Neurocomputing 131 (2014) 113–123

Contents lists available at ScienceDirect

Neurocomputing

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

Image-based classification of protein subcellular location patterns in human reproductive tissue by ensemble learning global and local features

Fan Yang [a], Ying-Ying Xu [a], Shi-Tong Wang [b], Hong-Bin Shen [a,c,*]

[a] *Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China*
[b] *School of Digital Media, Jiangnan University, Wuxi 214122, China*
[c] *Shanghai Key Laboratory for Reproductive Medicine, Shanghai 200025, China*

ARTICLE INFO

Article history:
Received 14 July 2013
Received in revised form
16 September 2013
Accepted 12 October 2013
Communicated by: L. Kurgan
Available online 11 November 2013

Keywords:
Bioimage processing
Subcellular location prediction
Local texture pattern
Feature selection
Ensemble classifier

ABSTRACT

The reproductive system is a specific system of organs working together for the purpose of reproduction. As one of the most significant characteristics of human cell, subcellular localization plays a critical role for understanding specific functions of mammalian proteins. In this study, we developed a novel computational protocol for predicting protein subcellular locations from microscope images of cells in human reproductive tissues. Three major steps are contained in this protocol, i.e., protein object identification, image feature extraction, and classification. We first separated protein and DNA staining in the images with both linear and non-negative matrix factorization separation methods; then we extracted protein multi-view global and local texture features including wavelet Haralick, local binary patterns, local ternary patterns, and the local quinary patterns; finally based on the selected important feature subset, we constructed an ensemble classifier with support vector machines for classifications. Experiments are performed on a benchmark dataset consisting of seven major subcellular classes in human reproductive tissues collected from human protein atlas. Our results show that the local texture pattern features play an important complementary role to global features for enhancing the prediction performance. An overall accuracy of 85% is obtained through current system, and when only confident classifications are considered, the accuracy can reach 99%. It is the first developed image based protein subcellular location predictor specifically for human reproductive tissue. The promising results indicate that the developed protocol can be applied for accurate large-scale protein subcellular localization annotations in human reproductive system.

1. Introduction

The human reproductive system is a unique system of organs working together for reproduction. This process is simply known as creating new life. At the cellular level, the reproductive system is composed of many cells consisting of thousands of proteins, which are the very basic biological molecules in cell. Moreover, the genome of human encodes tens of thousands of proteins, some of which are made in all cells and some of which are made only in specific cell types. To understand the biochemical context of specific functions of a protein, it is generally acknowledged that its subcellular location is one of the most crucial characteristics [1,2], as it can describe the behaviors of a protein under various cell conditions and is thus helpful in understanding the protein's functions. It is extremely important for a protein to appear at the right place at a right time to guarantee its normal functionalities, e.g. finding its correct interaction molecular partners. Protein mislocalizations may cause serious diseases, such as cancer [3,4], hence, annotating the protein subcellular locations in the human reproduction cell is one of the most important tasks of reproductive medicine studies.

Although the protein subcellular localizations could be identified by various biological experiments, which are both time-consuming and expensive. As a matter of course, it is highly desired for the researchers to develop automated protein subcellular localization classification systems with high accuracy. Current efforts on automatic predictions of protein subcellular localizations can be generally grouped into two groups depending on how to represent the protein targets: (1) 1-D sequence based approaches [5,6] and (2) 2-D image based classification systems

[*] Corresponding author at: Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China.
Tel.: +86 21 342 053 20; fax: +86 21 342 040 22.
E-mail address: hbshen@sjtu.edu.cn (H.-B. Shen).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.10.034

[4,7–9]. If the proteins are represented in 1-D amino acid sequence, e.g. through sequencing technology, then methods of the former category can be applied. The algorithms aim to predict protein subcellular localizations through the features extracted from amino acid sequences, such as amino acid composition and gene ontology information [5,6]. If the proteins are represented with 2-D images, e.g. through fluorescence microscopy, then the approaches in the second group are suitable, which accomplish this task by image processing algorithms. While the sequence of a certain protein is invariant in each human tissue and it cannot reveal the different distribution patterns in the cell of the same protein under different conditions, a growing number of researchers and institutions devote their efforts to the development of 2-D medical and bioimage-based classification systems.

