probability of new sequence based on given transition matrix. Maximum probability at transition matrix became justification for new sequence which could be classified to this class.

C. Probabilistic Neural Network (PNN)

Probabilistic Neural Network (PNN) is one of classification technique with four layer structures; such as input layer, pattern layer, summation layer and decision layer. Differs from Neural Network (NN), PNN does not have a backward process. Therefore, PNN is more efficient, for instance the performance of PNN in [9] could obtained 200.000 times faster than back-propagation.

Input layer is a first layer in PNN. In this research, input layer is a feature vector input with length of 400. This number was obtained from feature extraction using Markov chain from the value of transition matrix between 20 amino acid. Therefore there are 20x20 value of probability will be a input vector in PNN.

Pattern layer as a second layer represented pattern of classes. In this layer, training data would be grouped into three classes of protein family; 1-cysPrx_C, 4HBT and ABC_Tran. There were 75 fragments as training data for each class. Principally, in pattern layer, there is a process of calculating the similarity between new sequences and the sequences in the training data using gaussian kernel withσ as smoothing parameter refers to (6)

$$f(x) = exp\left(-\frac{(x-x_{Aj})^T(x-x_{Aj})}{2\sigma^2}\right) \quad (6)$$

Summation layer is a third layer in PNN. In this layer, we calculated summation of probability for testing data and third class in the pattern layer. Equation (7) is summation layer [9]

$$p(\omega_A)p(x/\omega_A) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d N}\sum_{i=1}^{N_A} exp\left(-\frac{(x-x_{A_i})^T(x-x_{A_i})}{2\sigma^2}\right) \quad (7)$$

where

p(ωA)	= probability of class A
p(x|ωA)	= conditional probability x in class A
xAi	= vector of ith training data in class A
d	= dimension of input vector
N	= number of training pattern all class
NA	= number of training pattern in class A
σ	= smoothing parameter

Decision pattern as fourth represent justification of new amino acid can be classified into one of three classes in pattern layer. This layer, we calculated the maximum value produced by summation layer. Structure of PNN can be seen in Figure 1.

Figure 1 show 400 inputs from feature extractions, three patter layers representing 3 classes of protein family, and three summation layers for each class Y as decision layer to justify classification results.

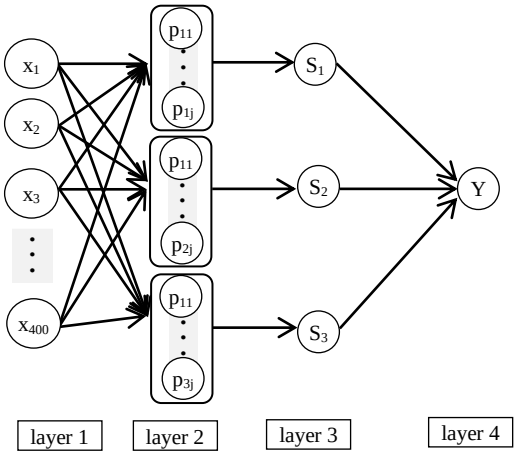

Fig. 1. Structure of PNN with 400 input vector input and 3 pattern layer

III. EVALUATION METHOD

Performance of classification method was measured using sensitivity and specificity for each protein family based on confusion matrix (see Figure 2).

	Predicted as *sequences* of protein family	Predicted not as sequences protein family
sequences of protein family	tp	fn
not sequences protein family	fp	tn

Fig. 2. Confusion matrix for each protein family

Where

tp : *true positive* (number of sequences protein family in actual that predicted as protein family)
tn : *true negative* (number of not sequences protein family in actual that predicted not as protein family)
fp : *false positive* (number of not sequences protein family in actual that predicted as protein family)
fn : *false negative*(number of sequences protein family in actual that predicted not as protein family)

Sensitivity measures positive proportion of new protein family which can be classified correctly refers to (8) while specificity measures negative proportion of new protein family which can be classified correctly refers to (9)

$$sensitivity = \frac{tp}{tp+fn} \quad (8)$$

$$specificity = \frac{tn}{tn+fp} \quad (9)$$

IV. RESULT AND DISCUSSION

A. Data Acquisition

A total of 300 protein family data set were obtained from protein family database at URL ftp://ftp.sanger.ac.uk/pub/databases/Pfam. There were three classes of protein family, 1-cysPrx_C, 4HBT and ABC_Tran. Table I show the detail description of the data

TABLE I
DETAIL DESCRIPTION OF PROTEIN FAMILY DATA SET

No	Name	Functional
1	1-cysPrx_C	Member of peroxiredoxin superfamily. To Protect cells from membrane oxidation.
2	4HBT	4HBT serves to catalyze biosynthetic process
3	ABC_Tran	Also called ABT Trasporter. Uses the ATP hydrolysis for translocation process some molecules where passing biological membranes

Studies in bioinformatics utilize sequence data obtained from FASTA format as well in protein family data. Data with format FASTA were separated by "<" followed by identifier, and then sequence of amino acid in the last. This is example of FASTA format data for 1-cysPrx_C class.

```
>B3ARY6_ECO57/154-186PF10417.3;1-
cysPrx_C;
AAQYVASHPGEVCPAKWKEGEATLAPSLDLVGK
```

B. Feature extraction

Markov chain was conducted to extract information from fragment of amino acid. Both of extraction types, the first order and the second order Markov chain, produced transition matrixes with dimensions of 20x20. Therefore, each matrix was transformed into vector with a length of 400 as PNN input.

