Talanta 147 (2016) 609–614

Contents lists available at ScienceDirect

Talanta

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

Multiplex protein pattern unmixing using a non-linear variable-weighted support vector machine as optimized by a particle swarm optimization algorithm

Qin Yang [a,b], Hong-Yan Zou [c], Yan Zhang [a], Li-Juan Tang [a,*], Guo-Li Shen [a], Jian-Hui Jiang [a], Ru-Qin Yu [a,*]

[a] State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China
[b] School of Physics and Optoelectronic Engineering, Yangtze University, Jingzhou 434023, China
[c] Key Laboratory of Luminescence and Real-Time Analytical Chemistry, Ministry of Education, College of Pharmaceutical Science, Southwest University, Chongqing 400715, China

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form
14 October 2015
Accepted 18 October 2015
Available online 21 October 2015

Keywords:
Protein distribution
Pattern unmixing
Support vector machine
Variable weight
Non-linear machine learning

ABSTRACT

Most of the proteins locate more than one organelle in a cell. Unmixing the localization patterns of proteins is critical for understanding the protein functions and other vital cellular processes. Herein, non-linear machine learning technique is proposed for the first time upon protein pattern unmixing. Variable-weighted support vector machine (VW-SVM) is a demonstrated robust modeling technique with flexible and rational variable selection. As optimized by a global stochastic optimization technique, particle swarm optimization (PSO) algorithm, it makes VW-SVM to be an adaptive parameter-free method for automated unmixing of protein subcellular patterns. Results obtained by pattern unmixing of a set of fluorescence microscope images of cells indicate VW-SVM as optimized by PSO is able to extract useful pattern features by optimally rescaling each variable for non-linear SVM modeling, consequently leading to improved performances in multiplex protein pattern unmixing compared with conventional SVM and other exiting pattern unmixing methods.

1. Introduction

Proteins are involved nearly every work in cells, and the genome of each living organism encodes a wide range of proteins. Some proteins are made in all cells of an organism and some are only made in particular cell types [1]. Accurate knowledge of protein subcellular localizations is important for elucidating their functions, understanding cellular processes as well as facilitating the identification of the drugs [2–4]. These multiplex proteins together with their dynamic features have attracted great interests because of their potential unique biological functions [5]. Nevertheless, it is common for proteins to exhibit intricate spatial distributions between several subcellular locations. Approximately 60% of human proteins have been recognized localizing in multiple organelles [6,7]. It is, thus, a hard and challenge task of determining, describing or predicting of protein distribution patterns within cells.

In earlier studies, the presence or absence of a specific protein in a subcellular organelle was predicted based on the knowledge of the biological role of proteins [8–10]. For example, Gene Ontology [11], an available tool for the unification of biology, is invoked to roughly explain the place in cells where a gene product is active [12], but it is unable to offer any further information about the proteins in every organelle regarding their amount or dynamic changes. Cell imaging techniques, including high-throughput fluorescence microscopy, have enabled large-scale collection of subcellular organelle images showing the distribution of fluorescently tagged proteins [13]. Relying on these fluorescence microscope images, subcellular protein location determining are also carried out via the examination by human experts, which has became the most common method in this field [5,14]. However, visual examination of multiplex proteins is not only quite labor intensive, but such a subjective approach also easily results in different interpretations from investigator to investigator. Besides, it lacks applicable terms capable of describing subtle distribution differences that proteins display.

Abbreviations: VW-SVM, variable-weighted support vector machine; SOF1, subcellular object features 1; AIC, Akaike Information Criterion; PSO, particle swarm optimization; ILS, individual learning strategy; CLS, collective learning strategy; R, correlation coefficient; RMSE, root mean squared error

* Corresponding authors.

http://dx.doi.org/10.1016/j.talanta.2015.10.047
0039-9140/

In a more objective manner, some efforts have been made in developing automated analysis techniques to interpret fluorescence microscope images in terms of the subcellular protein distribution patterns [15,16]. Previous studies have created automated classifiers and have demonstrated they can identify the patterns of all major subcellular locations with higher accuracy than visual analysis [17–19]. Automated systems also have been developed to learn what subcellular patterns are present in large collections of images without prior knowledge of the possible patterns and to quantify the amount of fluorescently tagged proteins that the patterns contain based on object extraction, feature calculation and object type learning [16]. Furthermore, recent studies have reported machine-learning approaches for estimating the amount of fluorescent signal in different subcellular organelles without extensive hand-tuning of algorithms [13,20]. In these studies, the mixing subcellular distribution patterns were approximately estimated by using single-location distribution patterns of a protein. The single-location distribution patterns, treated as fundamental patterns, were learned from the cell fluorescence images and described the frequencies of interested protein at each subcellular organelle [16]. Based on these fundamental patterns, the estimation of protein mixing patterns was realized via performing linear combinations of these fundamental patterns. Inversely, the fundamental patterns of a protein can also be resolved from its mixing patterns by solving linear equations. Although the estimated patterns resolved using linear unmixing would not be the real status of protein in subcelluar organelles, they still approximately informed the patterns of protein in an easily accessible manner. The significant advantages of this approach are it is cell type independent and requires only the acquisition of separate training images of fluorescent markers for each subcellular organelle [13].

