Pattern Recognition in High-Content Cytomics Screens for Target Discovery – Case Studies in Endocytosis

Lu Cao[1,*], Kuan Yan[1,*], Leah Winkel[2], Marjo de Graauw[2], and Fons J. Verbeek[1]

[1] Imaging & BioInformatics, Leiden Institute of Advanced Computer Science
[2] Toxicology, Leiden Amsterdam Centre for Drug Research,
Leiden University, Leiden, The Netherlands
{lucao,kyan,fverbeek}@liacs.nl, m.de.graauw@lacdr.leidenuniv.nl

Abstract. Finding patterns in time series of images requires dedicated approaches for the analysis, in the setup of the experiment, the image analysis as well as in the pattern recognition. The large volume of images that are used in the analysis necessitates an automated setup. In this paper, we illustrate the design and implementation of such a system for automated analysis from which phenotype measurements can be extracted for each object in the analysis. Using these measurements, objects are characterized into phenotypic groups through classification while each phenotypic group is analyzed individually. The strategy that is developed for the analysis of time series is illustrated by a case study on EGFR endocytosis. Endocytosis is regarded as a mechanism of attenuating epidermal growth factor receptor (EGFR) signaling and of receptor degradation. Increasingly, evidence becomes available showing that cancer progression is associated with a defect in EGFR endocytosis. Functional genomics technologies combine high-throughput RNA interference with automated fluorescence microscopy imaging and multi-parametric image analysis, thereby enabling detailed insight into complex biological processes, like EGFR endocytosis. The experiments produce over half a million images and analysis is performed by automated procedures. The experimental results show that our analysis setup for high-throughput screens provides scalability and robustness in the temporal analysis of an EGFR endocytosis model.

Keywords: phenotype measurement, image analysis, classification, EGFR endocytosis.

1 Introduction

In this paper we address the problem of deriving a phenotype of a cell in the context of time-lapse cytomics data; in particular we investigate the process of endocytosis and epidermal growth factor receptor (EGFR) signaling.

Enhanced epidermal growth factor receptor (EGFR) signaling triggers breast cancer cells to escape from the primary tumor and spread to the lung, resulting in poor disease prognosis. Moreover, it may result in resistance to anti-cancer therapy. In normal epithelial cells, EGFR signaling is regulated via endocytosis, a process that

[*] Equally contributed.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 330–342, 2011.

results in receptor degradation and thereby attenuation of EGFR signaling. However, in cancer cells the endocytosis pathway is often defective, resulting in uncontrolled EGFR signaling. Over the past years, RNA interference combined with fluorescence microcopy-based imaging has become a powerful tool to the better understanding of complex biological processes [18]. Such combined experiment often produces over half a million multi-channel images; manual processing of such data volume is impractical and jeopardizes objective conclusions. Therefore, an automated image and data analysis solution is indispensable. To date, analysis was done with simple extraction of basic phenotypes from EGFR images using tools such as BioApplication[6], ImageXpress[5] and QMPIA[1]. However, these tools are not suitable for a profound study of the dynamics behind EGFR endocytosis which requires more attention. From the existing literature [1, 5, 6, 12, 20, 21, 22] a generic model, defining four major episodes of EGF-induced EGFR endocytosis, can be distilled. Under basic conditions EGFR localizes at the plasma-membrane site, which is in our study defined as “plasma-membrane” (1). Upon binding of EGF to the receptor, EGFR is taken up into small vesicular structures (e.g. early endosomes), which is here defined as “vesicle” (2). Over time EGFR containing vesicles are transported to late endosomes localizing near the nuclear region and form into a larger complex defined here as “cluster” (3). In addition to this EGFR degradation route EGFR can partly also be transported back to the plasma-membrane. Using this dynamic model as the major guideline, the analysis of EGFR-regulation-related gene pathway could be linked to each stage of EGFR endocytosis. Instead of looking at one fixed time point, our current experimental design includes a series of time points at which images are captured. An image potentially contains a ratio in the three characteristic episodes in the EGFR endocytosis process. The image analysis solution should be able to extract basic phenotype measurements as well as to identify the stage of EGFR. In this paper, we illustrate the design and implementation of an automated setup for high-content image and data analysis which can properly capture EGFR dynamics and classify different EGFR phenotypes.

Our workflow for automated analysis solution is depicted in Figure 1. Each high throughput screen (HTS) experiment starts with the design of the experimental scheme, followed by the wet-lab experiment and high throughput microscopy-based imaging. Both experimental schemes and image data are organized and stored in a database. Subsequently, image analysis is used to extract phenotype measurements from these images and classifiers are introduced to recognize each phenotypic stage of EGFR. Finally, comprehensive conclusions are drawn based on comparisons of EGFR expression at each stage and time point.

