International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:4, No:5, 2010

Sequence-based Prediction of Gamma-turn Types using a Physicochemical Property-based Decision Tree Method

Chyn Liaw, Chun-Wei Tung, Shinn-Jang Ho, Shinn-Ying Ho

***Abstract*—**The γ-turns play important roles in protein folding and molecular recognition. The prediction and analysis of γ-turn types are important for both protein structure predictions and better understanding the characteristics of different γ-turn types. This study proposed a physicochemical property-based decision tree (PPDT) method to interpretably predict γ-turn types. In addition to the good prediction performance of PPDT, three simple and human interpretable IF-THEN rules are extracted from the decision tree constructed by PPDT. The identified informative physicochemical properties and concise rules provide a simple way for discriminating and understanding γ-turn types.

***Keywords*—**Classification and regression tree (CART), γ-turn, Physicochemical properties, Protein secondary structure.

I. Introduction

A γ-turn is defined as three consecutive residues with a hydrogen bond between the CO and NH groups of residues i and $i+2$ that largely influences the protein 3D structure. There are two types of γ-turns: the classic and inverse γ-turns. The classification of the classic and inverse γ-turns is determined by the values of dihedral angles. Their main-chain atoms are related by mirror symmetry.

The γ-turns often mediate the reversal of polypeptide chains and are important in the protein folding [1] and molecular recognition [2]. The major type of γ-turns is the inverse γ-turn that is less associated with the reversal of polypeptide chains. Compared to inverse γ-turns, classic γ-turns are rarely found and are frequently located in the end of loops of beta-hairpin [3]. The classic γ-turn gives rise to a 180 degree chain-reversal in proteins and is important for globular protein structure formation. Based on the hydrogen bonding patterns, the classic γ-turns can be further classified into four subclasses [1]. Inverse γ-turns comprise a large proportion of weak hydrogen bonds.

Previous works mainly focus on the prediction of γ-turns. A neural network method is firstly developed for γ-turn prediction using information of multiple sequence alignments and predicted secondary structures [3]. The subsequently improved predictions of γ-turns include a Markov Chains theory based method [4] and a support vector machine based method [5].

Recently, the problem of predicting γ-turn types is investigated by using a support vector machine based method [6] and a two-stage hybrid neural discriminate model [7]. Both of them are based on features of binary encoding of amino acids. Although high accuracy is obtained by their methods, however, as well recognized, the support vector machine and neural network based method are the so called black-box methods. It means that users are not able to know the decision rules. The binary encoding methods are also not human interpretable. Furthermore, the used secondary structure information calculated from the DSSP (definition of secondary structure of proteins) program [8] requires information from three-dimensional structures [7].

Due to the highly unbalanced occurrences of inverse and classic γ-turns, it is much difficult to predict γ-turn types. To interpretably predict γ-turn types, a physicochemical property-based decision tree (PPDT) method based on only sequence information is proposed to identify and analyze informative physicochemical properties for predicting γ-turn types. The proposed PPDT based on an efficient classification and regression tree (CART, [9]) is very simple and is able to provide human interpretable rules with good performance of AUC=0.674 that is comparable with previous studies.

A total of two informative physicochemical properties are identified by PPDT. The usability of the identified informative physicochemical properties for the prediction of γ-turn types is also demonstrated by using a simple Naïve Bayes classifier. A high performance with AUC=0.736 is obtained by applying a Naïve Bayes classifier with the identified two physicochemical properties to predict γ-turn types. Finally, three simple and human interpretable IF-THEN rules are extracted from the constructed decision tree. The rules provide a simple way to predict γ-turn types and give insights into the physicochemical effects on the differentiation of γ-turn types.

Chyn Liaw is with the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, 300 Taiwan (e-mail: chynliaw@gmail.com).

Chun-Wei Tung is with the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, 300 Taiwan (e-mail: cwtung@livemail.tw).

Shinn-Jang Ho is with the Department of Automation Engineering, National Formosa University, Yunlin 632, Taiwan (e-mail: sjho@sunws.nhit.edu.tw).

Shinn-Ying Ho is with the Institute of Bioinformatics and Systems Biology, Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 300 Taiwan. (corresponding author; phone: 886-3-5712121-56905; e-mail: syho@mail.nctu.edu.tw).

