Biological model representation and analysis

Cao, L.

Chapter 3

Hierarchical classification strategy for Phenotype extraction from epidermal growth factor receptor endocytosis screening

Based on:

L. Cao, M. de Graauw, K. Yan, L. Winkel, F.J. Verbeek. Hierarchical classification strategy for Phenotype extraction from epidermal growth factor receptor endocytosis screening. Journal of BMC bioinformatics, 2013 (submitted)

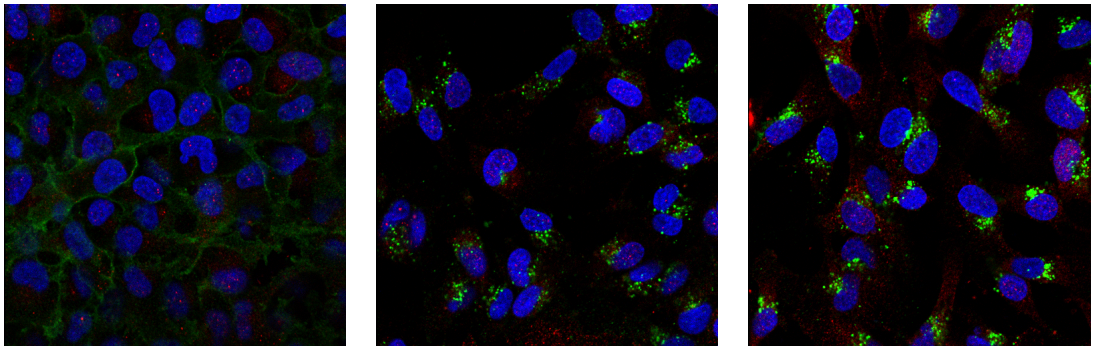
M. de Graauw, L. Cao, L. Winkel, M. H.A.M. van Miltenburg, S. E. le Dévédec, M. Klop, K. Yan, C. Pont, V. M. Rogkoti, A. Tijsma, A. Chaudhuri, R. Lalai, L. Price, F.J. Verbeek, B. van de Water. Annexin A2 depletion delays EGFR endocytic trafficking via cofilin activation and enhances EGFR signaling and metastasis formation. Oncogene, pages 1-10, 2013

Abstract: Endocytosis is regarded as a mechanism of attenuating the epidermal growth factor receptor (EGFR) signaling and of receptor degradation. There is increasing evidence becoming available showing that breast cancer progression is associated with a defect in EGFR endocytosis. In order to find related Ribonucleic acid (RNA) regulators in this process, high-throughput imaging with fluorescent markers is used to visualize the complex EGFR endocytosis process. Subsequently a dedicated automatic image and data analysis system is developed and applied to extract the phenotype measurement and distinguish different developmental episodes from a huge amount of images acquired from high-throughput imaging. For the image analysis, a phenotype measurement quantifies the important image information into distinct features or measurements. Therefore, the manner in which prominent measurements are chosen to represent the dynamics of the EGFR process becomes a crucial step for the identification of the phenotype. In the subsequent data analysis, classification is used to categorize each observation by making use of all prominent measurements obtained from image analysis. Therefore, a better construction of classification strategy will help to raise the performance level in our image and data analysis system. In this paper, we illustrate an integrated method employing wavelet-based texture measurements and a hierarchical classification strategy to further improve the recognition of phenotypic episodes of EGFR during endocytosis. Different strategies for normalization, feature selection and classification are evaluated. The performance estimation results show that our hierarchical classification scheme combined with wavelet-based texture measurements provides a notable improvement in the temporal analysis of EGFR endocytosis. The scheme could be further applied for the drug discovery to constrain the defect EGFR endocytosis process.

3.1 Background

The epidermal growth factor receptor (EGFR) is important for normal growth and function of breast tissue. Its signaling is regulated via endocytosis, a process that results in receptor degradation and thereby attenuation of EGFR signaling. In cancer cells, however, the endocytosis pathway is often found to be defective, resulting in an uncontrolled EGFR signaling. These enhanced EGFR signaling triggers breast cancer cells to escape from the primary tumor and spread to the lung, resulting in a poor prognosis for the disease progression. Moreover, it may result in complications like resistance to anti-cancer therapy.
From the literature [Geldner and Jürgens, 2006] a generic model of epidermal growth factor induced (EGF-induced) EGFR endocytosis can be distilled into four characteristic episodes. (1) Under normal conditions, EGFR is localized at the plasma-membrane site for internalization; this is in our study defined as the "plasma-membrane" episode. (2) Upon binding of EGF to the receptor, EGFR is taken up into small vesicular structures and starts sorting in early endosomes; in our study this is defined in our study as the "vesicle" episode. (3) Over time, EGFR containing vesicles are transported to late endosomes localized near the nuclear region and these form into a larger complex multi-vesicular body; in our study this defined as the "cluster" episode. (4) In final episode, EGFR is degraded in the lysosomes. In addition to this route, EGFR can also be partly transported back to the plasma-membrane sites. Using this dynamic model as the major guideline, the analysis of the EGFR-regulation-related genetic pathway could be linked to the analysis of the characteristic episodes in EGFR endocytosis. In this paper, we will focus on the analysis of this dynamic model but only the first three characteristic episodes as shown in Figure 3.1, since, in the final episode the EGFR signaling can not be traced.
Over the past years, RNA interference in combination with fluorescence microscopy-based imaging has become a powerful high-throughput tool for the visualization of complex EGFR endocytosis processes [Goldoni et al., 2006; Muniz Feliciano et al., 2013; Rappoport and Simon, 2009]. With these techniques, it becomes feasible to distinguish characteristic episodes and identify potential EGFR endocytosis regulators. However, it is impossible to perform manual processing of