In this paper, we limit the scope to image analysis and data analysis, some biology will be explained. Accordingly, the organization of this paper is divided into three major sections. In section 2, we introduce the methodology including image acquisition and image analysis; several innovative algorithms will be briefly introduced. After segmentation of the images, EGFR phenotype measurements are obtained. We will illustrate the categorization of phenotypic stages using feature selection and classification. The best combination pair is applied on image data to classify three phenotypic stages and construct a phenotype model. The experimental results are presented in section 3 with two case studies. The first case study tests our solution in identifying dynamic phenotype stages. The second study case examines robustness and scalability of our solution in analyzing a large number of phenotypes.

2 Methodology

Modern techniques in fluorescence microscopy allow visualizing various cell structures so that these can be specifically subject to analysis. Together with a computer-controlled microscope, a high-throughput image acquisition scheme, known as high-throughput screen (HTS), has become feasible. Depending on the biological question at hand, a HTS experiment may produce up to half million. Such a volume of images is beyond the capacity of manual processing and therefore, image processing and machine learning are required to provide an automated analysis solution for HTS experiments. In this section, we will introduce the image acquisition protocol followed by approaches for image analysis and data analysis.

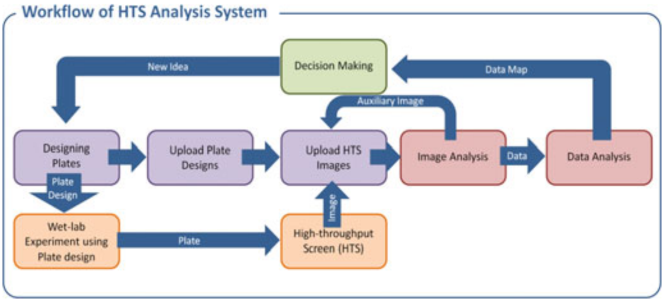

Fig. 1. Workflow of our HTS Analysis System

2.1 Image Acquisition

The workflow for data preparation includes three essential steps: (1) cell culturing, siRNA transfection and EGF exposure, (2) fluorescent staining of proteins of interest and (3) image acquisition. Here we use a design of an EGFR-regulation related siRNA screen to illustrate this workflow. In this design, cells are cultured in 96 well culture plate and transfected using Dharmafect smartpool siRNAs. Subsequently, the transfected cell population was exposed to epidermal growth factor (EGF) for a different duration of time. Cells are fixed at different time points and visualized with a confocal laser microscope (Nikon TE2000). Image acquisition automation was realized with an automated stage and an auto-refocusing lens controller. For each well, images are captured from ten randomly selected locations. For each image three channels are captured: (1) a red channel containing P-ERK expression staining (Cy3), (2) a green channel containing EGFR expression staining (Alexa-488) and (3) a blue channel containing a nuclear staining (Hoechst #33258). Upon completion of the acquisition process all images are uploaded to a database server for image analysis.

2.2 Image Analysis

High-Content Analysis. Basically, the image analysis procedure converts raw microscope images into quantifications characteristic to biological phenomena.

A number of steps are elaborated to achieve this purpose; starting from image acquisition, three steps are distinguished: (1) noise suppression, (2) image segmentation and (3) phenotype measurement. Image segmentation refers to the process of partitioning an image into multiple regions with the goal to simplify and/or change the representation of an image into something that is easier to analyze. For fluorescence microscopy cell imaging we specifically designed a segmentation algorithm: i.e. watershed masked clustering (WMC). The WMC algorithm (cf. Fig. 1d) [23] is an innovative and customized segmentation algorithm that serves different types of cytomics studies like dynamic cell migration analysis [24, 20, 9] and protein signaling modeling [19]. Due to the absence of an indicator for the cell border (cf. §2.1), a border reconstruction and interpolation algorithm is designed to provide artificial representations of the cell borders; i.e. the weighted Voronoi diagram based reconstruction (W-V) algorithm [19]. The W-V algorithm (cf. Fig. 1c) offers the possibility to measure both border-related signal localization [19] and protein expression in terms of continuity and integrity [19]; it does not require a complete cell border or cytoplasmic staining. Both binary mask and artificial cell border are used to derive a number of phenotype measurements for further data analysis.

