Subcellular Localisation of Proteins in Living Cells Using a Genetic Algorithm and an Incremental Neural Network

Marko Tscherepanow and Franz Kummert

Applied Computer Science, Faculty of Technology,
Bielefeld University, P.O. Box 100 131, D-33501 Bielefeld
Email: {marko, franz}@techfak.uni-bielefeld.de

Abstract. The subcellular localisation of proteins in living cells is a crucial means for the determination of their function. We propose an approach to realise such a protein localisation based on microscope images. In order to reach this goal, appropriate features are selected. Then, the initial feature set is optimised by a genetic algorithm. The actual classification of possible protein localisations is accomplished by an incremental neural network which not only achieves a very high accuracy, but enables on-line learning, as well.

1 Introduction

Location Proteomics, i.e. the automatic subcellular localisation of many or all proteins of a cell, has made considerable progress during the last decade [1]. By investigating fluorescence images of tagged proteins in living cells, essential information about their functions can be obtained. This knowledge is applicable for the simulation of cell behaviour which might facilitate the investigation of diseases and the development of novel drugs.

2 State of the art and new contribution

In comparison to the direct application of pixel intensities, the usage of numerical features has proven advantageous for the classification of fluorescence images showing tagged proteins [1, 2]. The feature sets proposed in the literature comprise, for instance, morphological data of binary image structures, Zernike moments and edge information. Wide-field microscope images are usually preprocessed by digital deconvolution in order to enhance the contrast.

Since unnecessary features adversely influence the result of the classification if too small a number of training samples is available and increase the computational effort, they should be removed. Several methods have been applied in order to achieve this goal [3]. At this, stepwise discriminant analysis (SDA) and a genetic algorithm have attained particularly good results. As classifiers, multilayer perceptrons (MLPs) [4] and support vector machines (SVMs) [3] are utilised frequently.

We propose an approach to protein localisation in living *Spodoptera frugiperda* cells (Sf9) which does not require digital deconvolution as a preprocessing step, thereby reducing the computational effort. In addition, we employ a classifier which has been developed for incremental learning. So, in principle, potential users can incorporate new data during the application. The relevance of features is determined by a genetic algorithm. In contrast to other approaches, here, no binary masking is performed $\big($cf. [3, 5]$\big)$. So, discontinuities in the optimisation function are avoided. Finally, the protein localisation is adapted for a cell recognition method introduced in [6]. Since automatic cell recognition constitutes a crucial precondition for performing an automated protein localisation, this connection enables a better collaboration enhancing the performance of the final complete system.

3 Methods

The protein localisations are classified based on three different types of features:

(i) Zernike moments [7] which are sensitive to the position of tagged proteins with respect to the surrounding cell,
(ii) granulometries [8] enabling the investigation of the shape and the size of protein accumulations directly using the image intensities,
(iii) and fractal features [9] allowing for the determination of the granularity and self-similarity at different scales.

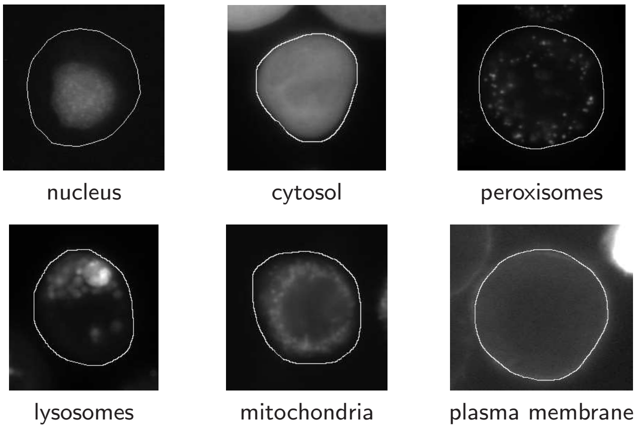
Instead of utilising classifiers which compute discrimination planes, we employ the simplified fuzzy ARTMAP (SFAM) [10], an incremental neural network. Its neurons span hyper-rectangular regions in the feature space – the categories. Their maximal size is determined by the vigilance parameter ρ. An input which is to be classified receives the label of the best-matching category enclosing it. Unknown inputs can easily be rejected, as they do not belong to an existing category. We have extended the subspace of known inputs by a small distance τ from each category so as to cope with slightly varying data.

