PROTEIN CLASSIFICATION BY ANALYSIS OF CONFOCAL MICROSCOPIC IMAGES OF SINGLE CELLS

Tanja Steckling[a], Olaf Hellwich[a], Stephanie Wälter[b], Erich Wanker[b]

[a] Technical University Berlin, Computer Vision and Remote Sensing, Sekr. FR 3-1, Franklinstr. 28/29, 10587 Berlin, Germany, Phone: +49-30-314-22796, Fax: +49-30-314-21104, e-mail: hellwich@fpk.tu-berlin.de
[b] Max-Delbrück-Centrum für Molekulare Medizin (MDC), Berlin-Buch, Robert-Rössle-Str. 10, 13092 Berlin, Germany

Commission WG V/3

KEY WORDS: image analysis, feature selection, classification, medical image processing, microscopic imagery

ABSTRACT:

Proteins being present in a living cell fulfil a certain task in the cell. As a consequence of its functionality a protein is located in certain parts of the cell. If it is made visible the resulting patterns can help to identify the protein, as the spatial distribution of the visible structures depends on the functionality of the protein inside of the cell and, therefore, characterises the protein. The cells used for the experiments were COS-1 cells typically allowing easy microscopic data takes as the cells are much larger than their nuclei. With the help of a suitable parameterisation the proteins can be automatically identified. In order to derive such a parameterisation, features describing the spatial structure of the protein are extracted. The stochastic behaviour of the features is of major importance for the performance of the method.

1. INTRODUCTION

A protein present in a cell can be made visible by a chemical treatment with antibodies. The spatial distribution of the visible structures depends on the functionality of a protein inside of the cell and characterises the protein. Therefore, it allows or at least helps to identify the protein. In this work a method to automatically classify proteins on the basis of single cell images is described.

The imagery of COS-1 cells used here has been acquired by fluorescence confocal microscopy. From a data take, i.e. a focus series of images, the image optimally showing the spatial distribution of the protein has been selected. A single cell extracted from such an image constitutes the input to the algorithm described.

In order to derive a parameterisation identifying proteins, features describing the spatial structure of the protein have to be extracted. An interactive classification of proteins by a human operator has shown that a classification accuracy of 95 to 100 % is possible. Similar classification accuracy can be achieved by an automatic analysis when suitable features are selected. As the consecutively following steps of the procedure and the facts being their basis, such as probability density distributions, their derivation from training data, the choice of a classification method, and the derivation of a classification decision, are well known, feature selection or feature reduction is the crucial step of the procedure. The importance of feature reduction corresponds to the fact that in human vision, particularly in deriving decisions from visual information, the large amount of data/information in images based on high spatial and radiometric resolution is first severely reduced before being extended again by associating knowledge, e.g. about objects and context, in order to derive new knowledge or decisions in a process of thinking (BECKER-CARUS, 1981).

Using our method, previously unknown proteins can be identified as long as the protein shows an individual spatial structure inside of the cell. With an automatic procedure, from a specific spatial structure conclusions with respect to the chemical role of the protein could be drawn, as the molecules appear where they are chemically active. This means that image analysis can provide a new method to proteomics research, possibly of efficiency previously unknown. It is our long term goal to derive and test such a method.

2. PREVIOUS WORK

BOLAND et al. (1997) describe a method to classify cellular protein localization patterns based on their appearance in fluorescence light microscope images. Numeric features were used as input values to either a classification tree or a neural network (BOLAND et al., 1998). MARKEY et al. (1999) developed methods for objectively choosing a typical image from a set of images, emphasizing cell biology. The methods include calculation of numerical features to describe protein patterns, calculation of similarity between patterns as a distance in feature space, and ranking of patterns by distance from the center of the distribution in feature space. The images chosen as most typical were in good agreement with the conventional understanding of organelle morphologies. MURPHY et al. (2000) describe an approach to quantitatively describe protein localization patterns and to develop classifiers able to recognize all major subcellular structures in fluorescence microscope images. Since fluorescence microscope images are a primary source of information about the location of proteins within cells, MURPHY et al. (2001) strive to build a knowledge-based system which can interpret such images in online journals. They developed a robot searching online journals to find fluorescence microscope images of individual cells. BOLAND & MURPHY (2001) used images of ten different subcellular patterns to train a neural network classifier. The classifier was able to correctly recognize an average of 83 % of the patterns. Fluorescence microscopy is the most common method used to determine subcellular location, e.g. VELLISTE & MURPHY (2002) have previously described automated systems recognizing all major