such a large volume of image data. This requires an automated method for the analysis of EGFR endocytosis [de Graauw et al., 2013].



(a) (b) (c)

Figure 3.1. Sample images of the 3 phenotypic groups: (a) Plasma-membrane, (b) Vesicle and (c) Cluster. Red channel is P-ERK expression staining (Cy3); green channel is EGFR expression staining (Alexa-488); blue channel is nuclear staining (Hoechst #33258).

In the solution presented here, a single-step multi-class classification solution is demonstrated to properly capture the EGFR dynamics which transforming along three characteristic episodes and classify different EGFR episodes. From earlier application we have identified some weaknesses require better solutions. First, the same subset of features is used to classify three episodes. Second, the flat classification ignores the existence of potential hierarchical relationships that may exist in the data set. Another shortcoming is including the average intensity in the phenotype measurements. A variation of fluorescent intensity in image datasets is always presented. Thus, a more advanced classification strategy is required.
For our observation from previous results, we found the vesicle and cluster have more morphological similarity with each other than with plasma-membrane episode. As a result, we designed a new hierarchical classification strategy [Silla and Freitas, 2011]. Hierarchical classification strategies are an efficient way to deal with complex classification problems. First of all, the problem is divided in an hierarchical manner where classes with higher similarity to each other are grouped together into a sub-class, resulting in a simplification of original problem [Kumar et al., 2002]. Each parent node in the hierarchical tree has an individual classifi-

cation scheme choosing related features and the best classifier to distinguish the child nodes. Specifically, this hierarchical classification strategy separates classifier training into two levels. For the first level, we train the first level a local classifier to distinguish the plasma membrane from endosome, the subgroup containing vesicle and cluster. Subsequently, we train a second level local classifier to separate vesicle from cluster. With this strategy, we can make use of prominence in the subsets of features and thereby improve the performance of the classifier noticeably. Meanwhile, instead of using average of intensity directly, we introduce a set of texture measurements including those from wavelet transform to describe the intensity characteristics in a more sophisticated way.

3.2 Methods

3.2.1 Cell material and preparation

The study of cell systems at the cellular level at large scale is called cytomics. The workflow for image data preparation in cytomics includes three essential steps: (1) cell culturing, (2) labeling, preparation for imaging and (3) image acquisition. In this paper we use a workflow of an EGFR-regulation related siRNA screening to illustrate this workflow. In this design, breast cells from human breast carcinoma cell line (HBL100) were cultured in 96 well culture plate and transfected using Dharmafect smartpool Small interfering RNAs (siRNAs). Subsequently, the transfected cell population was exposed to epidermal growth factor (EGF) for a specific duration of time. Cells were fixed at different time points and visualized as confocal slice with a confocal laser microscope (Nikon TE2000).

3.2.2 Image acquisition and processing

Automated image acquisition was realized with a controlled motion stage equipped with an auto-refocusing module. For each well, images were captured from ten randomly selected locations. For each image three channels were captured: (1) a red channel containing staining of Phospho-ERK(P-ERK) expression (Cy3), (2) a green channel containing EGFR expression staining (Alexa-488) and (3) a blue

channel containing a nuclear staining (Hoechst #33258).
The subsequent image processing consists of two major steps: noise suppression and image segmentation. 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 comprehensive components. For fluorescence microscopy cell imaging, we utilized a customized segmentation algorithm known as the watershed masked clustering (WMC [Yan and Verbeek, 2012a]). The WMC algorithm is an innovative segmentation algorithm that is particularly suitable for images in which the individual objects exhibit a variation in fluorescence. The WMC algorithm serves different types of cytomics studies like dynamic cell migration analysis [Roepstorff et al., 2008; Yan et al., 2009b] and protein signaling modeling [Qin et al., 2012a]. Output binary mask was used to derive a number of phenotypic measurements for further data analysis.