Phenotype Measurement. In the current experiment and imaging protocol, the phenotype measurements can be categorized into two subgroups: (1) basic measurements of the phenotypes covering shape descriptors and (2) the localization phenotype describing the assessment of the correlation between two information channels. The basic phenotype measurement [2, 11, 24] includes a series of shape parameters listed in Table 2. In addition to the basic phenotype measurement [2, 19, 11, 24], localization measurements can be derived for a specific experimental hypothesis; e.g. the expression ratio between protein channels or shape correlation between objects. The localization phenotypes are quantifications of comparative measurement between information channels such as relative structure-to-nucleus distance or structure-to-border distance [19]. In this paper, we will limit the scope of phenotype measurements to the set employed by the study on EGFR endocytosis. In Table 3 a list of EGFR screen based localization phenotypes is shown. On the basis of the phenotype measurements, objects are classified into phenotypic stages. For the assessment of significance statistical analysis is performed.

2.3 Data Analysis

The aim of the endocytosis study is to quantify the process of EGF-induced EGFR endocytosis in human breast cells and to identify proteins that may regulate this process. The EGFR endocytosis process can roughly be divided into three characteristic episodes: i.e. (1) at the onset EGFR is present at the *plasma-membrane*; (2) subsequently, small *vesicles* containing EGFR will be formed and transported from the plasma-membrane into the cytoplasm; and (3) finally, vesicles are gradually merging near the nuclear region forming larger structures or *clusters*. The characteristic episodes are the read-out for HTS. Based on this model it is believed that EGFR endocytosis regulators may be potential drug targets for EGFR-induced breast cancer. Studying each of the stages (cf. Fig. 3), i.e. plasma-membrane, vesicle and cluster, may provide a deeper understanding of the EGFR endocytosis process.

Table 2. Basic measurements for a phenotype (after segmentation to binary mask)

Feature Name	**Description**
Size	The size of object, aka as the surface area.
Perimeter	The perimeter of the object
Extension	Derived from 2nd-order invariants of the object [7, 24]
Dispersion	Derived from 2nd-order invariants of the object [7, 24]
Elongation	Derived from 2nd-order invariants of the object [7, 24]
Orientation	Derived from 2nd-order moments of the object [7, 24]
Intensity	Average intensity of all pixels belong to an object
Circularity	Area-to-perimeter ratio; higher compactness suggests a more smooth and less protrusive shape.
Semi-major axis length	Derived from 2nd-order moments of the object [7, 24]
Semi-minor axis length	Derived from 2nd-order moments of the object [7, 24]
Closest object distance	The distance to nearest neighbor of the object, the distance is measured similar to the border distance in Table 3
In nucleus	Boolean describing if the object is included in nucleus mask

Table 3. Localization measurement

Feature Name	**Description**
Nucleus distance	Distance between structure and nucleus, measured as the average distance between each pixel in an object and the mass center of the corresponding nucleus.
Border distance	Distance between structure and cell membrane, measured as the average distance between each object-pixel and the center of mass of the corresponding cell border (membrane).
Intactness	Overlap between structure expression and cell membrane divided by the total length of cell membrane

Phenotypic Sub Categorization. Here we introduce a profound explanation of the whole procedure employed in the phenotypic sub-categorization including the production of a training set and the procedure for the training of the classifier. The training set is derived from manually delineated outlines of each phenotypic group and subsequent training of a classifier distinguishing three different phenotypes. From two case studies the capability of our solution with respect to identifying characteristic episodes in the process under study stages as well as the scalability in describing different phenotypic groups, is assessed.

Preparation of the Training Set. Ground truth data were obtained by the outlines of the three characteristic episode groups, i.e. cell border/plasma-membrane, vesicle and cluster. These were separately delineated by biologists using our dedicated annotation software (TDR) with a digitizer tablet (WACOM, Cintiq LCD-tablet). From each outline a binary mask is created for each phenotypic stage. In Figure 4(b) the vesicle mask derived from a manually selected vesicle outline is shown. This mask is overlaid with the mask obtained from the WMC algorithm so as to extract the intersection set of two masks as shown in Figure 4(d). Finally, the phenotype measurements are computed with this mask. In similar fashion the ground truth datasets for the plasma-membrane and cluster groups are prepared.

