J Comput Aided Mol Des (2013) 27:91–103
DOI 10.1007/s10822-012-9628-0

EuLoc: a web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC

Tzu-Hao Chang · Li-Ching Wu · Tzong-Yi Lee · Shu-Pin Chen · Hsien-Da Huang · Jorng-Tzong Horng

Received: 23 July 2012 / Accepted: 17 December 2012 / Published online: 3 January 2013

Abstract The function of a protein is generally related to its subcellular localization. Therefore, knowing its subcellular localization is helpful in understanding its potential functions and roles in biological processes. This work develops a hybrid method for computationally predicting the subcellular localization of eukaryotic protein. The method is called EuLoc and incorporates the Hidden Markov Model (HMM) method, homology search approach and the support vector machines (SVM) method by fusing several new features into Chou's pseudo-amino acid composition. The proposed SVM module overcomes the shortcoming of the homology search approach in predicting the subcellular localization of a protein which only finds low-homologous or non-homologous sequences in a protein subcellular localization annotated database. The proposed HMM modules overcome the shortcoming of SVM in predicting subcellular localizations using few data on protein sequences. Several features of a protein sequence are considered, including the sequence-based features, the biological features derived from PROSITE, NLSdb and Pfam, the post-transcriptional modification features and others. The overall accuracy and location accuracy of EuLoc are 90.5 and 91.2 %, respectively, revealing a better predictive performance than obtained elsewhere. Although the amounts of data of the various subcellular location groups in benchmark dataset differ markedly, the accuracies of 12 subcellular localizations of EuLoc range from 82.5 to 100 %, indicating that this tool is much more balanced than other tools. EuLoc offers a high, balanced predictive power for each subcellular localization. EuLoc is now available on the web at http://euloc.mbc.nctu.edu.tw/.

Keywords Subcellular localization · Protein function · Eukaryote · Support vector machine

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9628-0) contains supplementary material, which is available to authorized users.

T.-H. Chang
Graduate Institute of Biomedical Informatics,
Taipei Medical University, Taipei, Taiwan
e-mail: kevinchang@tmu.edu.tw

L.-C. Wu · J.-T. Horng
Institute of Systems Biology and Bioinformatics,
National Central University, Chung-Li 320,
Taiwan
e-mail: Richard@mail.sybbi.ncu.edu.tw

J.-T. Horng
e-mail: horng@db.csie.ncu.edu.tw

T.-Y. Lee (✉)
Department of Computer Science and Engineering,
Yuan Ze University, Chung-Li 320, Taiwan
e-mail: francis@saturn.yzu.edu.tw

S.-P. Chen · J.-T. Horng
Department of Computer Science and Information Engineering,
National Central University, Chung-Li 320, Taiwan
e-mail: deluxe1007@gmail.com

H.-D. Huang
Institute of Bioinformatics and Systems Biology,
National Chiao Tung University, Hsin-Chu 300, Taiwan
e-mail: bryan@mail.nctu.edu.tw

H.-D. Huang
Department of Biological Science and Technology,
National Chiao Tung University, Hsin-Chu, Taiwan

J.-T. Horng
Department of Biomedical Informatics, Asia University,
Wufeng 413, Taiwan

Background

Most eukaryotic proteins are encoded in the nuclear genome, synthesized in the cytosol, and must be targeted to the correct subcellular compartments before they can have their biological effects. Each subcellular compartment contains particular proteins, including enzymes, to enable it to carry out its biological function. In addition, as is well known, proteins may simultaneously exist at, or move between, two or more different subcellular locations. Proteins with multiple locations or dynamic feature of this kind are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery. Knowing the subcellular localization of a protein helps in determining its function, because subcellular localization yields valuable information on the interaction partners, function and potential roles of a protein in the cell [1–3]. In recent years, the field of proteomics has expanded rapidly, and very many protein sequences have been recorded in databases. Despite technological advances, experimenting on the subcellular localization of proteins is time-consuming. Therefore an efficient computational method for predicting protein subcellular localization is becoming increasingly important.

As presented in Table 1, various computational methods have been developed to predict subcellular localization [4–39], and many machine-learning approaches have been adopted. They include those of support vector machines (SVM) [6, 8–10, 12–16, 18, 19, 30, 40–47], the use of neural networks [4, 5], the use of Bayesian networks [7, 9], text classification [17], fuzzy-nearest neighbors algorithm [20] and the *K*-nearest neighbor classifier [21–27]. Additionally, Chou et al. [33] developed several web-servers iLoc-Euk, iLoc-Hum [35], iLoc-Plant [36], iLoc-Gpos [34], iLoc-Gneg [39] and iLoc-Virus [37] by using multi-layer KNN approach to cope with the multiple location problems in eukaryotic, human, plant, Gram-positive, Gram-negative, and virus proteins, respectively.

Most approaches use the SVM as the predictor since the SVM is very useful in classifying proteins with diverse sequences and has therefore been widely adopted. In particularly, Chou's [31] pseudo-amino acid composition (PseAAC) for the feature representation of biological sequences have been widely used [38, 48–63]. The predictive performance of SVM depends sensitively on the sizes and diversity of the protein sequences used [64]. The SVM method tends to generate feature vectors that push the hyper-plane towards the side with fewer data [65], commonly resulting in reduced predictive accuracy for a class with fewer samples or less diversity. Therefore, subcellular localization is often hard to predict accurately using few data about the relevant proteins.

Some approaches use the GO annotation of a protein for predicting its subcellular localization [21–27] based on an assumption that proteins mapped onto the GO database space would be clustered in a way better reflecting their subcellular localizations [2]. Some approaches depend on a sequence homology search against a protein database to obtain additional information of a protein, including text annotations, localization annotations and Gene Ontology (GO) annotations, for making further prediction of its subcellular localization [8–10, 16, 19]. Several studies have already indicated that there is a close relationship between sequence similarity and identity in both subcellular localization and the signal peptide cleavage sites [66, 67]. For instance, Nair and Rost's large-scale analysis [66] shows that the subcellular compartment of a protein can be accurately inferred if the close homologs of experimentally verified localization can be found using HSSP distance [66], a measure for sequence similarity accounting for pairwise sequence identity and alignment length. Despite the usefulness of homology search approach, which infers subcellular localization of a protein based on subcellular localization of homologous sequence, in predicting subcellular localization, it performs poorly if no homologous sequence is found [14]. Therefore, other computational methods should be adopted in predicting the subcellular localization of proteins with no homologs. As presented in Table 1, different methods consider different features of a protein. They include the sequence-based features that are derived from the protein sequences, including N-/C-terminal amino acid sequences [4, 12, 13, 15, 18], amino acid compositions [8, 10–15, 20], general n-peptide compositions [6, 14, 16] and amphiphilic pseudo amino acid [21, 23–27], the biological features that are derived from the physio-chemical properties of amino acids [8, 14] or obtained using the detection models which are provided in biological databases, such as PROSITE, NLSdb and Pam [7, 9, 11, 15].