Considering non-linear machine learning techniques generally have better performance in solving complex issues than linear modeling ones, herein, a non-linear pattern unmixing method is proposed based on variable-weighted support vector machine (VW-SVM) [21] for the automated estimation of subcellular protein mixing patterns. SVM has been proofed to be a robust machine learning technique widely used in establishing both linear and non-linear models [22]. Coupled with particle swarm optimization (PSO) algorithm [23], VW-SVM was developed as a parameter-free modeling technique which enables the construction of a rational and self adaptive prediction model according to the performance of the total model [21]. Otherwise, rather than in the previous study it supposed mixing patterns are the linear combination of fundamental ones, in the present study it hypothesizes that any pattern of protein distribution can be resolved by a non-linear VW-SVM model trained using a data set made by other patterns, mixing one, fundamental one or both of them, thus, guaranteeing a more flexible strategy for unmixing of complex protein patterns.

2. Dataset

To demonstrate the feasibility of the non-linear machine learning method in protein pattern unmixing, VW-SVM as optimized by PSO is applied to the estimation of the subcellular protein location patterns. The cell image dataset was created by Murphy et al. using high-throughput automated microscopy [13]. It resulted in 64 images by labeling cells with varying mixtures of fluorophore-tagged mitochondrial and lysosomal probes (Mitotracker and Lysotracker) indicating the proteins locating in mitochondria and lysosomes. The image set is available at http://murphylab.web.cmu.edu/data/.

For image processing and feature extraction, briefly, binary images are first obtained by applying an automated threshold method to distinguishing probe-containing from nonprobe-containing pixels in all qualified images [13]. Then, each set of connected above-threshold pixels, which defined an object, is identified and described using a set of 11 variables (SOF1) to characterize the morphological and spatial properties of the object, as proposed in previous studies [16]. After that, object type learning is performed in two different methods for comparison. One is individual learning strategy (ILS), in which type learning is conducted on each protein location pattern from training or test set using *K*-means method. The value of k is set to be 11 in this strategy [13]. The other one is a collective learning strategy (CLS), in which type learning is performed on all the objects presenting in the training images. Then, the objects found in a training set are clustered using *K*-means to identify their types with the best value of *K* accessed under the Akaike Information Criterion (AIC) [24,25]. For test set, each protein object in an image is assigned to the cluster whose center was closest to it in the feature space. Based on the identified *K*-object types, each image is represented as a vector $\mathbf{x}=(x_1,\ldots, x_K)$ with x_k ($k=1,\ldots K$) defines as the frequency of kth-object type in that image. Taking into account that object SOF1 and the fluorescence intensity in an image also depicts important features of that image, therefore, besides object frequencies, these two sorts of image features, modified SOF1 and fluorescence intensity are also used to constitute the variable set of each image in the current study. Modified SOF1 values are obtained by average the eleven SOF1 features of all objects in an image and fluorescence intensity is the amount of fluorescence an image contains. Hence, for each image, it could be represented by a vector with $K+12$ variables.

The Jackknife test [12] is deemed as one of the most objective methods for assessing the performance of an algorithm. Therefore, 8-fold cross validation is adopted to demonstrate the performance of non-linear VW-SVM in protein pattern unmixing. The average results of 8-fold cross validation are reported.

3. Methods

3.1. Variable-weighted support vector machine (VW-SVM)

Support vector machine has been a widely data mining and modeling technique in the past decades [26–29]. Our previous study [21], automated data mining and modeling was realized in VW-SVM as optimized using PSO, in which it proposed flexible weighting of all variables rather than simply reserving or abandoning some variables during variable selection procedure considering that every variable in a dataset may contribute to a model more or less. It has demonstrated that variable weighting treatment is able to further greatly improve the performance of SVM when solving a non-linear regression problem [30]. The basic theory of VW-SVM is briefly interpreted as following.

Suppose $\mathbf{X}$ is a $P\times I$ matrix with P variables for I protein distribution patterns (images), and $\mathbf{y}$ is a vector representing the corresponding dependent variable for the I patterns. Herein, y_i is the pattern fraction or the probe concentration in a target cellular organelle for the Ith pattern. Contrasting with an ordinary SVM model, where variables in $\mathbf{X}$ are considered as making the same contribution to the regression model, in VW-SVM, $\mathbf{X}$ is left multiplied by a diagonal matrix diag($\mathbf{w}_a$),

$$\mathbf{X}_{\mathbf{W}_a} = diag\,(\mathbf{w}_a)\mathbf{X} \tag{1}$$

where $\mathbf{w}_a$ is a $P\times 1$ variable weighting vector with all the elements being non-negative values and diag($\mathbf{w}_a$) is a $P\times P$ matrix whose diagonals are the elements of $\mathbf{w}_a$. Consequently, VW-SVM model is