In the early stage, all the image-based subcellular location prediction systems suffered from the dilemma: the very limited amount of proteins with known location labels, insufficient detailed description of the locations, and the low resolution of the images [10]. Fortunately, significant progress has been made in recent years due to the development in relative areas such as biological knowledge, digital microscopy, tissue microarray technology and digital image processing techniques. These advancements make it possible to realize more accurate automatic analysis of the protein subcellular locations from immunohistochemistry and fluorescence microscopy images. Over the past fifteen years, the feasibility of automated analysis subcellular location patterns has been extensively demonstrated, e.g. by Murphy group [11], and many other groups [12,13]. These efforts can be generally grouped into the following categories:

(1) High quality benchmark dataset collections. Many groups have published their experimental research results in different journals. In order to solve the problem of collecting enough high quality images for training classification system, Murphy and Kou developed a system called subcellular location image finder (SLIF) for collecting images from on-line publications based on genome-wide determination [14].

(2) Image descriptor development. For a classification system, it is important to use proper descriptor to represent an image. The subcellular location features (SLFs) set has been proposed to describe protein subcellular distributions, which is composed of global texture features and demonstrated being robust to cell rotations and translation [10,11]. Besides the global texture features, the invariant locally binary patterns (LBP) has also been proposed for the classification of protein subcellular localization task [15].

(3) Classification approach effort. The accuracy of the whole classification system is also highly dependent on the learning algorithms. Considering of this, many efforts have been devoted to constructing an accurate learning model. For example, the back propagation neural network (BPNN) has been studied in [16]. Chen and Murphy proposed a new algorithm by combining graphical model with support vector machine (SVM) to improve the classification accuracy of subcellular patterns in multi-cell fluorescence microscope images [17]. Nanni et.al. studied the ensemble classifier by fusing different machine learning techniques for automated cell phenotype image classification, and obtained an outstanding accuracy on the 2D-Hela dataset [2]. For the multi-label protein subcellular location classification problem, Xu et.al has proposed the classifier chain model to incorporate the label dependency, which has been demonstrated effective [4].

(4) Organism specific studies. It has also been noticed that many specific prediction systems are reported for different organisms because of the organism-specific features and organization components, such as systems for human cell [11] and yeast cell [18].

Although much progress has been achieved, no system has been constructed for the reproductive tissue cells. The reproductive system is well-known for the unique embryonic stem cells of epigenetic features [19] that has distinguished characteristics from others calling for specific analysis systems. Because of this, current study aims to develop a specific prediction protocol in the human reproductive tissues. Beyond the global texture image descriptor applied in the existing studies, our system has also investigated some local texture features, including local binary pattern (LBP) and two variants of local ternary pattern (LTP) and local quinary pattern (LQP) features [20,21]. Both the feature selection outputs and the improvements of the classification accuracies demonstrate that the local texture features are important to distinguish different subcellular location classes. Our results also show that the ensemble of combining the diversities from different classifiers is also helpful for the performance improvement.

2. Datasets and methods

2.1. Datasets

The immunohistochemistry (IHC) images from the human protein atlas (HPA) (http://www.proteinatlas.org/), which is a bountiful source of location proteomics data, were employed in this study. Current release of HPA (version 10.0) contains 14,079 genes with protein expression profiles based on 17,298 antibodies corresponding to 46 different normal human tissues and 20 different cancer types. To avoid the organelles overlap problem in tissue image, tissue microarray technology (TMA) has been used to make tissue into many slides, and each slide is stained for a different protein [11]. Thus, a certain protein is shown in brown and DNA in purple in a stained slide. Then, pictures are taken by RGB camera for these slides and saved in HPA.

We have generated a collection of IHC images of human reproductive tissue from the HPA as our benchmark dataset, where the tissues images are generated using IHC based brightfield microscope. The confidences of proteins in HPA are scored based on two aspects: the validation score and the reliability score. The former is the confidence level of supporting the specificity of the antibody towards expected human target protein, e.g. IH and Western blot (WB) validation scores; the latter is set for proteins where two or more antibodies are available and used as a performance estimation of knowledge-based annotation of protein expression.

In this study, in order to collect high quality image samples, all of the proteins we selected are based on the combination of the validation score and the reliability score, i.e., supportive validation and high reliability. Table 1 summarizes the benchmark dataset, including the antibody identification numbers in HPA, the corresponding subcellular class, and the number of images of each protein. Namely, a set of 14 proteins of 353 images in 9 normal reproductive tissues belonging to 7 subcellular locations has been collected in this study. The seven subcellular locations are endoplasmic reticulum (ER), cytoskeleton, Golgi apparatus, mitochondria, nucleolus, nucleus, and vesicles, respectively.

2.2. Object identification

Each image in HPA contains many cells. Many biological image analysis approaches require segmentation of the whole image into regions of interest as a preprocessing step, which is recognized as a very challenging problem. Fortunately, it has been revealed that the classification accuracy would be still encouraging by the multi-cell protein images directly [11]. Considering of this, we will also follow the previous method by directly using the multi-cell image