This work presents a hybrid method, called EuLoc, to solve the aforementioned problem. It incorporates the Hidden Markov Model (HMM) method, homology search approach and the SVM method. EuLoc involves an SVM module to overcome the shortcoming of homology search approach in predicting the subcellular localization of the sequence which can only find low-homologous or non-homologous sequences in a protein subcellular localization annotated database, and involves an HMM module to overcome the shortcoming of SVM in predicting the subcellular localizations using few data on the relevant protein sequences. The SVM module considers many features of a protein sequence, including the sequence-based features, the biological features derived from PROSITE, NLSdb and Pfam, the post-transcriptional modification features and others. The total accuracy (TA) and location accuracy (LA)

Table 1 Comparison of features of various subcellular localization prediction tools

Tool name	Main method	Used features	Species: # of localizations	References
[29]	Covariant discriminant algorithm	Amino acid composition	Eukaryotes: 12	Chou and Elrod [29]
TargetP [4]	Neural network	N-terminal amino acid sequence	Plant: 4 Non-plant: 3	Emanuelsson et al. [4]
[31]	Covariant discriminant algorithm	Pseudo amino acid composition (PseAAC)	Eukaryotes: 12	Chou [31]
[30]	Covariant discriminant algorithm, SVM	Amino acid composition, functional domain composition	Eukaryotes: 12	Chou and Cai [30]
[5]	Neural network	Evolutionary and protein structure information	Eukaryotes: 4	Nair et al. [84]
PK method [6]	SVM	Gapped amino acid composition	Eukaryotes: 12	Park and Kanehisa [6]
PLST [7]	Bayesian network	InterPro motifs, specific membrane domain, co-occurrence of protem motifs/domain,	Human: 9	Scott et al. [7]
PSLpred [8]	PSI-BLAST, SVM	Sequence similarity, residues, dipeptides, physio-chemical properties	Gram-negative: 5	Bhasin et al. [8]
PSORTb [9]	Bayesian network, frequent subsequence based SVM, SCL-BLAST	Sequence similarity, PSORTb's SCL-BLAST module, PROSITE motifs and localization specific profiles, frequent subsequence, N-terminal signal peptide, membrane domain features	Gram-positive: 6 Gram-negative: 6	Gardy et al. [9]
LOCSVMPSI [10]	PSI-BLAST, SVM	Sequence similarity, position specific scoring matrix generated from profiles of PSI-BLAST, four-part amino acid composition	Eukaryotes: 4 (Swiss-Prot); 12 (PK data set)	Xie et al. [10]
pTARGET [11]	Pfam score, AAC score	Occurrence of protein functional domain, amino acid composition, Pfam motifs	Eukaryotes: 9	Guda [11]
MultiLoc [12]	Two-layer SVM	N-terminal targeting peptide, single anchor (SA), amino acid composition, motifs from PROSITE and NLSdb	Fungi: 9 Animal: 9 Plant: 10	Hoglund et al. [12]

BaCelLo [13]	SVMs in a decision tree	Whole sequence composition, the compositions of both the N- and C-termini	Fungi: 4;
			Animal: 4;
			Plant: 5
				Pierleoni et al. [13]

CELLOII [14]	Two-layer SVM	n-peptide, partitioned amino acid, g-gap dipeptide and local amino acid composition, physio-chemical properties	Eukaryotes: 12 Prokaryotes: 5	Hwang et al. [14]
SherLoc [15]	Two-layer SVM	N-terminal targeting peptide, single anchor (SA), amino acid composition, motifs from PROSITE and NLSdb, text from Pubmed abstract by its Swiss-Prot entry	Animal: 9 Plant: 10	Shatkay et al. [15]
PSLDoc [16]	PSI-BLAST, probabilistic latent semantic analysis (PSLA), SVM	Sequence similarity, gapped-dipeptides	Gram-negative: 5	Chang et al. [18]
[17]	Text classification	Synonyms from gene ontology (GO) and using GO hierarchy to generalize terms	Animal: 9 Plant: 10	Fyshe et al. [17]

ESLpred2 [18]	SVM	Evolutionary information, N-terminal sequence composition	Fungi: 4;
			Animal: 4;
			Plant: 5
				Garg and Raghava [18]

ProLoc-GO [19]	GO mining—an intelligent genetic algorithm with IGA and SVM classier	GO terms derived from the result of similar sequences of BLAST	Human: 12 Eukaryotes: 16	Huang et al. [19]

Table 1 continued

Tool name	Main method	Used features	Species: # of localizations	References
[20]	Fuzzy *k*-nearest neighbors algorithm	Amino acid composition	Eukaryotes: 4 Prokaryotes: 3	Nasibov and Kandemir-Cavas [20]
Cell-PLoc [21]-a package includes Euk-mPLoc [22] Hum-mPLoc [23] Plant-PLoc [24] Gpos-PLoc [25] Gneg-PLoc [26] Virus-PLoc [27]	An ensemble classifier by fusing K-nearest neighbor classifiers	GO annotation, PseAAC	Eukaryotes: 22 Human: 14 Plant: 11 Gram-positive: 5 Gram-negative: 8 Virus: 7	Chou et al. [22–27]
iLoc-Euk [33] ILoc-Hum [35] iLoc-Plant [36] iLoc-Gpos [34] iLoc-Gneg [39] iLoc-Virus [37]	Multi-layer KNN classifier, KNN, BLAST, PSI-BLAST	Sequence similarity, GO annotation, PseAAC, position-specific scoring matrix (PSSM)	Eukaryotes: 22 Human: 14 Plant: 12 Gram-positive: 4 Gram-negative: 8 Virus: 6	Chou et al. [33–37, 39]

of EuLoc are 90.5 and 91.2 %, respectively, and so the predictive performance is better than in other studies. The accuracies of 12 subcellular localizations of EuLoc range from 82.5 to 100 %, indicating that it is much more balanced than other tools. Therefore, EuLoc has a high and balanced capacity to predict each subcellular localization.

Methods

According to a recent comprehensive review [3], to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

Materials

The benchmark dataset used in the proposed method was taken from Park and Kanehisa [6], and is called the PK dataset. To remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25 % was imposed in [2, 21, 68, 69], to exclude those proteins from the benchmark datasets 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 benchmark dataset used here was taken from Park and Kanehisa [6] and that the main purpose of this paper is to show a different prediction approach. The subcellular localization information pertaining to the PK dataset was obtained by a keyword search of the CC-filed annotation in the Swiss-Prot database [70, 71]. The PK dataset contains 7,579 eukaryotic protein sequences in 12 subcellular localizations. They are chloroplast, cytoplasmic, cytoskeleton, endoplasmic reticulum (ER), extracellular, Golgi apparatus, lysosomal, mitochondrial, nuclear, peroxisomal, plasma membrane and vacuolar proteins. As presented in Additional file 1, the amounts of data for the subcellular location groups vary markedly.

To obtain other protein sequences with known subcellular localization, the protein sequences from the Swiss-Prot database release 53 were collected, and the annotated subcellular localization in the CC-fields was extracted. A total of 158,596 protein sequences exhibited known subcellular localization, and 87,675 protein sequences were eukaryotic. These sequences were used in the homological search process.

