BioTechnologia vol. 98(2) • pp. 85-96 • 2017
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPERS
http://doi.org/10.5114/bta.2017.68307

Prediction of protein subcellular localization using support vector machine with the choice of proper kernel

MD. AL MEHEDI HASAN*, SHAMIM AHMAD, MD. KHADEMUL ISLAM MOLLA
Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh

Abstract

The prediction of subcellular locations of proteins can provide useful hints for revealing their functions as well as for understanding the mechanisms of some diseases and, finally, for developing novel drugs. As the number of newly discovered proteins has been growing exponentially, laboratory-based experiments to determine the location of an uncharacterized protein in a living cell have become both expensive and time-consuming. Consequently, to tackle these challenges, computational methods are being developed as an alternative to help biologists in selecting target proteins and designing related experiments. However, the success of protein subcellular localization prediction is still a complicated and challenging problem, particularly when query proteins may have multi-label characteristics, i.e. their simultaneous existence in more than one subcellular location, or if they move between two or more different subcellular locations as well. At this point, to get rid of this problem, several types of subcellular localization prediction methods with different levels of accuracy have been proposed. The support vector machine (SVM) has been employed to provide potential solutions for problems connected with the prediction of protein subcellular localization. However, the practicability of SVM is affected by difficulties in selecting its appropriate kernel as well as in selecting the parameters of that selected kernel. The literature survey has shown that most researchers apply the radial basis function (RBF) kernel to build a SVM based subcellular localization prediction system. Surprisingly, there are still many other kernel functions which have not yet been applied in the prediction of protein subcellular localization. However, the nature of this classification problem requires the application of different kernels for SVM to ensure an optimal result. From this viewpoint, this paper presents the work to apply different kernels for SVM in protein subcellular localization prediction to find out which kernel is the best for SVM. We have evaluated our system on a combined dataset containing 5447 single-localized proteins (originally published as part of the Höglund dataset) and 3056 multi-localized proteins (originally published as part of the DBMLoc set). This dataset was used by Briesemeister et al. in their extensive comparison of multi-localization prediction system. The experimental results indicate that the system based on SVM with the Laplace kernel, termed LKLoc, not only achieves a higher accuracy than the system using other kernels but also shows significantly better results than those obtained from other top systems (MDLoc, BNCs, YLoc+). The source code of this prediction system is available upon request.

Key words: support vector machine, kernel, kernel selection, protein subcellular localization prediction, multi-label classification

Introduction

A biological cell is made up of many different compartments or organelles. These compartments are closed to one another and also have different functions. The proteins in the cell are responsible for most of the functions required for a cell's survival. A typical cell contains approximately one billion protein molecules that reside in many different compartments or organelles, usually termed "subcellular locations" (Chou and Shen, 2007a). Proteins can perform their appropriate functions when they are located in the right subcellular locations. Knowledge of the subcellular localization of proteins is important, because it (a) provides useful insights into their functions, (b) indicates how and in which kind of cellular environments they interact with one another and with other molecules, (c) helps in understanding the intricate

*Corresponding author: Department of Computer Science and Engineering, University of Rajshahi, Bangladesh; e-mail: mehedi_ru@yahoo.com

pathways that regulate biological processes at the cellular level and (d) helps in identifying and prioritizing drug targets during the process of drug development (Chou and Shen, 2010; Wang et al., 2011).

Although various experimental approaches have been developed for determining protein subcellular locations, most of those approaches are unfortunately costly and also time-consuming (Du and Xu, 2013). However, the number of newly discovered proteins has been growing exponentially, which in turn makes the subcellular localization prediction by purely laboratory tests prohibitively expensive (Wan et al., 2012). In this context, computational methods have been developed to help biologists in the selection of target proteins and in the design of related experiments. Moreover, computational methods are fast and can potentially predict locations for proteins whose actual locations have not yet been experimentally determined. Various methods for predicting subcellular localization of protein sequences have been extensively studied in the last decades, and researchers have developed increasingly numbers of new models to acquire better prediction performance (Yang et al., 2006).

Conventional methods for subcellular localization prediction can be roughly divided into sequence-based methods and annotation-based methods (Yang et al., 2006; Wan et al., 2012; Simha et al., 2014). Sequence-based predictors employ: 1) sequence-coded sorting signals (Bannai et al., 2002; Petsalaki et al., 2006), such as PSORT (Nakai and Kanehisa, 1991), WoLF PSORT (Horton et al., 2007), TargetP (Emanuelsson et al., 2000) and SignalP (Nielsen et al., 1997); 2) amino acid composition information (King and Guda, 2007), such as amino-acid compositions (AAC) (Nakashima and Nishikawa, 1994), amino-acid pair compositions (PairAA) (Nakashima and Nishikawa, 1994), gapped amino-acid pair compositions (GapAA) (Park and Kanehisa, 2003), and pseudo amino-acid composition (PseAA) (Chou and Cai, 2003); and 3) both information sources (Höglund et al., 2006; Horton et al., 2007). It should be noted that sequence-based methods are general in that they can be applied to any newly discovered proteins (Wan et al., 2013). However, their performance is usually poor, especially for datasets containing sequences with low-similarity.

Annotation-based predictors use information about functional domains and motifs (Chou and Cai, 2002; Scott et al., 2004), protein–protein interaction (Lee et al., 2008; Shin et al., 2009), homologous proteins (Mak et al., 2008; Lin et al., 2009), annotated Gene Ontology (GO) terms (Huang et al. 2008) such as Euk-OET-PLoc (Chou and Shen, 2006), Euk-mPLoc (Chou and Shen, 2007b), iLoc-Gneg (Xiao et al., 2011a), CELLO2GO (Yu et al., 2014) and Cell-PLoc 2.0 (Chou and Shen, 2010) and textual information from Swiss-Prot keywords (Nair and Rost, 2002; Lu et al., 2004) or PubMed abstracts (Brady and Shatkay, 2008; Fyshe et al., 2008). The annotation-based predictors often show higher accuracies than pure sequence-based predictors, although they are less robust when the protein is a newly discovered one and even if its close homologues are unknown (Briesemeister et al., 2010a). In fact, when the protein to be predicted is a newly discovered one, there is no existing annotation in the database. As a result, the prediction performance of annotation-based methods will be degraded. However, since the coverage of public annotation databases is increasing rapidly (Yang et al., 2006), so at least some annotation of the close homologs of a novel protein is expected to be available. This will reduce the above limitation of annotation-based methods.