subcellular structures in 2D fluorescence microscopic images. They have shown that pattern recognition accuracy is dependent on the choice of the vertical position of the 2D slice through the cell and that classification of protein localization patterns in 3D images results in higher accuracy than in 2D. Automated analysis of 3D images provides excellent distinction between two golgi proteins whose patterns are indistinguishable by visual examination. ROQUES & MURPHY (2002) describe the application of pattern analysis methods to the comparison of sets of fluorescence microscope images. MURPHY et al. (2002) report improved numeric features for pattern descriptions which are fairly robust to image intensity changes and different spatial resolutions. They validate their conclusions using neural networks. DANCKAERT et al. (2002) describe development and test of a classification system based on a modular neural network trained with sets of confocal focus series. The system performed well in spite of the variability of patterns between individual cells.

3. FEATURE EXTRACTION

In this work, to recognize proteins being active in a cell means to visually differentiate between their appearances in images. The latter depends on whether there are features which allow making a difference between them. This is valid for visual judgment by a human observer as well as for a pattern recognition algorithm. The criteria used by a human usually are directly related to known cell structure. For a pattern recognition algorithm, among the multitude of features which are present or can be defined in imagery, those have to be identified which help to separate different phenotypes of cells from each other in feature space. I.e., the pattern characterizing a protein has to be parameterized. In general, the parameters to be used have to describe the spatial distribution of the protein inside of the cell.

Prior to feature extraction from imagery, a laboratory procedure including chemical treatments of probes and microscopic image acquisition had to be established. First, antibodies had to be found allowing to stain the proteins making them – or the organelles as the locations of their activity – visible in the imagery. In addition to the protein investigated some organelles had to be stained to allow recognition of the most characteristic parts cell and, thereby, reference to the cell as such. Lamin was chosen as the marker of the membrane of the nucleus of the cell allowing separation of the nucleus form the cytoplasm, and Golgin97 was used to stain the golgi apparatus.

On the basis of the membrane of a cell's nucleus and golgi apparatus a reference system allowing translation and rotation invariant definition of features describing the proteins was defined. The centre of the nucleus is used as central reference point in the sense of the origin of a coordinate system. The cell is subdivided into sectors inside and outside the nucleus (Fig. 1). In the system of sectors the direction to the centre of the golgi apparatus is used as reference.

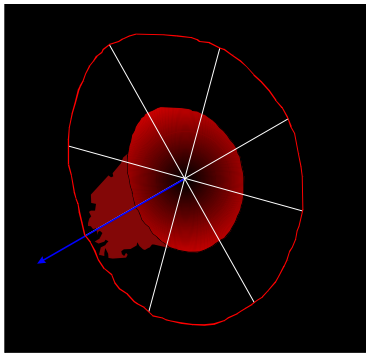

Fig. 1: Subdivision of the cell into sectors inside and outside the cell nucleus.

Ten proteins and corresponding antibodies were selected for the investigation. It was taken care to choose visually very different as well as rather similar proteins. Figs. 2 and 3 show Huntingtin and GIT as examples of visually similar proteins statistically varying.

