行政院國家科學委員會專題研究計畫 成果報告

建構可解讀模糊規則知識庫來預測與分析 DNA 結合的蛋白質
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 100-2221-E-009-130-
執 行 期 間 ：100 年 08 月 01 日至 101 年 07 月 31 日
執 行 單 位 ：國立交通大學生物科技學系（所）

計 畫 主 持 人：黃慧玲

計畫參與人員：大專生-兼任助理人員：陳力行
大專生-兼任助理人員：黃柏霖
博士班研究生-兼任助理人員：蔡明儒

報 告 附 件 ：出席國際會議研究心得報告及發表論文

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 101 年 10 月 31 日

中文摘要： DNA 結合區域/蛋白質在細胞機能中扮演著相當重要的角色，因其參與許多生物程序，如 DNA 的轉錄調控、修理、複製等，如果可以對於 DNA 與蛋白質如何作用對於日後研究與應用會有極大的幫助。因此以生物資訊的方式預測 DNA 結合區域/蛋白質便成為有趣且重要的議題。目前有許多研究使用不同的機械學習法作為預測方法，其中以支援向量分類器最多且效能最好。然而絕大多數以支援向量分類器做為預測 DNA 結合區域/蛋白質的方法，都使用到大量的特徵及數值作為分類依據，這些方法雖然具有良好的分類效能，但對於所學習的特徵資料，卻無法為生物學家提供良好的解讀性。因此，本計畫旨在建立可解讀的模糊邏輯規則，建立一套相關系統，以增進預測和有效分析 DNA 結合區域分類的物化特性知識。我們提供一個運用物化特性為特徵的演化式模糊規則分類器(iFRC)，做為預測 DNA 結合區域/蛋白質的系統。此系統主要是以智慧型基因演算法(IGA)來完成最佳化的特徵選擇，利用 IGA 系統化的挑選出最少的特徵、模糊規則數量最為分類依據，同時最佳化分類模型，以求得最高的辨識率。

本系統所建構的模糊規則用以分類 DNA 結合區域/蛋白質的平均正確率為 77.46%，測試的正確率達 83.33%。該結果說明本系統所建構出來的規則可信度極高。進一步分析經由本次研究結果，我們發現到在物化特性中以帶正電及凡德瓦利的特徵影響最大，而蛋白質和 DNA 之間的識別在第一步驟中，蛋白質和 DNA 的電荷之間的互補性，這項結果符合先前研究，且被認為是重要的。結果也同時顯示，胺基酸結合區的胜肽，其凡德瓦力值及所帶的正電荷皆高於非結合區。相較於之前的研究結果分生實驗的結果無法量化及其他生物資訊的方法無法解讀的缺點，我們提供較為精確的數值可供分子生物學實驗室作為實驗突變的參考，希望對日後分子生物研究及應用有所幫助。

中文關鍵詞： 去氧核醣核酸結合區域 特徵選擇 基因演算法 支援向量機 模糊邏輯規則 知識擷取 物化特性 特異位置分數矩陣 蛋白質功能預測

英文摘要： DNA-binding domains/proteins play essential roles in a cell, which are involved in transcription, replication, packaging, repair and rearrangement. Numerous prediction methods of DNA-binding domains/proteins were proposed by identifying informative features and designing effective classifiers. These researches reveal that the DNA-protein binding mechanism is complicated and existing accurate predictors such as support vector machine (SVM) with position specific scoring matrices (PSSMs) are regarded as black-box methods which are not easily interpretable for biologists. It is desirable to design predictors using interpretable features and classifiers, and the prediction results are explainable for knowledge acquisition. In this study, we propose an ensemble fuzzy rule base classifier consisting of a set of interpretable fuzzy rule classifiers (iFRCs) using informative physicochemical properties as features. In designing iFRCs, feature selection, membership function design, and fuzzy rule base generation are all simultaneously optimized using an intelligent genetic algorithm (IGA). IGA maximizes prediction accuracy, minimizes the number of features selected, and minimizes the number of fuzzy rules to generate an accurate and concise fuzzy rule base. Benchmark datasets of DNA-binding domains are used to evaluate the proposed ensemble classifier of 30 iFRCs. Each iFRC has a mean test accuracy of 77.46%, and the ensemble classifier has a test accuracy of 83.33%, where the method of SVM with PSSMs has the accuracy of 82.81%. The physicochemical properties of the first two ranks according to their contribution are positive charge and Van Der Waals volume. Charge complementarity between protein and DNA is thought to be important in the first step of recognition between protein and DNA. The amino acid residues of binding peptides have larger Van Der Waals volumes and positive charges than those of non-binding ones. The proposed knowledge acquisition method by establishing a fuzzy rule-based classifier can also be applicable to predict and analyze other protein functions from sequences.

英文關鍵詞： DNA-binding domains, feature selection, genetic algorithm, support vector machine, fuzzy rules, knowledge acquisition, physicochemical properties, position specific scoring matrix, protein function prediction

FRKAS: Knowledge Acquisition Using a Fuzzy Rule Base Approach to Insight of DNA-Binding Domains/Proteins

Hui-Lin Huang[1,2], Fang-Lin Chang[3], Shinn-Jang Ho[4], Li-Sun Shu[5], Wen-Lin Huang[6], and Shinn-Ying Ho[1,2*]

[1]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan

[2]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan

[3]Department of Anesthesiology, Tri-Service General Hospital, Taipei, Taiwan

[4]Department of Automation Engineering, National Formosa University, Yunlin, Taiwan

[5]Department of Information Management, Overseas Chinese University, Taichung, Taiwan

[6]Department of Multimedia Entertainment Science, Asia Pacific Institute of Creativity, Miaoli, Taiwan

*Corresponding author

Email addresses:

HLH: hlhuang@mail.nctu.edu.tw

FLC: 0324crab@pchome.com.tw

SJH: sjho@nfu.edu.tw

LSS: lssu@ocu.edu.tw

WLH: wenlinhuang2001@gmail.com

SYH: syho@mail.nctu.edu.tw

Abstract: DNA-binding domains/proteins play essential roles in a cell, which are involved in transcription, replication, packaging, repair and rearrangement. Numerous prediction methods of DNA-binding domains/proteins were proposed by identifying informative features and designing effective classifiers. These researches reveal that the DNA-protein binding mechanism is complicated and existing accurate predictors such as support vector machine (SVM) with position specific scoring matrices (PSSMs) are regarded as black-box methods which are not easily interpretable for biologists. It is desirable to design predictors using interpretable features and classifiers, and the prediction results are explainable for knowledge acquisition. In this study, we propose an ensemble fuzzy rule base classifier consisting of a set of interpretable fuzzy rule classifiers (iFRCs) using informative physicochemical properties as features. In designing iFRCs, feature selection, membership function design, and fuzzy rule base generation are all simultaneously optimized using an intelligent genetic algorithm (IGA). IGA maximizes prediction accuracy, minimizes the number of features selected, and minimizes the number of fuzzy rules to generate an accurate and concise fuzzy rule base. Benchmark datasets of DNA-binding domains are used to evaluate the proposed ensemble classifier of 30 iFRCs. Each iFRC has a mean test accuracy of 77.46%, and the ensemble classifier has a test accuracy of 83.33%, where the method of SVM with PSSMs has the accuracy of 82.81%. The physicochemical properties of the first two ranks according to their contribution are positive charge and Van Der Waals volume. Charge complementarity between protein and DNA is thought to be important in the first step of recognition between protein and DNA. The amino acid residues of binding peptides have larger Van Der Waals volumes and positive charges than those of non-binding ones. The proposed knowledge acquisition method by establishing a fuzzy rule-based classifier can also be applicable to predict and analyze other protein functions from sequences.

Keywords: DNA-binding domains, feature selection, genetic algorithm, support vector machine, fuzzy rules, knowledge acquisition, physicochemical properties, position specific scoring matrix, protein function prediction.

INTRODUCTION

DNA-binding domains are functional proteins in a cell, which play a vital role in various essential biological activities, such as DNA transcription, replication, packaging, repair and rearrangement [1]. These transcription factors are mainly DNA-binding proteins (DNA-BPs) coded by 2~3% of the genome in prokaryotes and 6~7% in eukaryotes [2]. DNA-BPs play a pivotal role in various intra- and extra-cellular activities ranging from DNA replications to gene expression control. The researches reveal that the DNA-protein recognition mechanism is complicated and there is no simple rule for this recognition problem [3].

Some researchers have increasingly interests in the prediction and analyse of DNA-BPs [4-6]. Stawiski *et al.* presented that DNA-binding proteins could be predicted using a neural network trained with features of secondary structures and charged patches [4]. Ahmad and Sarai found that net charge, net dipole moment and quadrupole moment could each distinguish binding and non-binding proteins with known structures well [5]. Kumar *et al.* proposed a method for predicting DNA-binding proteins using support vector machine (SVM) and position-specific scoring matrices (PSSMs) profiles [6]. The methods [4-6] can fairly analyze and predict DNA-binding proteins, but suffer from obtaining human-interpretable knowledge from sequences.

Leung et al. [7] focus on protein-DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs). A framework to discover associated TF-TFBS binding sequence patterns in the most explicit and interpretable form from TRANSFAC is proposed [7]. Recent mining on exact TF-TFBS-associated sequence patterns (rules) has shown great potentials and achieved very promising results [8]. The approximate rules reveal both the flexible and specific protein-DNA interactions accurately. Huang et al. [9] proposed a systematic approach Auto-IDPCPs to automatically identify a set of physicochemical and biochemical properties in the AAindex database to design SVM-based classifiers for predicting and analyzing DNA-binding domains/proteins from sequences. Auto-IDPCPs identified 23 features of properties from the AAindex database [10] belonging to five clusters such as hydrophobicity, secondary structure, charge, solvent accessibility, polarity, flexibility, normalized Van Der Waals volume, pK (pK-C, pK-N, pK-COOH and pK-a(RCOOH)), etc.

The trend in analyzing DNA-BPs is not only to predict binding proteins well but also to obtain knowledge for biological understanding and finding. It is desirable to design predictors using interpretable features and classifiers, and the prediction results are explainable for knowledge acquisition. Human thinking and reasoning frequently involve fuzzy information originating from inherently inexact human concepts and matching of similar rather than identical experiences. In many applications, rule-based classifiers are created starting from machine learning and fuzzy logic.

In this study, we propose an ensemble fuzzy rule base classifier consisting of a set of interpretable fuzzy rule classifiers (iFRCs) based on the 23 physicochemical properties as features [9]. Because the DNA-BPs have the property of natural clustering, fuzzy classifiers using a scatter partition of feature spaces often have a smaller number of rules than those using grid partitions. In designing iFRCs, feature selection, membership function design, and fuzzy rule base generation are all simultaneously optimized using an intelligent genetic algorithm (IGA) [11]. IGA maximizes prediction accuracy, minimizes the number of features selected, and minimizes the number of fuzzy rules to generate an accurate and concise fuzzy rule base.

A fuzzy rule-based knowledge acquisition system (FRKAS) using an ensemble fuzzy rule classifier consisting of 30 iFRCs is proposed for prediction and analyse of DNA-BPs. Each iFRCs has two fuzzy rules, one for binding and the other for non-binding prediction. The ensemble classifier using eight physicochemical properties performs well with a test accuracy of 83.33%, compared with an individual SVM with PSSMs (82.81%) [6] and SVM with 22 physicochemical properties (80.73%) [9]. The physicochemical properties of the first two ranks according to their contribution are positive charge and Van Der Waals volume. The amino acid residues of binding peptides have larger Van Der Waals volumes and positive charges than those of non-binding ones.

MATERIALS AND METHODS

The framework of the proposed FRKAS is given in Fig. 1. FRKAS uses an ensemble fuzzy rule classifier consisting of 30 interpretable fuzzy rule classifiers (iFRCs). The design aim of iFRCs is to generate an accurate and concise fuzzy rule base. The following sections present the used datasets, feature sets, the design of iFRCs, the ensemble fuzzy classifier and knowledge acquisition of DNA-binding domains.

Datasets

For comparisons with existing methods [6], [9], the same benchmark dataset DNAset, also called main dataset from Kumar et al. [6] was used to establish an ensemble fuzzy rule classifier. DNAset has 146 non-redundant DNA-binding domains (or protein chains) in which no two domains have the sequence identity of more than 25%. A non-redundant set of 250 non-binding domains was obtained from Stawiski et al. [4]. They used following criteria: 1) no two protein chains have similarity more than 25% and 2) the approximate size and electrostatics are similar to DNA-BPs. An independent data set DNAiset is additionally used, having 92 DNA-binding domains and 100 non-DNA-binding proteins [6].

Feature sets

The method Auto-IDPCPs [9] consists of three tasks: 1) clustering 531 vectors of physicochemical and biochemical properties in the AAindex database into 20 classes using a fuzzy c-means algorithm, 2) utilizing an efficient genetic algorithm based optimization method to select an informative feature set to represent sequences, and 3) analyzing the selected feature vectors to identify the related physicochemical properties which may affect the binding mechanism of DNA-BPs.