In addition to the above approaches, some researchers have developed hybrid prediction approaches (Chou and Shen, 2007b; Blum et al., 2009; Briesemeister et al., 2010a; Simha et al., 2014; Simha et al., 2015) which include both the sequence-based methods and the annotation-based methods.

Not only protein sequence information but also prediction algorithms could affect the accuracy of subcellular localization prediction (Li et al., 2012). To date, many computational techniques, such as the neural network (Zou et al., 2007), K-nearest neighbor (KNN) (Chou et al., 2006; Xiao et al., 2011b; He et al., 2012), fuzzy KNN (Gu et al., 2010), and Bayesian (Briesemeister et al., 2010a; Simha et al., 2014; Simha et al., 2015), have been introduced for the prediction of protein subcellular localization.

In recent times, a support vector machine (SVM) (Höglund et al., 2006; Li et al., 2012; Wan et al., 2012; Wan et al., 2014; Hasan et al., 2015,) has been extensively applied to provide potential solutions for the prediction of protein subcellular localization. However, the selection of an appropriate kernel and its parameters for a given classification problem influences the performance of the SVM. The reason for this is that different kernel functions construct different SVMs and affect the

generalization ability and learning ability of the SVM. However, there is no theoretical method for selecting the kernel function and its parameters. The literature survey has shown that most of the researchers have applied the radial basis function (RBF) kernel to build SVM based subcellular localization prediction system (Chou and Cai, 2002; Park and Kanehisa, 2003; Li et al., 2011; Wan et al., 2012; Wan et al., 2013) and have found the value of its parameter by using different techniques, such as trial and error, heuristics or grid search procedure (Wan et al., 2015). Surprisingly, still there are many other kernel functions which have not yet been applied in the protein subcellular localization prediction. However, the nature of this classification problem requires the application of different kernels for SVM to ensure an optimal result. This requirement motivated us to apply different kernel functions for SVM rather than simply using RBF in protein subcellular localization prediction, which, in turn, may provide better accuracy of the prediction system. At the same time, we tried to find out the parameter value to the corresponding kernel.

Moreover, as there exist multi-location proteins that can simultaneously reside at, or move between, two or more subcellular locations, recent studies have focused on predicting both single label and multi-label proteins (Wan et al., 2015). However, the consideration of the multi-label protein has been excluded in some studies (Shen et al., 2007). Identification of the multiple locations of a protein is important, because the translocation of proteins can serve some unique functions (Wan et al., 2015). Therefore, this article has also considered a multi-label prediction for predicting the subcellular localization of both single label and multi-label proteins.

The proposed system has been trained and tested on a dataset containing both single- and multi-localized proteins which has been used in the development and testing of the YLoc+ system (Briesemeister et al., 2010a) as well as MDLoc and BNCs systems (Simha et al., 2014; Simha et al., 2015) and derived from the Höglund dataset (Höglund et al., 2006) and the DBMLoc dataset (Zhang et al., 2008). As like other studies (Höglund et al., 2006; Shatkay et al., 2007; Blum et al., 2009; Briesemeister et al., 2010a; Simha et al., 2014; Simha et al., 2015), multiple runs of the 5-fold cross-validation have been performed in our work. The results clearly demonstrate the advantage of using Laplace kernel with SVM in protein subcellular localization prediction. The F_1-label score of 74% and overall accuracy of 70% obtained by LKLoc (SVM with Laplace kernel based system) are significantly better than the corresponding results obtained by the system using other kernels as well as other top existing classifiers (MDLoc, BNCs, YLoc+) when only multi-localized proteins were considered. In addition, in the case of both single and multi-localized proteins, LKLoc retained a higher overall accuracy than the system using other kernels, or the BNCs.

Materials and methods

Datasets

In our experiments, we used a combined dataset containing 5447 single-localized proteins, originally published as part of the Höglund dataset (Höglund et al., 2006) and 3056 multi-localized proteins, originally published as part of the DBMLoc set (Zhang et al., 2008). This combined dataset was initially constructed for an extensive comparison of multi-localization prediction systems by Briesemeister et al. (Briesemeister et al., 2010a). This dataset is already homology-reduced, i.e. the protein sequences from the Höglund dataset share no more than 30% sequence identity with each other; and at the same time, sequences from the DBMLoc dataset share less than 80% sequence similarity with one another. As it cannot be known a priori whether a protein may localize to a single or to multiple locations, we trained our system on the combined set of proteins, thus enabling it to handle the actual prediction task. We reported results using different evaluation metrics that obtained over the dataset containing both single- and multi-localized proteins for comparing our system to other published systems. As most of the results for other systems are only available for the set of multi-localized proteins only (Simha et al., 2015), we measured two sets of results from our trained system: one set was for the combined set of single and multi-localized proteins; and the other was for multi-localized proteins only. In cases where reports obtained on the combined set of single- and multi-localized proteins from other systems were available, we also made comparisons with our system. The 5447 single-localized proteins covered the following 9 locations (abbreviations and number of proteins per location are given in parentheses): cytoplasm (cyt, 1411 proteins), endoplasmic reticulum (ER, 198), extra cellular space (ex, 843), golgi apparatus (gol,

150), lysosome (lys, 103), mitochondrion (mi, 510), nucleus (nuc, 837), membrane (mem, 1238), and peroxisome (per, 157). The multi-localized proteins come from the following pairs of locations: cyt and nuc (cyt_nuc, 1882 proteins), ex and mem (ex_mem, 334), cyt and mem (cyt_mem, 252), cyt and mi (cyt_mi, 240), nuc and mi (nuc_mi, 120), ER and ex (ER_ex, 115), and ex and nuc (ex_nuc, 113). It should be noted that all the multi-location subsets used had over 100 representative proteins and this is currently the largest data set of proteins from multiple locations (Briesemeister et al., 2010b).