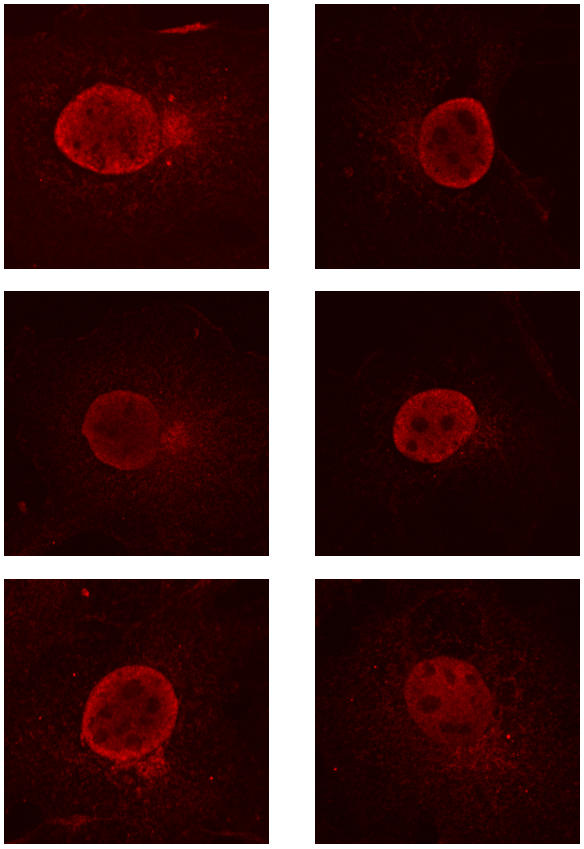

Fig. 2: Huntingtin

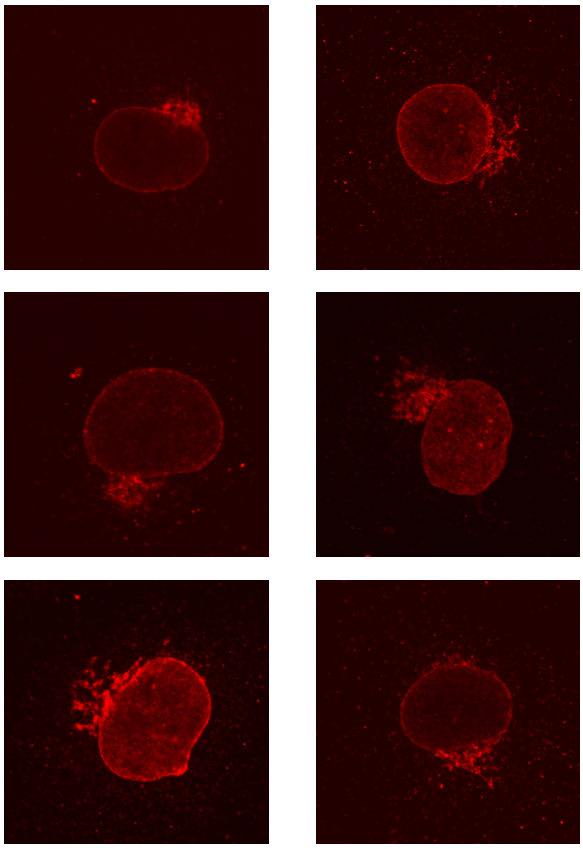

Fig. 3: GIT

Immune fluorescence imaging the probes in a confocal microscope, the red channel was used to image the protein, and the green channel to image the reference organelles, i.e. the cell nucleus membrane and the golgi apparatus. Fig. 4 shows an example of the colocalizations becoming visible.

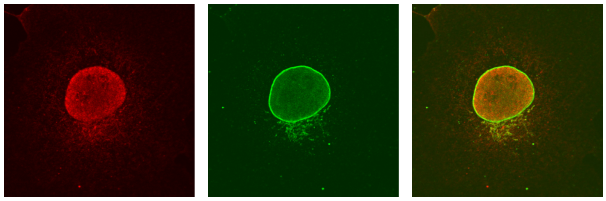

Fig. 4: Images of a cell prepared to show Huntingtin: the protein (or its marker HD1) in red, the reference organelles (markers Lamin and Golgin97) in green and the color image showing colocalizations (from left to right).

For each protein to be investigated ten images of different cells were acquired. In order to explore the statistic behaviour of the visual appearance, i.e. the feature vector of a protein, two visually similar proteins, Huntingtin and GIT (Figs. 2 and 3) were imaged 100 times. Half of the images were used as training data, the other half as test data.

