Available online at www.sciencedirect.com
ScienceDirect
BioSystems 90 (2007) 573–581
BioSystems
www.elsevier.com/locate/biosystems

ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features

Wen-Lin Huang [a], Chun-Wei Tung [b], Hui-Ling Huang [c], Shiow-Fen Hwang [a], Shinn-Ying Ho [b,d,*]

[a] *Institute of Information Engineering and Computer Science, Feng Chia University, Taichung, Taiwan*
[b] *Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*
[c] *Department of Information Management, Jin Wen Institute of Technology, Taipei, Taiwan*
[d] *Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan*

Received 12 October 2006; received in revised form 9 December 2006; accepted 1 January 2007

Abstract

Accurate prediction methods of protein subnuclear localizations rely on the cooperation between informative features and classifier design. Support vector machine (SVM) based learning methods are shown effective for predictions of protein subcellular and subnuclear localizations. This study proposes an evolutionary support vector machine (ESVM) based classifier with automatic selection from a large set of physicochemical composition (PCC) features to design an accurate system for predicting protein subnuclear localization, named ProLoc. ESVM using an inheritable genetic algorithm combined with SVM can automatically determine the best number m of PCC features and identify m out of 526 PCC features simultaneously. To evaluate ESVM, this study uses two datasets SNL6 and SNL9, which have 504 proteins localized in 6 subnuclear compartments and 370 proteins localized in 9 subnuclear compartments. Using a leave-one-out cross-validation, ProLoc utilizing the selected $m = 33$ and 28 PCC features has accuracies of 56.37% for SNL6 and 72.82% for SNL9, which are better than 51.4% for the SVM-based system using k-peptide composition features applied on SNL6, and 64.32% for an optimized evidence-theoretic k-nearest neighbor classifier utilizing pseudo amino acid composition applied on SNL9, respectively.

Keywords: Subnuclear localization; Support vector machine; k-Nearest neighbor; Prediction; Amino acid composition; Physicochemical property; Genetic algorithm

1. Introduction

Eukaryotic cells consist of two major parts, namely the nucleus and cytoplasm. The nucleus is a highly complex organelle that forms a package for cells and their corresponding regulatory factors (Heidi et al., 2001). The nucleus guides the life processes of cells by performing the following functions: (1) storing genes in chromosomes; (2) assembling genes into chromosomes to allow cell division; (3) transporting regulatory factors and gene products via nuclear pores; (4) producing messages (messenger ribonucleic acid or mRNA) that code for proteins; (5) generating ribosome in the nucleolus; and (6) organizing uncoiling of DNA to reproduce key genes (Heidi et al., 2001; Spector, 2001).

Many nuclear proteins participating in life processes tend to concentrate on subnuclear compartments.

* Corresponding author. Tel.: +886 35131405; fax: +886 35729288.
E-mail address: syho@mail.nctu.edu.tw (S.-Y. Ho).

0303-2647/$ – see front matter © 2007 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2007.01.001

The concentration within specific subnuclear compartments is important for the construction and function of the nucleus. Poor protein localization leads to human genetic diseases and cancers (Phair and Misteli, 2000). Therefore, accurately predicting protein subnuclear localization is crucial for understanding genome regulation and functions. Computational prediction methods for subnuclear localization from primary protein sequences are fairly economic in terms of identifying many nuclear proteins with unknown functions (Lei and Dai, 2005; Shen and Chou, 2005).

Many computational systems for predicting protein subcellular localization have emerged in the recent two decades (Nakai and Horton, 1999; Hua and Sun, 2001; Cai et al., 2002; Bhasin and Raghava, 2004; Szafron et al., 2004; Yu et al., 2004; Bhasin et al., 2005; Gardy et al., 2005; Nair and Rost, 2005; Sarda et al., 2005). Since these systems allow broad treatment of subcellular localization at the genomic level, they enable the prediction of particular subnuclear compartments. However, the predictors at the subnuclear level encounter more obstacles than those at the subcellular level (Lei and Dai, 2005). Two computational methods for predicting protein subnuclear localization have recently been proposed, namely an SVM-based system (Lei-SVM) using *k*-peptide composition encoded into the kernel function (Lei and Dai, 2005) and an optimized evidence-theoretic *k*-NN (OET-KNN) classifier using pseudo-amino-acid composition (Shen and Chou, 2005).

The high performance of most prediction methods arises mainly from the cooperation between informative features and efficient classifier design. Informative features from protein sequences have been investigated from factors such as amino acid composition (Nakai and Horton, 1999; Hua and Sun, 2001; Lei and Dai, 2005) and *k*-peptide encoding vectors (Nakai and Horton, 1999; Hua and Sun, 2001; Yu et al., 2004; Lei and Dai, 2005). The physicochemical properties of amino acids have recently been used to predict the subcellular localization of proteins. Bhasin et al. proposed ESLpred and PSLpred systems by using the features of amino acid composition, dipeptide composition and 33 different physicochemical properties averaged over the entire protein (Bhasin and Raghava, 2004; Bhasin et al., 2005). The 33 physicochemical properties were determined by analyzing the abundance of amino acids and variations in physicochemical properties from the P1 to P9 positions of TAP binders. Sarda et al. (2005) presented the pSLIP system by applying five top-ranking features of physicochemical properties according to the prediction accuracy of SVM using a single feature.

