BioSystems 98 (2009) 73–79

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Predicting protein subnuclear localization using GO-amino-acid composition features

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

[a] Department of Management Information System, Chin Min Institute of Technology, Miaoli, Taiwan
[b] Institute of Bioinformatics, National Chiao Tung University, Hsinchu, Taiwan
[c] Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan

ARTICLE INFO

Article history:
Received 18 February 2009
Received in revised form 10 June 2009
Accepted 26 June 2009

Keywords:
Gene Ontology
Subnuclear localization
Amino acid composition

ABSTRACT

The nucleus guides life processes of cells. Many of the nuclear proteins participating in the life processes tend to concentrate on subnuclear compartments. The subnuclear localization of nuclear proteins is hence important for deeply understanding the construction and functions of the nucleus. Recently, Gene Ontology (GO) annotation has been used for prediction of subnuclear localization. However, the effective use of GO terms in solving sequence-based prediction problems remains challenging, especially when query protein sequences have no accession number or annotated GO term. This study obtains homologies of query proteins with known accession numbers using BLAST to retrieve GO terms for sequence-based subnuclear localization prediction. A prediction method PGAC, which involves mining informative GO terms associated with amino acid composition features, is proposed to design a support vector machine-based classifier. PGAC yields 55 informative GO terms with training and test accuracies of 85.7% and 76.3%, respectively, using a data set SNL_35 (561 proteins in 9 localizations) with 35% sequence identity. Upon comparison with Nuc-PLoc, which combines amphiphilic pseudo amino acid composition of a protein with its position-specific scoring matrix, PGAC using the data set SNL_80 yields a leave-one-out cross-validation accuracy of 81.1%, which is better than that of Nuc-PLoc, 67.4%. Experimental results show that the set of informative GO terms are effective features for protein subnuclear localization. The prediction server based on PGAC has been implemented at http://iclab.life.nctu.edu.tw/prolocgac.

1. Introduction

The cell nucleus is a highly complex organelle that organizes the comprehensive assembly of genes and their corresponding regulatory factors. The nucleus guides life processes of cells by directing their reproduction, controlling their differentiation and regulating their metabolic activities. Many of the nuclear proteins participating in the life processes tend to concentrate on subnuclear compartments (Heidi et al., 2001). The knowledge of protein subnuclear localization can provide valuable clues about its molecular function, as well as the biological pathway in which it participates (Cocco et al., 2004).

The bulk of computation methods exist in literature for predicting protein subcellular localization and has achieved high accuracy (Bhasin and Raghava, 2004; Cai and Chou, 2004; Chou and Shen, 2006a,b; Huang et al., 2008; Nair and Rost, 2005; Nanni and Lumini, 2006; Pierleoni et al., 2006; Sarda et al., 2005), particularly systematically introduced in a recent review (Chou and Shen, 2007b), a step-by-step protocol paper (Chou and Shen, 2008) and a book chapter (Chou, 2009). However, the prediction of protein localization at subnuclear level is far more challenging (Lei and Dai, 2006). We have developed the first ProLoc system using SVM with automatic selection from physicochemical properties for this task using with considerable prediction accuracy (Huang et al., 2007b). In this work, we attempted to improve the performance of the system through the incorporation of information obtained from Gene Ontology (GO).

Gene Ontology, which is a controlled vocabulary of terms split into three related ontology consisting of molecular function, biological processes and cellular components (Ashburner et al., 2000), has been utilized to improve prediction of subcellular (Chou and Shen, 2006a,b; Huang et al., 2008) and subnuclear localization (Lei and Dai, 2006; Shen and Chou, 2007b). Additionally, GO annotation has been used for various sequence-based prediction tasks, such as grouping GO terms to improve the assessment of gene set enrichment (Lewin and Grieve, 2006); using GO with probabilistic chain graphs for (Carroll and Pavlovic, 2006; Wolstencroft et al., 2006); using GO for analyzing the mouse basic/Helix-Loop-Helix transcription factor family (Li et al., 2006), using GO for identify-

* Corresponding author at: Institute of Bioinformatics, National Chiao Tung University, 75 Po-Ai Street, Hsinchu, Taiwan.
E-mail address: syho@mail.nctu.edu.tw (S.-Y. Ho).

0303-2647/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2009.06.007

Table 1
The numbers of proteins within each subnuclear compartment in the data sets SNL_80 and SNL_35. There are nine essential GO terms corresponding to subnuclear compartments. M, B and C represent the three branches molecular function, biological process and cellular component, respectively. The number t of (t) in SNL_80L and SNL_35L represents the number of sequences which are correctly annotated by only one essential GO term.