Auto-IDPCPs [9] used a systematic approach to automatically identify a set of 23 properties for predicting and analyzing DNA-binding domains/proteins in the dataset DNAset. The 23 properties belonging to five clusters are used to design the proposed ensemble fuzzy rule classifiers, shown in Table 1.

Interpretable fuzzy rule classifiers (iFRCs)

High performance of iFRCs mainly arises from two aspects. One is to simultaneously optimize all parameters in the design of iFRCs where all the elements of the fuzzy classifier design have been transformed into parameters of a large parameter optimization problem. The other is to use an efficient optimization algorithm IGA which is a specific variant of the intelligent evolutionary algorithm

[11]. The intelligent evolutionary algorithm uses a divide-and-conquer strategy to effectively solve large parameter optimization problems. IGA is shown to be effective in the design of accurate classifiers with a concise fuzzy rule base using an evolutionary scatter partition of feature space [12].

Flexible membership functions

The classifier design of iFRCs uses flexible generic parameterized fuzzy regions which can be determined by flexible generic parameterized membership functions (FGPMFs) and a hyperbox-type fuzzy partition of feature space [12]. Each fuzzy region corresponds to a parameterized fuzzy rule. In this study, the value of each physicochemical property is normalized into a real number in the unit interval [0, 1]. An FGPMF with a single fuzzy set is defined as

$$\mu(x)=\begin{cases}0 & \text{if } x\le a \text{ or } x\ge d\\ \dfrac{x-a}{b-a} & \text{if } a<x<b\\ \dfrac{d-x}{d-c} & \text{if } c<x<d\\ 1 & \text{if } b\le x\le c\end{cases} \tag{1}$$

where $x \in [0, 1]$ and $a \le b \le c \le d$. The variables a, b, c and d determining the shape of a trapezoidal fuzzy set are the parameters to be optimized. It is well recognized that confining evolutionary searches within feasible regions is often much more reliable than penalty approaches for handling constrained problems [13]. Therefore, five parameters V^1, $V^2,\ldots,$ $V^5 \in [0,1]$ (without constraints among V^i) instead of a, b, c and d are encoded into a chromosome for facilitating IGA. Let an additional variable L=V^1 which determines location of the fuzzy set characterizing the occurrence of training patterns. When V^i are obtained, variables a, b, c, and d can be derived as follows: a=L-(V^2+V^3), b=L-V^3, c=L+V^4, and d=L+(V^4+V^5) where $b \le L \le c$. This transformation can always make the derived values of a, b, c and d feasible and reduce interactions among encoded parameters of the IGA's chromosomes. Some illuminations of FGPMF are shown in Fig. 2 [12].

Fuzzy rule and fuzzy reasoning method

The following fuzzy if–then rule base for n-dimensional classification problems are used in the design of iFRCs:

R_j : If x_1 is A_{j1} and . . . and x_n is A_{jn} then class CL_j with CF_j, $j = 1, . . . , N$.

where R_j is a rule label, x_i denotes a variable of physicochemical property, A_{ji} is an antecedent fuzzy set, C is a number of classes, $CL_j \in \{1, . . ., C\}$ denotes a consequent class label, CF_j is a certainty grade of this rule in the unit interval [0, 1], and N is a number of initial fuzzy rules in the training phase. In this study, C=2 (two classes for binding and non-binding), n=23 (initial number in the feature set to be selected), and N=3C (initial number in the rule set to be selected).

To enhance interpretability of fuzzy rules, linguistic variables in fuzzy rules can be used. Each variable x_i has a linguistic set U= {small, medium, large}. Each linguistic value of x_i equally represents 1/3 of the domain [0, 1]. An antecedent fuzzy set $A_{ji} \in A_U$ where A_U denotes a set of subsets of U. Examples of linguistic antecedent fuzzy sets are shown in Fig. 3. If x_i is A_{ji} representing {medium, large}, it means the value of x_i (physicochemical property) is belonging to the set of {medium, large}. If x_i is A_{ji} representing {small, medium, large}, it means the physicochemical property is

ALL, i.e., don't care.

In the training phase, all the variables CL_j and CF_j are treated as parametric genes encoded in a chromosome and their values are obtained using IGA. The following fuzzy reasoning method is adopted to determine the class of an input pattern $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ based on voting using multiple fuzzy if–then rules:

Step 1: Calculate score $S_{\text{Class}v}$ ($v = 1, \ldots, C$) for each class as follows:

$$S_{\text{Classv}} = \sum_{\substack{R_j \in FC \\ CL_j = Classv}} \mu_j(x_p) CF_j, \quad \mu_j(x_p) = \prod_{i=1}^{n} \mu_{ji}(x_{pi}), \tag{2}$$

where FC denotes the fuzzy classifier, and $\mu_{ji}(\cdot)$ represents the membership function of the antecedent fuzzy set A_{ji}.

Step 2: Classify x_p as the class with a maximal value of S_{Classv}.

Notably, x_p is classified into the binding or non-binding class for one iFRC. The final classification of x_p is determined using the proposed ensemble classifier consisting of 30 iFRCs in the study.

Chromosome representation of IGA

A chromosome consists of control genes for selecting useful features (physicochemical properties) and significant fuzzy rules, and parametric genes for encoding the membership functions and fuzzy rules. The control genes comprise two types of parameters. One is parameters for selecting features. The other is parameters for selecting fuzzy rules. The parametric genes determine variables of three types: $V_{ji}^{t} \in [0, 1]$, t=1, …, 5, for determining the antecedent fuzzy set A_{ji}, CL_j for determining the consequent class label of rule R_j, and $CF_j \in [0, 1]$ for determining the certainty grade of rule R_j , where j=1, ..., N and i=1, ..., n. A rule base with N fuzzy rules is represented as an individual. The detailed explanation of the chromosome representation and implementation can be referred to [12]. The design of an efficient fuzzy classifier is formulated as a large parameter optimization problem. Once the solution of IGA is obtained, an accurate classifier with a concise fuzzy rule base can be obtained.

Fitness function of IGA

We define the fitness function of IGA for designing iFRCs as follows:

$$\max Fit(FC) = ACC - W_r N_r - W_f N_f \tag{3}$$

where W_r and W_f are positive weights. In this study, the fitness function is used to optimize the three objectives: 1) to maximize the classification accuracy ACC, 2) to minimize the number N_r of fuzzy rules, and 3) to minimize the number N_f of selected features. For obtaining an easily-interpretable knowledge rule base for each iFRC, the smaller values of N_r and N_f are better. Therefore, we used large values of weights W_r = 0.2 and W_f = 0.1. Since the classification accuracy is not the first priority for iFRC, the ensemble fuzzy rule base classifier (EFRBC) consists of k (e.g., 30) iFRCs is necessary for obtaining high accuracy of predicting DNA-BPs.

Ensemble fuzzy rule base classifier (EFRBC)

There are three opinions for using an ensemble strategy [14]: 1) Statistical: the reason is related to lack of adequate data to properly represent the data distribution; 2)

Computational: the reason is the model selection problem, and 3) Representational: the reason is to address the cases when the chosen model cannot properly represent the sought decision boundary. In this study, the training dataset DNAset is relatively small, compared with the complex of recognition problems in the binding mechanism. For considering interpretability, the same model SVM is used to construct the EFRBC. Since the decision boundary of iFRCs is not complicated, the ensemble approach can advance the prediction accuracy.

The EFRBC is composed of *k*=30 iFRCs and a voting method.

(1) Classification of iFRCs: The prediction accuracy is highly related to the conciseness of the fuzzy rule base for every iFRC. The optimal design of iFRCs can simultaneously optimize the three objectives using a weighted sum approach: 1) to maximize prediction accuracy, 2) minimize the number of features selected, and 3) to minimize the number of fuzzy rules. However, the trade-off between prediction accuracy and conciseness of the rule base can be determined by tuning the weights W_r and W_f.

(2) Voting method: Different classification results of the query sequences will be obtained from the outputs of the *k* independent iFRCs, and then these results are integrated using the simple voting method.

$$\mathrm{VS}_j = \sum_{i=1}^{k} \tau, \quad \tau = \begin{cases} 1, & C_i = j \\ 0, & otherwise \end{cases}, \tag{4}$$

where *k* is the number of iFRC, *j*=1, 2, …, *C* is the class label, C_i is the predicted class label by *i*th iFRC. In this study, for a given query protein with *C*=2, the final class is determined by argmax {VS_1, VS_2}.

Four performance measurements were used to evaluate iFRC and EFRBC: sensitivity (SEN), specificity (SPE), accuracy (ACC), and Matthew's correlation coefficient (MCC), defined as follows: SEN = TP/(TP + FN), SPE = TN/(TN + FP), ACC = (TP+TN)/(TP+FP+TN+FN), and MCC = ((TP×TN)-(FN×FP))/SQRT ((TP+FN)(TN+FP)(TP+FP)(TN+FN)), where TP, TN, FP and FN are the numbers of true positive, true negative, false positive and false negative, respectively.

DNA-binding knowledge acquisition

This study proposes a knowledge acquisition approach based on the optimal design of fuzzy rule bases to insight of DNA-binding mechanism. The informative knowledge can be revealed from three aspects: 1) identified informative physicochemical properties, 2) rules of DNA-binding and non-binding mechanism, and 3) further analysis of binding mechanism using physicochemical properties.

Auto-IDPCPs [9] used a systematic approach to automatically identify a set of properties to design accurate SVM-based classifiers for predicting DNA-binding domains/proteins. By analysing the rules of EFRBC, we can further reduce the number of physicochemical properties with great contribution in predicting the binding mechanism. From the appropriate interpretations of fuzzy rules with linguistic variables, it is more understandable for biologists. The rule-based knowledge provides an effective approach to insight of DNA-binding domains.

An illustrating example of iFRC, shown in Fig. 4, and its explanation are given as follows. Fig. 4 shows two rules of one iFRC which uses tow features, H252 (PRAM900101, Hydrophobicity (Prabhakaran, 1990)) and H398 (ZIMJ680101, Hydrophobicity (Zimmerman et al., 1968)). The rules R1 and R2 with fuzzy sets are binding and non-binding rules, respectively. The descriptions of two rules are given as follows:

R1: if H252 is {medium, large} and H398 is {medium, large}, then binding with CF1=0.196.

R2: if H252 is {small, medium} and H398 is {small, medium}, then non-binding with CF2=0.549.

For example, a query sequence x_p has normalized values of H252 and H398, 0.4 and 0.3, respectively. The classification procedure using this iFRC is described as follows:

Step 1: Use equation (1) to calculate the values of membership functions $u_b()$ and $u_n()$ for binding and non-binding, respectively.

The value of $u_{b1}(0.4)$ is 0.537 for H252 and the value of $u_{b2}(0.3)$ is 1.0 for H398. The values of $u_{n1}(0.4)$ is 0.606 for H252 and the value of $u_{n2}(0.3)$ is 1.0 for H398. The binding value of $u_b(x_p)$ is $u_{b1}(0.4)$ x $u_{b2}(0.3)$=0.537 and the non-binding value of $u_n(x_p)=u_{n1}(0.4)* u_{n2}(0.3)$= 0.606.

Step 2: Use equation (2) to calculate the score for each class.
Because the non-binding score $S_{class} =u_n(x_p)$*CF2 =0.333 is larger than the binding score $S_{class} =u_b(x_p)$*CF1=0.105, the query sequence by using the single iFRC is classified into the non-binding class.

The fuzzy regions of binding and non-binding are illustrated in Fig. 4. The final classification of the query sequence using the proposed ensemble classifier is determined using the voting result of k=30 iFRCs. Generally, if the query sequence is located near the boundary of some fuzzy regions, the ensemble strategy can improve the prediction accuracy.

Results

The parameter settings of IGA [11] are N_{pop} = 20, P_c = 0.7, P_s =1−P_c, P_m = 0.01 and α = 15. Because the search space of the optimal design of iFRCs is proportional to the number N_p of parameters to be optimized, the stopping condition is suggested to use a fixed number 100N_p of fitness evaluations.

Prediction performance evaluation

The training samples with 23 properties in the dataset DNAset are represented as 23-dimensional feature vectors. This set of 23 physicochemical properties is identified by Auto-IDPCPs [9]. The dataset DNAiset was used for evaluating test performance of iFRCs and the ensemble classifier EFRBC. Due to the non-deterministic characteristic of genetic algorithms, the best iFRC with high training accuracy is selected for testing DNAiset from 30 runs. The average performance of 30 independent iFRCs in EFRBC is given in Table 2.

The SVM-based classifier with PSSMs has the training and test accuracies of 86.62% and 82.81%, respectively. The SVM-based classifier with 22 physicochemical properties identified by Auto-IDPCPs has high training accuracy 87.12% and a relatively small accuracy of 80.73%. The average performance of iFRCs has the training and test accuracies of 74.32% and 77.46%, respectively, without significant over-training problems. The average number of features is N_f

=1.34 and average number of rules is N_r =2.0. It reveals that the selected features are very effective and the rule bases are very concise. The test performance of EFRBC has a high test accuracy of 83.33%, sensitivity SEN=82.0%, specificity SPE=84.8%, and MCC=0.67, shown in Table 3. The ensemble strategy is effective for accurate prediction with an improvement of 5.87%.