The features to be used for classification are for instance statistical measures describing protein localisation inside of the nucleus, e.g. variance and entropy, edge segments appearing inside and outside of the nucleus, and the visibility of the golgi apparatus being attached to the nucleus. If these "features" are no numerical values directly, they have to be transformed into numeric measures such as edge length or strength. As the occurrence of the protein inside of a cell is a natural event more or less varying statistically, the statistic behaviour of the extracted features is of major importance for the performance of the method and, therefore, has to be taken into account by the algorithm, e.g., by using the probability density distributions of the features for classification.

The feature vector actually used includes the following features; c.f. (STECKLING & KLÖTZER, 2003; STECKLING et al., 2003):

1. White pixels: number of pixels whose grey value is greater than the average of all grey values of the image.

2. White segments: number of image segments fulfilling the same condition. A segment is defined as a four-connected neighbourhood (BOLAND & MURPHY, 2001).

3. Black segments: number of segments consisting of four-connected pixels with grey values lower or equal than the average of all grey values of the image (BOLAND & MURPHY, 2001).

4. Expectation value:

$$m = E(x) = \frac{1}{K}\sum_{k=1}^{K} x_k \qquad (1)$$

(BOLAND & MURPHY, 2001).

5. Energy: second angular moment

$$\sum_{i=0}^{N_{g-1}} \left(p_{x-y}(i)\right)^2 \qquad (2)$$

where x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_y(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$

(BOLAND & MURPHY, 2001; HARALICK et al., 1973).

6. Difference entropy:

$$-\sum_{i=0}^{N_{g-1}} p_{x-y}(i)\log\left\{p_{x-y}(i)\right\} \qquad (3)$$

(BOLAND & MURPHY, 2001; HARALICK et al., 1973).

7 Lines: number of segments extracted with a line extraction method.

4. CLASSIFICATION

A maximum likelihood classification was used. To illustrate the separability of the clusters of two visually similar proteins based on the feature vector defined in the previous section, Fig. 5 shows the sub-space of the three most informative features.

87 % of the test data samples were correctly classified (c.f. STECKLING et al., 2003). In contrast to this several individuals who were first shown the training data and who consecutively classified the test data achieved classification accuracies of 95 to 100 %. On the one hand, this result shows that the automatic classification procedure still has to be improved. On the other hand, the high classification rate of test persons who did not go through intensive training procedures indicates that it should be possible to reach this goal by an automatic procedure.

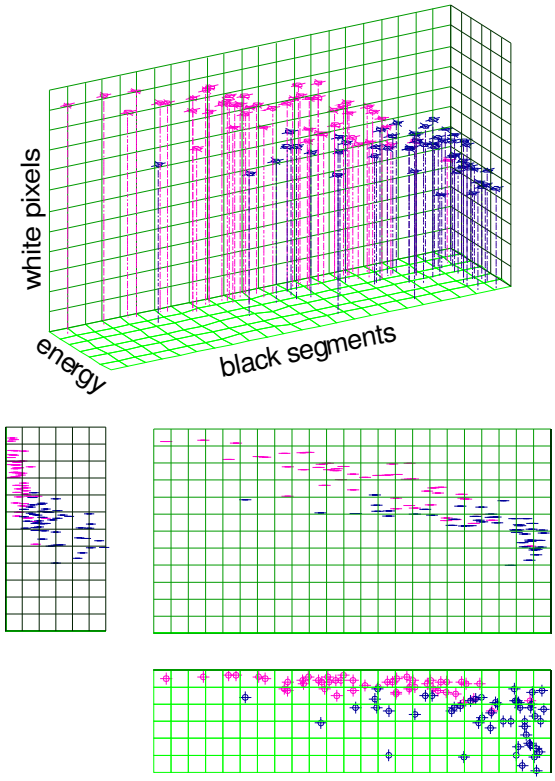

Fig. 5: Feature space (reduced to three dimensions).