Label	Compartment	Essential GO terms	Branch	SNL_80L	SNL_80T	SNL_35L	SNL_35T
1	Nuclear PML body	GO:0016605	C	8 (1)	5	8 (1)	4
2	Nuclear speckle	GO: 0016607	C	44 (32)	23	33 (24)	16
	Chromatin	GO:0000785	M	66 (9)	33	51 (7)	27
4	Nucleoplasm	GO:0005654	C	24 (2)	13	20 (3)	10
5	Nucleolus	GO:0005730	C	204 (83)	103	155 (35)	78
6	Heterochromatin	GO: 0000792	C	14 (1)	8	8 (0)	5
7	Nuclear envelope	GO:0005635	C	40 (13)	21	31 (10)	16
8	Nuclear matrix	GO:0016363	M	19 (17)	10	17 (14)	9
9	Nuclear pore complex	GO:0005643	C	52 (43)	27	48 (40)	25
Total				471 (201)	243	371 (134)	190

ing membrane proteins and their types (Cai et al., 2005), predicting the enzymatic attribute of proteins by hybridizing the gene product composition and pseudo amino acid composition (Cai et al., 2005), and predicting the transcription factor DNA binding preference (Qian et al., 2006). Querying a GO library to obtain GO terms requires the accession numbers of proteins. Therefore, the use of GO terms for solving sequence-based prediction problems is still worthy of study, especially when query protein sequences have no accession number or annotated GO term. Two ensemble classifiers Hum-PLoc (Chou and Shen, 2006a) and Euk-OET-PLoc (Chou and Shen, 2006b) directly use the accession numbers of known proteins to obtain GO terms, so they do not work for predicting novel proteins without known accession numbers. The GO-AA (Lei and Dai, 2006) utilizes the GO terms of their homologies that are retrieved by BLAST (Altschul et al., 1990) in predicting the subnuclear localization of novel proteins.

Most, but not all, eukaryotic protein sequences in the UniProtKB/Swiss-Prot database (Apweiler et al., 2004) have annotated GO terms. For example, the percentage of 2423 training proteins whose homologies are not annotated by GO terms is 3.96% (Huang et al., 2008). To predict the proteins that do not have annotated GO terms, existing GO-based prediction methods such as GO-AA (Lei and Dai, 2006), Euk-OET-PLoc (Chou and Shen, 2006b) and Hum-PLoc (Chou and Shen, 2006a), use two separate modules—one that uses GO terms as input features (called the GO-based classifier) and another that uses sequence-based features (called the sequenced-based classifier). The GO-based classifier is used for proteins with annotated GO terms. These proteins are represented as high-dimensional vectors of n binary features, where n is the total number of GO terms in the complete annotation set (a component of 1 indicates that the annotation is hit; otherwise, the component is 0). The sequence-based classifier is applied for proteins that have no corresponding GO terms.

This study proposes a prediction method PGAC for developing a single SVM-based classifier for sequence-based subnuclear localization prediction. First, BLAST is used to obtain homologies with known accession numbers from the query protein to retrieve GO terms. Each protein sequence had $\eta = n + 20$ GO-amino-acid composition (GAC) features, comprising 20 features of the conventional amino acid composition (AAC) and n GO terms. Subsequently, a feature mining algorithm, GACmining, which is an extension of GOmining (Huang et al., 2008), was proposed using an intelligent genetic algorithm (Ho et al., 2004a,b) with an SVM classifier to identify simultaneously a small number m of η GAC features and parameter settings of SVM, where $m \ll \eta$.

A data set SNL_35 of 561 subnuclear proteins with 35% sequence identity was established to evaluate the proposed prediction method. The data set SNL_35 was divided into two subsets, one for training (SNL_35L) and the other for independent test (SNL_35T), to avoid homolog bias and any overestimation of value of the methods. PGAC, when applied to the training data set SNL_35L, extracted $m = 75$ informative GAC features and yielded training and test accuracies of 85.7% and 76.3%, respectively. The Matthews correlation coefficient (MCC) (Hua and Sun, 2001; Huang et al., 2007b; Lei and Dai, 2006) performances were 0.749 and 0.668 for training and independent testing, respectively. Upon comparison with the existing method Nuc-PLoc which combines the amphiphilic pseudo amino acid composition of a protein with its position-specific scoring matrix (Shen and Chou, 2007b), PGAC yields a leave-one-out cross-validation accuracy of 81.1% (MCC = 0.691), which is better than Nuc-PLoc with 67.4% (MCC = 0.50) using SNL_80. The prediction server that is based on PGAC for protein subnuclear localization has been implemented at http://iclab.life.nctu.edu.tw/prolocgac.

2. Materials

2.1. Data Sets

A data set SNL_80 with 80% sequence identity obtained from another work (Shen and Chou, 2007b) has 714 protein sequences in nine subnuclear compartments. The proteins in the data set were screened strictly using the following rules: (1) sequences with a same subnuclear location (SUBCELLULAR LOCATION) in the CC field might be annotated with different terms so that several keywords were used for a same subcellular location, e.g. in search for nuclear envelope proteins, the keywords 'nuclear envelope', 'nuclear inner membrane' and 'nuclear outer membrane' were used; (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; (5) compartments with fewer than 10 proteins were eliminated, and (6) sequences with 80% identity were operated by a culling program (Shen and Chou, 2007b).

Some proteins can simultaneously exist at more than one location site. This kind of multiplex proteins may have special functions and hence are particularly interesting (Chou and Shen, 2007a; Shen and Chou, 2007a). However, the number of multiplex proteins in the existing nuclear protein database is not large enough to allow us to construct a statistically meaningful benchmark data set for studying multiplex nuclear proteins as done in (Shen and Chou, 2007a) for the eukaryotic and human protein systems. As a compromise, here let us just study the single-location nuclear proteins (rule 2, mentioned above). Nevertheless, by using the similar approach as elaborated in (Shen and Chou, 2007a), the current method can also be extended to deal with the multiplex nuclear proteins once more data for the nuclear proteins are available in future.