3.2.3 Phenotype measurement

A phenotype is considered as the composite of an organism's observable characteristics or traits: such as its morphology or development (cf Chapter 1). It is important for the detection of genetic variants in complex traits. Therefore, researchers should be aware of the theoretical importance of unbiased, reliable and replicable measurements [Johannsen, 1911]. In our previous work, we have already introduced amount of basic measurements and localization phenotype measures [Cao et al., 2011]. In order to attempt finding more prominent phenotype measurements to characterize the three EGFR phenotypes, two aspects were considered. On the one hand, the phenotype measurements should be representative and relevant. On the other hand, these measurements must be robust to small variations in fluorescent intensity, meaning that the measurements are scale-free and self-normalized.
Based on the empirical observations in a ground truth data set, several potential texture patterns in object intensities were identified to characterize EGFR episodes. For instance, the vesicle (1st episode) has a higher intensity in the central region and relatively lower intensity around the boundary. In contrast, the cluster episode (2nd episode) has a more evenly distributed intensity throughout

the region of interest. In addition, these three EGFR episodes could also present distinctively in different texture features. Therefore, we introduced several texture measurements to describe different phenotypical characteristics.

3.2.3.1 Texture measurement

Table 3.1. Phenotype measurements description

Feature Name	**Expression**	**Description**
std	$f_1 = \sqrt{\sum_i (i - mean)^2 H(i)}$	The standard deviation of intensity from all the pixels in a region.
Smoothness	$f_2 = 1 - \frac{1}{(1+f_1^2)}$	The relative smoothness of the intensity in a region. It is 0 for a region of constant intensity and 1 for a region with large excursion in the values of its intensity levels.
Skewness	$f_3 = \sum_i (i - mean)^3 H(i)$	The order moment about the mean. The departure from symmetry about the mean intensity. It is 0 for symmetric histograms, positive for histograms skewed to the right and negative for histograms skewed to the left.
Uniformity	$f_4 = \sum_i H^2(i)$	The sum of squared elements in Histogram. It reaches maximum when all intensity levels are equal and decreases from there.
Entropy	$f_5 = -\sum_i H(i) \log_2 H(i)$	The statistical measure of randomness.
i represents the intensity value. H(i) is the histogram of intensity. mean symbolizes the average intensity.		

The most frequently used approach for texture measurements are the First Order Statistics; these are derived from statistical properties of the intensity histogram

of an image [Bountris et al.]. We used standard First Order Statistics for each individual object as obtained from the segmentation; i.e. standard deviation of intensity, smoothness, skewness, uniformity and entropy. Definitions and formulations of these texture measurements are presented in Table 3.1.

3.2.3.2 Wavelet texture measurement

Table 3.2. Wavelet texture measurements

Feature Name	**Description**
H_mean	The average intensity of Horizontal detail from discrete wavelet transformation.
H_std	The intensity variation of Horizontal detail from discrete wavelet transformation.
H_Entropy	The statistical randomness of Horizontal detail from discrete wavelet transformation.
V_mean	The average intensity of Vertical detail from discrete wavelet transformation.
V_std	The intensity variation of Vertical detail from discrete wavelet transformation.
V_Entropy	The statistical randomness of Vertical detail from discrete wavelet transformation.
D_mean	The average intensity of Diagonal detail from discrete wavelet transformation.
D_std	The intensity variation of Diagonal detail from discrete wavelet transformation.
D_entropy	The statistical randomness of Diagonal detail from discrete wavelet transformation.

Recently, texture analysis based on the discrete wavelet transform (DWT) has shown to be feasible [Goldoni et al., 2006; Tsiaparas et al., 2012] as it turns out to be an efficient descriptor for phenotyping [Materka, 2001]. DWT provides a set of texture representations consisting of coefficients in different directions. We calculated our wavelet-based texture measurements by multiplying each direc-

tion detail with the binary mask obtained from the segmentation and calculating the mean, standard deviation and entropy of intensity for each labeled object in each direction details (see in Table 3.2). In this study, we include a biorthogonal wavelet because it has the property of exact reconstruction and it is an outstanding wavelet representation for image decomposition. After decomposition, it generated the coefficient matrices of the level-one approximation and horizontal, vertical and diagonal details. Subsequently, we reconstructed the level-one details respectively from the corresponding coefficients. In this way, we derived the texture details from three different directions on the same scale as the original image.

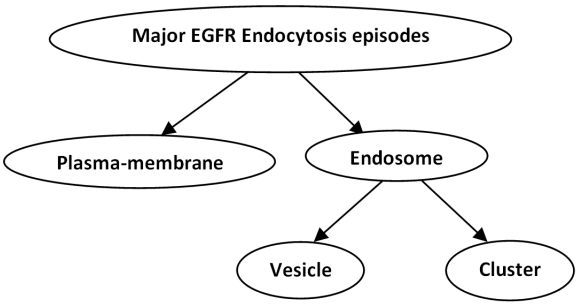


Figure 3.2. Hierarchical tree of EGFR Endocytosis Process.