II. Methods

A. Dataset

The used dataset for following analysis and prediction are obtained from previous studies [6], [7]. The IDs of 490 non-homologous proteins with less than 25% sequence identity are extracted and protein structure files are downloaded from PDB database. All protein structures were determined by X-ray crystallography with resolution better than 1.5Å. The program PROMOTIF is utilized to annotate γ-turn types. By removing four duplicated PDB IDs provided by the previous study [6] and one γ-turn with a modified amino acid, a slightly smaller dataset, which contains 1241 γ-turns consisting of 1145 inverse γ-turns and 96 classic γ-turns is used for the following analyses.

B. Physicochemical properties

Physicochemical properties play important roles in biomolecular recognition and protein folding. Being the most intuitive feature for biochemical reactions, a huge number of published bioinformatics studies use physicochemical properties for modeling and analyses [10], [11], [12], [13]. The amino acid indices (AAindex) database collects many published indices representing physicochemical properties of amino acids. For each physicochemical property, there is a set of 20 numerical values for amino acids. Currently, 544 physicochemical properties can be retrieved from the AAindex database of version 9.0 [14]. After removing physicochemical properties having the value 'NA' in the amino acid indices, 531 physicochemical properties are obtained for the following studies.

Given a peptide sequence of γ-turns, 531 physicochemical properties are used to represent each of the three residues. The number of total features vector for analysis are 1593 (531×3).

C. Physicochemical property-based decision tree method (PPDT)

Decision tree-based methods benefit from their simplicity and interpretability is wide used for interpretable analysis of various biological problems including the prediction of ubiquitylation sites [11] and protein stability [15]. In this study, a method named physicochemical property-based decision tree (PPDT) method is applied to predict and analyze γ-turn types. PPDT based on the famous classification and regression tree (CART) [9] to select informative physicochemical properties to build decision tree models. The CART is a non-parametric method that can deal with problems of limited priori information because it does not rely on any particular assumptions concerning the dependence type of the dependent variable Y on predictors X_i and statistical properties of the data. The system flow of PPDT method is shown in Fig. 1.

The construction of a PPDT tree includes two major steps: tree growing and tree pruning. The tree growing step recursively partitions dataset into two sub-datasets by utilizing a splitting rule based on an impurity measurement of Gini index. For a node t with estimated class probabilities $p(j\,|\,t),\ j=1,...,J$, where J is the total number of classes, its corresponding Gini index is defined as:

$$1-\sum_j p^2(j\,|\,t)\,. \qquad (1)$$

For each node t, the feature with largest impurity reduction is selected to split the dataset at node t. The splitting process proceeds until there is no significant purity gain by adding more nodes.

Applying a large tree for prediction can always result in an over-fitting problem. The tree pruning step remove nodes to achieve lowest misclassification rate of five-fold cross-validation. The pruned tree is expected to have better generalization ability with less over-fitting problem than a large tree.