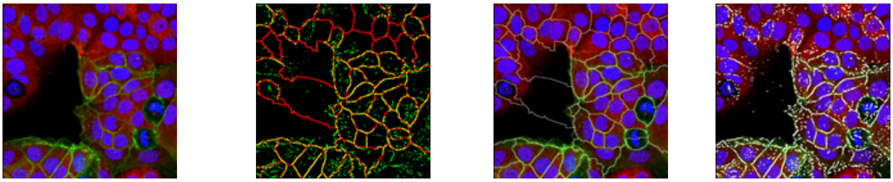

Fig. 2. (a) Original image: PERK (red), EGFR (green) and nucleus (blue), (b) Component definition: artificial cell border (red) and binary mask of protein expression (green), (c) cell border reconstruction : artificial cell border (W-V), (d) image segmentation: binary mask of EGFR channel by WMC

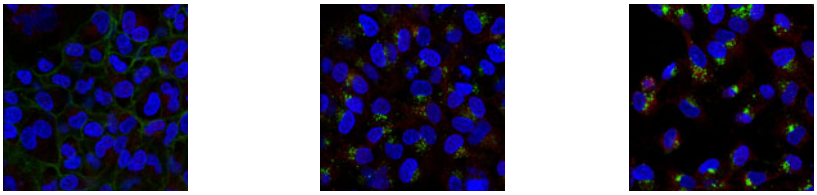

Fig. 3. Sample images of the 3 phenotypic groups with (a) Plasma-membrane, (b) Vesicle, (c) Cluster

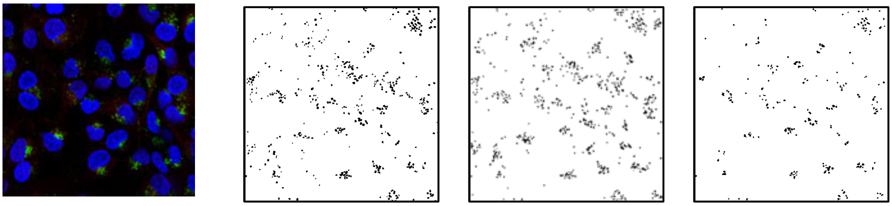

Fig. 4. Ground truth data production. (a) Original image, (b) manual mask, (c) WMC mask, (d) overlay of the mask.

The training dataset includes three characteristic episode groups with 2254 objects and 14 features. Given the huge differences in the feature ranges, it is necessary to normalize the dataset. Normalization is accomplished by shifting the mean of the dataset to the origin and scaling the total of variances of all features to 1. In this way the magnitude effect is successfully removed and the recognition accuracy can be significantly improved [17]. The normalized dataset is used for training of the EGFR classifier.

Feature Selection. First, it is crucial to make a selection of the probabilistic distance criterion for the discriminability estimation. For this we have chosen the Mahalanobis distance [14] since it takes the correlations among the variables into consideration and, in addition, it is scale-invariant. Other distance criteria, such as the Euclidean or Manhattan distance, are, more or less, related to the assumption that all features are independent and have an equal variance. We cannot be certain that all features in our dataset are independent and therefore the Mahalanobis distance is preferred.

Second, we have selected three representative search algorithms including parametric and non-parametric search algorithms; i.e. the branch and bound procedure [10], best individual N features and sequential backward selection [8]. Branch and bound is a top-down procedure, beginning with the set of variables and constructing a tree by deleting variables successively; i.e. an optimum searching procedure requiring the evaluation of partially constructed or approximate solutions without involving exhaustive search. Best individual N features procedure is the simplest suboptimal method for choosing the best N features by individually assigning a discrimination power estimate to each of the features in the original set. In some cases, especially if the features from original set are uncorrelated, this method results in a well-defined feature sets. Sequential backward selection is another suboptimal search algorithm. Variables are deleted one at a time until the required number of measurements remains [4]. The advantage of backward selection is its capability for global control during the feature selection.

Third, we choose three classifiers covering both linear and non-linear categories; i.e. the linear classifier (LDC), the quadratic classifier (QDC) and k-nearest neighbor classifier (KNNC). A linear classifier makes a classification decision based on the value of a linear combination of the characteristics [16]. If the data are strongly non-Gaussian, they can perform quite poorly relative to nonlinear classifiers [3]. A quadratic classifier, which is generalization of the linear classifier; it separates measurements of classes by a quadric surface. Finally, the k-nearest neighbor classifier classifies an object by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors. The k-nearest neighbor rule achieves a consistent high performance, without a priori assumptions about the distributions from which the training examples are drawn. Moreover, it is robust with respect to noisy training data and still effective if the training dataset is large. By permutation we obtained 9 pairs of combinations. The result of the error estimation is shown in Figure 5. An interesting characteristic can be observed in these plots. The weighted error of the quadratic classifier jumps abruptly when the number of features exceeds a certain threshold (10 for individual feature selection, 12 for branch & bound, and 5 for backward feature selection). This is caused (1) by including a feature with which it is hard to distinguish three phenotypic groups and (2) by the fact that the distribution of the three classes might be more properly classified by the linear and k-nearest neighbor classifier rather than quadratic classifier.