3.2.4 Production of ground truth data

Collecting objective and sufficient ground truth data is important for supervised classification. We use the ground truth data as our training set during the classifier training.

Ground truth data were obtained by outlining the three characteristic episode groups, i.e. plasma-membrane, vesicle and cluster. These were separately drawn by biologists using our annotation software (TDR) with a digitizer tablet (WACOM, Cintiq LCD-tablet)[Verbeek and Boon, 2002]. From each outline a binary mask was created for each phenotypic stage. The Figure 3.4(b) illustrates the vesicle mask derived from a manually selected vesicle outline. This mask was multiplied with the mask obtained from the WMC algorithm so as to extract the intersection (cf. Figure. 3.4(d)). Finally, the phenotype measurements were computed from these masks. The ground truth datasets for the plasma-membrane

and cluster episodes were prepared in a similar manner. The training dataset included the three characteristic episode groups with a total of 2254 objects and 25 features per object.

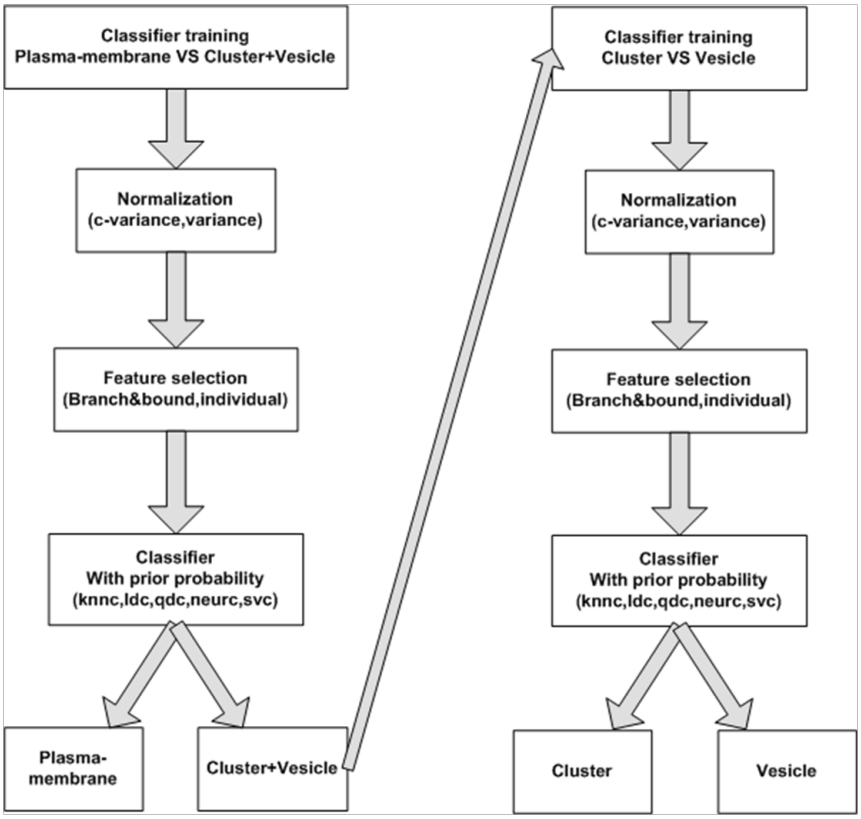
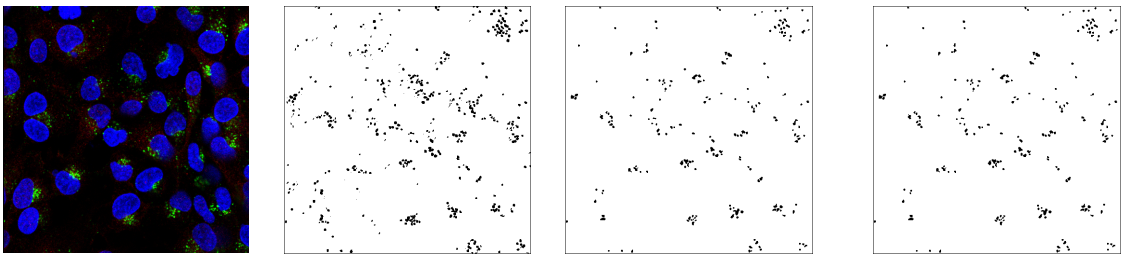


Figure 3.3. Hierarchical classification workflow.



(a) (b) (c) (d)

Figure 3.4. Ground truth data production. (a) Original image, (b) Manual mask, (c) WMC mask, (d) Multiply b and c.