As for efficient classifier design, SVM-based learning methods are shown to be effective for accurately predicting protein subcellular and subnuclear localizations from protein sequences (Hua and Sun, 2001; Cai et al., 2002; Bhasin and Raghava, 2004; Szafron et al., 2004; Yu et al., 2004; Bhasin et al., 2005; Gardy et al., 2005; Lei and Dai, 2005; Nair and Rost, 2005; Sarda et al., 2005). The extraction of informative features from a primary protein sequence using SVM is essential for designing an accurate system of predicting protein subnuclear localization. A well-designed SVM-based classifier for prediction aims to combine feature optimally according to feature correlation, parameter setting of SVM and cross-validation performance. However, conventional SVM does not provide internal detection of relevant-feature correlation, and appropriate setting of their control parameters is generally treated as another independent problem (Joachims, 2002).

The role of simultaneous optimization of feature selection and classifier design in high performance of predictors is well recognized (Brotherton et al., 1994; Ho et al., 2002; Ooi and Tan, 2003; Sun et al., 2004; Ho et al., 2006). Under consideration of such simultaneous optimization, this study proposes an SVM-based learning method with automatic selection from a large set of physicochemical composition (PCC) features to design an accurate system for predicting protein subnuclear localization, named ProLoc. The set of PCC features is derived from a protein sequence, and comprises 506 physicochemical properties obtained from an amino acid index (AAindex) database (Kawashima and Kanehisa, 2000) and 20 features of amino acid composition. Consequently, an evolutionary support vector machine (ESVM) method cooperated with an SVM classifier is used to automatically determine the best number m of PCC features and identify m out of 526 PCC features simultaneously.

The ESVM is measured from two datasets, SNL6 and SNL9, are used for comparison with the existing method Lei-SVM (Lei and Dai, 2005) and OET-KNN (Shen and Chou, 2005). SNL6 has 504 proteins localized in 6 subnuclear compartments, and SNL9 has 370 proteins localized in 9 subnuclear compartments. ESVM determined the best number of PCC features for SNL6 and SNL9 as $m = 33$ and 28, respectively. ProLoc utilizing SVM with the m selected PCC features, using the same leave-one-out cross-validation (LOOCV), yields accuracies of 56.37% for SNL6 and 72.82% for SNL9, which are better than 51.4% for Lei-SVM applied on SNL6, and 64.32% for OET-KNN applied on SNL9. The effectiveness of the selected PCC features to prediction accuracy can be quantified and

ranked for biological analysis based on factor analysis (Ho et al., 2004b).

2. Materials and methods

2.1. Datasets

Two datasets, SNL6 (Lei and Dai, 2005) and SNL9 (Shen and Chou, 2005), were used to evaluate the proposed prediction system ProLoc. SNL6 has 504 proteins localized in 6 subnuclear compartments, and SNL9 has 370 proteins localized in 9 subnuclear compartments. Table 1 shows the numbers of protein sequences within each subnuclear compartment in SNL6 and SNL9. The two datasets were obtained from the Nuclear Protein Database (Dellaire et al., 2003), which is a searchable database of information on proteins consisting of more than 2000 vertebrate proteins (mainly from mouse and human) in cell nuclei. Spector (2001) shows the known subnuclear compartments where proteins have been found.

The data about SNL6 proteins were extracted by a Perl script. The SNL6 proteins associated with more than one compartment were eliminated. SNL6 is a non-redundant dataset constructed from PROSET (Brendel, 1992) with low sequence identity (<50%). The SNL9 proteins were screened strictly using the following procedure: (1) sequences without a clear compartment description in the nucleus were eliminated; (2) only one of a group of protein sequences having the same name but from different species was included to avoid redundancy; (3) sequences annotated by multiple subnuclear compartments were eliminated; (4) sequences with fewer than 50 amino acid residues were eliminated; and (5) compartments with fewer than 10 proteins were eliminated (Shen and Chou, 2005). SNL6 and SNL9 have five common compartments, namely PML body, chromatin, nuclear diffuse, nucleolus, and nuclear splicing speckles, but different numbers of proteins.

Table 1
The numbers of protein sequences within each subnuclear compartment in the datasets SNL6 (six compartments) and SNL9 (nine compartments)

Label	Compartment	Number of sequences	
		SNL6	SNL9
1	PML body	38	40
2	Chromatin	61	59
3	Nucleoplasm (nuclear diffuse)	75	65
4	Nucleolus	219	115
5	Nuclear splicing speckles (or splicing factor enriched speckles)	56	15
6	Nuclear lamina	55	
	Heterochromatin		31
7	Nuclear pore (nuclear pore complex)		25
8	PcG body		10
9	Cajal body		11
Total		504	370