Feature Extraction. Feature extraction is another category to manage multi-dimensional features by reducing dimensionality of features trough combining. For the final result, we also tested the performance of the feature extraction combined with the three classifiers selected. As our starting point is a labeled training dataset, a supervised feature extraction method is most suitable. The Fisher mapping [4] was chosen as extraction method. Fisher mapping finds a mapping of the labeled dataset onto an N-dimensional linear subspace such that it maximizes the between-scatter over the within-scatter. It should be taken into account that the number of dimensions to map is less than the number of classes in the dataset. We have three phenotype classes and consequently the labeled dataset can only be mapped onto a 1D or 2D linear subspace. The result of the performance estimation is shown in Figure 6(a). In addition, in Figure 6(b,c,d), the scatter plots of mapped data with corresponding classifiers are shown.

Comparison of the Results. Each weighted classification error curve (cf. Fig. 5 and 6(a)) represents a combination of a feature selection/extraction method and a classifier algorithm. For each combination, we selected the lowest point value representing the best feature selection/extraction performance of the combination and, subsequently, compared the weighted error and standard deviation of each lowest point. The combination of branch and bound feature selection with k-nearest neighbor classifier has the lowest minimal value and relatively small standard deviation, as can be concluded from Table 4.

Table 4. Minimal value of Mean Weighted Errors and its Standard Deviation

	Individual		B&B		Backward		Fisher	
	min	σ	min	σ	min	σ	min	σ
LDC	0.0586	0.0093	0.0562	0.0098	0.0534	0.0109	0.0555	0.0105
QDC	0.0609	0.0119	0.0626	0.0117	0.0815	0.0113	0.0589	0.0125
KNNC	0.0502	0.0092	0.0450	0.0091	0.0535	0.009	0.0587	0.0124

The three selected features, derived from branch and bound feature selection with the best performance, are *closest object dist*, *object intensity* and *area*. The "*closest object dist*" is a distance measurement between an object and its nearest neighbor. It defines the local numerical density of an object. The cluster and vesicle categories usually have a much lower *closest object dist* since they tend to appear in clusters. The amount of fluorescence therefore directly relates to the amount of EGFR and can be measured as intensity at a certain spot. We suppose that plasma-membrane, vesicle or cluster are all composed of EGFR and the expression of EGFR is more evenly distributed in the plasma-membrane and gradually increases concentration in vesicle and cluster. Intensity represents the amount of EGFR and is significant. Size is undoubtedly the major feature for describing three characteristic episode groups. The results confirm our expectations. We have chosen the combination of branch and bound feature selection with k-nearest neighbor classifier as the best classifier for the case studies.

Statistical Analysis. We provide two case studies in order to sustain the performance of our solution. The first case study is aimed at a better understanding of EGFR endocytosis across time series. The EGFR endocytosis procedure is as follows: in the absence of EGF, EGFR localizes at the cell membrane (e.g. cell border localization). Upon EGF exposure, a portion of the plasma membrane containing EGFR is invaginated and pinched off forming a membrane-bounded vesicle. Some vesicles will be accumulated in clusters in the peri-nuclear region. As for the experimental design, the cells in separate wells are treated with EGF for a variable amount of time. In this way each well represents a fixed time point. After fixation, cells are stained and visualized. The images that have a clear representation of phenotype stage are carefully selected by a specialist. The result of image and data analysis based on selected images provides a notable capability of our solution on identifying the dynamics in the characteristic episodes. The source images include a total of 13 time points with 2 pairs of images each.

The second case study is on identification of mediators of EGFR endocytosis. The results demonstrate that our automated high-content analysis solution can properly describe different phenotypic groups and is capable to manage large quantities of phenotypes. Per culture plate ten images are acquired per well; i.e. 96×10 images are used in the image and data analysis. In order to evaluate the phenotype difference between wells, we calculate the number of each phenotypic group (vesicle, plasma-membrane, and cluster) per nucleus in each well. The plasma-membrane, representing the composed EGFR evenly distributed on the cell membrane, is always continuously linked between cells. The quantification is accomplished by calculating the pixels of plasma-membrane per nucleus.