The occurrence number of features and their descriptions in the 30 iFRCs are given in Table 4. There are eight features used in EFRBC. The 531 properties in the AAindex database were classified into six groups [9], [10]: 1) Alpha and turn propensities (A), 2) Beta propensity (B), 3) Composition (C), 4) Hydrophobicity (H), 5) Physicochemical properties (P), and 6) Other properties (O). Table 4 reveals that the two top-rank properties are the positive charge (H88, FAUJ880111) and normalized Van Der Waals Volume (P80, FAUJ880103).

Rule-based knowledge

Figure 5 shows seven iFRCs, a selected subset of 30 iFRCs, containing all the eight features in Table 4. We selected two iFRCs, iFRC_1 and iFRC_2, to illustrate the rules for DNA-binding mechanism. The iFRC_1 and iFRC_2 have training accuracies of 74.24% and 74.49%, the test accuracies of 72.40% and 82.81%, the feature numbers N_f of 2 and 1, and the rule numbers N_r of 2 and 2, respectively. The selected physicochemical properties are P80 (normalized Van Der Waals Volume), H88 (positive charge) and H355 (hydrophobicity). The fuzzy rules are linguistically interpretable as follows:

Fuzzy Classifier iFRC_1:

R1: if P80 is ALL and H355 is ALL, then binding with CF=0.161.

R2: if P80 is {small, medium} and H355 is {small, medium}, then non-binding with CF=0.576.

Fuzzy Classifier iFRC_2:

R1: if H88 is {medium, large}, then binding with CF=0.416.

R2: if H88 is {small}, then non-binding with CF=0.165.

Analysis of binding mechanism

The two top-rank features are positive charge (H88) and normalized Van Der Waals Volume (P80), shown in Table 4. A typical DNA-binding domain sequence in the training dataset DNAset, shown in Fig. 6(a), is used to have an insight into the binding mechanism. The sequence is the chain of the protein with PDBID 1WVL, a multimeric DNA-binding protein using Sac7d and GCN4 as templates, whose FASTA sequence is shown in Fig. 6(b).

We used the tool APBS [15] plugged in VMD 1.9 [16] to get the direct measurement of the charge distribution on 1WVL protein surface. The surface potential on 1WVL at neutral pH was calculated where the negatively and positively charged surfaces are shown in red and blue, respectively, shown as Fig. 7. The DNA-binding pocket, positively charged cavity, is visible in dark blue. Once the domain and DNA are assembled into clusters, hydrophobic molecules are held together by Van Der Waals interactions. Hydrophobic ridge of residues in the minor groove, the side-chain atoms of the hydrophobic ridge residues, is shown as gray spheres. Protein-DNA interaction resulting in the formation of salt bridges between cationic amino acid side chains and the phosphate backbone completely neutralize particular phosphate anions, eliminating repulsive interactions with fractional negative charges at neighboring phosphates [17].

Discussion

To avoid from overfitting the small-scale datasets in identifying physicochemical properties using an optimization approach, this study proposes a hybrid computational method of combining evidences by considering robust factors from the viewpoints of statistics and biological experiments from literature. It can be expected that the proposed method can effectively discover and rank more informative physicochemical and biochemical properties closely relative to the DNA-binding mechanism if the size of the training dataset is significantly increased. These discovered properties in predicting and analyzing the DNA-binding mechanism can be further investigated by biologists.

Conclusions

This study has proposed a systematic fuzzy rule based knowledge acquisition system (FRKAS) to predict and analyze DNA-binding domains/proteins. The merits of FRKAS can be summarized, described below.

1) The novel interpretable fuzzy rule classifiers (iFRCs) using informative physicochemical properties as features are proposed in this study. The features and classifiers are more helpful for understanding the binding mechanism rather than the SVM with PSSMs.

2) To obtain an accurate and concise fuzzy rule base, an intelligent genetic algorithm is utilized to optimize simultaneously the three objectives: maximizing prediction accuracy, minimizing the number of features selected, and minimizing the number of fuzzy rules. The designers can tune the weights in the weighted sum approach according to the preference to the three objectives.

3) Due to the small size of the training dataset, an ensemble classifier consisting of iFRCs is adopted to compensate the limitation, resulting in high-accuracy and robust performance.

4) The design of FRKAS considers both prediction accuracy and interpretability at the same time. FRKAS can also be applicable to predict and analyze other protein functions from sequences.

Acknowledgements

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 100-2221-E-009-130- and NSC 100-2221-E-009-143-.

References

[1] Gao, M., Skolnick, J., A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Comput Biol* 2009, *5*, e1000567.

[2] Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A., Brasseur, R., Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* 2005, *61*, 258-271.

[3] O'Flanagan, R. A., Paillard, G., Lavery, R., Sengupta, A. M., Non-additivity in protein-DNA binding. *Bioinformatics* 2005, *21*, 2254-2263.

[4] Stawiski, E. W., Gregoret, L. M., Mandel-Gutfreund, Y., Annotating nucleic acid-binding function based on protein structure. *J Mol Biol* 2003, *326*, 1065-1079.

[5] Ahmad, S., Sarai, A., Moment-based prediction of DNA-binding proteins. *J Mol Biol* 2004, *341*, 65-71.

[6] Kumar, M., Gromiha, M. M., Raghava, G. P., Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* 2007, *8*, 463.

[7] Leung, K.-S., Wong, K.-C., Chan, T.-M., Wong, M.-H., *et al.*, Discovering protein–DNA binding sequence patterns using association rule mining. *Nucleic Acids Research* 2010, *38 (19)*, 6324–6337.

[8] Chan, T. M., Wong, K. C., Lee, K. H., Wong, M. H., *et al.*, Discovering approximate-associated sequence patterns for protein-DNA interactions. *Bioinformatics* 2011, *27*, 471-478.

[9] Huang, H.-L., Lin, I.-C., Liou, Y.-F., Tsai, C.-T., *et al.*, Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *BMC Bioinformatics* 2011, *12 Suppl 1*.

[10] Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., *et al.*, AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008, *36*, D202-205.

[11] Ho, S. Y., Shu, L. S., Chen, J. H., Intelligent evolutionary algorithms for large parameter optimization problems. *IEEE Transactions on Evolutionary Computation* 2004, *8*, 522-541.

[12] Ho, S.-Y., Chen, H.-M., Ho, S.-J., Chen, T.-K., Design of accurate classifiers with a concise fuzzy rule base using an evolutionary scatter partition of feature space. *IEEE Trans. Systems, Man, and Cybernetics─Part B* 2004, *34*, 14.

[13] Michalewicz, Z., Dasgupta, D., Leriche, R. G., Schoenauer, M., Evolutionary algorithms for constrained engineering problems. *Comput Ind Eng* 1996, *30*, 851-870.

[14] Dietterich, T. G., *The Handbook of Brain Theory and Neural Networks*, The MIT Press, Cambridge 2002.

[15] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., McCammon, J. A., Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 2001, *98*, 10037-10041.

[16] Humphrey, W., Dalke, A., Schulten, K., VMD: visual molecular dynamics. *J Mol Graph* 1996, *14*, 33-38, 27-38.

[17] Strauss, J. K., Maher, L. J., 3rd, DNA bending by asymmetric phosphate neutralization. *Science* 1994, *266*, 1829-1834.

Figures

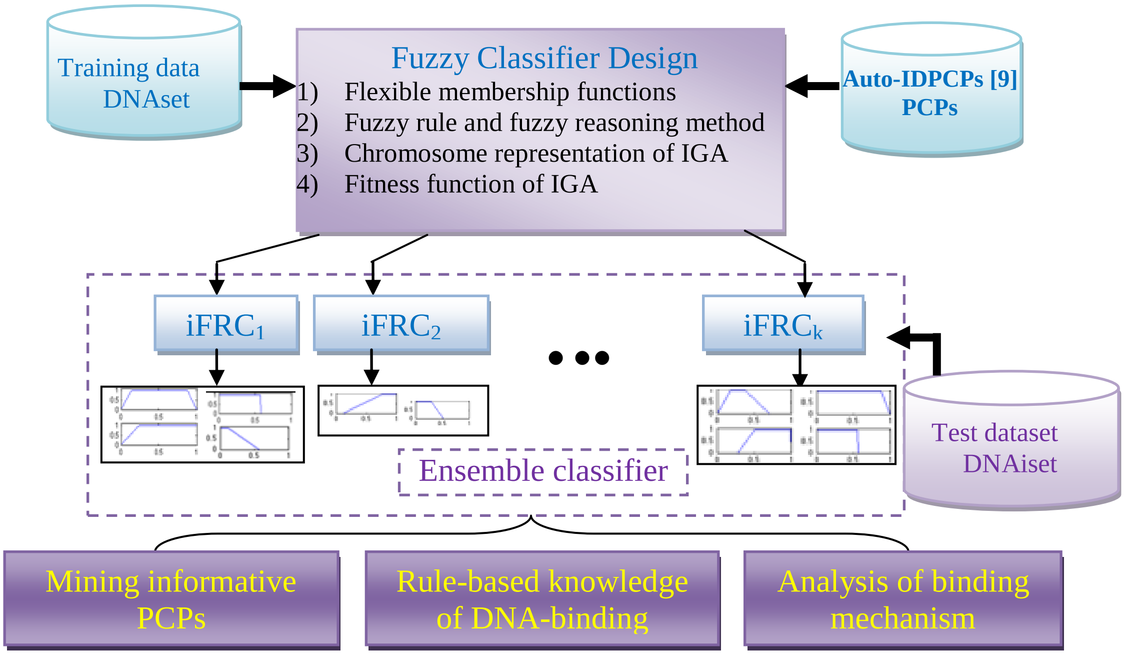

Training data
DNAset
Fuzzy Classifier Design
1) Flexible membership functions
2) Fuzzy rule and fuzzy reasoning method
3) Chromosome representation of IGA
4) Fitness function of IGA
Auto-IDPCPs [9]
PCPs
iFRC₁
iFRC₂
iFRCₖ
Ensemble classifier
Test dataset
DNAiset
Mining informative PCPs
Rule-based knowledge of DNA-binding
Analysis of binding mechanism

Figure 1. The framework of the proposed fuzzy rule-based knowledge acquisition system (FRKAS) based on physicochemical properties (PCPs).

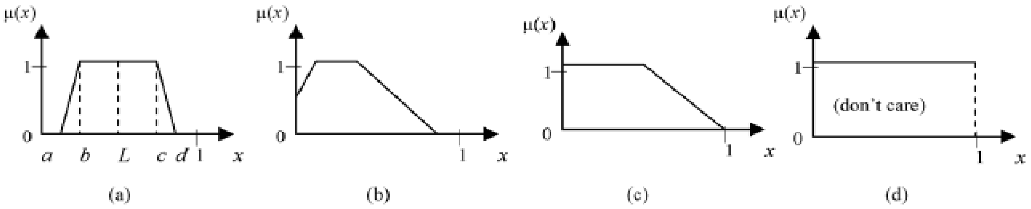


Figure 2. Illuminations of FGPMF: (a) a>0 and d< 1; (b) a<0<b, (c) b≦0; (d) b≦0 and c≧1.

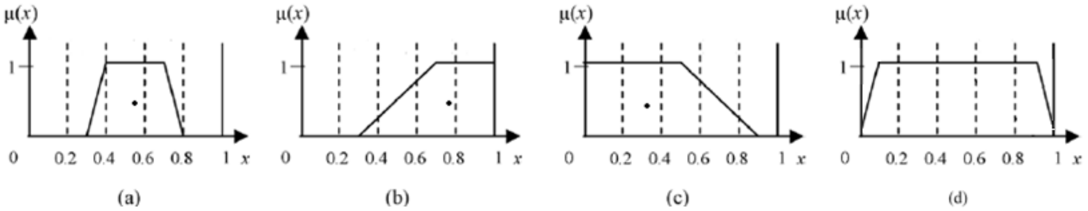


Figure 3. Examples of an antecedent fuzzy set A_{ji} with linguistic values (small, medium, large): (a) A_{ji} represents {medium}; (b) A_{ji} represents {medium, large}; (c) A_{ji} represents {small, medium} ; (d) A_{ji} represents {small, medium, large} or ALL.