2.2. ESVM-based prediction system ProLoc

Implementing the prediction system ProLoc for protein subnuclear localization by machine learning approach involves three essential tasks: (1) establishing a set of n potentially good candidate features from protein sequences; (2) determining the best number m of features and identify m out of n candidate features cooperated with SVM; and (3) designing an efficient SVM-based classifier with the selected features. The optimization of feature selection is a combinatorial optimization problem with a huge search space $C(n, m) = n!/(m!(n-m)!)$, which attempts to maximize a specified training accuracy by using a small number of selected features. The m features are selected by utilizing an inheritable genetic algorithm (IGA) that was proposed by Ho et al. (2004a). The inheritance mechanism makes IGA efficient in searching for a good solution S_{r+1} in the space $C(n, r+1)$ by inheriting a good solution S_r in the space $C(n, r)$, where $0 < r < n$. Let S_m be the most accurate solution among all investigated solutions S_r. For instance, r is in the range [15, 50] in this study. The proposed ESVM simultaneously selects r out of n features, and determines the corresponding SVM parameter setting using 10-fold cross-validation (10-CV) as an estimator of generalization ability in an independent run, where $r = 15, \ldots, 50$. Finally, the m selected features and the corresponding SVM classifier are employed to implement ProLoc.

2.2.1. Physicochemical composition features

A given protein sequence has $n = 526$ physicochemical composition (PCC) generated features, comprising 20 features of the conventional amino acid composition (AAC) and 506 physicochemical properties. Therefore, the sequence is represented as a 526-dimensional feature vector $\mathbf{P}$:

$$\mathbf{P} = [p1, \ldots, p20, p21, \ldots, p526]^{\mathrm{T}} \quad (1)$$

The AAC features reflect the normalized occurrence frequencies p_i of the 20 native amino acids where $i = 1, \ldots, 20$. The remaining 506 features are derived from the 506 physicochemical properties of AAindex by averaging over the protein sequence. All the features of $\mathbf{P}$ are rescaled in the range [0, 1] to apply SVM.

2.2.2. Evolutionary support vector machine (ESVM)

ESVM using IGA can efficiently determine a specified number r out of the $n = 526$ PCC features, and determine all parameter values of SVM simultaneously by maximizing the prediction accuracy. The N-fold cross-validation test provides a bias-free estimate of the accuracy at a much-reduced computational cost with comparison to LOOCV, and is considered an acceptable test for evaluating the prediction performance of an algorithm (Stone, 1974). In this study, $N = 10$. Consequently, the fitness function of IGA has a training accuracy of 10-CV. The multi-classification problem is solved by utilizing a series

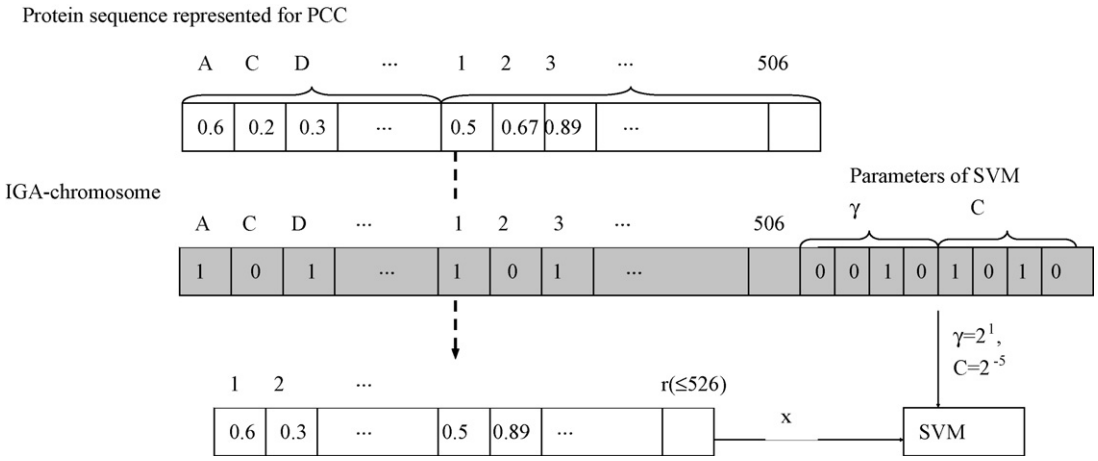

Fig. 1. A protein sequence is represented using a PCC feature vector. A potentially good feature set cooperated with SVM can be selected by the inheritable genetic algorithm (IGA) with the IGA-chromosome encoding method.

of binary classifiers of SVM$^{\text{Light}}$ (Joachims, 2002). A radial basis kernel function exp $(-\gamma||x^i - x^j||^2)$ is adopted, where x^i and x^j are training samples, and γ is a kernel parameter, is adopted for the SVM$^{\text{Light}}$.

Fig. 1 illustrates the IGA-chromosome encoding method for feature selection and parameter setting of SVM. The IGA-chromosome consists of $n = 526$ binary IGA-genes f_i for selecting informative features and two 4-bit IGA-genes for encoding the kernel parameter γ and a cost parameter C of SVM. The corresponding feature p_i is excluded from the SVM classifier if $f_i = 0$, and is included otherwise. This encoding method maps the 16 values of γ and C into $(2^{-12}, 2^{-11}, \ldots, 2^3)$ and $(2^{-10}, 2^{-11}, \ldots, 2^5)$, respectively.