3.2.5 Hierarchical classification strategy workflow

On the basis of first three characteristic episodes of the generic model of epidermal growth factor induced EGFR endocytosis, vesicle is single/early endosomes and cluster is clustered/late endosomes. These two episodes have more morphological similarity with each other compared to plasma-membrane. The vesicle and cluster episodes are located in the cytoplasma which have evenly distributed high intensity value and relatively circular shape. The plasma-membrane episode is located around cell membrane which has an elongated shape with a low and unevenly distributed intensity value. Therefore, we have constructed the three characteristic episodes into a hierarchical tree as shown in Figure 3.2. Subsequently, we use a local classifier per parent node approach to train a multi-class classifier for each parent node in the class hierarchy. From this approach one avoids the problem of making inconsistent predictions and takes into account the natural constrains of class membership [Davies et al., 2007; Neuwald et al., 2012; Silla and Freitas, 2011]. In this manner, both the best classifier and the most prominent features are selected for each parent node classifier so as to classify the dynamic model with three episodes in a better fashion. The workflow of the hierarchical classification strategy is shown in Figure 3.3. In our workflow, we normalized the dataset per feature, performed the feature selection, applied the classifier and calculated the weighted classification error in order to evaluate the performance of the classification. We look for the best combination of classification process according to the error estimation and use it for EGFR episode classification.

3.2.5.1 Feature normalization

The features can have a quite substantial difference in their dynamic range. Such is the case with the features that we use in this setup. Therefore, it is necessary to normalize our dataset. Feature normalization is required to approximately equalize ranges of the features and make them have roughly the same effect in the computation of similarity [Aksoy and Haralick, 2001]. The main advantage of normalization is to avoid attributes with larger numerical ranges to dominate over those with smaller numerical ranges. Another advantage is to avoid numeri-

cal complications during the computations. Because kernel values depend on the inner products of feature vectors, large attribute values might introduce numerical complications [Hsu et al., 2010].
We applied two types of normalization schemes to normalize the dataset. One standard normalization scheme was accomplished by shifting the mean of the dataset to the origin and scaling the total of variance for all features to 1, thereby, neglecting class priors. The other scheme was achieved by shifting the mean of the dataset to the origin and normalizing the average class variances (within-class). Class priors were taken into account. The concept of within-class covariance normalization for support vector machines (SVM) classifier was recently introduced [Hatch et al., 2006]. For the evaluation of the methods considered in this paper, we are evaluating these two normalization schemes and we are interested to see whether, in our case, within-class covariance normalization outperforms the standard normalization. We will benefit from the fact that the normalization avoids differences in numerical scales.

3.2.5.2 Feature selection

After normalization, we applied feature selection procedure. We did not use the supervised linear feature extraction method because of the limitation that the number of extracted features can, at most, be one less than the number of classes. Therefore, no more than this number of new features can be obtained. In our case, for each hierarchical step we can map all the features into a 1D linear subspace since we simply have two classes for each step. It would cause difficulties in the discrimination between two classes [Shadvar and Erfanian, 2010]. Furthermore, in our study, the feature extraction method is not outstanding as feature selection method compared to [Cao et al., 2011].
For feature selection a metric is required that considers strong correlation among the variables, therefore, the Mahalanobis distance [Mahalanobis, 1936b; Webb and Copsey, 2011] was chosen. Subsequently we selected two representative search algorithms: the branch and bound procedure [Land and Doig, 1960b] and best individual-N features. Branch and bound is a top-down procedure, beginning with the set of variables and constructing a tree by successively deleting vari-

ables. This feature selection method showed a robust and high performance in our previous study [Cao et al., 2011]. Best individual-N features procedure is a computationally efficient method for choosing the best N features by assigning a discrimination power estimate to each of the features in the original set. This method could have a well-defined feature set when the features are uncorrelated. We would use these two search algorithms in the feature selection part.

3.2.5.3 Prior probability setting

In probability theory and applications, the Bayes' theorem shows the relation between a conditional probability $P(A|B)$ and its reverse form $P(B|A)$, expressed as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Table 3.3. Prior probability comparison

C-variance(branch&bound)										
equal prior	knnc		ldc		qdc		neurc		svc	
	mean	std	mean	std	mean	std	mean	std	mean	std
1st step	0.0333	0	0.095	0.0224	0.0317	0.0075	0.0633	0.0149	0.0333	0
2nd step	0.0575	0.0335	0.0575	0.0335	0.15	0	0.055	0.0224	0.1025	0.0112
no prior	knnc		ldc		qdc		neurc		svc	
	mean	std	mean	std	mean	std	mean	std	mean	std
1st step	0.0181	0.0014	0.0365	0.0057	0.0221	0.0037	0.01	0.0031	0.0142	0.001
2nd step	0.0292	0	0.0357	0.0088	0.043	0.0121	0.0348	0.0065	0.0402	0.0069
with prior	knnc		ldc		qdc		neurc		svc	
	mean	std	mean	std	mean	std	mean	std	mean	std
1st step	0.0053	0.0011	0.0061	0.0038	0.0059	0.0028	0.0061	0.0029	0.0061	0.0038
2nd step	0.0214	0.0031	0.0321	0.0011	0.056	0.0019	0.0231	0.0048	0.0214	0.002