Some studies have shown that sequence similarity is useful when sequences share >25% identity in sequence-based prediction (Yu et al., 2006). To remove the homologous sequences from the benchmark data set, a cutoff threshold of 25% was imposed in (Chou and Shen, 2006a) to exclude those proteins from the benchmark data sets that have equal to or greater than 25% sequence identity to any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data for subnuclear proteins do not allow us to do so. Otherwise, the numbers of proteins for some subsets would be too few to have statistical significance. Therefore, we established another data set SNL_35 of 561 subnuclear proteins with 35% sequence identity using a culling program (Wang and Dunbrack, 2003) and SNL_80 to evaluate the proposed method. For the SNL_35, Table 1 shows that only 12 sequences were in Nuclear PML body compartment. Additionally, some GO annotations, corresponding to subnuclear compartments, are called *essential GO terms* for subnuclear localization prediction, such as GO:0005730 (Nucleolus), GO:0000785 (Chromatin) and GO:0005643 (Nuclear pore complex), as shown in Table 1.

Table 2
Results of GO annotation for all sequences in SNL_80L and SNL_35L.

Data set	Total GO terms n	Number of GO terms			Number of sequences annotated by g essential GO terms		
		Smallest	Largest	Mean	$g=0$	$g=1$	$g>1$
SNL_35L	677	0	43	10.7	186	162	23
SNL_80L	771	0	43	10.7	232	210	29

All of the proteins are divided randomly into two separated sets with sizes in the ratio 2:1, for training and independent testing, respectively. Table 1 presents the numbers of proteins within each subnuclear compartment in the SNL_80 and SNL_35. The accession numbers and sequences of the corresponding proteins in the training and testing data sets can be found at http://iclab.life.nctu.edu.tw/prolocgac.

2.2. Gene Ontology Annotation

The growth of Gene Ontology databases in size has increased the effectiveness of GO-based features. The newest version of the GO database (released on Dec. 2, 2008, http://www.geneontology.org/) contained 26,417 terms in the three branches of biological process, molecular function and cellular component. This study utilizes the GOA database, which includes GO annotations for non-redundant proteins from many species that are in the UniProtKB/Swiss-Prot database (Apweiler et al., 2004). The GOA database was downloaded directly from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/ (UniProt 63.0 released in June 2008).

The accession numbers of proteins are required in querying the GOA database to obtain their annotated GO terms. Considering novel query proteins, BLAST (Altschul et al., 1990, 1997) was used to obtain homologies with known accession numbers from the query protein to retrieve GO terms. The parameter e-value of BLAST is critical to the quality of the homologies and the number of candidate homologies. The best values of parameters h and e were determined from $h \in \{1, 2, \ldots, 5\}$ and $e \in \{10^{-1}, 10^{-2}, \ldots, 10^{-10}\}$ using a step-wise method with the k-nearest-neighbor classifier (Huang et al., 2007a, 2008). Table 2 shows the GO annotation results for all proteins in the SNL_35L and SNL_80L, where $(h, e) = (1, 10^{-9})$. The size of the complete set of all GO terms from the 371 subnuclear proteins of SNL_35L is $n = 677$. The smallest, largest and mean numbers of GO terms that were annotated for individual proteins were 0, 43 and 10.7, respectively.

To evaluate the prediction performance using the essential GO terms annotated alone, the numbers of sequences that were annotated with g essential GO terms were calculated. Tables 1 and 2 show that 162 out of 371 sequences are annotated by only one essential GO term ($g = 1$), and that 134 of these 162 sequences are correctly annotated. The other 209 (=186 + 23) sequences annotated with zero ($g = 0$) or more than one ($g > 1$) essential GO term cannot be effectively predicted using the essential GO terms alone. This finding indicates that essential GO terms are necessary but not sufficient for the design of accurate classifiers for the prediction of protein subnuclear localization.

3. Method

3.1. Proposed Mining Algorithm GACmining

The proposed PGAC was implemented based on a mining algorithm, GACmining which is extension of GOmining for feature selection (Huang et al., 2008). An analysis of the selected informative GO terms in the GO graph reveals that GOmining can consider the internal correlation within relevant features rather than individual features using an efficient global optimization method (Huang et al., 2008). GACmining uses an intelligent genetic algorithm (IGA, Ho et al., 2004b) associated with an inheritable mechanism, named IBCGA (Ho et al., 2004a,b) an SVM classifier to identify a small number, m, of a large number, η, of GAC features and parameter settings of SVM. The exploration of the m informative GAC features from η candidate GAC features is a combinatorial optimization problem $C(\eta, m)$ with a huge search space of size $C(\eta, m) = \eta!/(m!(\eta - m)!)$.

The leave-one-out cross-validation (LOOCV) is considered to be the most rigorous and objective test that can always yield a unique result for a given benchmark data set. Hence, LOOCV has been increasingly and widely used by investigators to examine the accuracy of various predictors (Chou and Shen, 2007b). Although bias-free, this test is very computationally demanding and is often impractical for large data sets. The N-fold cross-validation not only provides a bias-free estimation of the accuracy at a much reduced computational cost, but is also considered as an acceptable test for evaluating prediction performance of an algorithm (Stone, 1974). Therefore, GACmining uses the prediction accuracy of 10-fold cross-validation (10-CV) as the fitness function to perform IBCGA (Ho et al., 2004a,b) on the entire training sets of proteins under considering the computation cost.