In order to assess the importance of the n available features, each input $\underline{x}$ is multiplied by a weight vector $\underline{w}$. Its components w_i as well as the parameters ρ and τ are evolved by a genetic algorithm utilising rank-based selection in order to handle slight differences in the fitness values of the population. Furthermore, arithmetic cross-over and mutation for continuous-valued genes are employed [11]. The fitness $f(X)$ corresponds to the cross-validation accuracy $\mathrm{acc}(X)$ of the classifier X diminished by a punishment for large values of τ and high weights w_i. These punishments are scaled by the constants c_τ and c_w, respectively $\big($see (1)$\big)$.

$$f(X) = \mathrm{acc}(X) - c_\tau \cdot \tau - c_w \cdot \frac{1}{n} \sum_{i=0}^{n} w_i \qquad (1)$$

So, only the weights of features which are important for obtaining a good accuracy receive high values and the considered subspace is reduced. After a run

Fig. 1. Protein distributions in Sf9 cells: The white contours represent the surrounding cells which were manually extracted from corresponding bright-field images by biological experts



nucleus

cytosol

peroxisomes

lysosomes

mitochondria

plasma membrane

of the genetic algorithm, all weight vectors are normalised in such a way that the maximal component equals 1 in order to enable the usage of the possible input space $\big(\forall i \in \{1, \cdots, n\} : x_i \in [0,1]\big)$ and to avoid multiple solutions of the optimisation function resulting from scaling the occupied region of the feature space. By considering the weights of the final generation, conclusions about the relevance of features can be drawn, as these individuals are adapted to the task at hand.

4 Results

Our approach was evaluated on 972 images of single cells manually extracted from 99 bright-field micrographs taken in parallel with each fluorescence image. Here, six different protein locations were considered: nucleus (150 cells), cytosol (164 cells), peroxisomes (71 cells), lysosomes (222 cells), mitochondria (268 cells), and plasma membrane (97 cells). Protein distributions of these six classes are depicted in Fig. 1.

In addition to the manually segmented cells, 5368 cell images were generated automatically using an active contour approach [6]. This method yields segments which resemble the manually determined cells. As we plan to utilise it for cell recognition, the resulting segments are more likely to occur during an automated application of our protein localisation technique. In addition, the number of training samples is increased which alleviates the classification task.

After computing the features of every cell image, the resulting data set was split into ten disjoint parts. Five groups of eight data sets each were used for the determination of the input weights by our genetic algorithm. Here 100 generations with 100 individuals were applied after performing preliminary trials. Results from the literature confirm that these values are sufficiently high $\big($cf. [3]$\big)$.

Table 1. Confusion matrix for manually segmented cells. The table entries represent the number of cells from a specific class i (row) which were recognised as class j (column). A correct classification is characterised by equal labels i and j

	classification result						
cell compartment	(a)	(b)	(c)	(d)	(e)	(f)	unknown
nucleus (a)	145	3	0	0	1	1	0
cytosol (b)	0	156	0	4	3	1	0
peroxisomes (c)	0	0	63	2	5	1	0
lysosomes (d)	1	5	0	195	18	3	0
mitochondria (e)	0	0	2	14	250	2	0
plasma membrane (f)	0	0	2	3	5	87	0

The parameters c_τ and c_w were chosen in such a way that the fitness is mainly determined by the accuracy (c_τ=0.02 and c_w=0.1). At this, only slight variations of the fitness were intended, since the accuracy should not be decreased by the feature selection method.