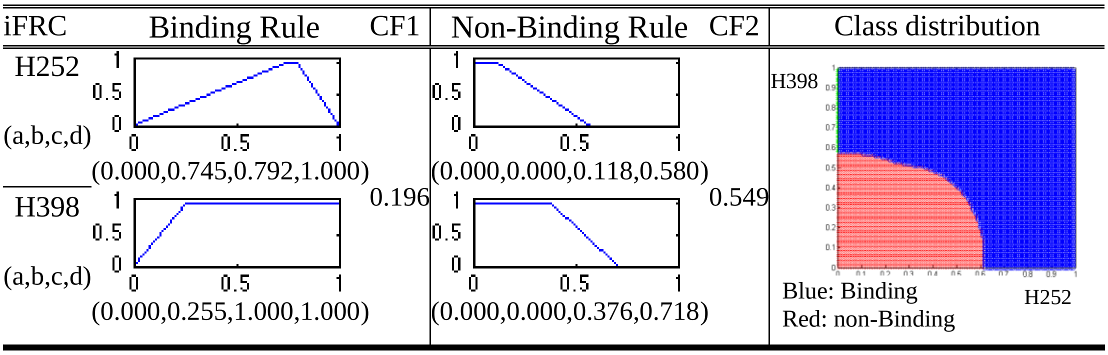


Figure 4. An illustrating example of iFRC uses tow features, H252 and H398.

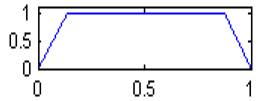
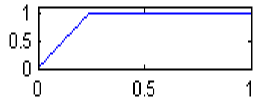
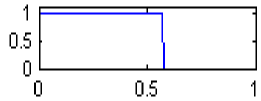
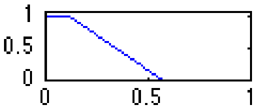
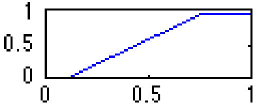
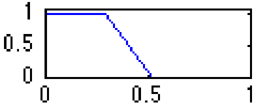
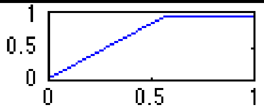
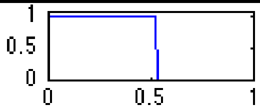
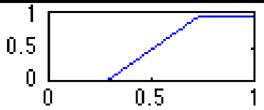
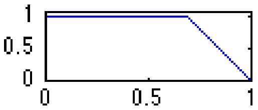
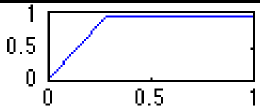
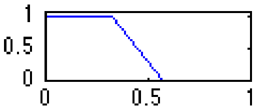
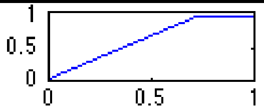
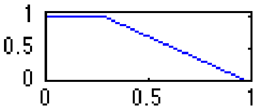
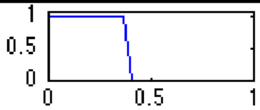
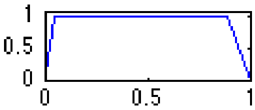
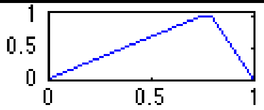
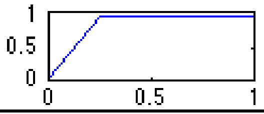
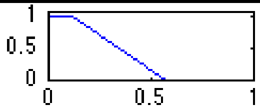
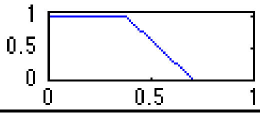
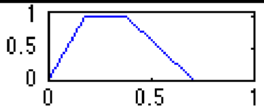
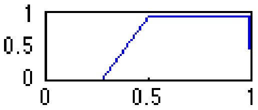
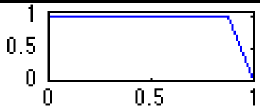
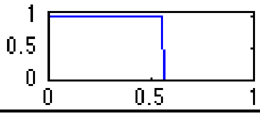
Classifier	PCP	Binding Rule	CF1	Non-Binding Rule	CF2
IFRC$_1$	P80 H355	 	R1 0.161	 	R2 0.576
IFRC$_2$	H88		R1 0.416		R2 0.165
IFRC$_3$	P80		R1 0.235		R2 0.969
IFRC$_4$	P80 A97	 	R1 0.949	 	R2 0.718
IFRC$_5$	H88 A237	 	R1 0.910	 	R2 0.286
IFRC$_6$	H252 H398	 	R1 0.196	 	R2 0.549
IFRC$_7$	A237 H482	 	R1 0.867	 	R2 0.153

Figure 5. Some typical iFRCs covering the eight physicochemical properties (PCPs) in Table 4.

```
MVKVKFKYKGEEKEVDTSKIKKVWRVGKMVSFTYDDNGKTGRGAVSEKDAPKELLDMLARAEREKK
```

(a)

```
>1WVL:A|PDBID|CHAIN|SEQUENCE
MVKVKFKYKGEEKEVDTSKIKKVWRVGKMVSFTYDDNGKTGRGAVSEKDAPKELLDMLARAEREKK
GVLKKLRAVENELH
>1WVL:B|PDBID|CHAIN|SEQUENCE
MVKVKFKYKGEEKEVDTSKIKKVWRVGKMVSFTYDDNGKTGRGAVSEKDAPKELLDMLARAEREKK
GVLKKLRAVENELH
>1WVL:C|PDBID|CHAIN|SEQUENCE
CCTATATAGG
>1WVL:D|PDBID|CHAIN|SEQUENCE
CCTATATAGG"
```

(b)

Figure 6. An illustrating example of PDBID 1WVL. (a) One domain sequence randomly selected from the training dataset. (b) The FASTA sequence of 1WVL.

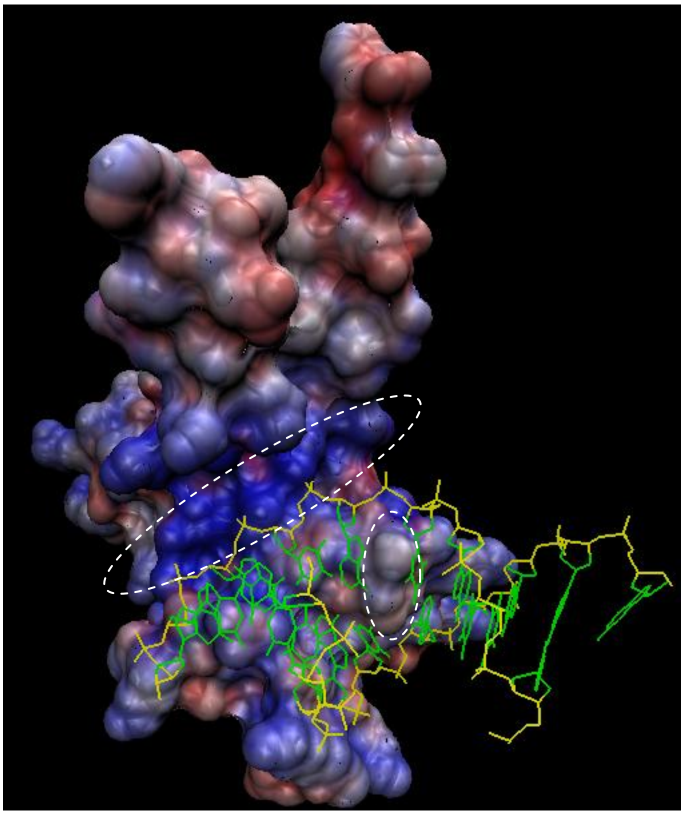


Figure 7. The molecular surface accessible for an electron donor protein. The negatively and positively charged surfaces are shown in red and blue, respectively. The figure was created by using VMD 1.9 [16]. Surface electrostatic potential was calculated by using the APBS tool [15]. The DNA-binding pocket on 1WVL is visible as dark blue, and hydrophobic ridge of residues in the minor groove is visible as gray spheres.

Tables

Table 1. Physicochemical properties (PCPs) in the five identified clusters C_{id} for analyzing DNA-binding domains, obtained from [9]

C_{id}	AAindex ID	PCP	C_{id}	AAindex ID	PCP
7	BHAR880101	Flexibility	10	FASG760105	pK-C
7	BURA740101	Secondary structure	10	JOND750102	pk- (-COOH)
7	CHOC760103	Solvent accessibility	10	RADA880108	Polarity
7	HOPT810101	Hydrophobicity	16	PRAM900101	Hydrophobicity
7	FAUJ880111	Charge	16	FUKS010104	Solvent accessibility
9	KARP850101	Flexibility	16	KUMS000103	Secondary structure
9	PALJ810115	Secondary structure	18	PONP800107	Solvent accessibility
9	ROSM880101	Hydrophobicity	18	GRAR740102	Polarity
9	KUHL950101	Solvent accessibility	18	FASG760104	pK-N
10	ZIMJ680101	Hydrophobicity	18	FAUJ880113	pK-a(RCOOH)
10	EISD860101	Solvent accessibility	18	FAUJ880103	Normalized van der Waals volume
10	GEIM800101	Secondary structure			

Table 2. The performance comparisons between the SVM and fuzzy rule based classifiers. The training dataset and test dataset are DNAset and DNAiset, respectively.

	DNAset			DNAiset
	Accuracy (%)	Feature no.	Rule no.	Accuracy (%)
SVM + PSSMs [6]	86.62	400	NA	82.81
SVM + PCPs [9]	87.12	22	NA	80.73
iFRCs	74.32	1.34	2.0	77.46
EFRBC	NA*	8	60	83.33

* The ensemble classifier EFRBC consisting of 30 iFRCs has no training accuracy

Table 3. Performances of the proposed FRKAS on DNAiset

Accuracy (%)	SEN (%)	SPE (%)	MCC
83.33	82.0	84.8	0.67

Table 4. The eight features used in the 30 iFRCs

No.	Feature ID	AAindex No.	Property
20	H88	FAUJ880111	Positive charge
12	P80	FAUJ880103	normalized Van Der Waals Volume
3	A237	PALJ810115	Secondary structure
2	A97	GEIM800101	Secondary structure
1	H252	PRAM900101	Hydrophobicity
1	H355	ROSM880101	Side chain hydropathy
1	H398	ZIMJ680101	Hydrophobicity
1	H482	KUHL950101	Solvent accessibility

A: Alpha and turn propensities. B: Beta propensity. C: Composition. H: Hydrophobicity. P: Physicochemical properties. O: Other properties [9].

Table 5. The rule-based knowledge of DNA-binding mechanism corresponding to the fuzzy classifiers, shown in Fig. 5

Fuzzy classifier	Rule-based knowledge				
iFRC$_1$	R1-b:	If P80 is ALL and H355 is ALL	then	binding	(CF= 0.161)
	R1-n:	If P80 is {small, medium} and H355 is {small, medium}	then	non-binding	(CF= 0.576)
iFRC$_2$	R2-b:	If H88 is {medium, large}	then	binding	(CF= 0.416)
	R2-n:	If H88 is {small}	then	non-binding	(CF= 0.165)
iFRC$_3$	R3-b:	If P80 is {medium, large}	then	binding	(CF= 0.235)
	R3-n:	If P80 is {small, medium}	then	non-binding	(CF= 0.969)
iFRC$_4$	R4-b:	If P80 is {medium, large} and A97 is ALL	then	binding	(CF= 0.949)
	R4-n:	If P80 is ALL and A97 is {small, medium}	then	non-binding	(CF= 0.718)
iFRC$_5$	R5-b:	If H88 is {medium, large} and A237 is {small, medium}	then	binding	(CF=0.910)
	R5-n:	If H88 is {small} and A237 is ALL	then	non-binding	(CF= 0.286)
iFRC$_6$	R6-b:	If H252 is {medium, large} and H398 is {medium, large}	then	binding	(CF= 0.196)
	R6-n:	If H252 is {small, medium} and H398 is {small, medium}	then	non-binding	(CF= 0.549)
iFRC$_7$	R7-b:	If A237 is {small, medium} and H482 is {medium, large}	then	binding	(CF= 0.867)
	R7-n:	If A237 is ALL and H482 is {small, medium}	then	non-binding	(CF= 0.153)

國科會補助教師出席國際會議結案心得報告

報告人姓名	黃慧玲	所屬學校 學系(所)	交通大學 生物科技學系
會議期間 及地點	2011/12/04 至 2011/12/07 南韓	補助項目 及金額	■ 機票費 ■ 註冊費 □ 生活費
會議名稱	（中文）2011 第二十二屆基因體資訊國際會議 （英文）2011 The 22nd International Conference on Genomic Informatics (GIW 2011)		
發表海報 論文題目	1. 應用於肌肉肝糖分解之新陳代謝路徑建模動態參數的最佳化方法 Optimization approach to estimation of kinetic parameters for modelling metabolic pathways of muscle glycogenolysis 2. 應用於選擇複雜資訊標記SNP的智慧型三目標基因演算法 Intelligent triple-objective genetic algorithm for selecting informative Tag SNPs		

報告內容：(1、參加會議經過；2、與會心得3、建議4、攜回資料)