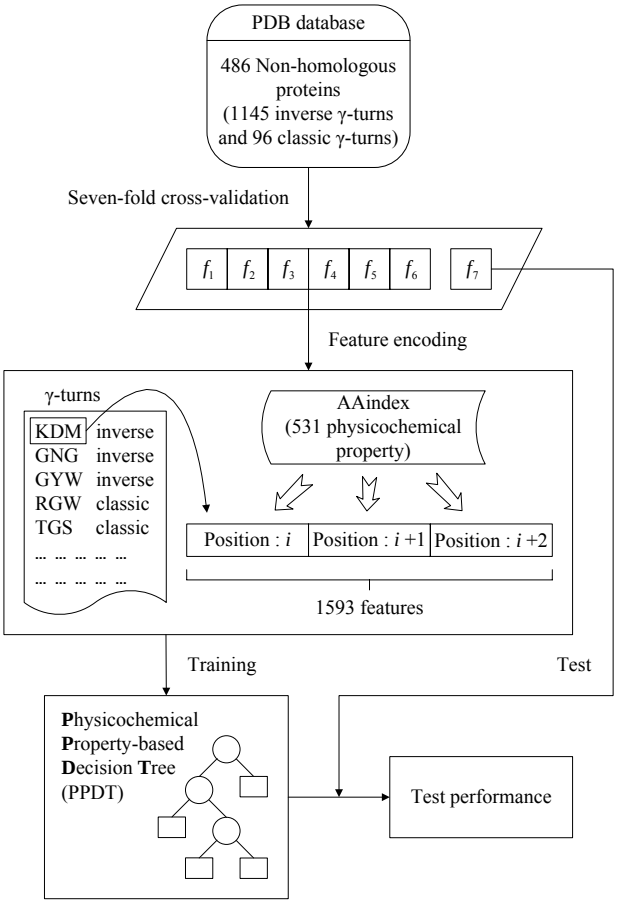

Fig. 1 System flow of the proposed PPDT method.

D. Performance evaluation

For comparison with previous methods, the same seven-fold cross-validation (7-CV) method is used to evaluate performances of PPDT. The procedure for applying 7-CV includes two steps as described in the follows. First, the whole dataset is randomly divided into seven data subsets with nearly equal sizes of inverse and classic γ-turn samples. Second, for each fold, six data subsets are combined into a training dataset

TABLE I
PERFORMANCES USING SEVEN-FOLD CROSS-VALIDATION

Methods	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC	AUC
SVM (binary encoding)	35.80	67.20	93.40	0.44	NA
LDA-NN (binary encoding)	38.24	98.14	93.80	0.46	NA
PPDT	40.63	97.64	93.23	0.46	0.674
Naïve Bayes classifier (using PPDT selected properties)	38.54	98.08	93.47	0.46	0.736

and the remaining one subset is used as test dataset. For convenience of performance representation, inverse and classic γ-turns are marked as positive and negative samples. Four measurements are calculated as following:

$$\text{Accuracy} = 100 \times \frac{(\text{TP} + \text{TN})}{N}, \quad (2)$$

$$\text{Sensitivity} = 100 \times \frac{\text{TP}}{(\text{TP} + \text{FN})}, \quad (3)$$

$$\text{Specificity} = 100 \times \frac{\text{TN}}{(\text{TN} + \text{FP})}, \quad (4)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}, \quad (5)$$

where TP, TN, FP, FN and *N* are numbers of true positives, true negatives, false positives, false negatives and total number of samples, respectively.

III. RESULTS

A. Prediction Performance

For predicting γ-turn types from peptide sequences, this study proposed a physicochemical property-based decision tree (PPDT) method, which utilizes the classification and regression tree (CART) method [9] to select informative physicochemical properties and create a decision tree model.

To evaluate the prediction performance of PPDT, seven-fold cross-validation is applied as following. First, the original dataset is randomly divided into seven sub-datasets with nearly equal size. Second, for each time of test, one fold is isolated as test dataset and the other six folds are used as training dataset for deriving a decision tree model. Finally, seven decision tree models are applied to predict the corresponding test datasets and the overall performance can be calculated.

A total of 30 runs of seven-fold cross-validations are applied to obtain statistically significant results. The highest, average and lowest AUC values of the 30 runs are 0.674, 0.668 and 0.662, respectively. The small value of standard deviation of 0.0026 shows the robustness of the PPDT methods.

Because previous studies use threshold-dependent performance measurements to evaluate their methods [6], [7], it is difficult to compare performances with different specificity or sensitivity levels. In order to easily compare PPDT with previous method, a threshold-independent ROC curve is applied for comparisons. As shown in Fig. 2, PPDT shows comparable performances with previous methods [6], [7]. In addition to the threshold-independent ROC curve, the threshold-dependent measurements of sensitivity, specificity, accuracy and MCC is also shown in Table I by selecting a specificity threshold nearly equal to 98.14% as reported by previous neural-network based method [7].

B. Interpretation of tree-based knowledge

In addition to good prediction performance, the knowledge obtained from constructed prediction models is especially important for better understanding the determination of γ-turn types. The tree-based PPDT method is able to provide interpretable decision rules, compared to black-box methods such as neural networks and support vector machines.