Biological input features of protein

In this study, we used a 30-dimensional feature vector of protein, similar to that used by Briesemeister et al. for YLoc+ and R. Ramanuja Simha for MDLoc and BNCs (Briesemeister et al., 2010a, 2010b; Simha et al., 2014; Simha et al., 2015). However, thirteen of these features were derived directly from the protein sequence, such as the length of the amino acid chain, the length of the longest very hydrophobic region, and the respective numbers of methionine, asparagine, and tryptophan, occurring in the N-terminus, etc. (Simha et al., 2014). Again, nine of these features were extracted from the pseudo-amino acid composition (Chou and Cai, 2003), which was based on certain physical and chemical properties of amino acid subsequences. The remaining 8 features came from two types of annotation based features. Here, one type of annotation-based features contained two features constructed using two distinct groups of PROSITE patterns, and the other type of annotation-based features contained six features extracted based on GO-annotations (Simha et al., 2014; Simha et al., 2015).

SVM classification

Consider the problem of separating a set of training vectors belonging to two separate classes, (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), where $x_i \in R^p$ and $y_i \in \{-1, +1\}$ is the corresponding class label, $1 \le i \le n$. The main task of this problem is to find a classifier with a decision function $f(x, \theta)$ such that $y = f(x, \theta)$, where y is the class label for x and θ is a vector of unknown parameters of the decision function. The support vector machine is a well-known classifier and it has been applied broadly in many classification problems. The SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires the solving of the following constraint problem (Vladimir, 1995):

$$\begin{aligned} &\text{minimize}_{w,\,b} && \frac{1}{2}\|w\|^2 \\ &\text{subject to} && y_i\left(w^T x_i + b\right) \ge 1, \qquad i = 1, 2, 3, \ldots, n \end{aligned} \tag{1}$$

To allow errors, the optimization problem now becomes:

$$\begin{aligned} &\text{minimize}_{w,\,b,\,\xi} && \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to} && y_i\left(w^T x_i + b\right) \ge 1 - \xi_i, \qquad i = 1, 2, 3, \ldots, n \\ & && \xi_i \ge 0, \qquad i = 1, 2, 3, \ldots, n \end{aligned} \tag{2}$$

Using the method of Lagrange multipliers, we can obtain the dual formulation which is expressed in terms of variables α_i (Scott et al. 2004; Yang et al. 2006; Wan et al. 2012):

$$\begin{aligned} &\text{maximize}_{\alpha} && \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ &\text{subject to} && \sum_{i=1}^{n} y_i\alpha_i = 0, \quad 0 \le \alpha_i \le C \\ & && \text{for all } \; i = 1, 2, 3, \ldots, n \end{aligned} \tag{3}$$

Finally, the linear classifier based on a linear discriminant function takes the following form:

$$f(x) = \sum_i^n \alpha_i y_i x_i^T x + b \tag{4}$$

In many applications, a non-linear classifier provides better accuracy. The naive way of making a non-linear classifier out of a linear classifier is to map our data from the input space X to a feature space F using a non-linear function $\varnothing$: $X \rightarrow F$. In space F, the optimization takes the following form using kernel function (Schölkopf and Smola 2002):

$$\begin{aligned} &\text{maximize}_{\alpha} && \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k\left(x_i, x_j\right) \\ &\text{subject to} && \sum_{i=1}^{n} y_i\alpha_i = 0, \quad 0 \le \alpha_i \le C \\ & && \text{for all} \quad i = 1, 2, 3, \ldots, n \end{aligned} \tag{5}$$

Finally, in terms of the kernel function the discriminant function takes the following form:

$$f(x) = \sum_i^n \alpha_i y_i k\left(x, x_i\right) + b \tag{6}$$

Kernel and its parameters selection

A kernel function and its parameter have to be chosen to build a SVM classifier (Schölkopf and Smola

2002; Hasan et al. 2014). In this study, four main kernels have been used to build SVM classifier. These are:

1) Linear kernel: $K(x_i, x_j) = \langle x_i, x_j \rangle$,

2) Polynomial kernel: $K(x_i, x_j) = \left(\langle x_i, x_j \rangle + 1\right)^d$, d is the degree of the polynomial.

3) Gaussian kernel: $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$, σ is the width of the function.

4) Laplace kernel: $K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\sigma}\right)$, σ is the width of the function.

Training an SVM finds the large margin hyperplane, i.e. sets the parameters α_i. The SVM has another set of parameters called hyperparameters: the soft margin constant, C, and any parameters the kernel function may depend on (width of a Gaussian kernel or degree of a polynomial kernel) (Ben-Hur and Weston 2010). The soft margin constant C adds a penalty term to the optimization problem. For a large value of C, a large penalty is assigned to errors/margin errors and creates force to consider points close to the boundary and decreases the margin. A smaller value of C allows to ignore points close to the boundary, and increases the margin.

Kernel parameters also have a significant effect on the decision boundary (Ben-Hur and Weston, 2010). The degree of the polynomial kernel and the width parameter σ of the Gaussian kernel or Laplace kernel control the flexibility of the resulting classifier. The lowest degree polynomial is the linear kernel, which is not sufficient when a non-linear relationship between features exists. Higher degree polynomial kernels are flexible to discriminate between the two classes with a sizable margin and greater curvature for a fixed value of the soft-margin constant. On the other hand in the Gaussian kernel or Laplace Kernel, for a fixed value of the soft-margin constant, for large values of σ the decision boundary is nearly linear. As σ decreases, the flexibility of the decision boundary increases and small values of σ lead to overfitting (Ben-Hur and Weston, 2010).

A question frequently posed by practitioners is "which kernel should I use for my data?". There are several answers to this question. The first is that it is, like most practical questions in machine learning, data-dependent, so several kernels should be tried. We typically follow the following procedure: we try a linear kernel first, and then see if we can improve on its performance using a non-linear kernel (Ben-Hur and Weston, 2010; Hasan et al., 2014).