The values of this matrix are filled by probability of occurrence of amino acid refers to (10).

$$P = \begin{bmatrix} P(G|G) & \cdots & P(G|W) \\ \vdots & \ddots & \vdots \\ P(W|G) & \cdots & P(W|W) \end{bmatrix} \quad (10)$$

P(G|G) means the occurrence probability of the Glisina amino acid if the previous amino acid is Glisina. P(G|W) means the occurrence probability of the Glisina amino acid if the previous amino acid is Triptofan and so on until 400 values of the probabilities were filled.

During classification process of protein family, 4-fold cross validation is conducted. Table II is scenario for conducting 4-cross validation.

TABLE II
SCENARIO 4-FOLD CROSS VALIDATION

subset	*fold*	*training data*	*testing data*
1	*fold* 1	*s2 s3 s4*	*s1*
2	*fold* 2	*s1 s3 s4*	*s2*
3	*fold* 3	*s1 s2 s4*	*s3*
4	*fold* 4	*s1 s2 s3*	*s4*

Therefore, there were 75 training data and 25 testing data for each class. Totally, for each subset or each fold, there were 225 training data and 75 testing data.

C. Comparison of two classification method

Average of sensitivity and specificity for all fold was calculated to evaluate both of classification techniques, PNN and join probability. Both values of sensitivity and specificity were divided into two methods of Markov chain order, first order and second order.

Figure 3 showed the comparison of sensitivity between PNN and joint probability with Markov assumption. The figure showed that sensitivity of PNN and joint probability in class 1-cysPrx and 4HBT were comparable. Nevertheless, in class ABC_tran, the sensitivity of PNN is 1.0 higher than the sensitivity ofthe joint probability (0.94). If the sensitivity is 1.0, it means that all of ABC_tran in actual data were classified correctly as ABC_tran.

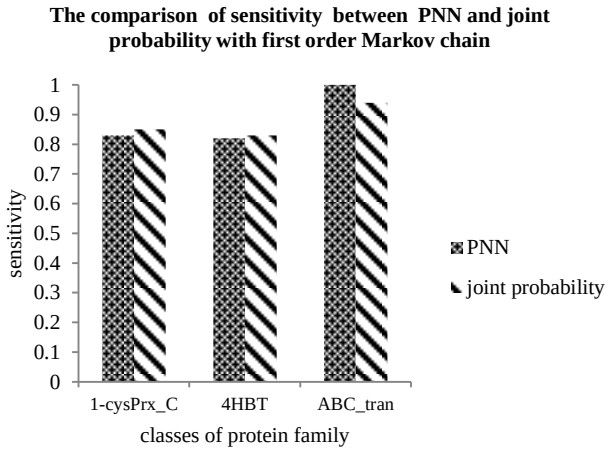

Fig. 3. Sensitivity for three classes protein family using two method of classification PNN and joint probability with first order Markov chain feature extraction

We also compared the sensitivity of both classifier PNN and joint probability using the second order Markov assumption. Figure 4 showed that all of ABC_tran were correctly classified pointed by the sensitivity value of 1.0. Moreover, for two other classes, the values of sensitivity are comparable. From Figure 3 and Figure 4, we could conclude that PNN was able to capture pattern for ABC_tran.

Using both feature extraction methods, the first order and the second order, we could always obtained maximum value of sensitivity. Showing that the sequences pattern of ABC_tran were different with two other classes.

Moreover, Figure 5, show that PNN could obtain the specificity value of 1.0 for protein family 1-cysPrx and 4HBT. The results show that all of data which were not belonging to the 1-cysPrx or 4HBT were classified into classes, except 1-cysPrx or 4HBT. In this case, value of false positive (fp) is 0 and true negative (tn) is 50 for class 1-cysPrx and 4HBT respectively. Whereas the specificity value of ABC_tran is similar.