A physicochemical property-based decision tree built on the whole dataset is shown in Fig. 3. The constructed decision tree is based on only two physicochemical properties (shown in Table II): AAindex IDs of CHAM820102 and WERD780101 represent the free energy of solution in water [16] and propensity to be buried inside [17], respectively. In order to easily interpret the decision tree, a set of three IF-THEN rules can be extracted from the obtained decision tree as shown in

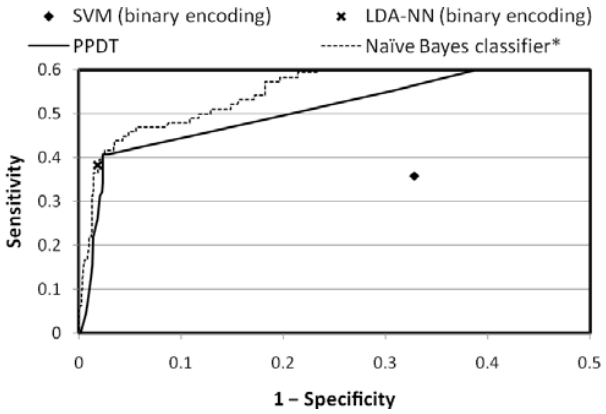

Fig. 2 Performance comparisons of PPDT and SVM and LDA-NN methods based on binary encoding

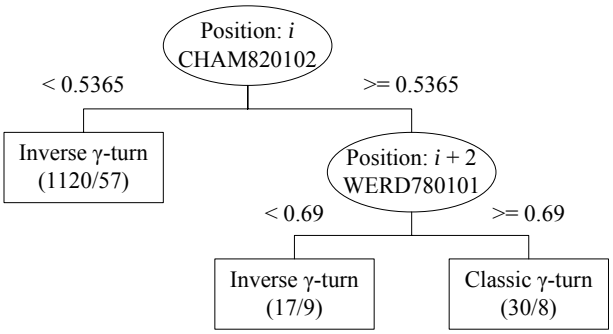

Fig. 3 Physicochemical property-based decision tree.

Table III.

The first rule means that given a γ-turn if the free energy of solution in water (kcal/mole) of the residue in position *i* is less than 0.5365 then it is an inverse γ-turn. The second rule means that given a γ-turn if the free energy of solution in water (kcal/mole) of the residue in position *i* is larger than or equal to 0.5365 and the propensity to be buried inside of the residue in position $i+2$ is less than 0.69 then it is an inverse γ-turn. The third rule means that given a γ-turn if the free energy of solution in water (kcal/mole) of the residue in position *i* is larger than or equal to 0.5365 and the propensity to be buried inside of the residue in position $i+2$ is larger than or equal to 0.69 then it is a classic γ-turn.

C. The usability of identified informative physicochemical properties

To demonstrate the usability of identified informative physicochemical properties for the prediction of γ-turn types, a simple Naïve Bayes classifier using PPDT selected informative physicochemical properties is applied to predict γ-turn types. The Naïve Bayes classifier (NB) based on the assumption of conditional independence is a relatively simple classification model [18]. Given a feature vector $\mathrm{X} = (x_1,...,x_n)$, the NB classifier calculates the probability of the given sample belongs to a certain class *C* that maximizes the likelihood $P(X \mid C) = P(x_1,\ldots,x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$.

In this study, a modified version of the NB classifier with kernel estimator that is available in WEKA [19] is applied to predict γ-turn types using the identified two physicochemical properties. As shown in Table I, the NB classifiers can predict γ-turn types with a high AUC value of 0.736 that is higher than the proposed PPDT method. The high performance of NB classifier shows the usability of PPDT selected informative physicochemical properties.

TABLE II
PHYSICOCHEMICAL PROPERTIES USED BY THE DECISION TREE TRAINED ON WHOLE DATASET

Position	AAindex ID	Description	Reference
i	CHAM820102	Free energy of solution in water, kcal/mole	[16]
i+2	WERD780101	Propensity to be buried inside	[17]

TABLE III
PHYSICOCHEMICAL PROPERTIES USED BY DECISION TREES TRAINED ON SEVEN-FOLD CROSS-VALIDATION PROCEDURES

Position	AAindex ID	Description	Reference
i	CHAM820102	Free energy of solution in water, kcal/mole	[16]
i	CHAM830103	The number of atoms in the side chain labeled 1+1	[20]
i	GEIM800104	Alpha-helix indices for alpha/beta-proteins	[21]
i	PRAM900104	Relative frequency in reverse-turn	[22]
i	ONEK900101	Delta G values for the peptides extrapolated to 0 M urea	[23]
i+1	GOLD730101	Hydrophobicity factor	[24]
i+2	MAXF760102	Normalized frequency of extended structure	[25]

To further demonstrate the discrimination ability of the informative physicochemical properties, a bubble chart representing the distribution of all γ-turns of our dataset in two PPDT selected physicochemical properties is shown in Fig. 4. By applying the concise tree, a sensitivity of 31.3% with a very high specificity level of 99.3% can be obtained for classic γ-turns.