The input of this algorithm is a training set of protein sequences that belong to nine classes. The output contains a subset of m selected GAC features and an SVM classifier with associated parameter settings. The following algorithm is used to solve the subnuclear localization prediction problems.

Step 1 (GO terms) Use BLAST to obtain the GO terms that are annotated for each training protein by retrieving the GOA database. Let n be the total number of GO terms that have ever appeared for all training proteins. For example, $n = 677$, including $d = 9$ essential GO terms for SNL_35L.

Step 2 (Sequence representation) The 20 AAC features and 9 essential GO terms are regarded as crucial GAC features. Represent each protein as a η-dimensional feature vector $P = [p_1, p_2, \ldots, p_\eta]$ consisting of the 20 AAC features and n GO terms, such that $\eta = 20 + n$.

Step 3 (Preparation of SVM) The multi-classification problem is solved by utilizing a series of binary classifiers of LIBSVM (Chang and Lin, 2001). 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. There are two parameters γ and a cost parameter C to be tuned in using the SVM. In this study, $\gamma \in \{2^{-7}, 2^{-6}, \ldots, 2^8\}$ and $C \in \{2^{-7}, 2^{-6}, \ldots, 2^8\}$.

Step 4 (Chromosome encoding) The IGA-chromosome consists of η binary IGA-genes f_i to select informative GAC features and two 4-bit IGA-genes for encoding γ and C. The corresponding feature p_i (the i-th GAC features) is excluded from the SVM classifier if $f_i = 0$, and is included if $f_i = 1$, where $f_i = 1$, $i = 1, \ldots, 20 + d$. Let m be the sum of f_i. Fig. 1 shows the protein representation and the IGA-chromosome encoding method.

Step 5 (Fitness function) The value of the fitness function is the prediction accuracy of 10-CV using the SVM classifier with the m selected GAC features, γ and C by decoding the IGA-chromosome.

Step 6 (Initial solution) Perform IBCGA to select r_{start} out of η GAC features, yielding the solution to the problem $C(\eta, r_{start})$. The details of the procedure and the parameter settings of IGA can be found in Table 3. The inheritance mechanism of IBCGA can advance the search for the solution to $C(\eta, r+1)$ by inheriting a good solution S_r to $C(\eta, r)$.

Table 3
The used control parameters of IGA.

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	7
Maximum generations, G_{max}	60

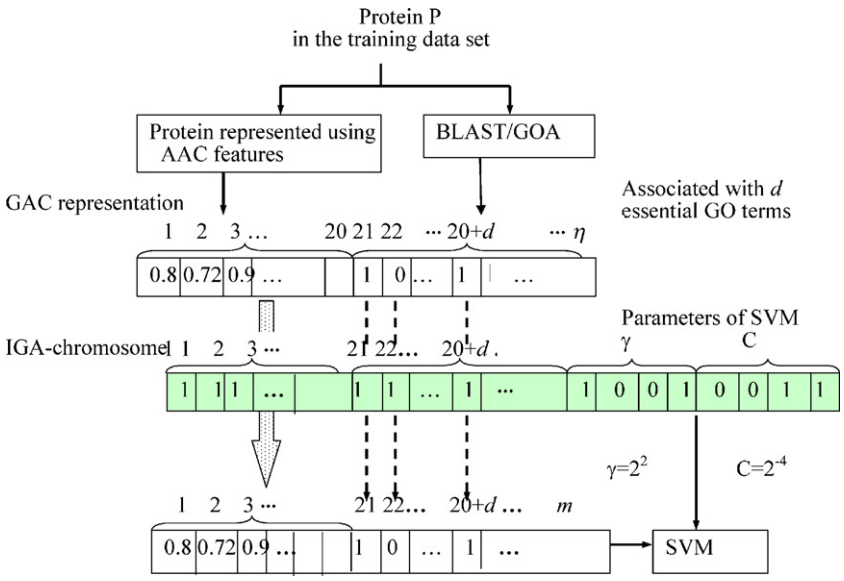

Fig. 1. Protein representation and IGA-chromosome encoding method.

Step 7 (Best solution) Obtain all solutions S_r from $r = r_{start} + 1$ to r_{end} one by one using IBCGA. For example, $r_{start} = 40$ and $r_{end} = 80$ in this study. Let S_m be the most accurate solution with m selected GO terms among all solutions S_r.

Step 8 (System uncertainty) Perform Steps 6 and 7 for N independent runs to obtain the best of N solutions, S_m, and the associated parameter setting of the SVM classifier. The best solution considers both high prediction accuracy and high mean of appearance frequency ratio, where the frequency ratio for each run was the percentage of its m_i selected GAC features in all N selected feature sets, $i = 1, \ldots N$.

The selection procedure is described as follows. First, collect the candidate solutions S_m from the N (e.g., 30) solutions that their frequency ratios are larger than the mean of frequency ratio. Secondly, obtain the best solution with the highest prediction accuracy from these candidate solutions S_m. For example, Fig. 2 shows that the best solution for SNL_35L was the 7th (i.e. $i = 7$) solution because its frequency ratio, 54.7%, is larger than the mean of the frequency ratio, 51.4%, and its training accuracy 87.3% is the best.