吾人發表的海報論文是在12/05~12/06二日下午的Hall B會議廳

1. 參加會議經過

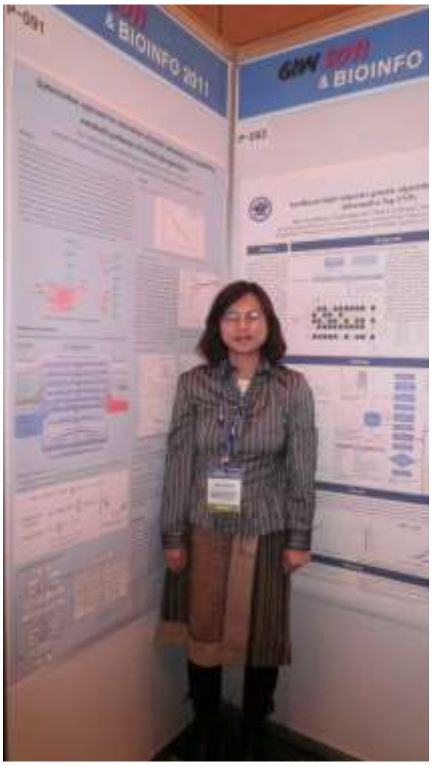
此次星期日由桃園機場搭釜山航空直飛釜山金海國際機場。我由釜山金海國際機場搭飯店公車抵達飯店，釜山溫度大約 5°C~-10°C，看著優美的景象，令疲勞的身體舒解些許。

今年GIW 2011與BIOINFO 2011二會議共同在海雲台大飯店在韓國釜山舉行這次的年度會議。本次會議有GIW與BIOINFO兩個議程可以選擇，而plenary talks是合併在一起舉行的。會議第一天開場是由MIT的David Bartel演講MicroRNAs and Their Regulatory Targets；第二場是由韓國生資中心(KOBIC)的Sanghyuk Lee演講有關Bioinformatics Research and Resources at KOBIC。第二天是西班牙國家癌症研究中心的Alfonso Valencia演講A Bioinformatics perspective of Cancer Personalize Genome Data；另一場是由東京大學Kiyoshi Asai演講Algorithms for RNA sequence analysis。最後一天是由首爾大學的Jeong-Sun Seo演講Genome-wide map of common and rare variants in Asian population using massively parallel DNA and RNA sequencing – Preliminary results from 1000 Asian Genome Project。

我此次發表兩篇poster，分別在會議第一天下午四點與會議第二天晚上六點進行。此次所發表的海報論文名稱為Optimization approach to estimation of kinetic parameters for modelling metabolic pathways of muscle glycogenolysis與Intelligent triple-objective genetic algorithm for selecting informative Tag SNPs。海報論文報告

內容豐富，各國學者彼此交換研究心得，獲得參加學者對吾人研究之正面印象，也讓其他學者多瞭解我們研究的方法與方向。

下圖一為本人至會議報到櫃臺所拍攝的照片；第二張圖片為本人在兩篇所發表之海報前的留影。



2、與會心得

感謝研發處與國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。本次會議GIW2011 為第二十二屆基因體資訊國際會議此次主辦單位是KSBSB (Korean Society for Bioinformatics and Systems Biology)。第一次的 GIW 會議成立于1990 年在東京和大多是在日本舉行了後續會議。2006 年以來，GIW 會面，成為真正的國際約亞洲-太平洋地區的國家每年舉行。現在它是亞洲協會生物資訊學 (AASBi) 的正式會議。今年，韓國社會生物資訊學和系統生物學 (KSBSB) 將 GIW 與

BIOINFO 二會議同時在海雲台大飯店在韓國釜山共同舉行這次的年度會議。今年大會共有五個全體會議講座、 5 小型座談會、 32 接受的論文和超過 120 海報。大會這次邀請五個世界著名的優秀演講者，其中每個人都是在自己的領域的領導者。因為他們會給一系列刺激全體會議講座，讓我們受益良多。透過這次會議將科學交流，和促進了解基因體資訊進展的良好機會。

個人覺得生物資訊這領域，由此次舉辦國韓國，這國家對生物資訊投入組織相當龐大，也可見他們對這領域的企圖心與團結。反觀台灣生物資訊投入與組織結構發展還需更努力。

3、建議

近年來國科會、教育部和學校積極鼓勵年輕研究人員，除鼓勵教師參與會議外，特別是博士班學生，參與大型國際會議，及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加生物資訊國際會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此交換研究心得，更可找到跨領域的學者國際合作，在跨領域的生物資訊研究更是重要。目前研究生已有多管道獲(部份)補助出席國際會議，建議繼續擴大進行。而國際化的學術交流是往後的趨勢，也能有所激勵國人學界能力與國際觀。

4、攜回資料

1. 期刊（電子檔安裝於贈送的隨身碟中）
2. 隨身袋一只。

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

報告人姓名	黃慧玲	所屬學校 學系(所)	交通大學 生物科技學系
會議期間 及地點	2012/5/17 至 2012/5/20 中國上海	補助項目 及金額	■ 機票費 □ 註冊費 ■ 生活費
會議名稱	（中文）2012 年第六屆生物資訊與生醫工程國際會議 （英文）The 2012 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2012)		
發表論 文題目	（中文）1.使用物化特性方法設計螢光蛋白質的預測器 2.使用計分卡設計醣結合蛋白質的預測器 （英文）1.Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method 2.Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method		

報告內容：(1、參加會議經過；2、與會心得；3、建議；4、攜回資料)

報告內容應包括下列各項：

一、　參加經過

近年來生物資訊愈來愈受重視，也越來越多專注在生物資訊的研討會如雨後春筍般在舉行。而本次參加的國際會議為2012年第六屆生物資訊與生醫工程國際會議(iCBBE 2012)，由國際電子電機學會IEEE Engineering in Medicine and Biology Society贊助舉辦。

我所發表之論文題目是「Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method」，主要是使用計算智慧的最佳化技術來從531個物化特性中挑選出一組最佳的小集合，結合support vector machine分類器，設計出螢光蛋白質的預測器，並分析這些蛋白質序列的物化特性在螢光發光所扮演的功能與腳色。另外一篇是「Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method」，主要是使用由智慧型基因演算法（Intelligent Gene Algorithm）從原始蛋白質序列計算相鄰氨基酸（di-peptide）的權重找出最佳化的計分卡（scoring card），設計出醣結合蛋白質的預測器，並分析這些蛋白質序列的物化特性在醣結合所扮演的功能與腳色。觀察發展的趨勢而言，生物資訊與生醫工程國際會議越來越受重視，而計算智慧是計算生物的重要技術。

今年大會共有4個Room，16個Section，包含口頭報告區、演講區、與海報展示。本次有四個重要的演講，包括（一）法國國家科學研究院Athel教授演講Metabolic modeling: A Necessary Tool for Biotechnology；（二)美國康乃爾大學Ann教授演講Fenton Oxidation of Contaminants using Nanomagnetite；(三）美國哈佛醫學院Chou教授演講有關An NMR view of membrane transporters: application to mitochondrial carriers；(四）以色列理工學院Daniel教授演講A general overview of medical robotics，令參與者可更深入瞭解此項領域的重大研究發展趨勢,提高了大家對這方面研究的瞭解。很榮幸我們的論文被接受口頭報告。此科學論壇，為助長生物資訊與生醫工程很重視分析工具、奈米元件與生醫應用的發展目前已經與眾多學術和科學界的領導組織共同合作，合作學術單位遍及義大利、美國、臺灣、中國、以色列、泰國、日本、摩洛哥、沙烏地阿拉伯、巴基斯坦、俄國、捷克、伊朗、德國、土耳其、馬來西亞、波蘭、愛爾蘭、英國、法國、卡達、加拿大、印尼、西班牙……等。本次投稿被iCBBE 2012接受之國際會議論文亦有機會被轉投至生物資訊與生醫工程國際期刊。本次會議參加的人員來自許多國家，包含大陸、美國、日本、澳洲、印度、新加坡、香港、韓國、台灣、馬來西亞、泰國……等等，其中中國學者人數明顯增加，大會安排讓來自各個國家的學者互相交流、聯誼，促進了與會學者日後的學術交流機會。其間大會於第二天安排午宴讓來自各個國家的學者互相交流、聯誼，期望能促進與會學者日後學術交流的機會。

我的口頭報告時段是19日早上8:30~12:00，我們這會場的發表主要是蛋白質和蛋白質體在生物資訊的發展，因此大部分都會使用機器學習法來進一步分析討論。我的論文報告主題是使用基因演算法的最佳化技術從531個物化特性挑選出一組最佳解，結合support vector machine分類器，設計出螢光蛋白質的預測器並進一步分析螢光蛋白質序列的物化特性在螢光發光所扮演的功能。

第二篇論文由共同作者李華錦博士後研究員上台發表，而我在台下參與全程。主題是使用計分卡（scoring card）設計醣結合蛋白質的預測器，介紹此計分卡由智慧型基因演算法算出最佳化的一份計分卡，因此，只要經過此簡單之計分卡即可以達到以往使用繁複計算方法才能達到的準確度，大幅降低一般研究生物資訊計算所需的複雜度。

由於生物資訊研究需求整合資訊、數學與生物領域的廣泛知識，可研究的題材也因此相當廣泛並且要有能整合跨領域知識的創意，所以聆聽來自眾多不同地區的學者的研究成果，可有效率地吸收新知。

圖一：論文發表中

圖二：論文發表後合影（左起黃文玲教授、徐禮燊教授、李華錦博士後研究員、何信瑩教授、何信瑋教授與我）

2、與會心得

感謝國科會補助參加國際會議之出國補助，使本人得以出席跨領域生物資訊國際會議，開拓眼界及促進國際觀。每次參加國計會議除了努力讓世界知道臺灣人在研究方面非常認真與相當有能力為心則。我對於本次會議重點 – 生物資訊與生醫工程特別感興趣，由其對很多研究主題及發展方向也關心，希望對提高台灣的學術聲望及研究能量提升有所貢獻。

關於跨領域的交流，由此次國際會議舉辦的大陸上海，他們對跨領域的企圖心及投入組織相當龐大，也可見他們對這領域的企圖心與團結。台灣對跨領域的投入還需要更加努力，我們可以從他們的經驗作為參考與借鏡，來促進國內的跨領域和產學合作的發展。

3、建議

近年來國科會、教育部和學校積極鼓勵年輕研究員，不僅是教師，還有特別是博士班學生，參與大型國際會議。及早進入研究領域的核心，吸取國際研究經驗，以提高國人的研究水準。參加國際生物資訊或是生醫工程相關會議對老師及學生是非常重要的，會議中不但可以得到相關研究的最新發展資訊，認識結交許多相關領域的學者，彼此做研究與心得交換，更可找到跨領域的國際合作夥伴，在跨領域的生物資訊研究更為重要。希望國家與學校單位能多在補助年輕學者出國，並且繼續擴大進行，以提升國內研究的品質。

4、攜回資料

Proceedings of the 2012 6th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2012), Shanghai, China, May 17-20, 2012. (含紙本與光碟)

Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method

Hua-Chin Lee[*1], Yi-Fan Liou[*1], Phasit Charoenkwan[*1], Shinn-Jang Ho[*2], Li-Sun Shu[*3], Shinn-Ying Ho[*1], and Hui-Ling Huang[*1*4]

Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan[*1]
Department of Automation Engineering, National Formosa University, Yunlin, Taiwan[*2]
Department of Information Management, Overseas Chinese University, Taichung , Taiwan[*3]
Corresponding author: 886-3-571-2121, ext: 56910; e-mail: hlhuang @mail.nctu.edu.tw [*4]

***Abstract*—Carbohydrate-binding proteins play a pivotal role in a variety of important biological recognition processes. Compared with most studies of predicting binding sites, very few studies investigate prediction of carbohydrate-binding proteins. This paper proposes a highly interpretable scoring card method for predicting carbohydrate-binding proteins. First, a large-scale data set of carbohydrate-binding proteins (CBPDB) collected from three up-to-date databases, CAZy, CGF and Swiss-Prot is utilized. The data set CBPDB consists of 2380 positive and negative proteins with sequence identity 25% by removing sequence redundancy. Secondly, we adopt a novel scoring card method by way of generating an optimized scoring card of dipeptides to predict the carbohydrate-binding proteins. The prediction performance is promising with an independent test accuracy of 78.67%. The dipeptide score is helpful in discovering motif. The scoring card provides an insight of further analyzing the binding mechanism between carbohydrate and proteins.**

Keywords—*Binding, carbohydrate, genetic algorithm, scoring card, protein, prediction.*

I. Introduction

Carbohydrates play an essential role in a variety of important biological recognition processes like infection, immune response, cell differentiation, and neuronal development. All of these biological phenomena may be regulated by the interaction of these carbohydrates with proteins [1-5]. Carbohydrate-binding proteins are becoming extremely useful in curing various illnesses. Experimental work for identifying carbohydrate-binding proteins is costly and time consuming. Therefore, effective computational methods for predicting carbohydrate-binding proteins are desirable.

It is vitally important to develop an automated and efficient method for timely identification of novel carbohydrate-binding proteins. However, some researches of using empirical rules [6] or machine learning methods [7] mainly focused on prediction and analysis of carbohydrate-binding sites of proteins that are already known as carbohydrate-binding proteins.