Multiclass multi-label classification using SVM

Support vector machines are formulated for two class single label problems. An extension to multiclass multi-label problems for SVM is not straightforward (Hasan et al., 2013; Hasan et al., 2014). We followed the Binary relevance method (BR) (Tsoumakas et al., 2009) to solve the multiclass multi-label problem. The binary relevance method (BR) (Tsoumakas et al., 2009) uses the one-against-rest strategy to convert a multi-label problem into several binary classification problems. Given a multi-label dataset with N class labels, the BR method trains one classifier for each class label. When training one classifier for each class label, the (BR) method annotates all of the training examples associated with that label as positive examples, while all remaining examples are regarded as negative examples (Wang et al., 2015). Given a test example, each classifier in BR will output a prediction score and BR will combine these scores into an N-dimensional score vector, where each score corresponds to a specific class label. The value of the score has two conditions, positive and negative: positive means the binary classifier predicts the test example belonging to the corresponding class label; and negative means it does not belong to the class label. Note that if all N scores are negative, the class label with the maximum score is assigned to the test example.

In accordance with the method discussed above, in order to predict the subcellular locations of datasets containing both single-label and multi-label proteins, N independent binary SVMs are trained, one for each location. Then, the subcellular location(s) of the i-th query protein will be predicted as:

$$M^*(x_i) = \bigcup_{j=1}^{N} \left\{ j : f_j(x_i) > 0 \right\} \tag{7}$$

Here, $M^*(x_i)$ is a predicted set that may have one or more elements, even it can be empty too, which enables us to make multi-label predictions. However, if Eq. 7 provides an empty class label, i.e. $M^*(x_i) = \varnothing$, in that case there will only be a single subcellular location of the query protein and that location will be given by

$$M^*(x_i) = \arg\max_{j=1} f_j(x_i) \tag{8}$$

Experimental setting

In a statistical prediction, there are three commonly used methods to derive the metric values for a predictor; these are the independent dataset test, a subsampling (e.g., *K*-fold cross validation) test, and a jackknife test (Chou and Shen, 2007a). These methods are often used for testing the accuracy of a statistical prediction algorithm. However, of these three methods, the jackknife test is deemed the most objective, because it can always yield a unique result for a given benchmark data set, as reported in a comprehensive review (Chou and Shen, 2007a). Although the jackknife test has been increasingly and widely adopted by investigators to examine the power of various prediction methods, it requires significant computational time for a larger dataset.

In this study, to save computational time, we used *K*-fold cross validation (subsampling) methods and compared the performance of LKLoc (SVM with Laplace kernel based system) with that of other systems (YLoc+ (Briesemeister et al., 2010a), Euk-mPLoc (Chou and Shen, 2007b), WoLF PSORT (Horton et al., 2007), and $\text{KnowPred}_{\text{site}}$ (Lin et al., 2009)) and the systems based on other kernels. The performance of YLoc+, Euk-mPLoc, WoLF PSORT, and $\text{KnowPred}_{\text{site}}$ on a large set of multi-localized proteins was studied comprehensively in (Simha et al., 2015). As the information about the exact 5-way splits of dataset used in previous studies has not been published, in order to validate the stability and the statistical significance of our results, we repeated the 5-fold cross-validation for 5 times. In each 5-fold cross-validation the given training samples are randomly partitioned into 5 mutually exclusive sets of approximately equal size and approximately equal class distribution. Finally, we reported the average results in this study.

All programs were run on a standard DELL Optiplex 390 machine with 8 GB RAM and a Core-i3 processor running at 3.30 GHz.

Evaluation metrics

The measurement of performance in a multi-label classification is more complicated than in the traditional single-label classification, as each example could be associated with multiple labels simultaneously. In this study, we used various types of adapted measures, such as multi-label accuracy and F_1 score proposed by Tsoumakas et al. (Tsoumakas et al., 2009), for our evaluation of the multi-label classification.

To formally define these evaluation measures, let D be a dataset containing m proteins and $S = \{s_1, s_2, ..., s_q\}$ be the set of q possible subcellular components in the cell. For a given protein P, let $M^P = \{s_i | I_i^P = 1, \text{where } 1 \le i \le q\}$ be the set of locations to which protein P localizes according to the dataset, and let $\hat{M}^P = \{s_i | \hat{I}_i^P = 1, \text{where } 1 \le i \le q\}$ be the set of locations that a classifier predicts for protein P, where $\hat{I}_i^P, I_i^P \in \{1, 0\}$. I_i^P or $\hat{I}_i^P$ takes the value 1 if P actually localizes s_i or is predicted to s_i, respectively. The multi-label accuracy and the multi-label F_1 score are computed as follows (Simha et al., 2014):

$$\text{Acc} = \frac{1}{|D|}\sum_{P \in D}\frac{\left|M^P \cap \hat{M}^P\right|}{\left|M^P \cup \hat{M}^P\right|} \text{ and}$$

$$F_1 = \frac{1}{|D|}\sum_{P \in D}\frac{\left|M^P \cap \hat{M}^P\right|}{\left|M^P\right| + \left|\hat{M}^P\right|} \text{ respectively.}$$

In our evaluation, we also used adapted measures of multi-label precision and recall denoted Pre_{si} and Rec_{si} to evaluate how well the classifier will classify proteins as localized or not localized to each individual location s_i, and these are defined as follows (Li et al., 2011):

$$\text{Pre}_{S_i} = \frac{1}{\left|\left\{P \in D \,|\, s_i \in \hat{M}^P\right\}\right|}\sum_{P \in D | s_i \in \hat{M}^P}\frac{\left|M^P \cap \hat{M}^P\right|}{\left|\hat{M}^P\right|}$$

$$\text{Rec}_{S_i} = \frac{1}{\left|\left\{P \in D \,|\, s_i \in M^P\right\}\right|}\sum_{P \in D | s_i \in M^P}\frac{\left|M^P \cap \hat{M}^P\right|}{\left|M^P\right|}$$

We used the terms Multilabel-Precision and Multilabel-Recall to refer to Pre_{si} and Rec_{si}, respectively. Here, Pre_{si} represents the ratio of the number of correctly predicted multiple locations to the total number of multiple locations predicted, and Rec_{si} represents the ratio of the number of correctly predicted multiple locations to the number of original multiple locations, for all the proteins that co-localize to location s_i (Simha et al., 2014). Therefore, high values of these measures for proteins that co-localize to the location s_i can be used to indicate that the sets of predicted locations that include location s_i are predicted correctly (Simha et al., 2014).