IGA with orthogonal array crossover (Ho et al., 2004a) performs well in exploring an enormous search space $C(n, r)$. The orthogonal array crossover based on orthogonal experimental design (Ho et al., 2004b) uses a divide-and-conquer strategy to solve large-scale parameter optimization problems. Furthermore, to speed up exploration, the orthogonal array crossover uses a systematic reasoning method instead of the conventional generate-and-go method by genetic algorithms. The proposed ESVM is concisely presented here. Ho et al. (2004a,b) have described in detail the merits of the inheritable mechanism and superiority of IGA.

ESVM generates the solutions S_r to design the SVM classifiers with r features by utilizing an inheritable mechanism in a single run, where $r = r_{\text{start}}, \ldots, r_{\text{end}}$. The ESVM algorithm optimizes the parameter values in an IGA-chromosome (individual) by using the training accuracy of 10-CV as a fitness function, which is described as follows:

Step 1: Initialization: Randomly generate an initial population of N_{pop} feasible individuals where the n binary parameters f_i have r 1s and $n - r$ 0s in an IGA-chromosome. Let $r = r_{\text{start}}$ and a generation index $g = 1$.

Step 2: Evaluation: Compute fitness values of all individuals in the population. Let I_{best} be the best individual in the population.

Step 3: Selection: Use the simple truncation selection that replaces the worst $P_s \cdot N_{\text{pop}}$ individuals with the best $P_s \cdot N_{\text{pop}}$ individuals to form a new population, where P_s is a selection probability.

Step 4: Crossover: Randomly select $P_c \cdot N_{\text{pop}}$ individuals including I_{best}, where P_c is a crossover probability. Perform orthogonal array crossover operations for all selected pairs of parents.

Step 5: Mutation: Apply a bit-inverse mutation operator to the population using a mutation probability P_m by keeping the n binary parameters in an individual having r 1s. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6: Termination test: If g is not equal to a maximal number $G_{\max}$ of generations, then set $g = g + 1$ and go to Step 2. Decode I_{best} to obtain a solution S_r. If $r = r_{\text{end}}$, then stop the algorithm.

Step 7: Inheritance: Reserve the best $N_{\text{pop}}/2$ individuals and randomly change one of the $n - r$ bits from 0 to 1 for each individual in the population. Randomly generate additional $N_{\text{pop}}/2$ individuals where the n binary parameters have $r + 1$ 1s and $n - r - 1$ 0s. Let $g = 1$ and $r = r + 1$. Go to Step 2.

Table 2 shows the parameter settings of ESVM. In this study, $r_{\text{start}} = 15$, $r_{\text{end}} = 50$, and $G_{\max} = 20$. The solution S_m with m selected features is the best among $S_{15}, S_{16}, \ldots, S_{50}$.

3. Results and discussion

3.1. Automatic feature selection

Two sets of candidate features are used to evaluate the PCC feature set performed with ESVM. One is the set of PCC features given in (1) and the other is an additional feature set named short physicochemical composition (SPC), which only consists of the 506

Table 2
The control parameters used in ESVM

Parameter	Value
Population size, N_{pop}	50
Selection probability, p_s	0.2
Crossover probability, p_c	0.8
Mutation probability, p_m	0.05
Factor number of orthogonal arrays (Ho et al., 2004a)	7
Maximum generations, G_{max}	20
Start number of selected features, r_{start}	15
End number of selected features, r_{end}	50

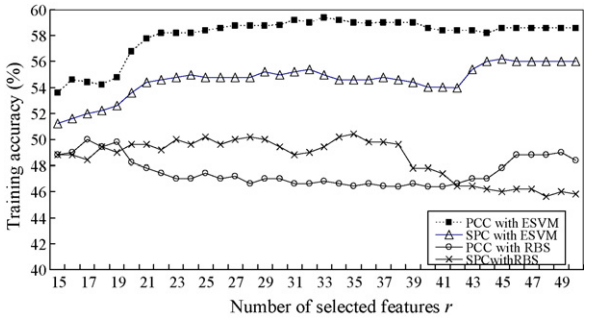

Fig. 2. Evaluations of the feature selection of ESVM and its associated PCC features, compared with the feature selection method RBS and a smaller feature set SPC using the dataset SNL6.

physicochemical properties. The proposed ESVM first determines r out of n features and the corresponding parameter setting of SVM, where $n = 506$ and 526 for SPC and PCC, respectively. The corresponding SVM classifier with the m selected features is then adopted to implement the prediction system ProLoc. Figs. 2 and 3 illustrate the accuracies of from performing 10-CV performed on SNL6 and SNL9, respectively. ESVM using PCC is much more accurate than ESVM using SPC for both SNL6 and SNL9. Additionally, Figs. 2 and 3 indicate that SVM with $m = 33$ and 28 PCC features has the best accuracy for SNL6 and SNL9, respectively.

