Journal of Theoretical Biology 340 (2014) 85–95

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Protein subcellular localization in human and hamster cell lines: Employing local ternary patterns of fluorescence microscopy images

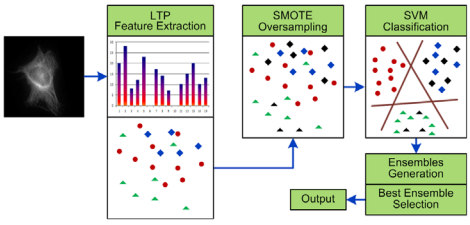
Muhammad Tahir [a], Asifullah Khan [a,*], Hüseyin Kaya [b]

[a] Department of Computer and Information Sciences, PIEAS, Islamabad, Pakistan
[b] Department of Biophysics, Faculty of Medicine, University of Gaziantep, 27310 Gaziantep, Turkey

HIGHLIGHTS

- LTPs exploit small variations in intensities of Human and Hamster protein images.
- SMOTE oversampling is utilized to increase the minority class samples.
- SVM shows significance performance improvement for balanced data.
- mRMR is not required for the performance improvement of LTPs.
- A web server is available online at http://111.68.99.218/Protein_SubLoc.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 May 2013
Received in revised form
9 July 2013
Accepted 15 August 2013
Available online 27 August 2013

Keywords:
Support vector machine
SMOTE
mRMR

ABSTRACT

Discriminative feature extraction technique is always required for the development of accurate and efficient prediction systems for protein subcellular localization so that effective drugs can be developed. In this work, we showed that Local Ternary Patterns (LTPs) effectively exploit small variations in pixel intensities; present in fluorescence microscopy based protein images of human and hamster cell lines. Further, Synthetic Minority Oversampling Technique is applied to balance the feature space for the classification stage. We observed that LTPs coupled with data balancing technique could enable a classifier, in this case support vector machine, to yield good performance. The proposed ensemble based prediction system, using 10-fold cross-validation, has yielded better performance compared to existing techniques in predicting various subcellular compartments for both 2D HeLa and CHO datasets. The proposed predictor is available online at: http://111.68.99.218/Protein_SubLoc/, which is freely accessible to the public.

1. Introduction

Protein is the crucial part of a cell in all living organisms to function properly. Among numerous characteristics, subcellular localization is the most important property of proteins (Chebira et al., 2007). Understanding the behaviour of individual protein is the key to the comprehension of various functions of cells in living organisms. A protein must reside in its natural localization to work properly. Hence, precise knowledge of protein subcellular localization allows to elucidate various protein functions (Lin et al., 2007). In addition, various cellular processes of hypothetical and newly revealed proteins can easily be described by the protein localization (Boland and Murphy, 2001; Murphy et al., 2000). Further, subcellular localization can aid in drug discovery (Nanni and Lumini, 2008). For instance, plasma membrane proteins and secreted proteins are easily reachable by drug molecules since they are located on the cell surface (Tscherepanow et al., 2008). Subcellular localization also helps in early diagnostics of various

* Corresponding author at: Postal address: Pattern Recognition Laboratory, Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, Nilore 45650, Islamabad, Pakistan.
Tel.: +92 51 2207380-84; fax: +92 51 2208070.
E-mail addresses: asif@pieas.edu.pk, khan.asifullah@gmail.com (A. Khan).

0022-5193/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.jtbi.2013.08.017

diseases. For example, aberrant subcellular localization has been observed in the cells affected by various diseases such as Alzheimer′s and cancer. In addition, the environment in which proteins function properly can also be determined by accurately finding their subcellular localization (Chen et al., 2006). In brief, accurate knowledge of protein localization is helpful in the identification as well as effectiveness of drugs.

Experimental classification of protein localization is time consuming and laborious (Murphy et al., 2000). Therefore, computational methods coupled with machine learning techniques are required to determine protein subcellular localizations. In this regard, researchers have endeavoured to develop numerous bioinformatics based prediction systems coupled with machine learning methods to localize a range of proteins (Chebira et al., 2007; Hamilton et al., 2007; Khan et al., 2008, 2011; Lin et al., 2007; Murphy et al., 2003; Nanni and Lumini, 2008; Nanni et al., 2010a; Zhang et al., 2009). Researchers have confirmed that many proteins have been found to be the part of a multi-label system in which they are able to reside in two or more subcellular locations simultaneously or travel across two or more subcellular location sites. This property of proteins, making them unique in their biological functionality, is of particular interest (Chou, 2013). In this connection, substantial efforts have been endowed for the last three years to localize multiplex proteins in addition to the singleplex proteins. In this regard, various prediction systems have been developed focussing on different organisms including animal (Lin et al., 2013), human (Chou et al., 2012), bacterial proteins (Wu et al., 2012; Xiao et al., 2011a), plant (Wu et al., 2011), virus (Xiao et al., 2011b), and Eukaryotic Proteins (Chou et al., 2011).