Standard precision and recall measures, denoted by $\text{Pre–Std}_{\text{si}}$ and $\text{Rec–Std}_{\text{si}}$, are used in this paper to evalu-

ate the correctness of predictions made for each location s_i and are computed as:

$$\text{Pre} - \text{Std}_{s_i} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Rec} - \text{Std}_{s_i} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP (true positives) denotes the number of proteins that localize to s_i and are predicted to localize to s_i, FP (false positives) denotes the number of proteins that do not localize to s_i but are predicted to localize to s_i, and FN (false negatives) denotes the number of proteins that localize to s_i but are not predicted to localize to s_i.

Additionally, the adapted measure of the F_1-label score used by Briesemeister et al. (Briesemeister et al., 2010a) for evaluating the performance of multi-location predictors was used in our evaluation and it is defined as:

$$F_1 - \text{label} = \frac{1}{|S|} \sum_{s_i \in S} \frac{2 \times \text{Pre}_{s_i} \times \text{Rec}_{s_i}}{\text{Pre}_{s_i} + \text{Rec}_{s_i}}$$

where S is the set of all locations.

Results and discussion

SVM model selection

In order to generate high performance SVM classifiers capable of dealing with real data, an efficient model selection is required. A grid-search technique was used to find the best model for SVM with different kernels in this work. Herein, this method selects the values of parameters considering the highest multi-label accuracy and then time, if more than one position in the search space has the same multi-label accuracy. Sequential minimization optimization with the following options in Matlab 2014b, shown in Table 1, was used to develop our system. According to the experimental setting, we performed 5 complete runs of the 5-fold cross-validation and each time we selected the best parameter of the classifier on the basis of the multi-label accuracy.

Table 1. Sequential Minimization Optimization Options

Option	Value
MaxIter	5 000 000
KernelCacheLimit	30 000

For the linear kernel based SVM, in order to find the parameter value C (penalty term for soft margin), we considered the values from 2^{-4} to 2^4 as our search space. The selected C of 5 complete runs of the 5-fold cross-validation on the combined set of single- and multi-localized proteins is shown in Table 2. Table 2 shows that on most occasions the best model is found for the values of $C = 2^{-4}$ or $C = 2^{-1}$. Finally, we used $C = 2^{-4}$ (using random selection between these two values) in all 5 complete runs of the 5-fold cross-validation and averaged our results in order to ensure unbiased model selection.

For a polynomial kernel based SVM, to find the parameter value C (penalty term for soft margin) and d, we considered the values from 2^{-4} to 2^4 for C and from 1 to 3 for d as our search space. The selected C and d of 5 complete runs of the 5-fold cross-validation on the combined set of single- and multi-localized proteins is shown in Table 2. Table 2 shows that on most occasions the best model was found for the values of $C = 2^1$ and $d = 3$. Finally, we used $C = 2^1$ and $d = 3$ in all 5 complete runs of the 5-fold cross-validation and averaged our results in order to ensure unbiased model selection.

For the radial basis function (RBF) kernel based SVM, to find the parameter value C (penalty term for soft margin) and σ (sigma), we considered the values from 2^{-8} to 2^8 for C and from 2^{-8} to 2^8 for sigma as our search space. The selected C and sigma of 5 complete runs of the 5-fold cross-validation on the combined set of single- and multi-localized proteins are shown in Table 2. Table 2 shows that on most occasions the best model is found for the value of $C = 2^1$ and $\sigma = 2^1$. Finally, we used $C = 2^1$ and $\sigma = 2^1$ in all complete runs of the 5-fold cross-validation and averaged our results in order to ensure unbiased model selection.

Again, for the Laplace kernel based SVM, to find the parameter value C (penalty term for soft margin) and σ (sigma), we considered the values from 2^{-8} to 2^8 for C and from 2^{-8} to 2^8 for sigma as our search space. The selected C and sigma of 5 complete runs of the 5-fold cross-validation on the combined set of single- and multi-localized proteins are shown in Table 2. Table 2 shows that on most occasions the best model is found for the value of $C = 2^8$ and $\sigma = 2^3$. Finally, we used $C = 2^8$ and $\sigma = 2^3$ in all 5 complete runs of the 5-fold cross-validation and averaged our results in order to ensure unbiased model selection.

Performance measure evaluations

In this section, we compare the performance of each kernel for SVM and also compare the performance of the

Table 2. Selected parameters of 5 complete runs of the 5-fold cross-validation on the combined set of single- and multi-localized proteins for each kernel based SVM (Linear, Polynomial, RBF, Laplace)

Number of complete runs	Linear kernel	Polynomial kernel		RBF kernel		Laplace kernel	
	C	C	d	C	σ	C	σ
1st	2^{-1}	2^{-2}	3	2^{1}	2^{1}	2^{8}	2^{3}
2nd	2^{-1}	2^{1}	3	2^{2}	2^{1}	2^{5}	2^{2}
3rd	2^{-4}	2^{-4}	3	2^{1}	2^{1}	2^{8}	2^{3}
4th	2^{-4}	2^{1}	3	2^{1}	2^{1}	2^{8}	2^{3}
5th	2^{-2}	2^{1}	3	2^{7}	2^{1}	2^{8}	2^{3}

Table 3A. Comparison of the results of multi-location prediction systems of different kernels, averaged over 5 complete runs of the 5-fold cross-validation applied on the combined set of single-localized and multi-localized proteins

	Linear	Polynomial	RBF	Laplace
F_1	0.613 (± 0.016)	0.677 (± 0.033)	0.810 (± 0.017)	0.829 (± 0.002)
Acc	0.498 (± 0.016)	0.602 (± 0.031)	0.764 (± 0.017)	0.786 (± 0.002)

Table 3B. Comparison of the results of multi-location prediction systems, averaged over 5 complete runs of the 5-fold cross-validation applied on the combined set of single-localized and multi-localized proteins

	LKLoc	BNCs
F_1	0.829 (± 0.002)	0.81 (± 0.01)
Acc	0.786 (± 0.002)	0.76 (± 0.01)