The hybrid method consists of three modules—the HMM module, the homology search module and the SVM module. Figure 1 is the analysis flowcharts of the three modules, respectively. Since the PK dataset was lack of the information of multiple locations, and our method did not deal with the case of multiplex proteins. The analytical and predictive processes are described below.

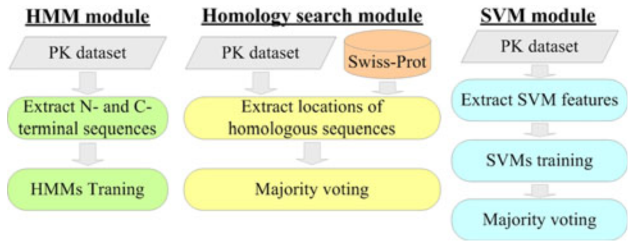

Fig. 1 Analysis flowchart of HMM module, homological module and SVM module

Feature extraction

To develop a powerful predictor for a protein system, one of the keys is to formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted [3]. To realize this, the concept of pseudo amino acid composition (PseAAC) was proposed [31] to replace the simple amino acid composition (AAC) for representing the sample of a protein. Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems (see, e.g., [50, 53–63, 72–74]). For various different modes of PseAAC, see [75]. According to a recent comprehensive review [3], the general form of PseAAC can be formulated as (see Eq. 6 of [3]):

$$P = [\Psi_1 \Psi_2 \cdots \Psi_u \cdots \Psi_\Omega]^T$$

where T is a transpose operator, while the subscript Ω is an integer and its value as well as the components Ψ_1, Ψ_2, … will depend on how to extract the desired information from the amino acid sequence of P. Here, we are to use a different feature extraction method to formulate the PseAAC.

The targeting signals in the N-terminal or C-terminal regions help in translocating some proteins into subcellular components [76–79], and they vary in length (from 15 to 70 amino acids) and primary sequence [80]. To analyze the characteristics of these regions, the N-terminal and C-terminal residues of each sequence in PK dataset are extracted using the different lengths, which are 20, 40, 60, 80, 100, 120, for further HMM analysis. In the homology search module, PSI-BLAST [81] searches for the sequences that are homologous to a PK dataset sequence against Swiss-Prot database release 53. Several representative keywords are generated for 12 subcellular localizations (Additional file 2), and used to search for the CC-field of a homologous sequences to extract the corresponding subcellular localizations. For example, if the keyword "plastid" is present in the CC-field, then the subcellular localization of the homologous sequence is annotated as "Chloroplast". Multiple subcellular localization annotations are acceptable for a single sequence. These annotated subcellular localizations of sequences that are homologous to a PK dataset sequence are extracted for further process of predicting subcellular localization.

Numerous sequence-based features are extracted from a protein sequence to construct 10 SVM modules, which are amino acid composition (ACC) of whole sequence (denoted as AAC module), ACC of four-part sequence (denoted as X4 module), ACC of N-terminal region, C-terminal region and centered window region (denoted as NCC module), and n-gapped dipeptide composition (n = 0, 1, 2, 3, 4, 5, 6; denoted as D0, D1, D2, D3, D4, D5 and D6 module, respectively). To compute the four-part sequence AAC, a protein sequence is separated into four parts of equal length and the AAC of each part is calculated. Therefore, four-part sequence AAC encodes a protein sequence as an 80 dimensional vector. The N-terminal AAC encodes a protein as a 60 dimensional vector based on the AACs of the first 40, 60 and 100 residues. The C-terminal AAC encodes a protein as a 60 dimensional vector based on the AAC of the last 50, 80 and 100 residues. The centered window AAC encodes a protein as a 100-dimensional vector, based on the AAC of the centered windows using 13, 15, 17, 19 and 21 residues. Thus, when using the general formulation of PseAAC to incorporate sequence-based features, we have $\Omega = 20, 400, 400, 400, 400, 400, 400, 400, 80$ and 220 for SVM module of AAC, D0, D1, D2, D3, D4, D5, D6, X4 and NCC, respectively.

Numerous biological features are used to construct 6 SVM modules, which are physio-chemical properties of the amino acid (denoted as Physio-chemical module), post-translational modification (PTM) of amino acid (denoted as PTM module), protein domains obtained from Pfam [82] (denoted as Pfam module), and motifs obtained from PROSITE [83] and NLSdb [84] (denoted as Motif module). Additionally, the biological features of PROFEAT [85], such as the structural and physio-chemical features, are also incorporated (denoted as PROFEAT300 and PROFEATF56 module). The physio-chemical properties of an amino acid are used to encode a protein as a 40-dimensional vector based on its classification (Additional file 3). Post-translational modification (PTM) is the chemical modification of a protein after its translation. It is one of the later steps in the biosynthesis of many proteins, and is related to protein functions and subcellular localizations [86–90]. Several PTM prediction tools, which are NetAcet [91], MASA [92], NetNGlyc [93], NetOGlyc [94], NetPhos [95], and SulfoSite [96], are incorporated to predict the substrates of N-acetyltransferase, the methylation site, the N-glycosylation sites, the mucin-type O-glycosylation, the phosphorylation sites and the sulfation sites of a protein sequence, respectively. The protein domains from the Pfam database and the sequence patterns from PROSITE and NLSdb are adopted to detect the biological domains and motifs of a protein. Thus, when using the general formulation of PseAAC to incorporate biological features, we have $\Omega = 300, 169, 825, 900, 40$ and 27 for SVM module of PROFEAT300, PROFEATF56, Motif, Physio-chemical and PTM, respectively. These sequence-based features and biological features are extracted for further SVM analysis.

Model training and evaluation

In the HMM module, the HMMER [97] is employed to build numerous HMMs with residues of different lengths in the N-terminal and the C-terminal region. Following the performance evaluation, three discriminative HMMs, which are constructed with 60, 100 and 100 residues, are found in the C-terminus of the cytoskeleton, the Golgi apparatus and the vacuole, respectively. They are used to predict the corresponding subcellular localization. In the homology search module, the subcellular localization of a protein is predicted using majority voting on the subcellular localizations of its three most homologous sequences. Since 4,826 proteins of PK dataset are recorded in Swiss-Prot database release 53, the subcellular localization of identical sequence of a query protein is not collected for voting. The iteration parameter is set to one and the E-value is set to 0.001 to perform PSI-BLAST. In the SVM module, the LIBSVM [98] package is adopted and the Radial Basis Function (RBF) kernel employed to construct 16 SVM models with sequence-based features and biological features. Ten sequence-based SVM models are constructed based on the features of the AAC of whole sequence, the n-gapped dipeptide composition, the AAC of the four-part sequence, and the AAC of the N-terminal region, the C-terminal region and the centered window region. Between zero and six gaps are used in the dipeptide composition and therefore seven SVM models are generated. The AAC from the N-terminal region, the C-terminal region and the centered window region are combined and incorporated into a single SVM model. The other six biological based SVM models are constructed using the physio-chemical properties, the PROFEAT features, the PTM features, the domains from Pfam [82], and the motifs from PROSITE [83] and NLSdb [84]. Some biological features obtained from PROFEAT are incorporated into two SVM models. Finally, the SVM module predicts the subcellular localization of a protein using majority voting on the predicted subcellular localizations of 16 SVM models.