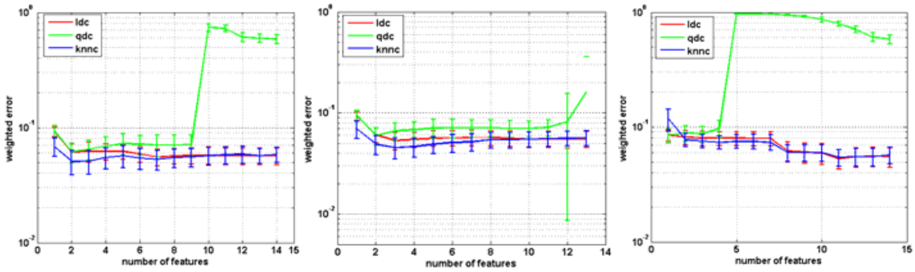

Fig. 5. Weighted classification error curves, with (a) Individual feature selection, (b) Branch and bound feature selection and (c) Backward feature selection

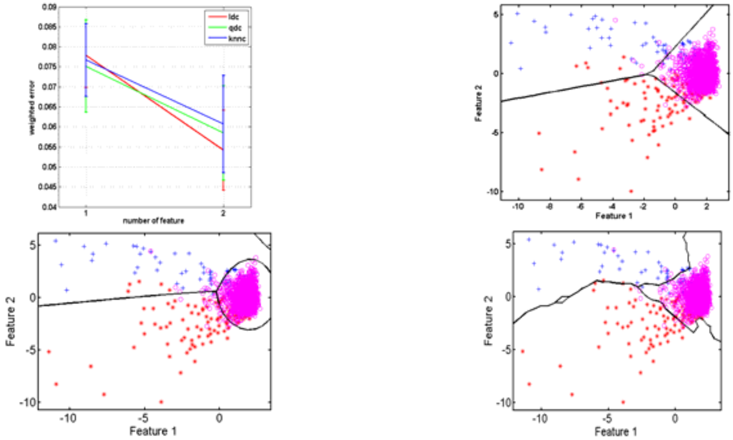

Fig. 6. Results of feature extraction: (a) Weighted classification error curve of Fisher feature extraction, (b) Fisher feature extraction with Linear Discriminant Classifier, (c) Fisher feature extraction with Quadratic Discriminant Classifier, (d) Fisher feature extraction with K-Nearest Neighbor Classifier

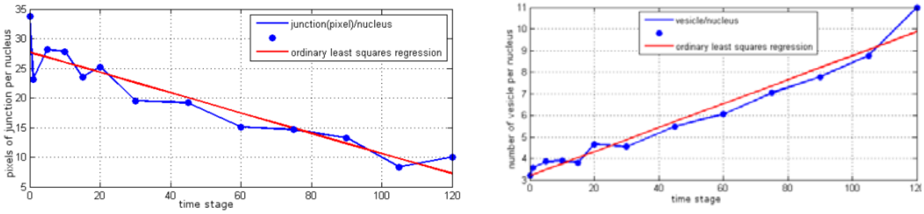

Fig. 7. Average number of plasma-membrane (a) and vesicle (b) per nucleus

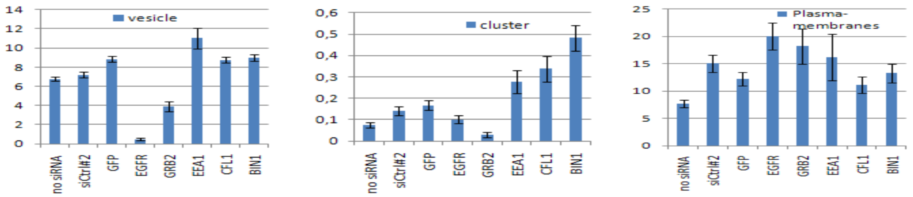

Fig. 8. (a) Number of vesicles per nucleus, (b) Number of clusters per nucleus (c) Plasma-membranes (pixel) per nucleus

An analysis with both the Jarque-Bera [9] and Lillie test [13] established that over 80% of our measurement data of the composition vesicle/plasma-membrane/cluster is not normally distributed. We, therefore, use the Kolmogorov-Smrinov test [15] with siCtrl#2 as control sample to identify significant changes in EGFR endocytosis.

3 Experimental Results

3.1 Dynamic Phenotype Stage

The results shown in Figure 7a illustrate that the amount of EGFR localized at the plasma-membrane (e.g. number of plasma-membranes, expressed as pixel/nucleus) decreases over time. This fits with the EGFR endocytosis process during which EGF exposure causes a gradual EGFR re-distribution from the plasma-membrane into vesicles. Meanwhile, the number of vesicles per nucleus increases caused by the formation vesicles as illustrated in Figure 7b. These graphs indicate the trend of the endocytosis process and are representative to illustrate phenotype stage dynamics.