3.2. Prediction Using SVM

For each query protein, the BLAST with $(h, e) = (1, 10^{-9})$ is first performed on the Swiss-Prot database to obtain its homologies with known accession numbers. Subsequently, the obtained accession numbers were used to retrieve the corresponding k GO terms, GO:1, GO:2, ..., GO:k. The query protein is represented as an m-dimensional GAC feature vector $[p_1, p_2, \ldots, p_m]$ as an input to the SVM classifier. The first 20 elements p_i are the AAC features. The remaining $m - 20$ binary features denote the corresponding informative GO terms, where $p_i = 0$ if the informative GO term is not in the set of the k GO terms; otherwise, $p_i = 1$.

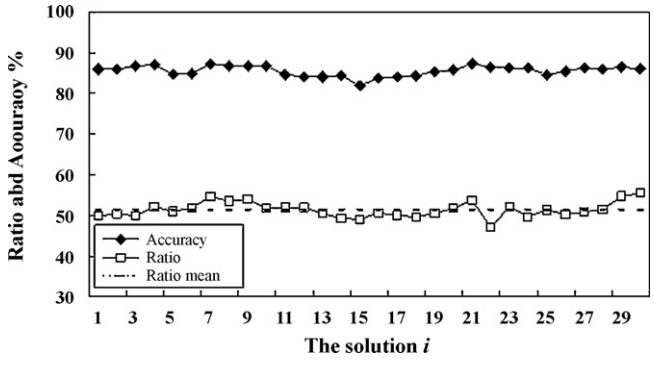

Fig. 2. The frequency ratios of the solutions $i = 7$–13, 21, 23, 28–30 are larger than the mean frequency ratio 51.4% for SNL_35L. Among these solutions, the 7th has the highest prediction accuracy.

4. Results and Discussion

4.1. Effectiveness of Feature Selection

To evaluate a candidate set of r informative GAC features accompanied with the SVM parameters, the prediction accuracy of 10-CV serves as a fitness function of IGA. Fig. 3 shows the training accuracies of PGAC from $r = 40, 41, \ldots, 80$, were higher than those of SVM-RBS for SNL_35L and SNL_80L, where SVM-RBS performed by using SVM with a number r of selected informative GAC features by the rank-based selection (RBS) method (Li et al., 2004; Tung and Ho, 2007). One previous work in ProLoc-GO (Huang et al., 2008) and ProLoc (Huang et al., 2007b) showed that this univariate method RBS is inferior to the multivariate feature selection by IGA for selecting GO terms and physicochemical properties, respectively.

The RBS method is described below. First, each of all η GAC features (for example, $\eta = 697$ for SNL_35L) was ranked according to the accuracy of SVM with the estimated single feature, where the best values of parameters (C, γ) were determined using a step-wise approach where $\gamma \in \{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$ and $C \in \{2^{-7}, 2^{-6}, \ldots, 2^{8}\}$. The top-ranking 80 features α_i, $i = 1, \ldots, 80$ are then picked, and the top-ranking 40 features with $r = 40$ are used as an initial feature set $\{\alpha_1, \ldots, \alpha_{40}\}$. Consequently, the feature set with size $r + 1$ is incrementally created by adding the best feature $\alpha_r + 1$ (having the highest accuracy of SVM using 10-CV) from the remaining $80 - r$ features into the current feature set.

Additionally, to measure the effectiveness of the selected m informative GAC features and associated SVM parameters (C, γ), the classification methods SVM and C5.0 (Quinlan, 2003) using the 20 AAC features, n GO terms and η GAC features (without feature selection), were also evaluated in terms of the prediction accuracy of 10-CV using SNL_35L and SNL_80L. The best values of parameters

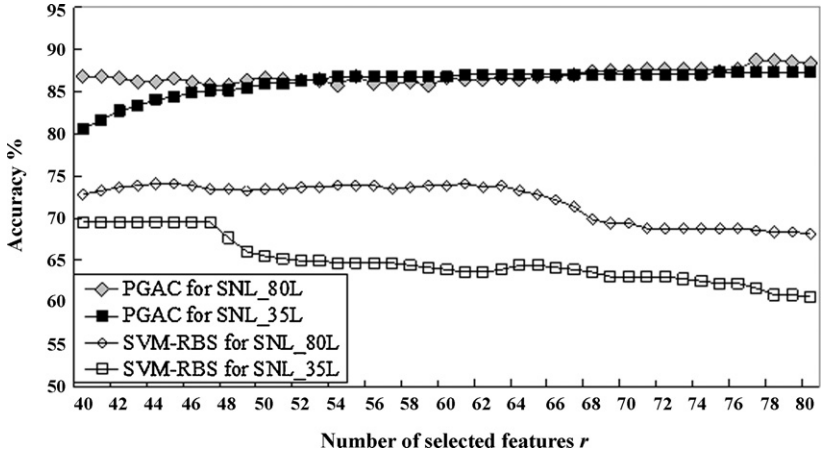

Fig. 3. Training accuracies of PGAC and SVM-RBS performed by using SVM with a number r of selected informative GAC features.

Table 4
Performance comparison uses prediction accuracy (%) of 10-CV.