A prior probability P(A) is the probability distribution of A before the specific condition is taken into account [Haldane, 1948; Witten and Frank, 2005]. It means to attribute uncertainty rather than randomness to the quantity under investigation. A prior is often a purely subjective assessment of an expert. In order to obtain this prior knowledge, we chose a group of images with 6 different time stages and manually counted the number of the three characteristic

episodes. Subsequently, we calculated the ratio between plasma-membrane and the subgroup (cluster and vesicle) as 0.0526 and the ratio between cluster and vesicle as 0.0556. This ratio is ascertained by biologists through observation. In addition, we verified the performance of this prior probability with simply no prior probability and with equal prior probability. The within-class covariance normalization was selected for this optimization scheme. Subsequently, the weighted error of different classifiers was calculated after branch and bound feature selection. The results in Table 3.3 show that the performance is increased when the prior probability is included.

3.2.5.4 Classifier

The classifiers were selected on their ability to cover both linear and non-linear categories; i.e. the linear classifier (LDC), the quadratic classifier (QDC), the k-nearest neighbor classifier (KNNC), the support vector machine classifier (SVC) and the neural network classifier (NEURC). The linear classifier makes a classification decision based on the value of a linear combination of the characteristics [Mitchell, 1997]. Compared to nonlinear classifiers, the performance of linear classifier is less preferred in data that are strongly non-Gaussian distributed [Devroye et al., 1996]. The quadratic classifier is generalized form of the linear classifier that separates classes on the basis of a quadratic hyperplane. The k-nearest neighbor classifier distinguishes an object by majority voting of its neighbors, with the object being assigned to the class most common amongst its neighbors. The Support Vector Machine (SVM) is primarily a classifier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. [Cortes and Vapnik, 1995] Key to the SVM is the use of kernels, the absence of local minima, the sparseness of the solution and the capacity control obtained by optimizing the margins [Cristianini and Shawe-Taylor, 2000].
In a neural network, units (neurons) are arranged in layers and these layers convert an input vector into some output. Each unit takes an input, applies a (often nonlinear) function to it and then passes the output on to the next layer [Zhang, 2000]. There are many artificial neural network (ANN) models; i.e., Feed-Forward

Networks, Radial Basis Function Networks, Recurrent Networks, etc. The advantage of neural networks is two-fold. First, neural networks are data driven self-adaptive methods. The flexibility is created by the combination of different nodes with related kernels. Second, they are universal functional approximators in which neural networks can approximate any function with arbitrary accuracy. The disadvantage of neural networks is that they are notoriously slow and is very difficult to determine the optimal number of kernel types, layers and nodes [SangitaB and Deshmukh, 2011]. In this study we used the biologically inspired feed-forward neural network with a single hidden layer. The feed-forward neural network is defined as a unit feeding its output to all the units in the next layer, but there is no feedback to the previous layer. It is the simplest form of artificial neural network and it can yet limit the complexity of network calculation.

3.2.5.5 Analysis

In order to find a panel of classifiers to address the categorization of our dynamic model with three episodes, we included two normalization schemes, two feature selection methods and five classifiers. We have presented the results of a weighted error estimation. The classification are the results of a weighted error estimation procedure as implemented in PRTools [Duin et al., 2007]. The results are depicted in Figure 3.5, 3.6. For the first classifier training (plasma-membrane vs. subgroup of cluster and vesicle), we observe in Figure 3.5, the weighted error of the linear classifier increases abruptly when the number of features exceeds a certain threshold (three for branch and bound and eight for individual feature selection method). The weighted error of the k nearest-neighbor classifier, however, is more stable and equals the lowest point in the branch and bound feature selection group. As for the second classifier training (cluster vs. vesicle), we noticed, in Figure 3.6, the weighted error of quadratic classifier performs worst. We created scatter plots of mapped data with a linear classifier and a quadratic classifier so as to indicate the reasons of the worst performance with the quadratic classifier; depicted in Figure 3.7. The error line of the support vector machine classifier shows quite stable; it obtains the lowest values in the group of branch and bound feature selection method. Nevertheless, in the individual feature se-