Someya et al. [8] first clarified the definition carbohydrate-binding proteins and then constructed positive and negative datasets. Using both informative features and an appropriate classifier is essential to design an effective method for predicting carbohydrate-binding proteins based on the primary sequence only. They developed a carbohydrate-binding protein prediction system by using support vector machines (SVMs) [9], where the prediction of carbohydrate-binding proteins was formulated as a binary classification problem. Someya et al. [8] trained the SVM with three different encoding methods: a direct encoding method (AA-20), and two grouping methods (Levitt-6 and Someya-7). The SVM-based method with AA-20 performs well with a leave-one-out accuracy of 87% for the sequence with a sequence identity 35%.

Kumar et al. [10] developed SVM modules for distinguishing between cancer and non-cancerlectin proteins by using dipeptide composition, split composition, position specific scoring matrix (PSSM) profiles and PSSM with 14 PROSITE domains as input features.

The merits of this study are to 1) utilize a large-scale data set of carbohydrate-binding proteins collected from three up-to-date databases, CAZy, CGF and Swiss-Prot, 2) propose an easily interpretable method rather than the black-box-like SVM for biologists, and 3) obtain a robust performance with accuracy of 78.67% on carbohydrate-binding proteins with sequence identity 25%.

II. Materials

A. Datasets

Carbohydrate binding proteins are obtained from the Consortium for Functional Glycomics (CFG) database and Carbohydrate Active Enzyme (CAZy) database. All the records from these two databases are served as positive protein samples which can bind carbohydrate. The Gene Ontology (GO) annotation terms about carbohydrate-binding functions are obtained from the GOA database. The GO term, carbohydrate binding function, and its child terms are collected. Finally, the number of GO terms which are defined as carbohydrate binding function is 778.

To obtain the negative dataset consisting of non-carbohydrate binding proteins, the Swiss-Prot database of release 2011_06, is also used. The Swiss-Prot database is divided into positive and negative datasets using GO terms mentioned above. If the GO terms of polypeptides in Swiss-Prot contain any GO terms defined as carbohydrate binding protein, the sequences are classified as positive samples. All sequences obtained from the three databases, CAZy, CGF and Swiss-Prot. The sequences from CAZy and CFG are carbohydrate-binding protein. Sequences from Swiss-Prot consist of carbohydrate- and non- carbohydrate-binding proteins.

The positive dataset is composed of 57330 polypeptides while the negative dataset is composed of 405046 polypeptides. USEARCH [11] is used to remove the sequence redundancy of the dataset. The threshold of USEARCH is set to 25%. After treating with USEARCH, the positive dataset contains 2380 polypeptides and the negative dataset contains 49647 polypeptides.

To avoid the unbalanced problem, the previous method [12] is used. The size of the negative dataset randomly chosen is equal to that of the positive dataset. Finally, the completed dataset contains 2380 polypeptides, shown in Table 1. The dataset is equally divided into the training and test datasets.

Table 1. The numbers of sequences in the dataset CBPDB

stage	positive	negative
Original	57330	405046
Identity threshold 25%	2380	49647
Chosen negative sequences	2380	2380
Final dataset CBPDB	1190	1190

III. Methods

A novel scoring card method for predicting carbohydrate-binding proteins is proposed.

A. Construction of a scoring card

Figure 1 shows the data structures and experimental flow chart of the proposed scoring card method. The scoring card in arrow figure stands for the average of the statistic scoring cards.

The CBPDB dataset was equally divided into two parts, one for training and the other for independent test (outer loop in Fig. 1). Furthermore, the training data set is split into ten parts randomly (inner loop in Fig. 1) for ten-fold cross-validation. Therefore, the method can obtain ten validation results and ten statistic scoring cards. The statistic procedure of the scoring card method is described as follows:

1) Separate the data set into two classes of carbohydrate- and non-carbohydrate-binding proteins and calculate 400 dipeptide amounts for each class.
2) Due to the variance of sequence lengths in the two classes, the number of each dipeptide in one certain class is divided by the total number of dipeptides in this class.
3) A dipeptide in the carbohydrate-binding class got +1 score; otherwise, a dipeptide in the non-carbohydrate-binding class got -1 score. So the 400 scores in a scoring card can be derived from summation of all the dipeptide scores in two classes.
4) Normalized the scores as positive numbers into the range from 0 to 1000 in the scoring card.

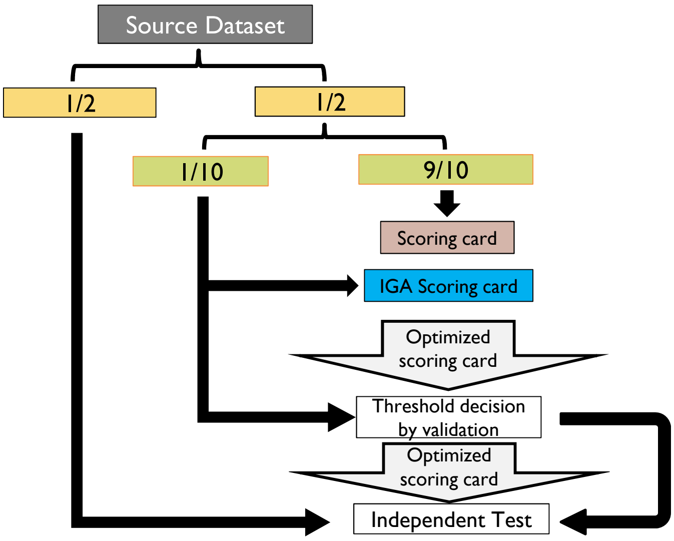


Fig. 1 Outline of the scoring card method

B. Threshold value determination

In order to find a best threshold to assort the data in two classes, the validation data in the inner loop were used. The amounts of 400 dipeptides of all samples in the validation data set were counted, and then the amounts were multiplied by the counterparts of dipeptide number in the scoring card. Finally, sum the 400 numbers up and the summation is divided by the sequence length to obtain a score of a sequence. The threshold with the highest accuracy in validation is chosen as the threshold value to classify two classes in the independent test.

In the test, the amounts of 400 dipeptides in a sample were multiplied by the counterparts of dipeptide number in the scoring card. Sum the 400 numbers up and the summation is divided by the sequence length of a protein to obtain a score for one sample. Equation 1 shows the calculation process. In Eq. 1, i and j = A, C, ..., Y, the 20 amino acids, W is the weighting of dipeptides of the test sample, and S is the score in the scoring card, and L is protein length. Fig.1 illustrates the process of test sample multiplied by the counterparts of dipeptide number in the scoring card. Then the query protein can be classified according to the threshold determined in the validation step.

$$\text{Sample score } = \frac{\sum_{i,j=1}^{20} W_{ij} \cdot S_{ij}}{L} \qquad (1)$$

C. *Scoring card*

The ten statistics scoring cards in inner loop were averaged to one scoring card (the arrow illustration in Fig.1). Afterward the average scoring card was evaluated by ten validation datasets. Ten best thresholds would be derived from ten validations, and then the average of the ten thresholds was used in the classification of independent test. In the final step, there only had one scoring card, one threshold and one test result.

D. *IGA-scoring card*

The scoring card is further optimized by an intelligent genetic algorithm (IGA) [13] [14]. IGA utilizes an orthogonal array (OA) [15] that was used in the crossover operation to bring the better children than the traditional crossover method, and it can efficiently obtain a high-quality solution set. Orthogonal array is a fractional factorial array, which assures a balanced comparison of levels of any factor.

Ten validation data sets were used for evaluating the fitness function. The algorithm of generating the IGA-scoring card is described as follows:

Step 1: Initial population

The half of initial population in the IGA-scoring card consists of the ten statistic scoring cards in an inner loop, and the other ten individuals in the initial population were randomly generated from scores 0 to 1000. Therefore, the initial population comprises Npop individuals totally and Npop＝20.

Step 2: Evaluation

The fitness function of every individual was appraised via AUC from TPR and FPR.

$$\mathrm{TPR} = \mathrm{TP}/\ (\mathrm{TP}+\mathrm{FN}) \qquad (2)$$

$$\mathrm{FPR} = \mathrm{FP}/\ (\mathrm{FP}+\mathrm{TN}) \qquad (3)$$

Where TP is true positive and FP is false positive. In the ROC curve, X axle is FPR and Y axle is TPR. The validation data were used to find the best threshold according to the highest accuracy and it can get both TPR and FPR values from each threshold. TPR and FPR were used to draw the ROC curve and evaluate the fitness function for every individual.

Step 3: Selection

Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4: Crossover

Select Pc•Npop parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents where Pc is the crossover probability. Pc＝0.9.

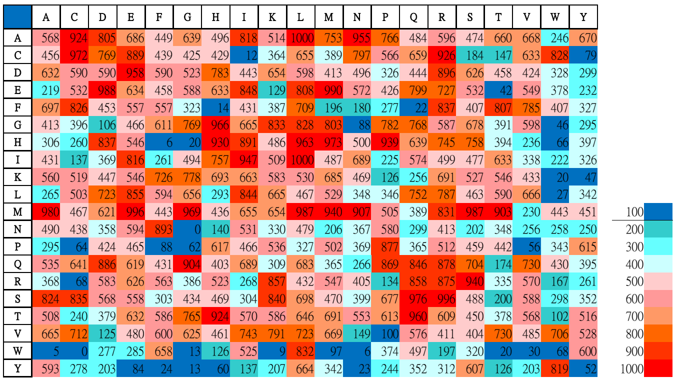
Step 5: Mutation

Apply the real number mutation operator to the randomly mutated the gene from 0 to 1000 if the generated digit < Pm, where Pm is the mutation probability. Pm＝0.01.

Step 6: Termination

When the IGA come to 100 generations is the stop condition and output the best individual as IGA-scoring card. Otherwise, go to Step 2.

This flow chart is divided into training and test processes. In the training part of the beginning, it used the training dataset to build a scoring card, then the scoring card is optimized by IGA with the validation dataset to get the best scoring card and threshold value. In the test process, the input sample could obtain a score through the scoring card.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	568	924	805	686	449	639	496	818	514	1000	753	955	766	484	596	474	660	668	246	670
C	456	972	769	889	439	425	429	12	364	655	389	797	566	659	926	184	147	633	828	79
D	632	590	590	958	590	523	783	443	654	598	413	496	326	444	896	626	458	424	328	299
E	219	532	988	634	458	588	633	848	129	808	990	572	426	799	727	532	42	549	378	232
F	697	826	453	557	557	323	14	431	387	709	196	180	277	22	837	407	807	785	407	327
G	413	396	106	466	611	769	966	665	833	828	803	88	782	768	587	678	391	598	46	295
H	306	260	837	546	6	20	930	891	486	963	973	500	939	639	745	758	394	236	66	397
I	431	137	369	816	261	494	757	947	509	1000	487	689	225	574	499	477	633	338	222	326
K	560	519	447	546	726	778	693	663	583	530	685	469	126	256	691	527	546	433	20	47
L	265	503	723	855	594	656	293	844	665	467	529	348	346	752	787	463	590	666	27	342
M	980	467	621	996	443	969	436	655	654	987	940	907	505	389	831	987	903	230	443	451
N	490	438	358	594	893	0	140	531	330	479	206	367	580	299	413	202	348	256	258	250
P	295	64	424	465	88	62	617	466	536	327	502	369	877	365	512	459	442	56	343	615
Q	535	641	886	619	431	904	403	689	309	683	365	266	869	846	878	704	174	730	430	395
R	368	68	583	626	563	386	523	268	857	432	547	405	134	858	875	940	335	570	167	261
S	824	835	568	558	303	434	469	304	840	698	470	399	677	976	996	488	200	588	298	352
T	508	240	379	632	586	765	924	570	586	646	691	553	613	960	609	450	378	568	102	516
V	665	712	125	480	600	625	461	743	791	723	669	149	100	576	411	404	730	485	706	528
W	5	0	277	285	658	13	126	525	9	832	97	6	374	497	197	320	20	30	68	600
Y	593	278	203	84	24	13	60	137	207	664	342	23	244	352	312	607	126	203	819	52

Fig. 2 The heat map of the IGA-scoring card

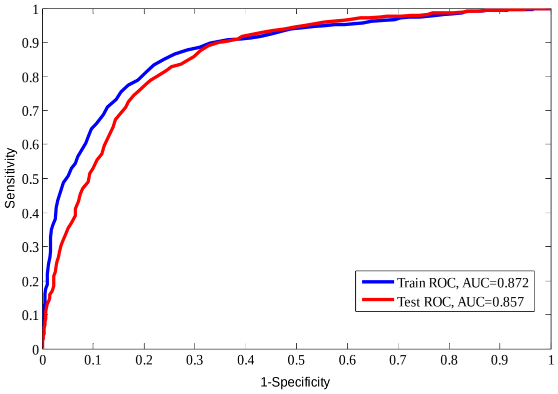


Fig. 3 The training and test ROC curves

IV. Results

A. *Scoring card results*

Figure 2 shows the best IGA-scoring card (with the highest training accuracy) of 150 generation optimizations from 25 independent runs. Figure 3 depicts the training and test ROC curves of the number-24 scoring card method. From the curves, the AUC of training and test are 0.872 and 0.857 and the accuracies of the training and independent test are 80.67% and 78.67%, respectively.