IV. CONCLUSION

The γ-turns are important protein secondary structures that are important in protein folding and molecular recognition. The prediction and analysis of γ-turn types can provide better understanding of the underlying mechanism. This study proposes a physicochemical property-based decision tree (PPDT) method to interpretably predict γ-turn types.

With its simplicity, the PPDT performs so well that is comparable with the black-box methods of support vector

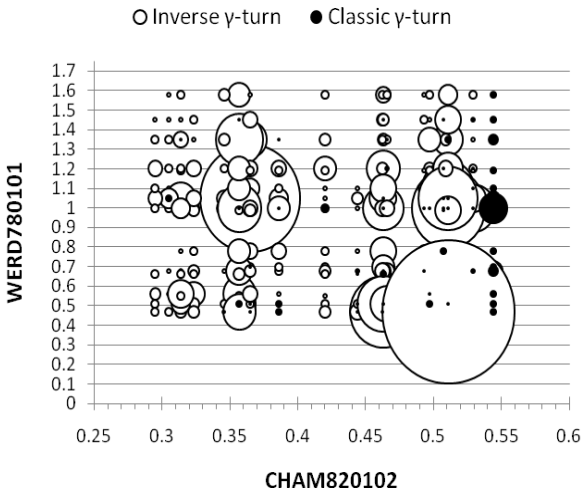

Fig. 4 The distribution of γ-turns in two PPDT selected informative physicochemical properties.

machines and neural networks. In addition to the high performance of PPDT for predicting γ-turn types, the knowledge obtained from PPDT can be extracted and represented as simple IF-THEN rules. Finally, we use a simple Naïve Bayes classifier to demonstrate the usability of PPDT identified informative physicochemical properties for discriminating γ-turn types.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 96-2628-E-009-141-MY3 and NSC 98-2627-B-009-004-.

REFERENCES

[1] E. Milner-White, B. M. Ross, R. Ismail, K. Belhadj-Mostefa, and R. Poet, "One type of gamma-turn, rather than the other gives rise to chain-reversal in proteins," *J Mol Biol,* vol. 204, pp. 777-82, Dec 5 1988.

[2] I. Alkorta, M. Suarez, R. Herranz, R. Gonzalez-Muniz, and M. Garcia-Lopez, "Similarity Study on Peptide?-turn Conformation Mimetics," *J Mol Model,* vol. 2, pp. 16-25, 1996.

[3] H. Kaur and G. P. Raghava, "A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment," *Protein Sci,* vol. 12, pp. 923-9, May 2003.

[4] K. Guruprasad, S. Shukla, S. Adindla, and L. Guruprasad, "Prediction of gamma-turns from amino acid sequences," *J Pept Res,* vol. 61, pp. 243-51, May 2003.

[5] X. Hu and Q. Li, "Using support vector machine to predict beta- and gamma-turns in proteins," *J Comput Chem,* vol. 29, pp. 1867-75, Sep 2008.

[6] S. Jahandideh, A. S. Sarvestani, P. Abdolmaleki, M. Jahandideh, and M. Barfeie, "gamma-Turn types prediction in proteins using the support vector machines," *J Theor Biol,* vol. 249, pp. 785-90, Dec 21 2007.

[7] S. Jahandideh, S. Hoseini, M. Jahandideh, A. Hoseini, and F. M. Disfani, "Gamma-turn types prediction in proteins using the two-stage hybrid neural discriminant model," *J Theor Biol,* vol. 259, pp. 517-22, Aug 7 2009.

[8] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers,* vol. 22, pp. 2577-637, Dec 1983.