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test. However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset. The reasons are as follows. (1) For the independent dataset test, although all the samples used to test the predictor are outside the training dataset used to train it so as to exclude the "memory" effect or bias, the way of how to select the independent samples to test the predictor could be quite arbitrary unless the number of independent samples is sufficiently large. This kind of arbitrariness might result in completely different conclusions. For instance, a predictor achieving a higher success rate than the other predictor for a given independent testing dataset might fail to keep so when tested by another independent testing dataset. (2) For the subsampling test, the concrete procedure usually used in literatures is the fivefold, sevenfold or tenfold cross-validation. The problem with this kind of subsampling test is that the number of possible selections in dividing a benchmark dataset is an astronomical figure even for a very simple dataset, as elucidated in [21] and demonstrated by Eqs. 28–30 in [3] Therefore, in any actual subsampling cross-validation tests, only an extremely small fraction of the possible selections are taken into account. Since different selections will always lead to different results even for a same benchmark dataset and a same predictor, the subsampling test cannot avoid the arbitrariness either. A test method unable to yield a unique outcome cannot be deemed as a good one. (3) In the jackknife test, all the samples in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining samples. During the process of jackknifing, both the training dataset and testing dataset are actually open, and each sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem as mentioned above for the independent dataset test and subsampling test can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset. Accordingly, the jackknife test has been increasingly and widely used by those investigators with strong math background to examine the quality of various predictors (see, e.g., [49–53, 72, 99, 100]). However, to reduce the computational time, we adopted the independent testing dataset cross-validation in this study as done by many investigators with SVM as the prediction engine.

In this work, the same validation procedures are used to determine predictive performance as are applied in earlier works [6, 10, 14]. The performance of SVM models are evaluated by five-fold cross-validation. Specificity (SP), sensitivity (SN), total accuracy (TA) [6], location accuracy (LA) [6], and Matthew's correlation coefficient (MCC) are utilized to evaluate the performance of classification. They are defined as

$$\mathrm{SN} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN});\ \mathrm{SP} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP});\mathrm{TA} = \frac{\sum_{i=1}^{k} T_i}{N};\mathrm{LA} = \frac{\sum_{i=1}^{k} P_i}{k}$$

$$\text{where } P_i = \frac{T_i}{n_i};$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}},$$ where

TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives [101]; N is the total number of proteins in the data set ($N = 7{,}579$); k is the number of subcellular locations ($k = 12$); n_i is the number of proteins in each subcellular localization i, and T_i is the number of correctly predicted proteins (true positives) in each subcellular localization i. The value of MCC is one for a perfect prediction, zero for a completely random prediction and −1 for a perfectly inverse correlation.

Predictive process

As presented in Fig. 2, the prediction flow of EuLoc involves the HMM module, the homology search module and the SVM module. First, the HMM module is applied to detect the subcellular localization of the cytoskeleton, Golgi apparatus or vacuole. If the HMM module cannot recognize the sequence, then the sequence is dispatched to the homology search module. If no homologous sequence with known subcellular localization is identified by the homology search module, then the sequence is dispatched to the SVM module for final prediction.

Results

To find the discriminative HMM of each subcellular localization, numerous HMMs with N-terminal and C-terminal sequences of different lengths are constructed in the HMM module. Table 2 presents the predictive performance and collected region of the most discriminative HMM of each subcellular localization. For example, the best discriminative HMM of the cytoskeleton and endoplasmic reticulum (ER) are constructed with the last 60 C-terminal residues and the first 100 N-terminal residues of protein sequences, respectively. The HMMs of cytoskeleton, Golgi apparatus and vacuole are the most discriminative, with MCC values of 0.99, 0.99 and 1, respectively. Since they have high discriminative ability, these three HMMs are adopted as the predictors for cytoskeleton, Golgi apparatus and vacuole, and incorporated into the first predictive process herein.

Table 3 presents the predictive performance of the homology search module. The TA of this module is 86.1 % when the iteration parameter is set to one and the E-value is set to 0.001 in PSI-BLAST. If the iteration parameter is increased to 2, then the TA declines to 82.3 % (Additional file 4). Therefore, the iteration parameter is set to one, and for 398 protein sequences of the PK dataset, no homologous sequence can be found with a known subcellular localization. The analysis results show that cytoplasmic and extracellular prediction reaches greatest sensitivities of 90.7 and 90.9 %, respectively, in the homologous search module.

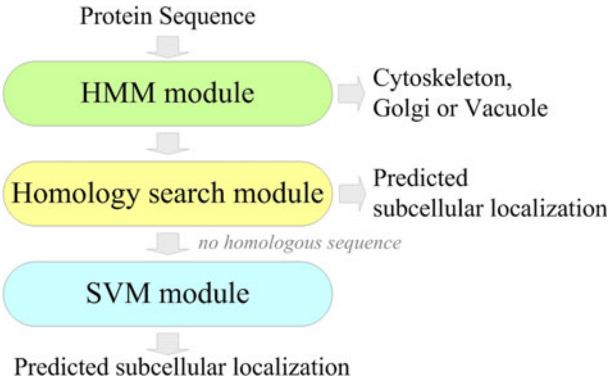

Fig. 2 Prediction flowchart of EuLoc

Table 4 presents the predictive performances of ten sequence-based SVM models. AAC stands for amino acid composition, D for dipeptide composition, where the number after the D denotes the length of the gap between the two residues; X4 stands for the AAC of the four-part sequence, and NCC stands for the AAC of the N-terminal region, the C-terminal region and the centered window region. Analysis result shows that the TA value of AAC is 67.6 % and the TA values of gapped-dipeptide composition are about 72 %. X4 has the best predictive performance with 75.5 % TA in sequence-based SVM models.