The heat map reveals that the dipeptides have widely-distributed binding abilities. The dipeptides LA and LI have the highest score 1000 and the dipeptides CW and GN have the lowest score 0. From the support of knowledge in the scoring card, the motif discovering objective can be achieved more easily.

By using the scoring card optimized by IGA, the AUC can be significantly improved and hence increases the training and test accuracies. The high performance arises mainly from IGA and it can efficiently obtain a high-quality solution set. So the performance in the scoring card method optimized by IGA can get high performance.

V. CONCLUSIONS

The proposed method provides a more easily and intuitive way to predict the carbohydrate-binding protein than any other method, like SVM. Moreover, the scoring card method which utilizes the 400 scores as weights is derived from the protein sequence of dipeptide, and further efficiently optimized by the intelligent genetic algorithm.

It can efficiently analyze the dipeptide feature through the interpretable score feature in scoring card, and directly predict the class in the problem which influenced by dipeptide or protein sequence. Hence, the window threshold function in scoring card method gives the advantage to select the samples with stable prediction accuracy and also provides stable independent test experiments.

REFERENCES

[1] Calvin F. ROFF, Paul R. ROSEVEAR, John L. WANG and Robert BARKER, Identification of carbohydrate-binding proteins from mouse and human fibroblasts, Biochem. J. (1983) 211, 625-629

[2] Wei-Yao Chou, Wei-I Chou, Tun-Wen Pai, Shu-Chuan Lin, Ting-Ying Jiang, Chuan-Yi Tang and Margaret Dah-Tsyr Chang, Feature-incorporated alignment based ligand-binding residue prediction for carbohydrate-binding modules, Bioinformatics, Vol. 26 no. 8 2010, pages 1022–1028

[3] Alisdair B. BORASTON, David N. BOLAM, Harry J. GILBERT and Gideon J. DAVIES, Carbohydrate-binding modules: fine-tuning polysaccharide recognition, Biochem. J. (2004) 382, 769–781

[4] H. Hashimoto, Recent structural studies of carbohydrate-binding modules, Cellular and Molecular Life Sciences, 63 (2006) 2954–2967

[5] A. Malik, A. Firoz, V. Jha, and S. Ahmad. PROCARB: A database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools. Advances in Bioinformatics. Vol. 2010, Article ID 436036, 9 pages.

[6] C. Shionyu-Mitsuyama, T. Shirai, H. Ishida, and T. Yamane. An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. Protein Engineering, 2003, vol. 16, no. 7, pp. 467–478.

[7] A. Malik and S. Ahmad. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Structural Biology, 2007, vol. 7, article 1.

[8] Seizi Someya et. al.. Prediction of carbohydrate-binding proteins from sequences using support vector machines. Advances in Bioinformatics. Vol. 2010, Article ID 289301, 9 pages.

[9] C. C. Chang, and, C. J. Lin (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[10] R. Kumar, B. Panwar, J. S Chauhan and G. PS Raghava. Analysis and prediction of cancerlectins using evolutionary and domain information. BMC Research Notes 2011, 4:237.

[11] Xiaojing Yua, Jianping Caob, Yudong Caic, Tieliu Shia and Yixue Lia. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. Journal of Theretical Biology, 240 (2006), pp.175-184

[12] Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. Bioinformatics Advance Access published August 12, 2010

[13] S.-Y. Ho, L.-S. Shu, and J.H. Chen, Intelligent evolutionary algorithms for large parameter optimization problems. Ieee Transactions on Evolutionary Computation, 2004. 8(6): p. 522-541.

[14] S. Y. Ho, J. H. Chen, and M. H. Huang, Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans Syst Man Cybern B Cybern, 2004. 34(1): p. 609-20.

[15] C.-W. Tung and S.-Y. Ho, POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics, 2007. 23(8): p. 942-9.

Designing predictors of bioluminescence proteins using an efficient physicochemical property mining method

Hui-Ling Huang*1, Yi-Fan Liou*1, Hua-Chin Lee*1, Wen-Lin Huang*2, and Shinn-Ying Ho*1,*3
Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan*1
Department of Management Information System, Asia Pacific Institute of Creativity, Miaoli, Taiwan*2
Corresponding author: 886-3-571-2121, ext: 56909; e-mail: syho @mail.nctu.edu.tw*3

***Abstract*— Bioluminescence proteins are becoming increasingly important in a variety of research fields such as in situ imaging and the study of protein-protein interactions in vivo, and increased spectral variety of bioluminescent reporters is needed for further progress. The existing method BLProt using support vector machine (SVM) and physicochemical properties to predict bioluminescence proteins. The BLProt method identified the most prominent features using various filter approaches, ReliefF, infogain, and mRMR. BLProt utilized 100 features to achieve a training accuracy of 80% and test accuracy of 80.06%. Physicochemical properties are well recognized to be effective in designing various predictors for understanding the functions and characteristics of proteins. In this study, we propose an efficient method for designing predictors of bioluminescence proteins using a small set of informative physicochemical properties obtained by using an inheritable bi-objective genetic algorithm. The benchmark datasets were used to evaluate the proposed method using SVM and informative physicochemical properties as the features. The prediction accuracy of independent test is 81.79% using 15 properties. From the analysis of informative physicochemical properties, some knowledge of bioluminescent problems can be revealed. The proposed physicochemical property mining method can be used conveniently as the core for designing predictors for various types of bioluminescent problems.**

Keywords* — Bioluminescent protein, genetic algorithm, SVM physicochemical properties, prediction

I. Introduction

Bioluminescence is a light producing process. The basic two factors included in this process are luciferase and luciferin, which are the catalytic enzyme and its substrate respectively. Work on bioluminescence is actively pursued at all levels, such as naturalist or phtochemist, due to it abnormal characters. The visible light, generated from luciferase, is emitted at room temperature while light often can be generated at extreme high temperature causing violent oxidation of some objects. The actual emission of bioluminescence is the extremely rapid final process of usually multistep reaction. Most often, the excited state of luciferin is excited by electron or photon [1].

Bioluminescence provides an ideal tool to solve scientific problems. Previous studies [2] are already renowned for the preparation and application of an extended series of radiometric ion-sensitive indicators and a number of sophisticated reporter molecules based on fluorescence resonance energy transfer (FRET). In order to generate genetically encoded FRET probes which are suitable for radiometric measurements, more fluorophores are need to be discovered or generated.

However, the biofunction of those bioluminescence proteins are quite alike, they do not share strongly homologous. Many orgasms use different proteins which have different mechanisms to generate light [3]. Bioluminescence proteins are becoming increasingly important in a variety of research fields such as in situ imaging and the study of protein-protein interactions in vivo, and increased spectral variety of bioluminescent reporters is needed for further progress.

Beside the bioluminescent characters, some characters are also interesting. First, the luciferins are extremely hydrophobic macro molecules. To catalyze the molecules, the catalytic sites must be very different to tune the catalytic orientation between the enzymes and subtracts. Secondly, the bioluminescence light in some live orgasms, like firefly, is regulated. The GFP does not have a significant regulation structure like the C-terminal ball-chain structure of voltage-dependent gate channel on neuron. But some regulation mechanisms still occur for this purpose [4]. Third, the bioluminescence does not share homologous but they have a similar function. Understanding physicochemical properties of the bioluminescence proteins may help improve the applications of bioluminescence proteins.

Kandaswamy et al. [5] proposed an accurate prediction method BLProt that uses a support vector machine (SVM) and physicochemical properties to predict bioluminescence proteins. BLProt used a training dataset consisting of 300 bioluminescence proteins and 300 non-bioluminescence proteins, and an independent test dataset consisting of 141 bioluminescence proteins and 18202 non-bioluminescence proteins. To identify the most prominent features, they carried out feature selection with three different filter approaches, ReliefF, infogain, and mRMR. For the aim of designing accurate prediction methods, the major concern is to identify feature vectors with high discrimination abilities for classifying positive and negative samples. Their feature selection method suffers from a large set of candidate features.

We investigate the optimal design of predictors for bioluminescence proteins from amino acid sequences using both informative features and an appropriate classifier. Furthermore, we obtain a set of informative physicochemical properties which can advance prediction performance. Physicochemical properties extracted from protein sequences were utilized as effective features in recent years. Our previous work Auto-IDPCPs [6] is an SVM based classifier with automatic feature selection from a large set of physicochemical composition features to predict DNA-binding domain/protein. The POPI method used physicochemical properties as efficient features to predict peptide immunogenicity [7]. The prediction method UbiPred [8] mined informative physicochemical properties from protein sequences to identify promising ubiquitylation sites.

The informative physicochemical properties of amino acids indices selected in this study were used as features in designing SVM classifiers. An efficient algorithm inheritable bi-objective genetic algorithm (IBCGA) was used to select significant features which could discriminate the two classes of proteins. The feature sets selected by IBCGA were analyzed carefully to reveal the fundamental differences existed between bioluminescence proteins and non-bioluminescence proteins. In conclusion, we proposed a novel prediction method combining the informative physicochemical properties of amino acid and SVM to solve the prediction problem of bioluminescence proteins.

II. METHOD

We propose a novel method using the physicochemical properties for predicting bioluminescence proteins (PBLP). The identification of an effective feature set of physicochemical properties is mainly derived by using an inheritable bi-objective genetic algorithm (IBCGA) [9]. The IBCGA mines informative physicochemical properties and tune parameter settings of SVM simultaneously while maximizing 5-fold cross validation (5-CV) accuracy.

A. Datasets

The bioluminescence proteins (BLPs) extracted from Martinetz et al. Pfam database are used to obtain the seed proteins of BLPs. To enrich the dataset, PSI-BLAST with stringent threshold (E value 0.01) is carried out to search against the non-redundant sequence database. Then, CD-hit are performed to remove the sequences with identity >= 40% in the collected dataset. After all, a total 441 bioluminescence proteins are kept as positive dataset. The statistic of the training and test sets is shown in Table 1.

There are 300 BLPs randomly selected from the 441 positive samples and are served as training samples. The others are served as test samples. There are 300 non-BLPs also randomly picked from seed proteins of Pfam protein families. These proteins, served as negative samples, are unrelated to BLPs.

The negative testing dataset is composed of the seed proteins of non-BLPs Pfam protein families. All sequences contained in the training dataset have less than 40 residues are removed. Finally, the test dataset is composed of 141 BLPs and 18202 non-BLPs.

Table 1. The statistic of the training/test sets.

dataset	*Number of BLPs*	*Number of non-BLPs*
Training	300	300
Test	141	18202

B. Support Vector Machine

Support vector machine (SVM) is a learning model dealing with binary classification problems. SVM constructs a binary classifier by finding a hyperplane to separate two classes with a maximal distance between margins of two classes consisting of support vectors. In order to make linear separation of samples easier, SVM uses one of various kernel functions to transform the samples into a high-dimensional search space. In this work, the commonly-used radial basis function is applied to nonlinearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|), \gamma > 0 \quad . \qquad (1)$$

The kernel parameter *γ* determines how the samples are transformed into a high-dimensional search space. The cost parameter *C*>0 of SVM adjusts the penalty of total error. These two parameters *C* and *γ* must be tuned to get the best prediction performance.

For multi-class classification problems, 'one-against-one' strategy is applied to transform the multi-class problem into several binary classification problems. Given *h* classes, there are *h*(*h*−1)/2 classifiers constructed and each one trains the samples from two classes. A voting strategy is applied to give a final prediction for test samples. In this study, *h*=2 and the used SVM is obtained from LIBSVM package version 2.81 [10].

C. Inheritable Bi-objective Genetic Algorithm

Selecting a minimal number of informative features while maximizing prediction accuracy is a bi-objective 0/1 combinatorial optimization problem. An efficient inheritable bi-objective genetic algorithm [11] is utilized to solve this optimization problem. IBCGA consists of an intelligent genetic algorithm [12] with an inheritable mechanism. The intelligent genetic algorithm uses a divide-and-conquer strategy and an orthogonal array crossover to efficiently solve large-scale parameter optimization problems. In this study, the intelligent genetic algorithm can efficiently explore and exploit the search space of C(*n*, *r*). IBCGA can efficiently search the space of C(*n*, *r* ± 1) by inheriting a good solution in the space of C(*n*, *r*) [11]. Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run where *r* is specified in an interesting range such as [5, 20].

The proposed chromosome encoding scheme of IBCGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters, where the gene and chromosome are commonly-used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this paper. The GA-chromosome consists of n=531 binary GA-genes b_i for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If b_i=0, the i^{th} property is excluded from the SVM classifier; otherwise, the i^{th} property is included. This encoding method maps the 16 values of γ and C into $\{2^{-7}, 2^{-6} \ldots, 2^{8}\}$.