3.2 Phenotype Classification

We validated our automated high throughput image analysis using siRNA-mediated knock-down of several known EGFR endocytosis regulators (e.g. siGrb2, siEEA1, siCFL) To this end images were selected from WT cells (not treated with siRNA), control siRNA treated cells (siCtrl#2 and siGFP), siEGFR treated cells and three target siRNAs. In Figure 8a-c the comparison of selected results with three phenotypic groups is shown. In Figure 9 some sample images are depicted to check the correctness of our solution for phenotype description. Our analysis shows that cells treated with siCtrl#2 resemble non-treated WT cells, while siGFP differs significantly; indicating that siCtrl#2 is the best control for further analysis. As expected, siEGFR showed decreased levels of all three classifiers since treatment of cells with siEGFR results in > 90% knock-down of EGFR. In addition, siGrb2, siEEA1 and siCFL behaved as expected. These results demonstrate that the automated high throughput analysis could be used for large scale siRNA screening.

A comprehensive overview of the results of a complete experiment is shown in the heatmaps depicted in Figure 10. The data are derived from a siRNA screen of more than 200 potential regulators of EGFR endocytosis. The y-axis represents different siRNA targets (regulators) and the x-axis represents the features plus the number of different phenotypic groups.

4 Conclusions

This paper provides an efficient solution to analyze the high-throughput image data sets on the level of protein location. The experimental results of both case studies show that our automated analysis procedure can be involved in the identification of the characteristic episodes in the EGFR process and provides a set of robust and precise phenotypic descriptions. From the case studies it is illustrated that our solution is suitable for a robust analysis of different phenotypes in a siRNA based HTS. Furthermore, the whole process, from image segmentation, phenotypic quantification to classification, is part of a successfully automated procedure. Our solution can be easily extended to cope with studies utilizing fluorescence microscopy.

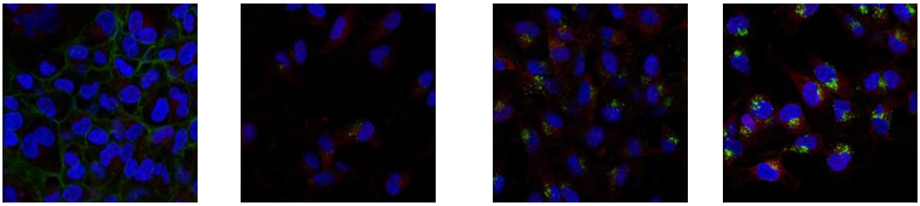

Fig. 9. Sample images, (a) no siRNA no EGF, (b) EGFR, (c) GRB2, (d) BIN1 (> response)