Table 3C. Per-location based results, averaged over 5 complete runs of the 5-fold cross-validation applied on the combined dataset

		cyt (3785)	nuc (2952)	ex (1405)	mem (1824)	mi (870)
Rec_{si}	LKLoc	0.829 (± 0.008)	0.818 (± 0.015)	0.818 (± 0.005)	0.800 (± 0.002)	0.753 (± 0.004)
	MDLoc	0.825 (± 0.009)	0.830 (± 0.010)	0.780 (± 0.020)	0.822 (± 0.012)	0.773 (± 0.013)
	BNCs	0.795 (± 0.011)	0.784 (± 0.017)	0.737 (± 0.022)	0.780 (± 0.014)	0.730 (± 0.025)
Pre_{si}	LKLoc	0.833 (± 0.004)	0.843 (± 0.005)	0.886 (± 0.009)	0.864 (± 0.001)	0.868 (± 0.005)
	MDLoc	0.819 (± 0.013)	0.822 (± 0.014)	0.864 (± 0.020)	0.872 (± 0.014)	0.861 (± 0.024)
	BNCs	0.809 (± 0.018)	0.832 (± 0.013)	0.912 (± 0.019)	0.900 (± 0.012)	0.885 (± 0.023)
$Rec - Std_{si}$	LKLoc	0.890 (± 0.003)	0.764 (± 0.027)	0.837(± 0.009)	0.740 (± 0.004)	0.713 (± 0.009)
	MDLoc	0.867 (± 0.015)	0.808 (± 0.021)	0.715 (± 0.030)	0.842 (± 0.017)	0.719 (± 0.028)
	BNCs	0.861 (± 0.014)	0.736 (± 0.031)	0.652 (± 0.024)	0.805 (± 0.017)	0.664 (± 0.034)
$Pre - Std_{si}$	LKLoc	0.855 (± 0.004)	0.814 (± 0.008)	0.907(± 0.009)	0.831(± 0.003)	0.868 (± 0.003)
	MDLoc	0.854 (± 0.014)	0.783 (± 0.020)	0.839 (± 0.028)	0.882 (± 0.014)	0.843 (± 0.026)
	BNCs	0.840 (± 0.011)	0.786 (± 0.026)	0.906 (± 0.022)	0.900 (± 0.015)	0.873 (± 0.034)

best performed kernel with that of the existing location prediction systems. We trained our system using the combined dataset and measured two set of results, one for a combined set of single and multi-localized proteins and one for multi-localized proteins only. Herein, all the values of all metrics of our system are the average result of 5 complete runs of the 5-fold cross-validation. Moreover, standard deviations of each metric of 5 complete runs of the 5-fold cross-validation are shown in parentheses.

Table 4A. Multi-location prediction results, averaged over 5 complete runs of the 5-fold cross-validation, for multi-localized proteins only

	LKLoc	MDLoc	BNCs	YLoc+	Euk-mPLoc	WoLF PSORT	$KnowPred_{site}$
F_1-label	0.741 (± 0.004)	0.71 (± 0.02)	0.66 (± 0.02)	0.68	0.44	0.53	0.66
Acc	0.700 (± 0.010)	0.68 (± 0.01)	0.63 (± 0.01)	0.64	0.41	0.43	0.63

Table 4B. Per-location based results, averaged over 5 complete runs of the 5-fold cross-validation, for multi-localized proteins only

		cyt (2374)	nuc (2115)	mem (586)	ex (562)	mi (360)
Rec_{si}	LKLoc	0.750 (± 0.014)	0.776 (± 0.017)	0.557 (± 0.009)	0.590 (± 0.008)	0.527 (± 0.006)
	MDLoc	0.750 (± 0.012)	0.776 (± 0.014)	0.527 (± 0.022)	0.547 (± 0.035)	0.519 (± 0.026)
	YLoc+	0.712 (± 0.009)	0.728 (± 0.011)	0.543 (± 0.018)	0.573 (± 0.026)	0.536 (± 0.031)
Pre_{si}	LKLoc	0.934 (± 0.003)	0.944 (± 0.0006)	0.870 (± 0.013)	0.917 (± 0.008)	0.868 (± 0.014)
	MDLoc	0.911 (± 0.008)	0.929 (± 0.008)	0.807 (± 0.036)	0.833 (± 0.044)	0.832 (± 0.042)
	YLoc+	0.893 (± 0.010)	0.924 (± 0.008)	0.764 (± 0.029)	0.740 (± 0.053)	0.765 (± 0.033)
$Rec - Std_{si}$	LKLoc	0.849 (± 0.004)	0.700 (± 0.034)	0.615 (± 0.020)	0.440 (± 0.009)	0.431 (± 0.017
	MDLoc	0.817 (± 0.021)	0.746 (± 0.028)	0.588 (± 0.042)	0.385 (± 0.058)	0.388 (± 0.062)
	YLoc+	0.786 (± 0.020)	0.684 (± 0.015)	0.614 (± 0.042)	0.401 (± 0.037)	0.429 (± 0.060)
$Pre - Std_{si}$	LKLoc	0.950 (± 0.002)	0.929 (± 0.001)	0.867 (± 0.013)	0.921 (± 0.006)	0.829 (± 0.013)
	MDLoc	0.942 (± 0.009)	0.904 (± 0.014)	0.794 (± 0.039)	0.830 (± 0.046)	0.784 (± 0.057)
	YLoc+	0.935 (± 0.009)	0.914 (± 0.014)	0.730 (± 0.047)	0.771 (± 0.055)	0.670 (± 0.055)

Table 3A shows comparisons of the F_1 score and the accuracy obtained by each kernel used in SVM for the combined dataset. The table shows that SVM with Laplace kernel based system, termed LKLoc, performs better than other kernels. In addition, Table 3B shows comparative studies of the F_1 score and the accuracy obtained by LKLoc with those obtained by other multi-location predictors applied on the combined dataset (BNCs as reported in Table 2 of Ramanuja Simha et al. (Simha et al., 2014)). It is clear from this Table that LKLoc provides better accuracy than the existing systems.