From the literature survey, it is observed that sequence based methods cover more subcellular location sites compared to image-based methods. In addition, singleplex and multiplex proteins are mostly covered by sequence based techniques. For example, a benchmark dataset utilized in Chou et al. (2011) covers 22 subcellular locations. Similarly, another benchmark dataset reported in Lin et al. (2013) has 20 subcellular location sites. Likewise, 14 subcellular locations of human proteins are reported in Chou et al. (2012). Due to the wider coverage of protein location sites, the applications of sequence-based methods are more likely. However, the current work is based on the fluorescence microscopy images; therefore, the singleplex proteins are targeted in order to simplify the treatment. The proposed method covers 10 and 8 subcellular locations in HeLa and CHO datasets.

Literature survey has revealed that both individual and ensemble classifiers have been employed in conjunction with various feature extraction strategies to accurately predict subcellular localization (Chen et al., 2006; Hu and Murphy, 2004). A model has been developed in which a random subspace of local binary and ternary patterns with high variance is selected. Reduced dimensionality is achieved through Neighbourhood Preserving Embedding. Further, support vector machine (SVM) is trained using the reduced dimensionality space (Nanni et al., 2010b). Similarly, an adaptive multi-resolution approach has been proposed in which Haralick textures and morphological features are extracted at the sub-bands. The predictions at different sub-bands are obtained by utilizing *k*-means algorithm and weighting (Srinivasa et al., 2006). In another approach, a random subspace of Levenberg–Marquardt neural networks and a variant of the AdaBoost learning algorithms are trained using hybrid feature sets. The decisions of the two ensembles are fused together through sum rule (Nanni et al., 2010c). Similarly, a model based on back propagation neural network has been reported, which employs Haralick textures, Zernike moments, and morphological features for protein subcellular localization (Murphy et al., 2003). Likewise, an SVM based model is proposed, which utilizes Zernike moments, Threshold Adjacency Statistics (TASs), and hybrid feature space of TASs and Haralick textures (Hamilton et al., 2007). An Artificial Neural Network based prediction system has been proposed, which utilizes Haralick, morphological and Zernike moment based features in multi-resolution subspaces. Final decision is made through weight assignment (Chebira et al., 2007).

The existing approaches have shown great improvement in achieving higher accuracies; however, we need systems, which are capable of achieving nearly 100% accuracy particularly, in diagnosing cancer like diseases. In addition, the nature of available unbalanced data should also be taken into account to enhance the performance of the classifier. In our previous work, we have focussed on the same problem exploiting different spatial and transform domain features (Tahir et al., 2012). The ensemble classification based on different SVM kernels has achieved 99.7% accuracy for the 2D HeLa dataset. However, we did not take into account the unbalanced nature of the data. The principal focus of this study is to develop a model that is reliable, efficient, and highly accurate even in the presence of unbalanced data keeping the feature space as small as possible. In order to have small and well discriminative feature space, Local Ternary Patterns (LTPs) (Nanni et al., 2010b) have been utilized for the classification of protein subcellular localization images. Further, Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) has been employed to balance the datasets. The feature selection technique; minimum Redundancy Maximum Relevance (mRMR) has also been utilized to know whether the feature space has any overlapping information. Polynomial SVM of degree 2 has been employed to nonlinearly transform the input features to make them linearly separable. The proposed novel combination of LTPs, SMOTE, and SVM performed well for protein classification compared to the existing ensemble classification techniques, especially, due to the exploitation of discriminative capability of LTPs as well as availability of balanced feature space constructed using SMOTE. The discrimination power of LTPs is better because of its less sensitive behaviour in homogenous image regions.

The rest of the paper is structured as follows. Section 2 discusses materials and methods. Section 3 describes the performance measures used in this paper. Section 4 is devoted to the analysis of experimental results. Section 5 draws the conclusion at the end.

2. Materials and methods

In this section, we present the datasets, the feature extraction and post-processing techniques as well as the classifier adopted to develop the proposed model.

2.1. Datasets

We have tested the performance of proposed model using 2D HeLa and Chinese Hamster Ovary (CHO) datasets from Murphy′s Lab (Murphy, 2004) and AIIA lab (Lin et al., 2007), respectively. In 2D HeLa dataset, there are 862 images distributed in ten distinct classes including ActinFilaments, Endosome, ER, Golgi Giantin, Golgi GPP130, Lysosome, Microtubules, Mitochondria, Nucleolus, and Nucleus. Sample image from each class is shown in Fig. 1. Image distribution, per class in the original and oversampled datasets, is illustrated in Fig. 2.

On the other hand, CHO dataset has 668 protein images grouped in eight different categories, which include Actin, ER, Golgi, Microtubule, Mitochondria, Nucleolus, Nucleus, and Peroxisome. Some sample images are illustrated in Fig. 3. Comparison of the synthetic samples and original samples are shown in Fig. 4.