Table 5 presents the predictive performance of six biological based SVM models. Three hundred top-ranked PROFEAT features are collected to construct PROFEAT300SVM because the overall predictive performance reaches greatest accuracy when these features are used to construct SVM model (Additional file 5). These top-ranked features are extracted by using the WEKA [102]. Two PROFEAT families, F5 and F6, are related to the structural and physio-chemical features, and are used to build the PROFEATF56 SVM model. The TA values of PROFEAT300 and PROFEATF56 are 71.8 and 68.9 %, respectively. The TA values of the motif SVM model, the Pfam SVM model, the physio-chemical property SVM model and the PTM SVM models are 62.5, 70.6, 68.3 and 45.3 %, respectively. Although the overall accuracy of the PTM SVM model is not high as that of the other models, the post-translational modification (PTM) features seem to be more strongly related to the nuclear and plasma

Table 2 Predictive performance of the most discriminative HMM of each subcellular localization and the corresponding region for constructing HMM

Subcellular localization (# of proteins)	Region	Number of residues	MCC	Sensitivity (%)	Specificity (%)
Chloroplast (671)	N-terminus	60	0.49	41	98
Cytoplasmic (1,241)	–	–	–	–	–
Cytoskeleton (40)	C-terminus	60	0.99	98	100
ER (114)	N-terminus	100	0.52	57	99
Extracellular (861)	N-terminus	60	0.70	81	95
Golgi apparatus (47)	C-terminus	100	0.99	98	100
Lysosomal (93)	N-terminus	100	0.81	66	100
Mitochondrial (727)	N-terminus	40	0.30	35	94
Nuclear (1,932)	N-terminus	40	0.29	51	79
Peroximal (125)	–	–	–	–	–
Plasma membrane(1,674)	N-terminus	60	0.47	41	96
Vacuole (54)	C-terminus	100	1	100	100

For example, the best discriminative HMMs of cytoskeleton and ER are established from the last 60 C-terminal residues and the first 100 N-terminal residues of the protein sequences, respectively. The contents in cytoplasmic and peroximal are recorded as dashes since no discriminative HMM that corresponds with either of these two subcellular localizations is found

Table 3 Predictive performance of homological search module using iteration parameter of one and E-value of 0.001 in PSI-BLAST

Subcellular localization (# of proteins)	Sensitivity (%)	Specificity (%)	MCC
Chloroplast (671)	87.2	98.7	0.86
Cytoplasmic (1241)	90.7	96.9	0.85
Cytoskeleton (40)	10.0	1	0.32
ER (114)	89.5	99.8	0.89
Extracellular (861)	90.9	98.2	0.87
Golgi apparatus (47)	19.1	1	0.44
Lysosomal (93)	87.1	99.8	0.87
Mitochondrial (727)	80.7	98.5	0.81
Nuclear (1932)	88.2	95.4	0.83
Peroximal (125)	84.8	99.8	0.87
Plasma membrane (1674)	84.3	95.7	0.80
Vacuole (54)	48.1	1	0.69
Total accuracy, TA	86.1		
Location accuracy, LA	71.7		

membrane than other subcellular localizations and provide useful information that can be used to predict the proteins of these localizations. Table 6 presents the predictive performance of the SVM module which determines the subcellular localization of a protein using majority voting on the predicted subcellular localizations of 16 SVM models. The TA and LA values of the SVM module are 82.6 and 60.6 %, respectively, which are much higher than those of individual SVM models.

Finally, EuLoc incorporates the HMM module, the homology search module and the SVM module into an integrated predictive process, and the TA is then increased to 90.5 % (Table 7). In prediction, EuLoc outperforms previous methods applied to the PK dataset, such as CELLO II (90.3 %) [14], the PK method (78.2 %) [6] and LOCSVMPSI (83.5 %) [10]. The predictive accuracy of EuLoc when applied to plasma membrane (91 %) is a little weaker than the CELLO II (96.1 %), the LOCSVMPSI (94.7 %) or the PK method (92.2 %). However, EuLoc remarkably improves upon the predictive accuracy in most subcellular localizations, especially in the cytoskeleton (97.5 %), the Golgi apparatus (97.9 %), the peroximal (84.8 %) and the vacuole (100 %). The predictive accuracies of 12 subcellular localizations of EuLoc range from 82.5 to 100 %, and the LA of EuLoc is 91.2 %. This value is much better than those of CELLO II (83.4 %), the LOCSVMPSI (67.5 %) or the PK method (57.8 %). These results show that EuLoc has a favorable and balanced predictive capacity in each subcellular localization.

Discussion

Since the difference among the data sizes of the subcellular location groups in the PK dataset vary greatly (Additional file 1), by up to 48 times, both TA and LA must be considered in the evaluation of the predictive performance. TA has been extensively utilized to measure predictive performance. However, TA can be easily optimized at the cost of accuracy for small groups. Therefore, LA is used to balance predictive performance between large and small groups.

As stated above, the predictive performance of SVM depends sensitively on the sizes and diversity of protein

Table 4 Predictive performance of ten sequence-based SVM models

Subcellular localization (# of proteins)	Sensitivity (%)									
	AAC	D0	D1	D2	D3	D4	D5	D6	X4	NCC
Chloroplast (671)	46.8	64.2	61.8	67.2	65.6	63.6	63.9	62.3	68.9	66.0
Cytoplasmic (1241)	66.7	61.2	60.9	63.5	66.1	64.5	63.0	65.4	69.9	70.4
Cytoskeleton (40)	22.5	67.5	62.5	67.5	70.0	70.0	70.0	62.5	57.8	27.5
ER (114)	13.2	50.0	53.5	45.6	49.1	44.7	50.8	44.7	52.6	36.8
Extracellular (861)	67.3	74.6	76.2	79.7	78.7	76.7	77.5	78.3	77.7	80.4
Golgi apparatus (47)	0.0	19.1	10.6	12.8	17.0	12.8	10.6	8.5	38.3	0.0
Lysosomal (93)	28.0	57.0	58.1	49.5	57.0	54.8	55.9	49.4	54.8	32.3
Mitochondrial (727)	32.7	45.8	45.0	43.7	43.2	40.4	41.0	42.5	56.4	58.5
Nuclear (1932)	84.0	81.5	80.4	82.3	82.5	81.6	81.2	83.2	82.4	83.7
Peroximal (125)	0.0	29.6	32.0	26.4	25.6	29.6	21.6	24.8	28.8	12.8
Plasma membrane (1674)	89.0	90.1	90.3	89.8	89.1	89.5	89.6	88.9	90.7	84.5
Vacuole (54)	0.0	37.0	25.9	25.9	25.9	20.4	24.1	27.8	24.1	7.4
Total accuracy, TA	67.6	71.9	71.5	72.8	72.9	71.8	71.6	72.3	75.5	73.5
Location accuracy, LA	37.5	56.5	54.8	54.5	55.8	54.1	54.1	53.2	58.5	46.7