Table 3C shows comparative study of the results of per-location predictions applied on the combined dataset of both single- and multi-localized proteins obtained by LKLoc and those obtained by MDLoc and BNCs (Simha et al., 2014, Simha et al. 2015). It is obvious from the Table 3C, that in most of the cases the precision values provided by LKLoc are somewhat higher than those obtained by MDLocs and on the other hand, the recall provided by LKLoc has a little bit variation (up and down) than those of MDLocs.

Table 4A shows the comparisons of the F_1-label score and the accuracy obtained by the best performing kernel (Laplace kernel in this case) with those obtained by other multi-location predictors for multi-localized proteins only (MDLoc, BNCs, YLoc+, Euk-mPLoc, WoLF PSORT and $KnowPred_{site}$ as reported in Table 1 of Ramanuja Simha et al. (Simha et al., 2015)). It can be noted here that all the predictors mentioned above used the same set of multi-localized proteins. The table shows that the prediction based on SVM with the Laplace kernel or LKLoc performs better than the existing top-systems, including MDLoc, YLoc+, and BNCs which have the best performance reported so far.

Table 4B shows the per-location prediction results for multi-localized proteins obtained by LKLoc compared with those systems reported by MDLoc (Simha et al., 2015). Since the per-location predictions for the other systems (BNCs, Euk-mPLoc, WoLF PSORT and $KnowPred_{site}$) are not publicly available, we could not show those findings. In Table 4B, the results are shown for the five locations with the largest number of associated

proteins. However, for each location s_i, we show Multi-label-Precision (Pre_{si}) and Multilabel-Recall (Rec_{si}) as well as standard precision ($Pre–Std_{si}$) and recall ($Rec–Std_{si}$). The results show that in almost all of the cases, these four measures obtained from LKLoc are significantly higher than those obtained for all protein locations using MDLoc and YLoc+.

Conclusion

In this research work, we evaluated the performance of different kernels for SVM used in the protein subcellular localization prediction in terms of various measures. The results indicate that the performance of the SVM classification mainly depends on the type of kernels and their parameters. Moreover, the obtained results also justify the motivation of this work, i.e. only a single kernel cannot be considered blindly for the SVM used in protein subcellular localization prediction when optimal performance is desired. Our results show that the LKLoc provides better performance than other kernel based systems and other existing top classifiers (MDLoc, BNCs, YLoc+). Research in protein subcellular localization prediction using the SVM approach is still demanding due to its better performance. The research community working on the SVM based classification will benefit from the results of this study. In the future, we will try to improve the performance of the classifier by considering other information such as location inter-dependencies in addition to feature information. Moreover, we shall make efforts in our future work to provide a web-server for the method presented in this paper. To sum up, we believe that these findings may be extended to other biological problems.

References

Bannai H., Tamada Y., Maruyama O., Nakai K., Miyano S. (2002) *Extensive feature detection of N-terminal protein sorting signals*. Bioinformatics 18(2): 298-305.

Brady S., Shatkay H. (2008) *EpiLoc: a (working) text-based system for predicting protein subcellular location*. Pacific Symp. Biocomput. 13: 604-615.

Blum T., Briesemeister S., Kohlbacher O. (2009) *MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction*. BMC Bioinformatics 10(1): 1.

Ben-Hur A., Weston J. (2010) *A user's guide to support vector machines*. Meth. Mol. Biol. 609: 223-239.

Briesemeister S., Rahnenführer J., Kohlbacher O. (2010a) *Going from where to why – interpretable prediction of protein subcellular localization*. Bioinformatics 26(9): 1232-1238.

Briesemeister S., Rahnenführer J., Kohlbacher O. (2010b) *YLoc – an interpretable web server for predicting subcellular localization*. Nucl. Acids Res. 38(suppl 2): W497-W502.

Chou K.C., Cai Y.D. (2002) *Using functional domain composition and support vector machines for prediction of protein subcellular location*. J. Biol. Chem. 277(48): 45765-45769.

Chou K.C., Cai Y.D. (2003) *Prediction and classification of protein subcellular location-sequence order effect and pseudo amino acid composition*. J.Cell Biochem. 90(6): 1250-1260.

Chou K.C., Shen H.B. (2006) *Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers*. J. Proteome Res. 5(8): 1888-1897.

Chou K.C., Shen H.B. (2007a) *Recent progress in protein subcellular location prediction*. Anal. Biochem. 370(1): 1-16.

Chou K.C., Shen H.B. (2007b) *Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites*. J. Proteome Res. 6(5): 1728-1734.

Chou K.C., Shen H.B. (2010) *Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms*. Nat. Sci. 2(10): 1090.

Du P., Xu C. (2013) *Predicting multisite protein subcellular locations: progress and challenges*. Exp. Rev. Proteom. 10(3): 227-237.

Emanuelsson O., Nielsen H., Brunak S., von Heijne G. (2000) *Predicting subcellular localization of proteins based on their N-terminal amino acid sequence*. J. Mol. Biol. 300(4): 1005-1016.

Fyshe A., Liu Y., Szafron D., Greiner R., Lu P. (2008) *Improving subcellular localization prediction using text classification and the gene ontology*. Bioinformatics 24(21): 2512-2517.

Gu Q., Ding Y.S., Jiang X.Y., Zhang T.L. (2010) *Prediction of subcellular location apoptosis proteins with ensemble classifier and feature selection*. Amino Acids 38(4): 975-983.

Höglund A., Dönnes P., Blum T., Adolph H.W., Kohlbacher O. (2006) *MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition*. Bioinformatics 22(10): 1158-1165.

Horton P., Park K.J., Obayashi T., Fujita N., Harada H., Adams-Collier C.J., Nakai K. (2007) *WoLF PSORT: protein localization predictor*. Nucl. Acids Res. 35(suppl 2): W585-W587.

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(1): 1.

He J., Gu H., Liu W. (2012) *Imbalanced multi-modal multi-label learning for subcellular localization prediction of human proteins with both single and multiple sites*. PloS One 7(6): e37155.

Hasan M.A.M., Nasser M., Pal B. (2013) *On the KDD'99 Dataset: support vector machine based intrusion detection system (IDS) with different kernels*. Intern. J. Electron. Commun. Comp. Eng. 4(4): 1164-1170.