Method	Feature (no.)	SNL_35L(C,γ)	SNL_80L (C, γ)
		n = 677, m = 75	n = 771, m = 78
SVM	AAC (20)	55.8 (2^2,2^{-1})	58.2 (2^4, 2^{-3})
	GO terms (n)	81.1 (2^3, 2^{-5})	83.4 (2^3, 2^{-6})
	GAC ($\eta = 20 + n$)	81.4 (2^3, 2^{-5})	84.3 (2^3, 2^{-6})
C5.0	AAC (20)	52.8	55.0
	GO terms (n)	79.5	83.0
	GAC ($\eta = 20 + n$)	78.4	83.4
ProLoc-GAC	Selected GAC (m)	87.3 (2^4, 2^{-4})	88.7 (2^4, 2^{-1})

γ and C in the SVM-based classifiers were determined using a stepwise approach. GACmining when applied to SNL_35L and SNL_80L extracted $m = 75$ and 78 informative GAC features, where $(C, \gamma) = (2^4, 2^{-4})$ and $(2^4, 2^{-1})$, respectively.

Table 4 shows that the GO term features outperformed the AAC features and performed similarly to the GAC features when SVM and C5.0 were applied to both data sets SNL_35L and SNL_80L. The AAC features that are included in the GAC feature set are especially useful in predicting the proteins that do not have annotated GO terms. The accuracies of SVM and C5.0 when applied to SNL_35L were slightly worse than those when applied to SNL_80L because of the low sequence identity of the former. The best classifier other than PGAC is the SVM with the η GAC features, yielding accuracies 81.4% and 84.3% when used on SNL_35L and SNL_80L, respectively. Using GACmining for feature selection, PGAC improved the accuracies to 87.3% and 88.7% for SNL_35L and SNL_80L, respectively.

Table 6
Prediction accuracies preformed on SNL_80 using leave-one-out cross-validation (LOOCV).

Method	Features	SNL_80 (LOOCV)
ProtLock	AAC	36.6%
SVM	AAC	48.9%
OET-KNN	Pse-AAC	55.6%
Nuc-PLoc	Fusion of PsePSSM and Pse-AAC	67.4%
ProLoc-GAC	GAC	81.1%

4.2. Performance of PGAC

The Matthews correlation coefficient (MCC) (Matthews, 1975) is typically employed to evaluate the performance on unbalanced data sets. Table 5 shows detailed results for individual subnucelar compartments that consist of MCC and the accuracy of leave-one-out cross-validation (LOOCV) when applied to SNL_35, using $m = 75$ selected GAC features. The MCC performances of PGAC were 0.749 and 0.668 for SNL_35L and SNL_35T, respectively, and the corresponding overall accuracies were 85.7% and 76.3%. Additionally, the accuracy for each single subnucelar compartment was correlated with the number of proteins within the compartment. For example, the training and testing accuracies for a single compartment with a large sample size, such as Chromatin, Nucleolus and Nuclear pore complex, are relatively high, compared with those for other compartments with small sample sizes.

PGAC, using SVM with $m = 78$ informative GAC features, was applied to the whole data set SNL_80 to compare it with existing prediction methods. Table 6 presents the results of the performance comparison in terms of the LOOCV accuracy. ProtLock (Cedano et al., 1997) and SVM (Vapnik, 1998) using the AAC features only have

Table 5
Accuracies (%) and MCC preformed on SNL_35.

Label	Compartment	LOOCV	Independent test
		SNL_35L	SNL_35T
1	Nuclear PML body	37.5 (0.422)	75.0 (0.517)
2	Nuclear speckle	84.8 (0.864)	75.0 (0.819)
	Chromatin	84.3 (0.709)	74.1 (0.573)
4	Nucleoplasm	55.0 (0.668)	30.0 (0.303)
5	Nucleolus	94.8 (0.833)	80.8 (0.674)
6	Heterochromatin	25.0 (0.496)	40.0 (0.627)
7	Nuclear envelope	71.0 (0.831)	62.5 (0.777)
8	Nuclear matrix	88.2 (0.937)	77.8 (0.946)
9	Nuclear pore complex	97.9 (0.976)	100.0 (1.000)
Overall accuracy % (MCC)		85.7 (0.749)	76.3 (0.668)

Table 7
The $m = 75$ informative GAC features selected from SNL_35L. The features in bold style are essential GO terms.