For further evaluating the feature selection method of ESVM, a rank-based selection (RBS) method (Li et al., 2004) cooperated with the SVM^{Light} using the best values of C and γ was applied, where C $\{2^{-12}, 2^{-11}, \ldots, 2^{3}\}$ and γ $\{2^{-10}, 2^{-11}, \ldots, 2^{5}\}$. Each feature in the tested feature set was first ranked according to the accuracy of SVM with the evaluated single feature. The top-rank 50 features a_i, $i = 1, \ldots, 50$ were then picked, and the top-rank 14 features were used as an initial feature set $\{a_1, \ldots, a_{14}\}$ and. Let the size of the current feature set be r, where $r = 14$ initially. The feature set with size $r + 1$ is incrementally established by adding the best feature b_{r+1} from the remaining $50 - r$ features into the current feature set. The feature b_{r+1} from the current feature set can derive the highest SVM prediction accuracy, among all $50 - r$ using 10-CV. are shown in Figs. 2 and 3 show the prediction accuracies of SVM using the feature set with size r, $r = 15, \ldots, 50$, applied on SNL6 and SNL9, respectively. These two figures reveal that the feature selection of ESVM is better than that of RBS, where using PCC or SPC feature. The high performance of the feature selection in ESVM is due to the consideration of the internal relevant-feature correlation in global optimization using IGA.

To understand further the performance of inheritance mechanism in the feature selection of ESVM, we observe the convergence of IGA in selecting the feature set with size $r + 1$ based on the feature set with size r was studied. Fig. 4 illustrates the convergence performance of IGA at $r = 33$ and $r = 28$ on SNL6 and SNL9, respectively, in terms of a training accuracy of 10-CV and a generation index g from 1 to $G_{max} = 20$. The convergence stabilizes when $g = 13$. Analytical results reveal that (1) $G_{max} = 20$ generations are sufficient for IGA to obtain a satisfactory solution and (2) the inheritance mechanism makes IGA

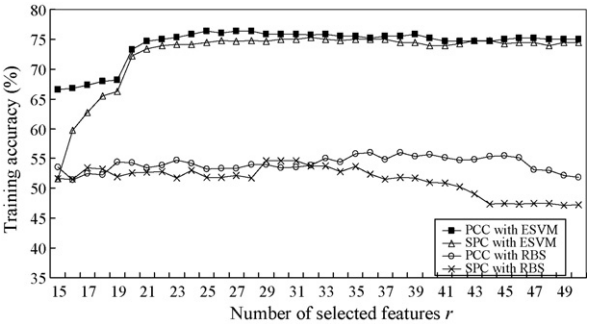

Fig. 3. Evaluations of the feature selection of ESVM and its associated PCC features, compared with the feature selection method RBS and a smaller feature set SPC using the dataset SNL9.

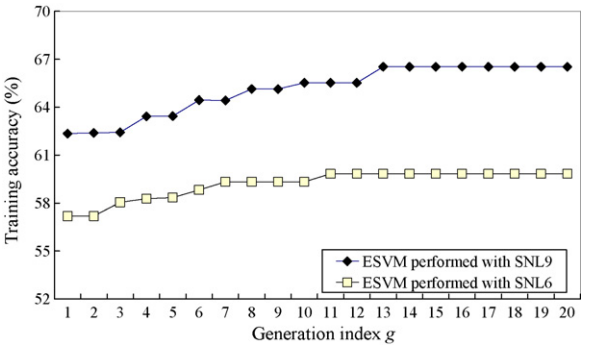

Fig. 4. The training accuracies are obtained by recording the best solution I_{best} of ESVM at $r = 33$ and 28 using SNL6 and SNL9, respectively.

efficient in exploring the search space $C(n, r+1)$, based on the exploration result of $C(n, r)$ using a small number (20) of generations (Ho et al., 2004a).

3.2. Factor analysis of selected features

The effect of several factors can be efficiently studied simultaneously by orthogonal experimental design with both orthogonal array and factor analysis (Ho et al., 2004b). The factors are the parameters that affect the evaluation function (i.e., the fitness function of IGA), and a setting (or a discriminative value) of a factor is regarded as a level of the factor. The main effect reveals the individual effect of a factor. The factor analysis using the orthogonal array's tabulation of experimental results allow the rapid estimation of the main effect of a specific factor, without the fear of distortion of results by the effects of other factors. The most effective factor has the largest main effect difference between two main effects corresponding to the two levels of an evaluated factor. In this study, the two levels of one factor are the inclusion ($f_i = 1$) and exclusion ($f_i = 0$) of a PCC feature (f_i) in the feature selection using IGA. Therefore, the factor analysis can evaluate the effects of individual factors on the evaluation function, rank the most effective factors and determine the best level for each factor to optimize the evaluation function (Ho et al., 2004b).