lection group, the error of neural network classifier always evaluates as the best performance in terms of magnitude of the error.
For a global observation, we selected the minimal mean error from all weighted error with different feature dimensions. This value represents the best performance of the combination between feature selection and classifier. In Table 3.4, the standard deviation of each minimal mean value is shown. The combination of branch and bound feature selection with k-nearest neighbor classifier has the lowest minimal value and relatively small standard deviation for the first classifier training with both normalization schemes. For the second classifier training, both the combination of branch and bound feature selection with support vector machine classifier and the combination of individual feature selection with neural network classifier have the same lowest minimal value and comparatively small standard deviation in two normalization schemes. In order to make a final combination more general and with a lower feature dimension, we choose the combination of branch and bound feature selection with k-nearest neighbor classifier for first step classification with variance normalization scheme. For the second step, the combination of branch and bound feature selection with support vector machine classifier is chosen with variance normalization scheme. The reason for the selection of branch and bound feature selection method is the existence of correlated features on our feature set which causes the low performance of the best individual-N feature selection method. The branch and bound feature selection method guarantees the optimal feature subset without explicitly evaluating all possible feature subsets because the criterion function fulfils the monotonicity condition [Somol et al., 2004]. The K-nearest neighbor classifier as the selection of the first step is a simple but efficient classifier for the basic recognition problem such as two classes classification problem. On the other hand, a higher value of K provides smoothness which reduces the vulnerability to noise in the training data. For the second step, support vector machine is selected for its flexibility in threshold choosing since its function is non-parametric. Additionally, support vector machine has the ability of maximizing the generalization because it is trained to maximize the margin.

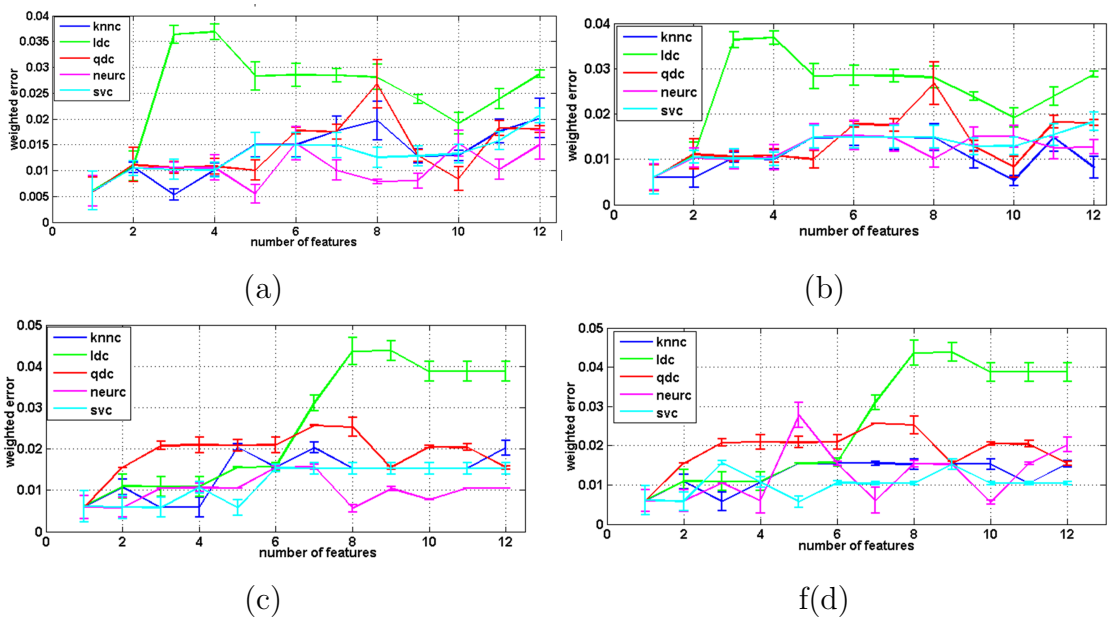


Figure 3.5. First classifier training (plasma-membrane VS subgroup of cluster and vesicle). (a) Branch&bound feature selection method with standard variance normalization (b) Branch&bound feature selection method with within-class variance normalization (c) Individual feature selection method with standard variance variance normalization (d) Individual feature selection method with within-class variance normalization.

The prominent features we choose for the first step are long axis, D-entropy and standard deviation of intensity. For the second step we choose the top five features; i.e. skewness, entropy, H-entropy, closest object Dist, area. The evaluation of the feature selection performance is shown in Table 3.5. We do the evaluation by calculating the probability of each feature being selected by the feature selection method. These highly selected features reflect the phenotype changes between three characteristic episodes. For example, for the step 1, Long Axis is chosen most of the time because the plasma membrane is around the membrane which tends to have an elongated shape than other two episodes. In step 2, the intensity entropy is selected because clusters have a flatter region than vesicles which results in a lower entropy value. In Figure 3.8 and Figure 3.9 scatter-plots are shown for both hierarchical steps. The evaluation results tell us that there is no relevance in using large amounts of features. Just a few will contribute to

the final results. We can also see both steps choose the wavelet texture features that we introduced in this research; these clearly resulted in a better performance compared to the previous classifier scheme.