The evaluation occurred based on the remaining groups of two data sets averaging the results (five-fold cross-validation). In order to determine the fitness value of a classifier, eight-fold cross-validation was applied in each group (see Section 3). During the evaluation, manually and automatically obtained samples were distinguished, since the manually segmented cells are more biologically relevant. At this, an accuracy of 92% was achieved. Table 1 shows the corresponding confusion matrix. For the automatically determined samples, accuracies up to 94% were reached.

In order to reduce the dimensionality of the feature space, a computation of the mean weight vector over all individuals of the final generation occurred. Then, inputs with mean weights smaller than a threshold τ_w were rejected (see Fig. 2). Using a value of τ_w=0.9, the number of required features n could be decreased from 64 to 19.2 on average without impairing the classification results. Higher values of τ_w resulted in considerably reduced accuracies.

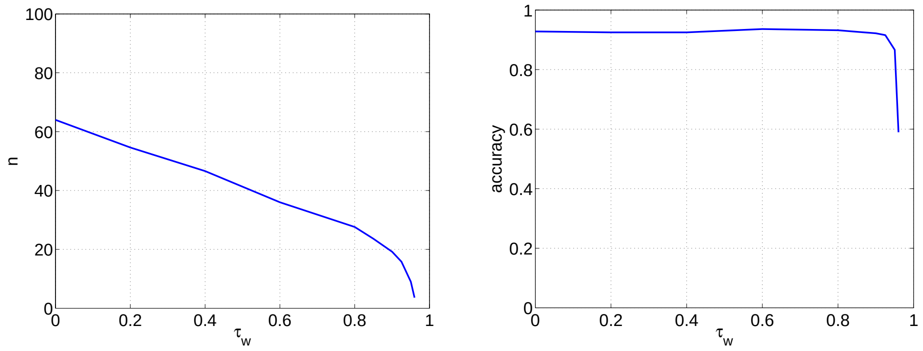
5 Discussion

We have proposed an approach to the localisation of proteins with a high accuracy. The number of the employed features, which were chosen with respect to the task at hand, was significantly reduced by means of a genetic algorithm. In contrast to known approaches, our method enables on-line learning and does not require optical deconvolution. Furthermore, it can be applied in an automated context, since it is adapted for an automatic cell recognition method.

References

1. Chen X, Velliste M, Murphy RF. Automated interpretation of subcellular patterns in fluorescence microscope images for location proteomics. Cytometry 2006;69A:631–640.

Fig. 2. Accuracy with respect to the manually segmented cells and number of required features n, which decreases if the threshold τ_w is rising or the accuracy remains high for values of τ_w up to 0.9



2. Murphy RF, Velliste M, Porreca G. Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images. Journal of VLSI Signal Processing 2003;35:311–321.
3. Huang K, Velliste M, Murphy RF. Feature reduction for improved recognition of subcellular location patterns in fluorescence microscope images. Procs SPIE 2003;4962:307–318.
4. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 2001;17(12):1213–1223.
5. Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK. Dimensionality reduction using genetic algorithms. IEEE Trans on Evolutionary Computation 2000;4(2):164–171.
6. Tscherepanow M, Zöllner F, Kummert F. Classification of segmented regions in brightfield microscope images. Procs ICPR 2006;3:972–975.
7. Khotanzad Alireza, Hong YawHua. Invariant Image Recognition by Zernike Moments. IEEE Trans on Pattern Analysis and Machine Intelligence 1990;12(5):489–497.
8. Soille P. Morphological Image Analysis: Principles and Applications. Springer; 2003.
9. Wu CM, Chen YC, Hsieh KS. Texture features for classification of ultrasonic liver images. IEEE Trans on Medical Imaging 1992;11(2):141–152.
10. Vakil-Baghmisheh MT, Pavešić N. A fast simplified fuzzy ARTMAP network. Neural Processing Letters 2003;17(3):273–316.
11. Engelbrecht AP. Fundamentals of Computational Swarm Intelligence. John Wiley & Sons; 2005.