Rank by MED	GO term	Branch	MED	Rank by MED	GO term	Branch	MED
1	**GO:0016607**	C	312.9	39	V	AAC	36.9
2	**GO:0005643**	C	193.8	40	Y	AAC	35.8
3	**GO:0016363**	M	188.4	41	T	AAC	35.8
4	GO:0065002	B	176.5	42	GO:0000079	B	35.3
5	A	AAC	175.5	43	GO:0051457	B	35.3
6	**GO:0005730**	C	142.0	44	GO: 0000027	B	34.2
7	**GO:0005635**	C	90.3	45	GO:0018991	B	33.2
8	GO:0005789	C	81.7	46	GO: 00063 66	B	32.6
9	GO:0005637	C	81.1	47	**GO:0016605**	C	32.1
10	GO:0005886	C	73.0	48	D	AAC	31.5
11	R	AAC	72.5	49	GO:0030869	C	31.0
12	GO:0017056	M	66.6	50	K	AAC	31.0
13	N	AAC	62.3	51	GO: 0006917	B	29.9
14	**GO:0000792**	C	55.8	52	GO:0000480	B	28.8
15	GO:0000151	C	53.6	53	GO: 0000209	B	26.7
16	GO: 003 0674	M	52.6	54	H	AAC	24.0
17	GO: 0042692	C	52.6	55	L	AAC	24.0
18	GO:0005823	C	49.3	56	GO:0006378	M	22.4
19	GO:0001682	B	49.3	57	Q	AAC	18.6
20	GO: 0006612	B	48.2	58	GO:0000793	C	14.3
21	GO: 0003 702	M	47.2	59	GO: 0040007	B	11.6
22	GO: 003 0496	C	47.2	60	GO:0000775	C	11.1
23	GO:0008267	M	46.1	61	S	AAC	10.5
24	**GO:0005654**	C	46.1	62	I	AAC	10.0
25	GO:0005652	C	45.0	63	GO:0030686	M	9.4
26	GO: 0042254	B	43.9	64	GO: 0003 701	M	5.1
27	GO: 0007140	B	43.9	65	**GO:0000785**	M	5.1
28	GO: 0018024	M	43.4	66	M	AAC	5.1
29	GO:0045132	B	42.9	67	GO:0000781	M	4.0
30	GO:0005732	B	42.3	68	GO:0009987	B	3.5
31	GO: 0003 73 5	M	41.8	69	GO:0016586	C	3.0
32	GO:0046580	B	41.8	70	W	AAC	2.4
33	GO:0048666	B	40.7	71	E	AAC	1.9
34	GO:0008168	M	40.7	72	GO:0043596	C	1.9
35	G	AAC	39.6	73	P	AAC	0.8
36	GO:0001837	B	39.6	74	C	AAC	0.8
37	GO:0007005	B	38.5	75	F	AAC	0.3
38	GO:0005681	C	36.9				

accuracies 36.6% and 48.9%, respectively. PGAC has the highest accuracy of 81.1% and MCC = 0.691, which is better than ProtLock (Cedano et al., 1997) and SVM (Vapnik, 1998); it is also better than 55.6% of OET-KNN, using the Pse-AAC features (Shen and Chou, 2005) and 67.4% (MCC = 0.50) of Nuc-PLoc, which combines the Pse-AAC feature of a protein and its position-specific scoring matrix (Shen and Chou, 2007b). The results reveal that the informative GAC features are effective for predicting protein subnuclear localization.

4.3. Informative GAC Features

The quantified effectiveness of individual GAC features in prediction is valuable in characterizing subnuclear localization. Orthogonal experimental design with factor analysis can be used to examine the individual effects of GAC features according to the main effect difference (MED) (Ho et al., 2006; Tung and Ho, 2007). The factors are the parameters (GAC features) that manipulate the evaluation function, and a setting of a parameter is regarded as a level of the factor. In this study, the two levels of one factor are the inclusion and exclusion of the *i*-th GO term in the feature selection using IGA. The factor analysis can quantify 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. The GAC feature with the largest MED is the most effective in predicting subnuclear localization. The $m = 75$ informative GAC features, consisting of 20 AAC features, nine essential GO terms and 46 instructive GO terms are ranked by MED and described in Table 7. The 46 instructive GO terms, comprising 14 GO terms from the molecular function branch, 21 terms from the biological process branch, and 11 terms from the cellular component branch are denoted 14(M), 21(B) and 11(C), respectively.

Table 7 shows that the essential GO term GO:0016607 (Nuclear speckle), with the largest MED (=312.9), is the most effective feature among the 75 informative GAC features. The instructive (non-essential) GO term with the largest MED (=176.5) is GO:0065002 (Intracellular protein transport across a membrane, rank 4). Among the AAC features, amino acid A is the best with rank 5 and MED = 175.5. The top ten features are five essential GO terms, four instructive GO terms and one AAC feature, revealing that the essential GO terms and the instructive GO terms are more effective than the AAC features.

The GO terms near the root of the GO graphs are considered to be more generic while those near the leaves are more specific. Considering the high correlation among numerous GO terms in the GO graph, the feature selection algorithm GACmining has two advantages; (1) the consideration of a set of informative GO terms simultaneously, rather than individual GO terms, and (2) the reduction of the search space for candidate instructive GO terms by using the essential GO terms. The position relationships between instructive and essential GO terms in the GO graph indicate that at least 21(B) of the 46 instructive GO terms were not offspring of any essential GO term, because none of the nine (=7(C) + 2(M)) essential GO terms originated in the branch of biological process (Tables 1 and 7). This scenario reveals that the nine essential GO terms are indispensable features, but are not very effective when used alone. The 46 instructive GO terms are helpful in increasing predictive accuracy.

5. Conclusions

This study not only investigated the prediction of protein subnuclear localization by studying the features of GO annotation, but also developed a generalized method for deriving a GO-based feature set to be used with a specified classifier such as SVM to predict the functions or properties of protein sequences. A single-classifier prediction method PGAC was proposed to predict protein subnuclear localization. The SVM classifier used informative GAC features, consisting of 20 AAC features, nine essential GO terms and a number of instructive GO terms that were selected by the proposed feature mining algorithm, GACmining.