Each of these $m = 33$ or $m = 28$ PCC features is treated as a factor. The main effect of features on the fitness function can be estimated using the orthogonal experimental design. Figs. 5 and 6 display the features selected by ESVM using SNL6 and SNL9, respectively, ranked according to main effect difference (MED). For SNL6, five of the $m = 33$ features are from AAC, and 28 features are physicochemical properties. The five components of AAC (C, K, M, R, G) make a significant contribution, as illustrated in Fig. 5. The physicochemical property with the highest rank (MED = 34.47) is the

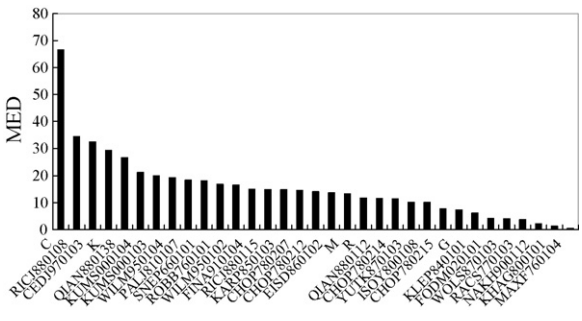

Fig. 5. The $m = 33$ features selected by ESVM with SNL6. The features are ranked according to the main effect difference (MED).

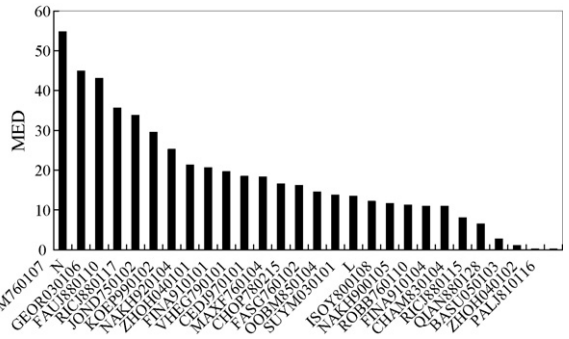

Fig. 6. The $m = 28$ features selected by ESVM with SNL9. The features are ranked according to the main effect difference (MED).

identity RICJ880108 of AAindex, which corresponds to a relative preference value at N5. Table 3 presents the detailed descriptions of the remaining selected features.

The $m = 28$ features using SNL9 comprise two components of AAC (N, L) and 26 physicochemical properties, as listed in Fig. 6. The physicochemical property with the highest rank (MED = 54.80) is the identity LEVM760107 of AAindex, which corresponds to a van der Waals parameter epsilon. Table 4 lists the detailed descriptions of the remaining selected features. With a comparison between Tables 3 and 4, there are five common informative features: CHOP780215, FINA910104, ISOY800108, MAXF760104, and RICJ880115 whose corresponding main effect differences are shown in Table 5. Besides the features with high ranks, the common features working for two different datasets have high reliability for further verification by biologists.

The number of common PCC features selected by ESVM using SNL6 and SNL9 is not large for the following three reasons: (1) the numbers of training samples for both SNL6 and SNL9 are not large enough; (2) the numbers of compartments for two datasets are not equal; and (3) only five of the nine compartments in SNL9 are common to those in SNL6. The proposed method ESVM using PCC features from protein sequences is effective for small-scale training datasets by considering avoidance of overtraining. To obtain stable features for prediction of protein subnuclear localization from novel proteins, it is better to adopt a larger training dataset due to high overlap of sample distributions.

3.3. Performance comparison of ProLoc

ProLoc utilizes the m PCC features and the established SVM classifier in the training phase. For comparison, the LOOCV performance is used to evaluate ProLoc. For SNL6, the overall prediction accuracy 56.37% for ProLoc using $m = 33$ PCC features is better

Table 3
Selected features by utilizing ESVM with the dataset SNL6

ID of AAindex	Description
RICJ880108	Relative preference value at N5
CEDJ970103	Composition of amino acids in membrane proteins (percent)
QIAN880138	Weights for coil at the window position of 5
KUMS000104	Distribution of amino acid residues in the alpha-helices in mesophilicsproteins
KUMS000103	Information measure for C-terminal helix
WILM950104	Hydrophobicity coefficient in RP-HPLC, C18 with 0.1%TFA/2-PrOH/MeCN/H_2Os
PALJ810107	Normalized frequency of alpha-helix in all-alpha class
SNEP660101	Principal component I
ROBB760101	Information measure for alpha-helix
WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H_2O
FINA910104	Helix termination parameter at posision $j+1$
RICJ880115	Relative preference value at *C*-cap
KARP850103	Flexibility parameter for two rigid neighbors
CHOP780207	Normalized frequency of C-terminal non-helical region
CHOP780212	Normalized frequency of C-terminal non-helical region
EISD860102	Atom-based hydrophobic moment
QIAN880112	Weights for alpha-helix at the window position of 5
CHOP780214	Frequency of the third residue in turn
YUTK870103	Activation Gibbs energy of unfolding, pH7.0
ISOY800108	Normalized relative frequency of coil
CHOP780215	Frequency of the fourth residue in turn
KLEP840101	Net charge
FODM020101	Propensity of amino acids within pi-helices
WOLS870103	Principal property value $z3$
RACS770103	Side chain orientational preference
NAKH900112	Transmembrane regions of mt-proteins
KHAG800101	The Kerr-constant increments
MAXF760104	Normalized frequency of left-handed alpha-helix