The robustness of the classification primarily depends on the statistic behaviour of the feature vector which is not only determined by the visual appearance of the proteins, i.e. the differences between the spatial structures of individual cells, but also by the variations caused by the chemical preparation of the cells and the conditions under which the imagery was acquired. Therefore, successful application of the method proposed here requires well-controlled laboratory procedures.

5. REFERENCES

BECKER-CARUS, C. (1981): *Grundriß der Physiologischen Psychologie*, HEIDELBERG.

BOLAND, MICHAEL V.; MARKEY, M.K.; MURPHY, ROBERT F.; (1997): Classification of Protein Localization Patterns Obtained via Fluorescence Light Microscopy, Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1997, pp. 594-597.

BOLAND, MICHAEL V; MARKEY, M. K.; MURPHY, ROBERT F.; (1998): Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images. Cytometry 33: 366-375, 1998.

BOLAND, MICHAEL V.; MURPHY, ROBERT F.; (2001): A Neural Network Classifier Capable of Recognizing the Patterns of all Major Subcellular Structures in Fluorescence Microscope Images of HeLa Cells, Bioinformatics 17, 1213-1223, 2001.

DANCKAERT, A.; GONZALEZ-COUTO, E.; BOLLONDI, L.; THOMPSON, N.; HAYES, B.; (2002): Automated Recognition of Intracellular Organelles in Confocal Microscope Images, Traffic, vol. 3, no 1, pp. 66-73, January 2002.

HARALICK, R. M.; SHANMUGAN, R.; DINSTEIN, I.; (1973): Textural Features for Image Classification, IEEE Trans Sys. Man Cyb., vol. SMC-3, no. 6, pp. 610-621, 1973.

MARKEY, M. K.; BOLAND, MICHAEL V; MURPHY, ROBERT F.; (1999): Towards Objective Selection of Representative Microscope Images. Biophys. J. 76:2230-2237, 1999.

MURPHY, ROBERT F.; BOLAND, MICHAEL V.; VELLISTE, MEEL; (2000). Towards a Systematics for Protein Subcellular Location: Quantitative Description of Protein Localization Patterns and Automated Analysis of Fluorescence Microscope Images. Proc Int Conf Intell Syst Mol Biol (ISMB 2000) 8: 251-259, 2000.

MURPHY, ROBERT F.; VELLISTE, MEEL; YAO, JIE; PORRECA, GREGORY; (2001): Searching Online Journals for Fluorescence Microscope Images Depicting Protein Subcellular Location Patterns, Proceedings of the 2nd IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2001), pp. 119-128.

MURPHY, ROBERT F.; VELLISTE, MEEL; PORRECA, GREGORY; (2002): Robust classification of subcellular location patterns in fluorescence microscope images, Proseedings of the 2002 IEEE International Workshop on Neural Networks for Signal Processing (NNSP 12), , pp. 67-76, Proceedings of the 12th IEEE Workshop on 4.-6. September 2002.

ROQUES, E.J.S.; MURPHY, ROBERT F.; (2002): Objective evaluation of differences in protein subcellular distribution. Traffic 3: 61-65, 2002.

STECKLING, TANJA; KLÖTZER, HARTMUT; (2003): Objekterkennung und Modellierung zellulärer Strukturen aus mikroskopischen Bildern; Diplomarbeit, Technische Universität Berlin, 2003.

STECKLING, TANJA; KLÖTZER, HARTMUT, SUTHAU, TIM; WÄLTER, STEPHANIE; WANKER, ERICH; HELLWICH, OLAF; (2003): Objekterkennung und Modellierung zellulärer Strukturen aus mikroskopischen Bildern; 23. Jahrestagung der DGPF, Bochum 2003.

VELLISTE, MEEL; MURPHY, ROBERT F; (2002): Automated Determination of Protein Subcellular Locations from 3D Fluorescence Microscope Images. Proceedings of the 2002 IEEE International Symposium on Biomedical Imaging (ISBI 2002), pp. 867-870.