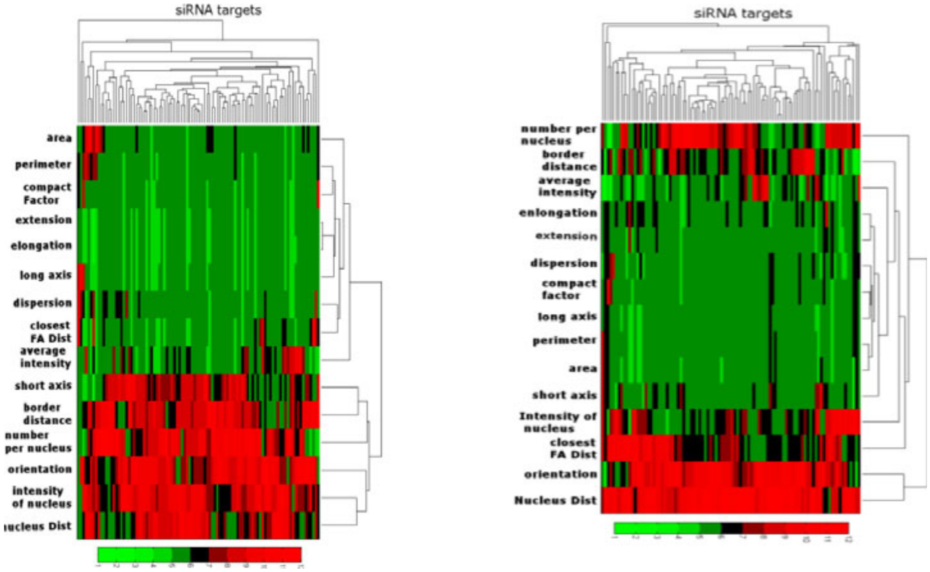

Fig. 10. (a) Vesicle p-value heat map (b) Plasma-membrane p-value heat map

References

1. Collinet, C., Stöter, M., Bradshaw, C.R., Samusik, N., Rink, J.C., Kenski, D., et al.: Systems survey of endocytosis by multiparametric image analysis. Nature 464, 24–249 (2010)
2. Damiano, L., Le Dévédec, S., Di Stefano, P., Repetto, D., Lalai, R., Truong, H., Xiong, J.L., Danen, E.H., Yan, K., Verbeek, F.J., Attanasio, F., Buccione, R., van de Water, B., Defilippi, P.: p140Cap suppresses the invasive properties of highly metastatic MTLn3-EGFR cells via paired cortactin phosphorylation. Oncogene 30 (2011)
3. Devroye L, Lugosi G.: A probabilistic theory of pattern recognition (1999)
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press (1990)
5. Galvez, T., Teruel, M.N., Heo, W.D., Jones, J.T., Kim, M.L., Liou, J., et al.: siRNA screen of the human signaling proteome identifies the PtdIns(3,4,5)P3-mTOR signaling pathway as a primary regulator of transferrin uptake. Genome Biology 8(7), 142 (2007)
6. Ghosh, R.N., DeBiasio, R., Hudson, C.C., Ramer, E.R., Cowan, C.L., Oakley, R.: Quantitative cell-based high-content screening for vasopressin receptor agonists using transfluor technology. Journal of biomolecular screening: the official journal of the Society for Biomolecular Screening 10(5), 47–84 (2005)
7. Hu, M.K.: Visual pattern recognition by moment invariants. Information Theory 8(2), 179–187 (1962)
8. Jain, A.K., Duin, P.W.: Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)
9. Jarque, C.M., Bera, A.K.: Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6(3), 255–259 (1980)
10. Land, A.H., Doig, A.G.: An Automatic Method of Solving Discrete Programming Problems. Econometrica 28(3), 497–520 (1960)
11. Le Dévédec, S., Yan, K., de Bont, H., Ghotra, V., Truong, H., Danen, E., Verbeek, F.J., van de Water, B.: Systems microscopy approaches to understand cancer cell migration and metastasis. Science 67(19), 3219–3240 (2010)
12. Li, H., Ung, C.Y., Ma, X.H., Li, B.W., Low, B.C., Cao, Z.W., et al.: Simulation of crosstalk between small GTPase RhoA and EGFR-ERK signaling pathway via MEKK1. Bioinformatics (Oxford, England) 25(3), 358–364 (2009)
13. Lilliefors, H.W.: On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. Journal of the American Statistical Association 64, 38–389 (1969)
14. Mahalanobis, P.C.: On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 2, 49–55 (1936)
15. Massey, F.J.: The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association 46(253), 68–78 (1951)
16. Mitchell, T.: Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression (2005)
17. Okun, O.: Feature Normalization and Selection for Protein Fold Recognition. In: Proc. of the 11th Finnish Artificial Intelligence Conference, pp. 207–221 (2004)
18. Pelkmans, L., Fava, E., Grabner, H., Hannus, M., Habermann, B., Krausz, E., et al.: Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 436(7047), 78–86 (2005)
19. Qin, Y., Stokman, G., Yan, K., Ramaiahgari, S., Verbeek, F.J., de Graauw, M., van de Water, B., Price, L.: Cyclic AMP signalling protects proximal tubular epithelial cells from cisplatin-induced apoptosis via activation of Epac. British Journal of Pharmacology (in Press, 2011)

20. Roepstorff K, Grøvdal L, Grandal M, Lerdrup M, Deurs B van.: Endocytic downregulation of ErbB receptors: mechanisms and relevance in cancer. Histochemistry and cell biology. 129(5):563–78(2008)
21. Tarcic, G., Boguslavsky, S.K., Wakim, J., Kiuchi, T., Liu, A., Reinitz, F., et al.: An unbiased screen identifies DEP-1 tumor suppressor as a phosphatase controlling EGFR endocytosis. Current Biology 19(21), 88–98 (2009)
22. Ung, C.Y., Li, H., Ma, X.H., Jia, J., Li, B.W., Low, B.C., et al.: Simulation of the regulation of EGFR endocytosis and EGFR-ERK signaling by endophilin-mediated RhoA-EGFR crosstalk. FEBS Letters 582(15), 2283–2290 (2008)
23. Yan K, Verbeek FJ.: Watershed Masked Clustering Segmentation in High-throughput Image Analysis (submitted, 2011)
24. Yan, K., Le Dévédec, S., van de Water, B., Verbeek, F.J.: Cell Tracking and Data Analysis of in Vitro Tumour Cells from Time-lapse Image Sequences. In: VISSAPP, vol. 1, pp. 281–286 (2009)