Hasan M.A.M., Nasser M., Pal B., Ahmad S. (2014) *Support vector machine and random forest modeling for intrusion detection system (IDS)*. J. Int. Learn. Syst. Appl. 6(1): 45.

Hasan M.A.M., Nasser M., Pal B., Ahmad S., Molla M.K.I. (2015) *Prediction of multi-label protein subcellular location using support vector machine with proper kernel selection*. Second International Conference on Theory and Application of Statistics 32.

King B.R., Guda C. (2007) *ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes*. Genome Biol. 8(5): R68.

Lu Z., Szafron D., Greiner R., Lu P., Wishart D.S., Poulin B., Eisner R. (2004) *Predicting subcellular localization of proteins using machine-learned classifiers*. Bioinformatics 20(4): 547-556.

Lee K., Chuang H.Y., Beyer A., Sung M.K., Huh W.K., Lee B., Ideker T. (2008) *Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species*. Nucl. Acids Res. 36(20): e136-e136.

Lin H.N., Chen C.T., Sung T.Y., Ho S.Y., Hsu W.L. (2009) *Protein subcellular localization prediction of eukaryotes using a knowledge-based approach*. BMC Bioinformat. 10(15): 1.

Li L.Q., Kuang H., Zhang Y., Zhou Y., Wang K.F., Wan Y. (2011) *Prediction of eukaryotic protein subcellular multilocalisation with a combined KNN-SVM ensemble classifier*. J. Comput. Biol. Bioinform. Res. 3(2): 15-24.

Li L., Zhang Y., Zou L., Li C., Yu B., Zheng X., Zhou Y. (2012) *An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity*. PLoS One 7(1): e31057.

Mak M.W., Guo J., Kung S.Y. (2008) *PairProSVM: protein subcellular localization based on local pairwise profile alignment and SVM*. IEEE/ACM Trans. Comput. Biol. Bioinform. 5(3): 416-422.

Nakai K., Kanehisa M. (1991) *Expert system for predicting protein localization sites in gram negative bacteria*. Proteins 11(2): 95-110.

Nakashima H., Nishikawa K. (1994) *Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies*. J. Mol. Biol. 238(1): 54-61.

Nielsen H., Engelbrecht J., Brunak S., Heijne G.V. (1997) *A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites*. Int. J. Neural Syst. 8(05n06): 581-599.

Nair R., Rost B. (2002) *Inferring sub-cellular localization through automated lexical analysis*. Bioinformatics 18(suppl. 1): S78-S86.

Park K.J., Kanehisa M. (2003) *Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs*. Bioinformatics 19(13): 1656-1663.

Petsalaki E.I., Bagos P.G., Litou Z.I., Hamodrakas S.J. (2006) *PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization*. Genom. Proteom. Bioinform. 4(1): 48-55.

Schölkopf B., Smola A.J. (2002) *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Scott M.S., Thomas D.Y., Hallett M.T. (2004) *Predicting subcellular localization via protein motif co-occurrence*. Genome Res. 14(10a): 1957-1966.

Shatkay H., Höglund A., Brady S., Blum T., Dönnes P., Kohlbacher O. (2007) *SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data*. Bioinformatics 23(11): 1410-1417.

Shen H.B., Yang J., Chou K.C. (2007) *Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction*. Amino Acids 33(1): 57-67.

Shin C.J., Wong S., Davis M.J., Ragan M.A. (2009) *Protein-protein interaction as a predictor of subcellular location*. BMC Syst. Biol. 3(1): 1.

Simha R., Shatkay H. (2014) *Protein (multi-) location prediction: using location inter-dependencies in a probabilistic framework*. Algorithms Mol. Biol. 9(1): 1.

Simha R., Briesemeister S., Kohlbacher O., Shatkay H. (2015) *Protein (multi-) location prediction: utilizing interdependencies via a generative model*. Bioinformatics 31(12): i365-i374.

Tsoumakas G., Katakis I., Vlahavas I. (2009) *Mining multi-label data. Data mining and knowledge discovery handbook*. Springer US.

Vladimir N.V. (1995) *The nature of statistical learning theory*. Springer-Verlag New York.

Wang X., Li G.Z., Liu J.M., Zhao R.W. (2011) *Multi-label learning for protein subcellular location prediction*. Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference 282-285.

Wan S., Mak M.W., Kung S.Y. (2012) *mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines*. BMC Bioinformat. 13(1): 1.

Wan S., Mak M.W., Kung S.Y. (2013) *GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition*. J. Theor. Biol. 323: 40-48.

Wan S., Mak M.W., Kung S.Y. (2014) *HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins*. PloS One 9(3): e89545.

Wan S., Mak M.W., Kung S.Y. (2015) *mPLR-Loc: An adaptive decision multi-label classifier based on penalized logistic regression for protein subcellular localization prediction*. Anal. Biochem. 473: 14-27.

Wang X., Zhang J., Li G.Z. (2015) *Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble*. BMC Bioinformatics 16(Suppl 12): S1.

Xiao X., Wu Z.C., Chou K.C. (2011a) *A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites*. PloS One 6(6): e20592.

Xiao X., Wu Z.C., Chou K.C. (2011b) *iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites*. J. Theor. Biol. 284(1): 42-51.

Yang W.Y., Lu B.L., Yang Y. (2006) *A comparative study on feature extraction from protein sequences for subcellular localization prediction*. Computational Intelligence and Bioinformatics and Computational Biology, 2006 IEEE Symposium 1-8.

Yu C.S., Cheng C.W., Su W.C., Chang K.C., Huang S.W., Hwang J.K., Lu C.H. (2014) *CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation*. PloS One 9(6): e99368.

Zou L., Wang Z., Huang J. (2007) *Prediction of subcellular localization of eukaryotic proteins using position-specific profiles and neural network with weighted inputs*. J. Genet. Genomics 34(12): 1080-1087.

Zhang S., Xia X., Shen J., Zhou Y., Sun Z. (2008) *DBMLoc: a Database of proteins with multiple subcellular localizations*. BMC Bioinformatics 9(1): 127.