Table 4
Selected features by utilizing ESVM with the dataset SNL9

ID of AAindex	Description
LEVM760107	van der Waals parameter epsilon
GEOR030106	Linker propensity from medium dataset (linker length is between six and 14sresidues)
FAUJ880110	Number of full non-bonding orbitals
RICJ880117	Relative preference value at C''
JOND750102	p*K* (–COOH)
KOEP990102	Beta-sheet propensity derived from designed sequences
NAKH920104	AA composition of EXT2 of single-spanning proteins
ZHOH040101	The stability scale from the knowledge-based atom-atom potential
FINA910101	Helix initiation parameter at posision $i-1$
VHEG790101	Transfer free energy to lipophilic phase
CEDJ970101	Composition of amino acids in extracellular proteins (percent)
MAXF760104	Normalized frequency of left-handed alpha-helix
CHOP780215	Frequency of the fourth residue in turn
FASG760102	Melting point
OOBM850104	Optimized average non-bonded energy per atom
SUYM030101	Linker propensity index
ISOY800108	Normalized relative frequency of coil
NAKH900105	AA composition of mt-proteins from animal
ROBB760110	Information measure for middle turn
FINA910104	Helix termination parameter at posision $j+1$
CHAM830104	The number of atoms in the side chain labelled $2+1$
RICJ880115	Relative preference value at *C*-cap
QIAN880128	Weights for coil at the window position of -5
BASU050103	Interactivity scale obtained by maximizing the mean of correlations over pairs of sequences sharing the TIM barrel fold
ZHOH040102	The relative stability scale extracted from mutation experiments
PALJ810116	Normalized frequency of turn in alpha/beta class

than the 51.4% for the Lei-SVM method using *k*-peptide composition (Lei and Dai, 2005). Table 6 displays the results of ProLoc for every compartment, revealing that ProLoc performs best in three compartments, worst in one, and ties with Lei-SVM in two compartments.

For SNL9, ProLoc using the $m=28$ PCC features has an accuracy 72.82% using LOOCV, which is better than 64.32% for the OET-KNN utilizing a pseudo-amino acid composition (Shen and Chou, 2005). Table 7 presents the detailed results of ProLoc using $m=28$ PCC features for every compartment. The accuracies of all compartments are larger than 60%, except for the four compartments nucleolus, chromatin, PcG body, and Cajal body having 25, 15, 10, and 10 training samples, respectively. The low accuracies of these four compartments arise from two aspects: (1) the numbers of these four compartments are fairly small and (2) the fitness function of IGA is the overall accuracy without considering the accuracy of an individual compartment. Notably, the fitness function of IGA can be adaptively modified according to the preference of predictor designers.

Table 5
Main effect differences (MEDs) of the common features selected by ESVM using SNL6 and SNL9

ID of AAindex	MED	
	SNL6	SNL9
CHOP780215	7.729404	16.238091
FINA910104	15.015682	10.956711
ISOY800108	10.145092	11.645027
MAXF760104	0.545101	16.582489
RICJ880115	14.937252	6.548813

Table 6
Prediction accuracies (%) of performance comparisons on the dataset SNL6

Compartment	Lei-SVM	ESVM
PML body	10.5	18.42
Nuclear splicing speckles	33.9	26.79
Chromatin	21.3	21.31
Nuclear diffuse	28.0	42.67
Nucleolus	83.1	90.32
Nuclear lamina	36.4	36.37
Overall prediction accuracy	51.4	56.37

ProLoc utilizing the m PCC features selected by a general-purpose method ESVM performs better for two different datasets, than the two existing methods where each designed for one dataset. Simulation results reveal that the automatic feature selection of ESVM is efficient, and the selected PCC features employed in SVM are effective.

3.4. Discussion

ProLoc benefits from ESVM by extracting informative features from a large set of candidate PCC features cooperating with SVM without using domain knowledge. The automatic feature selection and parameter tuning of SVM embedded in ESVM are simultaneously optimized by IGA. IGA having the advantages of inheritance mechanism and an intelligent evolutionary algorithm (Ho et al., 2004b) can efficiently identify a small set of informative features to establish an SVM-based classifier. Factor analysis also allows the feature selection of ESVM to rank the most effective features.

Because the number of training samples is usually much smaller than the number of candidate features, which can be obtained from protein sequences, that multiple different sets may have the same small number of selected features may happen. To cope with this problem of system uncertainty, the quantized information such as main effect difference and rank of selected features are valuable for further verification by biologists. Therefore, ESVM can serve as an adaptive feature extractor for solving the prediction problems involving system uncertainty. ESVM is an efficient bioinformatic tool, and can be treated as the core for designing various prediction systems for novel proteins using only the protein sequence information. For example, ESVM can be utilized for designing a prediction system for protein subcellular localization.

Table 7
Prediction accuracies (%) of performance comparisons on the dataset SNL9

Compartment	OET-KNN	ESVM
PML body	NA	62.50
Chromatin	NA	40.00
Nuclear diffuse	NA	74.13
Nucleolus	NA	50.00
Nuclear splicing speckles	NA	94.78
Heterochromatin	NA	74.19
Nuclear pore complex	NA	67.69
PcG body	NA	30.00
Cajal body	NA	30.00
Overall prediction accuracy	64.32	72.82

NA: not available.