The application of PGAC to SNL_35 using 75 informative GAC features yielded training and test accuracies of 85.7% and 76.3%, respectively. PGAC yielded an LOOCV accuracy of 81.1%, which is better than that of Nuc-PLoc, 67.4%, the current best method applied to SNL_80. The high performance of PGAC arises mainly from the proposed algorithm, GACmining, an extension of GOmining for GO term selection. Both GOmining and GACmining performed well in selecting informative GO terms for predicting subcellular and subnuclear localization. Domain knowledge is embedded into the classifier design using the essential GO terms. We believe that the proposed approach to classifier design can be widely adopted in the sequence-based prediction of protein functions/characteristics using informative GO terms.

Acknowledgements

The authors would like to thank the National Science Council of Taiwan for financially supporting this research under the contract numbers NSC 97-2218-E-243-002, NSC 96-2628-E-009-141-MY3 and NSC 97-2627-B-009-005.

Contributions: W.L. Huang designed the system, implemented programs, participated in manuscript preparation and carried out the detail study. C.W. Tung, W.L. Huang and H.L. Huang designed the system and implemented programs. S.Y. Ho and W.L. Huang conceived the idea of this work. Additionally, S.Y. Ho supervised the whole project and participated in manuscript preparation. All authors have read and approved the final manuscript.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSIBLAST:a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32, D115–D119.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25–29.

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

Cai, Y.D., Chou, K.C., 2004. Predicting 22 protein localizations in budding yeast. Biochem. Biophys. Res. Commun. 323, 425–428.

Cai, Y.D., Zhou, G.P., Chou, K.C., 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J. Theor. Biol. 234, 145–149.

Carroll, S., Pavlovic, V., 2006. Protein classification using probabilistic chain graphs and the Gene Ontology structure. Bioinformatics 22, 1871–1878.

Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. J. Mol. Biol. 266, 594–600.

Chang, C.C., Lin, C.J., 2001. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chou, K.C., Shen, H.B., 2006a. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.

Chou, K.C., Shen, H.B., 2007a. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.

Chou, K.C., Shen, H.B., 2007b. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat. Protoc. 3, 153–162.

Chou, K.C., Shen, H.B., 2006b. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteome Res. 5, 1888–1897.

Chou, K.C., 2009. Automated prediction of protein attributes and its impact to biomedicine and drug discovery. In: Alterovitz, G., Benson, R., Ramoni, M.F. (Eds.), Automation in proteomics and genomics: an engineering case-based approach, Harvard-MIT interdisciplinary special studies courses, Chap. 5. Wiley & Sons, Ltd., West Sussex, UK, pp. 97–143.

Cocco, L., Manzoli, L., Barnabei, O., Martelli, A.M., 2004. Significance of subnuclear localization of key players of inositol lipid cycle. Adv. Enzyme Regul. 44, 51–60.

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

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

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

Huang, W.L., Chen, H.M., Hwang, S.F., Ho, S.Y., 2007a. Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. BioSystems 90, 405–418.

Huang, W.L., Tung, C.W., Ho, S.W., Hwang, S.F., Ho, S.Y., 2008. ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. BMC Bioinformatics 9, 80.

Huang, W.L., Tung, C.W., Huang, H.L., Hwang, S.F., Ho, S.Y., 2007b. ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. BioSystems 90, 573–581.

Lei, Z., Dai, Y., 2006. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. BMC Bioinformatics, 491–590.

Lewin, A., Grieve, I., 2006. Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. BMC Bioinformatics 7, 426.

Li, J., Liu, Q., Qiu, M., Pan, Y., Li, Y., Shi, T., 2006. Identification and analysis of the mouse basic/Helix-Loop-Helix transcription factor family. Biochem. Biophys. Res. Commun. 350, 648–656.

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.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405, 442–451.

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

Nanni, L., Lumini, A., 2006. An ensemble of K-local hyperplanes for predicting protein-protein interactions. Bioinformatics 22, 1207–1210.

Pierleoni, A., Martelli, P.L., Fariselli, P., Casadio, R., 2006. BaCelLo: a balanced subcellular localization predictor. Bioinformatics 22, 408–416.

Qian, Z., Cai, Y.D., Li, Y., 2006. A novel computational method to predict transcription factor DNA binding preference. Biochem. Biophys. Res. Commun. 348, 1034–1037.

Quinlan, J.R., 2003. C5.0 Online Tutorial. http://www.rulequest.com.

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

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.

Shen, H.B., Chou, K.C., 2007a. Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem. Biophys. Res. Commun. 355, 1006–1011.

Shen, H.B., Chou, K.C., 2007b. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering. Des. Select. 20, 561–567.

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

Tung, C.W., Ho, S.Y., 2007. POPI: Predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. Bioinformatics 23, 942–949.

Vapnik, V., 1998. Statistical Learning Theory. Wiley-Interscience, New York.

Wang, G.L., Dunbrack Jr., R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591.

Wolstencroft, K., Lord, P., Tabernero, L., Brass, A., Stevens, R., 2006. Protein classification using ontology classification. Bioinformatics 22, e530–538.

Yu, C.S., Chen, Y.C., Lu, C.H., Hwang, J.K., 2006. Prediction of protein subcellular localization. Proteins 64, 643–651.