Table 5 Predictive performance of six biological based SVM models

Subcellular localization (# of proteins)	Sensitivity (%)					
	PROFEAT300	PROFEATF56	Motif	Pfam	Physio-chemical property	PTM
Chloroplast (671)	63.9	60.8	36.2	53.8	40.5	27.6
Cytoplasmic (1241)	69.5	68.3	55.1	75.3	64.8	42.2
Cytoskeleton (40)	35.0	27.5	32.5	5.0	30.0	10.0
ER (114)	31.8	25.4	59.6	49.1	14.9	6.1
Extracellular (861)	88.4	86.3	59.9	71.2	63.9	32.9
Golgi apparatus (47)	10.6	10.6	14.9	36.2	0.0	2.1
Lysosomal (93)	19.4	15.1	43.0	58.1	20.4	6.5
Mitochondrial (727)	66.2	60.3	34.0	43.2	34.0	15.3
Nuclear (1932)	81.3	78.6	72.9	75.0	81.8	59.4
Peroximal(125)	24.0	15.2	43.2	55.2	0.0	2.4
Plasma membrane (1674)	73.5	70.7	86.2	86.7	86.7	69.4
Vacuole (54)	9.2	1.9	31.5	48.2	3.7	0.0
Total accuracy, TA	71.8	68.9	62.5	70.6	68.3	45.3
Location accuracy, LA	47.7	43.4	47.4	54.7	36.7	22.8

sequences used [64]. Therefore, subcellular localization is commonly hard to predict well when few data on the proteins are available. As presented in Table 7, the three previous SVM methods, CELLOII, the PK method and LOCSVMPSI, perform subcellular localization using many data on the protein sequences with an accuracy of more than 90 %, but subcellular localization using fewer data on the protein sequences can be even under 20 % accurate. The SVM method is apparently optimized for large subcellular location groups in the PK dataset, such as the plasma membrane (22 % of the PK dataset) and the nuclear (25 %), at the cost of accuracy for small subcellular location groups, such as the cytoskeleton (0.5 %), the Golgi apparatus (0.6 %) and the vacuole (0.7 %). To solve with this problem, this work presents a hybrid method to reduce the trade-off between TA and LA. The presented method is more balanced than earlier methods. For example, although the cytoskeleton is a small subcellular location group in the PK dataset, which contains only the 40 protein sequences, the accuracy of EuLoc when applied to the cytoskeleton is 97.5 %.

Some proteins are well known to be translocated into the subcellular components because of the targeting signals in the N-terminal or C-terminal regions [76–79]. The targeting signals are pieces of information, which are contained in a polypeptide chain or in a fold protein, to enable proteins to be transported to the suitable subcellular component. Therefore, numerous HMMs with sequences of

Table 6 Predictive performance of SVM module which determines the subcellular localization of a protein using majority voting on the predicted subcellular localizations of 16 SVM models

Subcellular localization (# of proteins)	Sensitivity (%)	Specificity (%)	MCC
Chloroplast (671)	79.7	98.6	0.81
Cytoplasmic (1241)	80.7	96.5	0.78
Cytoskeleton (40)	67.5	99.9	0.78
ER (114)	52.6	99.8	0.65
Extracellular (861)	89.4	97.6	0.84
Golgi apparatus (47)	6.4	99.9	0.16
Lysosomal (93)	55.9	99.8	0.69
Mitochondrial (727)	62.3	98.0	0.67
Nuclear (1932)	90.3	93.9	0.82
Peroximal (125)	28.0	99.8	0.45
Plasma membrane (1674)	93.6	94.6	0.85
Vacuole (54)	20.4	99.9	0.39
Total accuracy, TA	82.6		
Location accuracy, LA	60.6		

different lengths in the N-terminal or C-terminal regions are established for each subcellular localization to analyze these pieces of information. Three discriminative HMMs are established using the C-terminal regions of the cytoskeleton, the Golgi apparatus and the vacuole. Additional file 6 presents the sequence logo of these regions, generated using WebLogo [103, 104]. Information from these regions greatly increases the predictive capacity associated with these three subcellular localizations, as presented in Table 2, and the MCC values of the HMM of cytoskeleton, Golgi apparatus and vacuole are 0.99, 0.99 and 1, respectively. Due to their high discriminative abilities, they are incorporated into the first predictive process to improve low predictive accuracy of SVM module in these three small subcellular location groups, which are cytoskeleton (67.5 %), golgi apparatus (6.4 %) and vacuole (20.4 %). Only one sequence in the cytoskeleton and one in the Golgi apparatus fail to be detected because they are shorter than the corresponding HMM building lengths of 60 and 100.

Table 4 indicates that the proteins in various subcellular localizations may be associated with different sequence-based features. The proteins in the nucleus (84 %) and the plasma membrane (89 %) seem to be more strongly related to the AAC of the whole sequence, and the proteins in cytoplasmic (70.4 %), extracellular (80.4 %) and mitochondrial (58.5 %) seem to be more strongly related to the AAC of the N-terminal region, the C-terminal region and the centered window region.

As presented in Table 5, the predictive performance of the biological based SVM models is poorer than that of the sequence-based SVM models—especially the motif SVM model and the PTM SVM model—because numerous proteins cannot be identified by any PROSITE, NLSdb or PTM features, or only a few such features on these proteins can be identified, increasing the difficulty of SVM prediction.

Only 30 % of nuclear proteins are estimated to have an NLS [105]. In the PK dataset, 1,335 out of 1,932 nuclear protein sequences matched the nuclear localization signal from NSLdb, while only 82 nonnuclear protein sequences matched this signal. Therefore, the nuclear localization signal is a very useful feature for distinguishing nuclear sequences from other subcellular localization sequences. Another discriminative signal is the ER retention signal from PROSITE. Of 144 ER protein sequences in the PK dataset, 53 matched the ER retention signal, and no other

Table 7 Comparison of predictive performance of proposed method with those of other predictive tools using PK dataset

Subcellular localization (# of proteins)	Predictive accuracy (%)			
	Our method	CELLO II (Hybrid) [14]	LOCSVMPSI [10]	PK method [6]
Chloroplast (671)	87.5	90.0	76.5	72.3
Cytoplasmic (1241)	**91.4**	84.4	76.4	72.2
Cytoskeleton (40)	**97.5**	80.0	60.0	58.5
ER (114)	**89.5**	80.7	61.4	46.5
Extracellular (861)	93.0	93.5	89.4	78.0
Golgi apparatus (47)	**97.9**	74.5	46.8	14.6
Lysosomal (93)	87.1	87.1	62.4	61.8
Mitochondrial (727)	**82.5**	80.5	68.2	57.4
Nuclear (1932)	92.3	94.5	91.5	89.6
Peroximal (125)	**84.8**	74.4	41.6	25.2
Plasma membrane (1674)	91.0	96.1	94.7	92.2
Vacuole (54)	**100**	64.8	40.7	25.0
Total accuracy, TA	**90.5**	90.3	83.5	78.2
Location accuracy, LA	**91.2**	83.4	67.5	57.9

Numbers in bold font represent that our method performs the best predictive accuracy in specific subcellular localization as compared with other methods subcellular localization protein matched this signal. The ER retention signal provides useful information, and so the accuracy of the prediction of ER using the motif SVM model exceeds that using other biological based SVM models.

A protein domain is a part of a protein sequence. It is a structure with biological functions, and can exist independently of the rest of the protein sequence. Several protein domains, such as the transmembrance domain, have been experimentally proven to be involved in protein translocation or to be necessary for its retention in particular location [106–108]. Therefore, the protein domains of Pfam are utilized herein to construct the Pfam SVM model. In this model, the subcellular localizations of cytoplasmic (75.3 %), extracellular (71.2 %), lysosomal (58.1 %), plasma membrance (86.7 %) seem to be more strongly related to Pfam domains.