4. Conclusions

Computational prediction of protein subnuclear localization from primary protein sequences is crucial for understanding genome regulation and functions. Support vector machine (SVM) based learning methods are shown to be effective for predicting protein subcellular and subnuclear localizations. Extraction of informative features cooperating with SVMs plays an important role in designing an accurate system for predicting protein subnuclear localization.

This study proposes an ESVM learning method with automatic feature selection from physicochemical composition features to design an accurate system named ProLoc for predicting protein subnuclear localization. To discover potentially good informative features, ESVM utilizes 526 candidate features from protein sequences, comprising 20 features of amino acid composition and 506 physicochemical properties, taken from the AAindex database. ESVM utilizing an inheritable genetic algorithm combined with SVM can automatically determine the best number m of PCC features and identify m out of 526 PCC features simultaneously. The feature selection of ESVM can also rank the most effective features when accompanied by factor analysis.

To evaluate the efficiency of ESVM, two datasets SNL6 and SNL9 were used to compare two existing methods. Simulation results show that ProLoc performs well when using ESVM, which can select a small set of physicochemical composition features with a high prediction accuracy. Using a leave-one-out cross-validation, ProLoc utilizing the selected $m = 33$ and 28 PCC features has accuracies of 56.37% and 72.82%, respectively, which is better than the 51.4% of the existing SVM-based system using k-peptide composition features applied to SNL6, and 64.32% of an optimized evidence-theoretic k-nearest neighbor classifier utilizing pseudo amino acid composition applied to SNL9,

respectively. A web site for ProLoc has also been (set up OR established) to accurately predict protein subnuclear localization across SNL6 and SDN9 (available at http://iclab.life.nctu.edu.tw/proloc).

References

Bhasin, M., Garg, A., Raghava, G.P.S., 2005. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21, 2522–2524.

Bhasin, M., Raghava, G.P.S., 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res. 32, W414–W419.

Brotherton, T.W., Simpson, P.K., Fogel, D.B., Pollard, T., 1994. Classifier design using evolutionary programming. In: Sebald, A.V., Fogel, L.J. (Eds.), Proceedings for the Third Annual Conference on Evolutionary Programming. World Scientific, Singapore, pp. 68–75.

Brendel, V., 1992. PROSET—a fast procedure to create non-redundant sets of protein sequences. Math. Comput. Modell. 16, 37–43.

Cai, Y.D., Liu, X.J., Xu, X.B., Chou, K.C., 2002. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J. Cell. Biochem. 84, 343–348.

Dellaire, G., Farrall, R., Bickmore, W.A., 2003. The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. Nucleic Acids Res. 31, 328–330.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., Brinkman, F.S.L., 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21, 617–623.

Heidi, G.E.S., Gail, K.M., Kathryn, N., Lisa, V.F., Rachel, F., Graham, D., Javier, F.C., Wendy, A.B., 2001. Large-scale identification of mammalian proteins localized to nuclear sub-compartments. Hum. Mol. Genet. 10, 1995–2011.

Ho, S.-Y., Hsieh, C.-H., Chen, H.-M., Huang, H.-L., 2006. Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. BioSystems 85, 165–176.

Ho, S.-Y., Chen, J.-H., Huang, M.-H., 2004a. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern.—Part B 34, 609–620.

Ho, S.-Y., Liu, C.-C., Liu, S., 2002. Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. Pattern Recognit. Lett. 23, 1495–1503.

Ho, S.-Y., Shu, L.-S., Chen, J.-H., 2004b. Intelligent evolutionary algorithms for large parameter optimization problems. IEEE Trans. Evol. Comput. 8, 522–541.

Hua, S., Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721–728.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers.

Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Res. 28, 374.

Lei, Z., Dai, Y., 2005. An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 6, 291–298.

Li, T., Zhang, C., Ogihara, M., 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20, 2429–2437.

Nair, R., Rost, B., 2005. Mimicking cellular sorting improves prediction of subcellular localization. J. Mol. Biol. 348, 85–100.

Nakai, K., Horton, P., 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. 24, 34–35.

Ooi, C.H., Tan, P., 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics 19, 37–44.

Phair, R.D., Misteli, T., 2000. High mobility of proteins in the mammalian cell nucleus. Nature 404, 604–609.

Sarda, D., Chua, G., Li, K.-B., Krishnan, A., 2005. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics 6, 152–163.

Shen, H.B., Chou, K.C., 2005. Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem. Biophys. Res. Commun. 337, 752–756.

Spector, D.L., 2001. Nuclear domains. J. Cell Sci. 114, 2891–2893.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. 36, 111–147.

Sun, Z., Bebis, G., Miller, R., 2004. Object detection using feature subset selection. Pattern Recognit. 37, 2165–2176.

Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Poulin, B., Eisner, R., Lu, Z., Anvik, J., Macdonell, C., Fyshe, A., et al., 2004. Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Res. 32, W365–W371.

Yu, C.S., Lin, C.J., Hwang, J.K., 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on *n*-peptide compositions. Protein Sci. 13, 1402–1406.