The feature vector for training the SVM classifier is obtained from decoding a GA-chromosome using the following steps. Consider a given DNA-PBs sequence. At first, the index vectors for all selected physicochemical properties are constructed from AAindex for each amino acid. Feature vector of a peptide consists of the selected features whose values are obtained by averaging the values in their corresponding index vectors. Finally, all values of the feature vectors are normalized into [-1, 1] for applying SVM.

Fitness function is the only guide for IBCGA to obtain desirable solutions. The fitness function of IBCGA is the 5-CV overall accuracy. IBCGA with the fitness function f(X) can simultaneously obtain a set of solutions, X_r, where $r=r_{start}$, $r_{start}+1$, …, r_{end} in a single run. The algorithm of IBCGA with the given values r_{start} and r_{end} is described as follows:

Step 1) (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary GA-genes have r 1’s and n-r 0’s where $r = r_{start}$.

Step 2) (Evaluation) Evaluate the fitness values of all individuals using f(X).

Step 3) (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4) (Crossover) Select $p_c \cdot N_{pop}$ parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents where p_c is the crossover probability.

Step 5) (Mutation) Apply the swap mutation operator to the randomly selected $p_m \cdot N_{pop}$ individuals in the new population where p_m is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6) (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r. Otherwise, go to Step 2). In this study, the stopping condition is to perform 40 generations.

Step 7) (Inheritance) If $r < r_{end}$, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

D. *Prediction Method PBLP*

The selected m physicochemical properties and the associated parameter set of SVM by using PBPL are used to implement the computational system and analyze the physicochemical properties to further understand the BLPs. Since the PBPL is a non-deterministic method, it should make more effort to identify an efficient and robust feature set of informative physicochemical properties in five aspects. The procedure is as the following steps:

Step 1 : We prepare the independent data sets where each set is used as the training data set of 5-CV.

Step 2 : PBPL is performed R independent runs for each of independent data sets. In this study, R = 30. There are total 30 sets of m physicochemical properties for each of independent data sets.

Step 3 : Choose the set of selected physicochemical properties with a maximal accuracy.

PBLDs will automatically determine a set of informative physicochemical properties and an SVM-model for prediction bioluminescent and non- bioluminescence proteins.

III. Results

A. *Results of training and test datasets*

The training data sets contain 300 positive and 300 negative samples. The sequence similarity of the training data set is smaller than 40%. We performed 30 independent runs of PBPL to select robust feature set which could improve the performance of SVM classifier on discriminating the two classes of proteins. The highest training accuracy of 30 PBPL runs was 84.11% and its corresponding test accuracy was 81.79%. (Table 2).

Table 2. Results of the training and independent test by BLProt and PBLP.

	Method	Specificity (%)	Sensitivity (%)	Accuracy (%)	Feature subset
Train	**BLProt**	84.21	74.47	80.00	100
	PBLP	79.25	84.11	84.5	15
Test	**BLProt**	74.47	84.21	80.06	100
	PBLP	81.89	68.79	81.79	15

B. *Selected a small set of physicochemical properties.*

The quantified effectiveness of individual physicochemical properties on prediction is useful to characterize the PBLP mechanism by physicochemical properties. Orthogonal experimental design with factor analysis can be used to estimate the individual effects of physicochemical properties according to the value of main effect difference (MED) [7, 12]. The property with the largest value of MED is the most effective in predicting BLPs.

According to MED, the 15 informative properties are ranked and their descriptions are shown in Table 3 and Fig. 1. The most effective property with MED=16.16668 is RACS820111 denoting “Differential geometry and polymer conformation. Conformational and nucleation properties of individual amino acids”.

Table 3. The highest accuracy with selected m = 15 feature set

Feature ID	AAindex ID	Description
8	BHAR880101	Positional flexibilities of amino acid residues in globular proteins
13	BROC820102	The isolation of peptides by high-performance liquid chromatography using predicted elution positions
18	BUNA790103	1H-nmr parameters of the common amino acid residues measured in aqueous solutions of the linear tetrapeptides H-Gly-Gly-X-L-Ala-OH
95	FINA910104	Physical reasons for secondary structure stability: alpha-helices in short peptides
107	GEIM800111	Amino acid preferences for secondary structure vary with protein class
202	NAKH920101	The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins
223	PALJ810101	Protein secondary structure
310	RACS820111	Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids
380	VENT840101	Hydrophobicity parameters and the bitter taste of L-amino acids
439	PARS000102	Protein thermal stability: insights from atomic displacement parameters (B values)
473	MITS020101	Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces
475	TSAJ990102	The packing density in proteins: standard radii and volumes
489	PUNT030101	A knowledge-based scale for amino acid membrane propensity
491	GEOR030101	An analysis of protein domain linkers: their classification and role in protein folding
502	ZHOH040103	Quantifying the effect of burial of amino acid residues on protein stability

IV. Discussion

The merits of the proposed method are twofold: 1) a small set of informative physicochemical properties is identified for predicting bioluminescence proteins (PBLP) with promising accuracy, and 2) the small set of informative physicochemical properties can be more easily interpretable. The existing method BLProt with a test accuracy of 80.06% has been proved to be more accurate than BLAST and HMM using 100 features. The proposed method PBLP achieves a higher test accuracy of 81.79% using only 15 physicochemical properties for predicting bioluminescence proteins.

The identified feature sets from 30 independent runs of PBLP are very robust. The appearance frequency of each identified cluster in the 30 runs is shown in Fig. 3. From the statistic result, the clusters 7, 9, 10 and 16 with very high selection frequencies are more informative for predicting bioluminescence proteins. The selected clusters of the 30 runs are very similar in terms of cluster ID from 20 clusters. The most effective property with RACS820111 is belonging to the 10th cluster with Beta propensity in six groups.

PBLP is an efficient approach to selecting informative physicochemical properties for SVM classifier. With the IBCGA-selected features, the prediction accuracy of our method is better than the existing method. This method can be also applied to other sequence-based prediction problems.

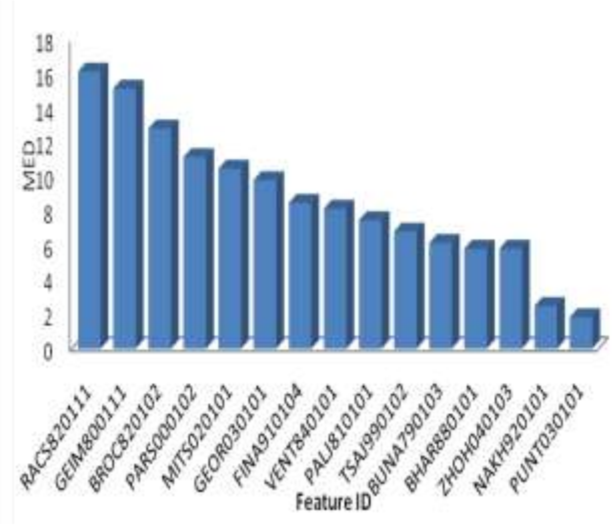


Figure 1. The rank of the selected feature set with the highest training accuracy is analyzed by MED analysis.

References

[1] Wilson T. 1995. Comments on the mechanisms of chemi- and bioluminescence. Photochem.Photobiol.62:601–6

[2] Heim, R., and Tsien, R.Y. (1996). Engineering greenfluorescent protein for improved brightness, longerwavelengths and fluorescence resonance energytransfer.Curr. Biol. 6, 178–182.

[3] Cubitt AB, Heim R, Adams SR, Boyd AE,Gross LA,Tsien RY. 1995. Understanding,improving and using green fluorescent proteins.Trends Biochem. Sci. 20:448–55

[4] Johnson CH, Knight MR, Kondo T, Masson P,Sedbrook J, et al. 1995. Circadian oscillationsof cytosolic and chloroplastic free calcium inplants. Science 259:1863–65

[5] K. K. Kandaswamy, G. Pugalenthi, M. K. Hazrati, K.-U. Kalies, and T. Martinetz. BLProt: Prediction of bioluminescent proteins based on Support Vector Machine and Relief feature selection. *BMC Bioinformatics*, 2011.

[6] Huang, H.-L., Lin, I.-C., Liou, Y.-F., Tsai, C.-T., et al., Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. BMC Bioinformatics 2011, 12 Suppl 1.

[7] Chun-Wei Tung and Shinn-Ying Ho, “POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties,” *Bioinformatics*, vol. 23, no. 8, pp. 942–949, 2007.

[8] Chun-Wei Tung and Shinn-Ying Ho, “Computational identification of ubiquitylation sites from protein sequences,” *BMC Bioinformatics*, vol. 9:310, July 2008.

[9] JR Quinlan. *C4.5: programs for machine learning*. In. San Mateo, CA: Morgan Kaufmann. 1993.

[10] C. C. Chang, and, C. J. Lin (2001) *LIBSVM: a library for support vector machines.* Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[11] S.-Y. Ho, *et al.*,“Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications,” *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 34, pp. 609-620, 2004a.

[12] Ho, S.Y., Shu, L.S., Chen, J.H. 2004. Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Transactions on Evolutionary Computation 8, 522‑541

Subject: Acceptance Notification for Your Paper Submitted to iCBBE 2012
From: "submit@icbbe.org" <submit@icbbe.org>
Date: 2011/12/18 下午 07:02
To: "huachinlee@nctu.edu.tw" <huachinlee@nctu.edu.tw>

Dear author,

Thanks for your submission to the iCBBE 2012. We are pleased to inform you that your paper:

ID: 72769

TITLE: Prediction of Carbohydrate-Binding Proteins Using a Scoring Card Method

AUTHOR(S): Hua-Chin Lee, Shinn-Ying Ho, Hui-Ling Huang

has been accepted as a full paper for the final program. Congratulations!

All accepted papers in iCBBE 2012 will be published by IEEE and indexed by Ei Compendex and ISTP.

The registration deadline is on **December 28, 2011**. You should finish the three steps below before the deadline or you will be deemed to withdraw your paper:

Step 1: Making Registration Payment.

Step 2: Uploading Your Paper to IEEE Website.

Step 3: Providing Registration Information

Please visit the conference website www.icbbe.org/icbbe2012Submission/index.aspx , enter your username and password to login the “Author Registration”, you can get more information under the column of “Registration Instructions”.

We are looking forward to seeing you in Shanghai!

Best Regards,

iCBBE Organizing Committee

THE 22ND INTERNATIONAL CONFERENCE ON GENOME INFORMATICS

GIW2011

DECEMBER 5~7, 2011

The Official Conference of the Association of Asian Societies for Bioinformatics(AASBi)

November 17, 2011

Dear. Hui-Ling Huang
Institute of Bioinformatics and Systems Biology
National Chiao Tung University
Taiwan

It is our pleasure to inform you that your paper entitled, “Optimization approach to estimation of kinetic parameters for modelling metabolic pathways of muscle glycogenolysis” and “Intelligent triple-objective genetic algorithm for selecting informative Tag SNPs” have been accepted for the 22nd International Conference on Genome Informatics (GIW2011) to be held in Busan, Korea, December 5–7, 2011.

It is also our great honor to invite you to give an oral presentation at the conference.

GIW is the longest running international bioinformatics conference. The first GIW was held at Tokyo in 1990. Since then, GIW has been held annually around the countries in Asia-Pacific region. This year's GIW is the 22nd anniversary. Korean Society for Bioinformatics and Systems Biology (KSBSB) is honored to host GIW2011. We are delighted to give you a warm welcome to Busan, Korea.

Congratulations, we look forward to seeing you at GIW2011.

Sincerely Yours,

Sanghyuk Lee

Sanghyuk Lee, Ph.D.
President, Korean Society for Bioinformatics and Systems Biology (KSBSB)
Conference Chair, The 22nd International Conference on Genome Informatics (GIW2011)
Director, Korean Bioinformation Center (KOBIC)

國科會補助計畫衍生研發成果推廣資料表

日期:2012/10/28

國科會補助計畫	計畫名稱: 建構可解讀模糊規則知識庫來預測與分析DNA結合的蛋白質
	計畫主持人: 黃慧玲
	計畫編號: 100-2221-E-009-130- 學門領域: 生物資訊

無研發成果推廣資料

100 年度專題研究計畫研究成果彙整表

計畫主持人：黃慧玲			計畫編號：100-2221-E-009-130-				
計畫名稱：建構可解讀模糊規則知識庫來預測與分析 DNA 結合的蛋白質							
成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	1	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

🟥	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：■已發表 □未發表之文稿 □撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）

本計畫主要集中在建立可解讀的模糊邏輯規則，以增進預測和分析的去氧核醣核酸鍵結區域分類的知識。利用機器學習提供生物實驗學者在 DNA-binding 方面重要結果:1)一組物化特性的組合，2)一套演化式模糊規則分類器的原型，以及 3)一組簡潔又帶有知識且具有高度預測性的模糊規則。

本計畫已發表會議論文 1 篇與 SCI 期刊論文 1 篇。