[9] L. Breiman, *Classification and regression trees*: Chapman & Hall/CRC, 1984.

[10] C.-W. Tung and S.-Y. Ho, "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics,* vol. 23, pp. 942-9, Apr 15 2007.

[11] C.-W. Tung and S.-Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics,* vol. 9, p. 310, 2008.

[12] W.-L. Huang, C.-W. Tung, H.-L. Huang, S.-F. Hwang, and S.-Y. Ho, "ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *Biosystems,* Jan 4 2007.

[13] K.-T. Hsu, H.-L. Huang, C.-W. Tung, Y.-H. Chen, and S.-Y. Ho, "Analysis of physicochemical properties on prediction of R5, X4, and R5X4 HIV-1 coreceptor usage," *Int J Biol Life Sci,* vol. 5, pp. 208-15, 2009.

[14] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: amino acid index database, progress report 2008," *Nucleic Acids Res,* vol. 36, pp. D202-5, Jan 2008.

[15] L.-T. Huang, M. M. Gromiha, and S.-Y. Ho, "iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations," *Bioinformatics,* vol. 23, pp. 1292-3, May 15 2007.

[16] M. Charton and B. I. Charton, "The structural dependence of amino acid hydrophobicity parameters," *J Theor Biol,* vol. 99, pp. 629-44, Dec 21 1982.

[17] D. H. Wertz and H. A. Scheraga, "Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule," *Macromolecules,* vol. 11, pp. 9-15, Jan-Feb 1978.

[18] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter,* vol. 11, pp. 10-18, 2009.

[20] M. Charton and B. I. Charton, "The dependence of the Chou-Fasman parameters on amino acid side chain structure," *J Theor Biol,* vol. 102, pp. 121-34, May 7 1983.

[21] M. Geisow and R. Roberts, "Amino acid preferences for secondary structure vary with protein class," *Int J Biol Macromol,* vol. 2, pp. 387-389, 1980.

[22] M. Prabhakaran, "The distribution of physical, chemical and conformational properties in signal and nascent peptides," *Biochem J,* vol. 269, pp. 691-6, Aug 1 1990.

[23] K. T. O'Neil and W. F. DeGrado, "A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids," *Science,* vol. 250, pp. 646-51, Nov 2 1990.

[24] D. E. Goldsack and R. C. Chalifoux, "Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins," *J Theor Biol,* vol. 39, pp. 645-51, Jun 1973.

[25] F. R. Maxfield and H. A. Scheraga, "Status of empirical methods for the prediction of protein backbone topography," *Biochemistry,* vol. 15, pp. 5138-53, Nov 16 1976.

Chyn Liaw received the BS degree in Applied Mathematics, Tatung University, Taipei, Taiwan, in 2007. She is currently a direct PhD student at the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan. Her research interests include bioinformatics, machine learning and data mining.

Chun-Wei Tung received the BS degree in Biology, National Cheng Kung University, Tainan, Taiwan, in 2005. He is currently a PhD candidate at the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan. His research interests include immunoinformatics, bioinformatics, machine learning and data mining.

Shinn-Jang Ho received the B.S. degree in power mechanic engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1983 and the M.S. and Ph.D. degrees in mechanical engineering from National Sun-Yat-Sen University, Kaohsiung, Taiwan, in 1985 and 1992, respectively. He is currently an Associate Professor in the Department of Automation Engineering at National Huwei Institute of Technology, Huwei, Yulin, Taiwan. His research interests include optimal control, fuzzy systems, genetic algorithms, and system optimization.

Shinn-Ying Ho received the BS, MS, and PhD degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1984, 1986, and 1992, respectively. From 1992 to 2004, he was with the Department of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan. He is currently the vice dean of College of Biological Science and Technology, National Chiao Tong University, and a professor in the Department of Biological Science and Technology and the Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan. He serves on the Editorial Boards of International Journal of Applied Metaheuristic Computing (IJAMC), Theoretical Biology Insights and Biomedical Engineering and Computational Biology. His research interests include evolutionary algorithms, soft computing, image processing, pattern recognition, bioinformatics, data mining, machine learning, computer vision, fuzzy classifier, large-scale parameter optimization problems, and system optimization. He is a member of the IEEE.