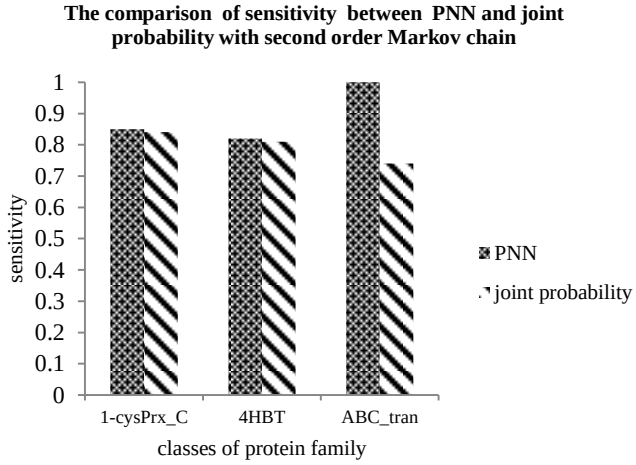

Fig. 4. Sensitivity for three classes protein family using two method of classification PNN and joint probability with second order Markov chain feature extraction

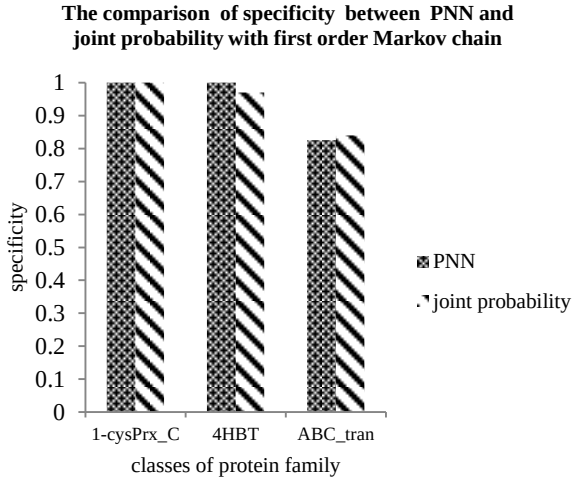

Fig. 5. Specificity for three classes protein family using two method of classification PNN and joint probability with first order Markov chain feature extraction

The specificity values were also calculated for second order Markov chain as well as those of the sensitivity. Figure 6 show that the specificity value of PNN are 1.0, 1.0, and 0.83 for 1-cycPrx, 4HBT and ABC_tran, respectively. Whereas the specificity values were 1.0, 0.85, and 0.84. Compared to Figure 6, the specificity of using second order was not better than that of using first order Markov chain in feature extraction process.

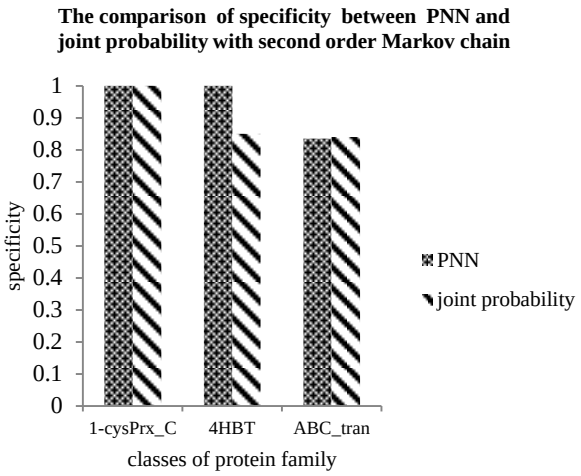

Fig. 6. Specificity for three classes protein family using two method of classification PNN and joint probability with second order Markov chain feature extraction

V. CONCLUSION AND FUTURE WORK

Sequences of amino acid as primary structures have information which can be extracted by Markov chain. Classification of protein family can be conducted using PNN or joint probability. Generally, the classification performance of PNN was slightly better than that of using joint probability. First order markov chain delivers more information in feature extraction. Classification of protein family can also utilize chemical, physical information and combined with statistical information for feature extraction method in the future.

REFERENCES

[1] Polanski A, Kimmel M. *Bioinformatics*. New York: Springer-Verlag. 2007.

[2] Wu CH, Huang H, Yeh L Lai-Su and Barker W.C. "Protein family classification and functional annotation". *Computational Biology and Chemistry*, vol. 27, Nov 2003, pp. 37-47.

[3] Eskin E, Grundy W.N, Singer Y. "Protein Family Classification using Sparse Markov Transducers (Published Conference Proceedings style)," in ISMB-00, 2000, pp. 134-145.

[4] Ryabko B, Usotskaya N. "Application of information-theoretic tests for the analysis of DNA sequences based on Markov chain models". *Computational Statistics and Data Analysis*, vol. 53, July 2009, pp. 1861-1872.

[5] Ching W.K, Ng M.K. *Markov Chain Models, Algorithm and Applications*. Springer Science + Business Media, New York. USA. 2006.

[6] RaoNageswara PV, T Uma Devi, DsvkgKaladhar, GR Sridhar, AllamAppaRao. 2009. "A Probabilistic Neural Network Approach for Protein Superfamily Classification". Journal of Theoretical and Applied Information Technology, vol 16 No.1 pp 101-105.

[7] Wu X, Lu F, Wang B and Cheng J. "Analysis of DNA Sequence Pattern Using Probabilistic Neural Network Model". *Journal of Research and Practice in Information Technology*, vol. 37, No. 4, Nov 2005, pp. 353-363.

[8] F Kunin V, Cases I, Enright A.J, De Lorenzo V, Ouzounis C.A. 2003. Myriads of protein families, and still counting. Cambridge (UK): The European Bioinformatics Institute.

[9] Spetch DF. 1990."Probabilistic neural network". *Neural Network*. 3(1)109-118.