In the PTM SVM model, the nuclear (59.4 %) and plasma membrane (69.4 %) seem to be more strongly related to PTM features. As suggested in previous studies, PTM is critically involved in the translocation of proteins to different subcellular locations, especially in the plasma membrane [86, 89, 109]. Protein localized in the nucleus is associated with phosphorylation at serine and threonine residues [110]. In Swiss-Prot release 53, 2,898 out of 6,253 nuclear proteins were annotated as having PTM sites, and 1,791 and 296 proteins were annotated as having phosphoserine and phosphothreonine sites, respectively. As presented in Additional file 7, phosphorylation occurs in more than half of all nuclear PTM proteins.

To determine whether the low predictive performance of the PTM SVM model is cause of the predictive power of the PTM prediction tool, another PTM SVM model was built using the PTM sites recorded in dbPTM [110]. Of 7,579 proteins in the PK dataset, 1,436 were recorded as having PTM data. The overall accuracy of the this PTM SVM model is 51.4 %, which is a little improved than the PTM SVM model was built using the predicted PTM sites derived from PTM prediction tools (45.3 %). However, the model is still not good as other models. As presented in Additional file 8, only one fifth of proteins are annotated as having PTM features, and therefore the number of PTM features may be too few for prediction. Interestingly, 69.9 and 66.7 % of lysosomal and vacuole proteins, respectively, are recorded as having PTM sites. Despite the low accuracy of the individual biological based SVM model, each model can somewhat increase the overall accuracy of the SVM module.

Conclusions

This work presents a hybrid method for the computational prediction of the subcellular localization of eukaryotic protein, called EuLoc. It incorporates the HMM method, homology search approach and the SVM method with sequence-based features and biological features. The SVM module overcomes the shortcoming of the homology search approach in predicting the subcellular localization of a sequence which can only find low-homologous or non-homologous sequences in a protein subcellular localization annotated database, and the HMM modules overcomes the shortcoming of SVM in predicting subcellular localizations using few data on protein sequences. Different features in a protein sequence are considered, including the amino acid composition, the motifs from PROSITE and NLSdb, the domains from Pfam and the PTM features. The TA and LA of EuLoc are 90.5 and 91.2 %, respectively, revealing that EuLoc is more accurate than CELLOII, the PK method and LOCSVMPSI. Although the amounts of data in the subcellular location groups in the PK dataset vary greatly, the accuracies of 12 subcellular localizations in the proposed method range from 82.5 to 100 %, and so the method is much more balanced than those in previous works. In some subcellular localizations, such as that in plasma membrane, the proposed method does not predict as accurately as other approaches. However, the proposed method improves the predictive accuracy for most of subcellular localizations—especially in the cytoskeleton, Golgi apparatus, the peroximal and the vacuole. The LA of the proposed method (91.2 %) markedly exceeds that of CELLO II (83.4 %), the PK method (57.8 %) and the LOCSVMPSI (67.5 %). Results of this study demonstrate that the proposed hybrid method has a favorable and balanced predictive ability in each subcellular localization. Hence EuLoc is a useful tool for computationally predicting the subcellular localization of eukaryotic protein. It is our intention to enhance EuLoc in the future by extending our method to deal with more subcellular localizations, such as the 22 subcellular localizations of Euk-mPLoc [22], for providing a more precise and comprehensive prediction of protein subcellular localization.

Acknowledgments The authors would like to thank the National Science Council of the Republic of China, No. NSC 101-2628-E-155-002-MY2, 99-2221-E-008-083-MY3, NSC 101-2311-B-009-003-MY3 and NSC 100-2627-B-009-002. This work was supported in part by the UST-UCSD International Center of Excellence in Advanced Bioengineering sponsored by the Taiwan National Science Council I-RiCE Program under Grant Number: NSC 101-2911-I-009-101, and Veterans General Hospitals and University System of Taiwan (VGHUST) Joint Research Program under Grant Number: VGHUST101-G5-1-1. This work was also partially supported by MOE ATU.

References

1. Nakai K (2000) Adv Protein Chem 54:277
2. Chou KC, Shen HB (2007) Anal Biochem 370:1
3. Chou KC (2011) J Theor Biol 273:236

4. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) J Mol Biol 300:1005
5. Nair R, Rost B (2003) Proteins 53:917
6. Park KJ, Kanehisa M (2003) 19:1656
7. Scott MS, Thomas DY, Hallett MT (2004) Genome Res 14:1957
8. Bhasin M, Garg A, Raghava GP (2005) Bioinformatics 21:2522
9. Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS (2005) Bioinformatics 21:617
10. Xie D, Li A, Wang M, Fan Z, Feng H (2005) Nucleic Acids Res 33:W105
11. Guda C (2006) Nucleic Acids Res 34:W210
12. Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O (2006) Bioinformatics 22:1158
13. Pierleoni A, Martelli PL, Fariselli P, Casadio R (2006) Bioinformatics 22(14):E408
14. Yu CS, Chen YC, Lu CH, Hwang JK (2006) Proteins 64:643
15. Shatkay H, Hoglund A, Brady S, Blum T, Donnes P, Kohlbacher O (2007) Bioinformatics 23:1410
16. Chang JM, Su EC, Lo A, Chiu HS, Sung TY, Hsu WL (2008) Proteins 72(2):693
17. Fyshe A, Liu Y, Szafron D, Greiner R, Lu P (2008) Bioinformatics 24:2512
18. Garg A, Raghava GP (2008) BMC Bioinform 9:503
19. Huang WL, Tung CW, Ho SW, Hwang SF, Ho SY (2008) BMC Bioinform 9:80
20. Nasibov E, Kandemir-Cavas C (2008) Comput Biol Chem 32:448
21. Chou KC, Shen HB (2008) Nat Protoc 3:153
22. Chou KC, Shen HB (2007) J Proteome Res 6:1728
23. Shen HB, Chou KC (2007) Biochem Biophys Res Commun 355:1006
24. Chou KC, Shen HB (2007) J Cell Biochem 100:665
25. Shen HB, Chou KC (2007) Protein Eng Des Sel 20:39
26. Chou KC, Shen HB (2006) J Proteome Res 5:3420
27. Shen HB, Chou KC (2007) Biopolymers 85:233
28. Nakashima H, Nishikawa K (1994) J Mol Biol 238:54
29. Chou KC, Elrod DW (1999) Protein Eng 12:107
30. Chou KC, Cai YD (2002) J Biol Chem 277:45765
31. Chou KC (2001) Proteins 43:246
32. Zhou GP, Doctor K (2003) Proteins 50:44
33. Chou KC, Wu ZC, Xiao X (2011) PLoS ONE 6:e18258
34. Wu ZC, Xiao X, Chou KC (2012) Protein Pept Lett 19:4
35. Chou KC, Wu ZC, Xiao X (2012) Mol BioSyst 8:629
36. Wu ZC, Xiao X, Chou KC (2011) Mol BioSyst 7:3287
37. Xiao X, Wu ZC, Chou KC (2011) J Theor Biol 284:42
38. Mei S (2012) J Theor Biol 310:80
39. Xiao X, Wu ZC, Chou KC (2011) PLoS ONE 6:e20592
40. Lee TY, Chen YJ, Lu CT, Ching WC, Teng YC, Huang HD (2012) Bioinformatics 28:2293
41. Lee TY, Lin ZQ, Hsieh SJ, Bretana NA, Lu CT (2011) Bioinformatics 27:1780
42. Lee TY, Chen YJ, Lu TC, Huang HD (2011) PLoS ONE 6:e21849
43. Lee TY, Bretana NA, Lu CT (2011) BMC Bioinformatics 12:261
44. Lee TY, Bo-Kai Hsu J, Chang WC, Huang HD (2011) Nucleic Acids Res 39:D777
45. Lee TY, Hsu JB, Lin FM, Chang WC, Hsu PC, Huang HD (2010) J Comput Chem 31:2759
46. Wong YH, Lee TY, Liang HK, Huang CM, Wang TY, Yang YH, Chu CH, Huang HD, Ko MT, Hwang JK (2007) Nucleic Acids Res 35:W588
47. Huang HD, Lee TY, Tzeng SW, Horng JT (2005) Nucleic Acids Res 33:W226
48. Qiu JD, Huang JH, Shi SP, Liang RP (2010) Protein Pept Lett 17:715
49. Chen C, Shen ZB, Zou XY (2012) Protein Pept Lett 19:422
50. Gu Q, Ding YS, Zhang TL (2010) Protein Pept Lett 17:559
51. Li LQ, Zhang Y, Zou LY, Zhou Y, Zheng XQ (2012) Protein Pept Lett 19:375
52. Zia Ur R, Khan A (2012) Protein Pept Lett 19:890
53. Mohabatkar H, Mohammad Beigi M, Esmaeili A (2011) J Theor Biol 281:18
54. Zeng YH, Guo YZ, Xiao RQ, Yang L, Yu LZ, Li ML (2009) J Theor Biol 259:366
55. Chen C, Chen L, Zou X, Cai P (2009) Protein Pept Lett 16:27
56. Ding H, Luo LF, Lin H (2009) Protein Pept Lett 16:351
57. Zhou XB, Chen C, Li ZC, Zou XY (2007) J Theor Biol 248:546
58. Georgiou DN, Karakasidis TE, Nieto JJ, Torres A (2009) J Theor Biol 257:17
59. Yu LZ, Guo YZ, Li YZ, Li GB, Li ML, Luo JS, Xiong WJ, Qin WL (2010) J Theor Biol 267:1
60. Jiang XY, Wei R, Zhang TL, Gu Q (2008) Protein Pept Lett 15:392
61. Li FM, Li QZ (2008) Protein Pept Lett 15:612
62. Lin H, Ding H, Guo FB, Zhang AY, Huang J (2008) Protein Pept Lett 15:739
63. Zhang GY, Li HC, Gao JQ, Fang BS (2008) Protein Pept Lett 15:1132
64. Han L, Cui J, Lin H, Ji Z, Cao Z, Li Y, Chen Y (2006) Proteomics 6:4023
65. Veropoulos K, Cristianini N, Campbell C (1999) Proceedings of the international joint conference on artificial intelligence (IJCAI99), workshop ML3, p 55
66. Nair R, Rost B (2002) Protein Sci 11:2836
67. Nielsen H, Engelbrecht J, von Heijne G, Brunak S (1996) Proteins 24:165
68. Chou KC, Shen HB (2010) PLoS ONE 5:e9931
69. Chou KC, Shen HB (2010) PLoS ONE 5:e11335
70. UniProt C (2008) Nucleic Acids Res 36(Database issue):D190
71. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, Roechert B, Bairoch A (2005) C R Biol 328:882
72. Esmaeili M, Mohabatkar H, Mohsenzadeh S (2010) J Theor Biol 263:203
73. Mohabatkar H (2010) Protein Pept Lett 17:1207
74. Lin H (2008) J Theor Biol 252:350
75. Chou KC (2009) Curr Proteomics 6:262
76. Carrie C, Giraud E, Whelan J (2009) FEBS J 276:1187
77. Millar AH, Whelan J, Small I (2006) Curr Opin Plant Biol 9:610
78. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S (2002) Bioinformatics 18:298
79. von Heijne G (1990) Curr Opin Cell Biol 2:604
80. Hurtley SM (1996) Protein targeting. Oxford University Press, Oxford
81. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389
82. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL (2002) Nucleic Acids Res 30:276
83. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002) Briefings Bioinform 3:265
84. Nair R, Carter P, Rost B (2003) Nucleic Acids Res 31:397
85. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ (2006) Nucleic Acids Res 34:W32
86. Solito E, Christian HC, Festa M, Mulla A, Tierney T, Flower RJ, Buckingham JC (2006) Faseb J 20:1498
87. Jensen LJ, Gupta R, Blom N, Devos D, Tamames J, Kesmir C, Nielsen H, Staerfeldt HH, Rapacki K, Workman C, Andersen CA, Knudsen S, Krogh A, Valencia A, Brunak S (2002) J Mol Biol 319:1257
88. Mizushima S (1984) Mol Cell Biochem 60:5
89. Eichler J (2001) Eur J Biochem 268:4366

90. Pal-Bhowmick I, Vora HK, Jarori GK (2007) Malar J 6:45
91. Kiemer L, Bendtsen JD, Blom N (2005) Bioinformatics 21(7):1269
92. Shien DM, Lee TY, Chang WC, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD (2009) J Comput Chem 30(9):1532
93. Gupta R, Jung E, Brunak S (2004) [online] Available http://www.cbs.dtu.dk/services/NetNGlyc/
94. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S (1998) Glycoconj J 15:115
95. Blom N, Gammeltoft S, Brunak S (1999) J Mol Biol 294:1351
96. Chang WC, Lee TY, Shien DM, Hsu JB, Horng JT, Hsu PC, Wang TY, Huang HD, Pan RL (2009) J Comput Chem 30(15):2526
97. Eddy SR (1998) Bioinformatics 14:755
98. Chang CC, Lin CJ (2001) Software available at http://www.csie. ntu. edu. tw/cjlin/libsvm 80:604
99. Zakeri P, Moshiri B, Sadeghi M (2011) J Theor Biol 269:208
100. Nanni L, Lumini A, Gupta D, Garg A (2011) IEEE/ACM Trans Comput Biol Bioinform 9(2):467
101. Jiawei Han MK (2006) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco
102. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco
103. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) Genome Res 14:1188
104. Schneider TD, Stephens RM (1990) Nucleic Acids Res 18:6097
105. Cokol M, Nair R, Rost B (2000) EMBO Rep 1:411
106. Schaecher SR, Diamond MS, Pekosz A (2008) J Virol 82:9477
107. Ladd AN, Cooper TA (2004) J Cell Sci 117:3519
108. Hirata T, Okabe M, Kobayashi A, Ueda K, Matsuo M (2009) Biosci Biotechnol Biochem 73(3):619
109. Eisenhaber B, Eisenhaber F (2007) Curr Protein Pept Sci 8:197
110. Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH (2006) Nucleic Acids Res 34:D622