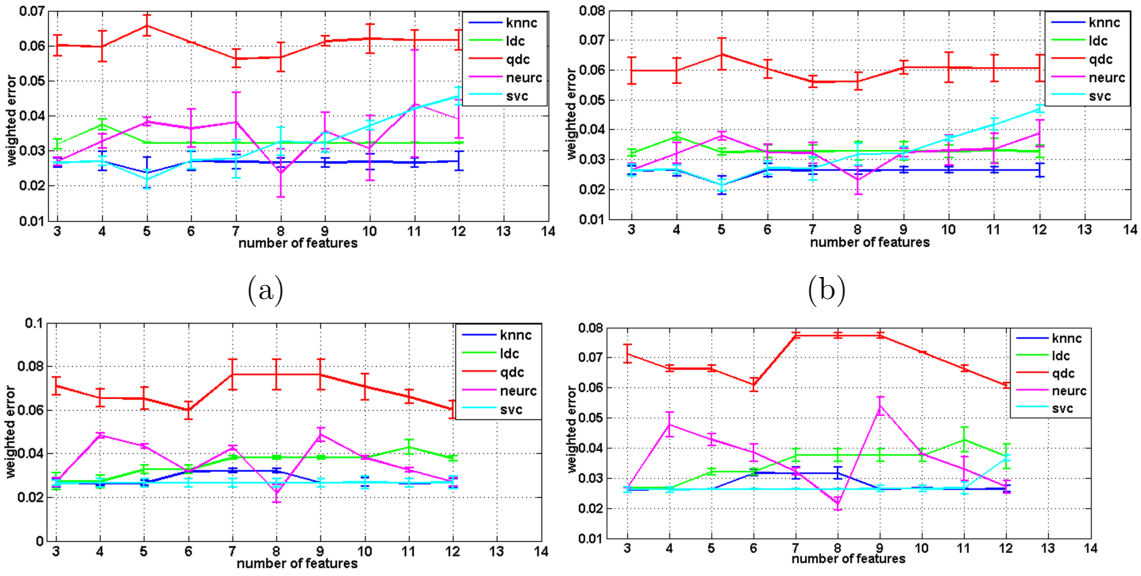


Figure 3.6. Second classifier training (cluster VS vesicle). (a) Branch&bound feature selection method with standard variance normalization, (b) branch&bound feature selection method with within-class variance normalization, (c) Individual feature selection method with standard variance variance normalization, (d) Individual feature selection method with within-class variance normalization.

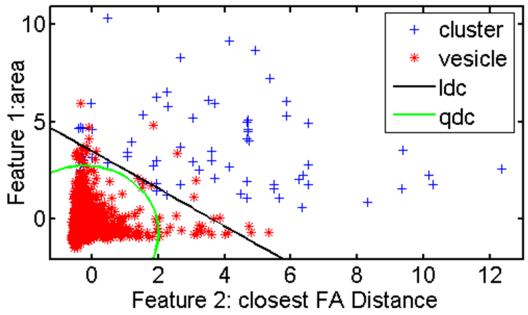


Figure 3.7. Scattered plot of training data with ldc and qdc.

Table 3.4. Weighted error comparison

1st step	knnc		ldc		qdc		neurc		svc	
C-V	mean	std	mean	std	mean	std	mean	std	mean	std
B&B	0.0053	0.0011	0.0061	0.0038	0.0059	0.0028	0.0061	0.0029	0.0061	0.0038
IND	0.0058	0.0023	0.0061	0.0038	0.0059	0.0028	0.0055	0	0.0057	0.0014
2nd step	knnc		ldc		qdc		neurc		svc	
C-V	mean	std	mean	std	mean	std	mean	std	mean	std
B&B	0.0214	0.0031	0.0321	0.0011	0.056	0.0019	0.0231	0.0048	0.0214	0.002
IND	0.0261	0	0.0267	0	0.0607	0	0.0214	0.0021	0.0261	0
1st step	knnc		ldc		qdc		neurc		svc	
VAR	mean	std	mean	std	mean	std	mean	std	mean	std
B&B	0.0053	0.0011	0.0061	0.0038	0.0059	0.0028	0.0055	0.0018	0.0061	0.0038
IND	0.0058	0.0023	0.0061	0.0038	0.0059	0.0028	0.0056	0	0.0058	0.0019
2nd step	knnc		ldc		qdc		neurc		svc	
VAR	mean	std	mean	std	mean	std	mean	std	mean	std
B&B	0.0237	0.0044	0.0319	0.0015	0.0563	0.0027	0.0236	0.0069	0.0218	0.0028
IND	0.0263	0	0.0272	0.0029	0.0598	0.004	0.0218	0.004	0.0265	0.001

C-V represents c-variance. B&B represents branch & bound. IND represents individual. VAR represents variance.

Table 3.5. Feature selection performance

Features	Step 1
Long Axis	100
Int Std	100
D_entropy	87
H_entropy	7
V_entropy	5
Smoothness	1
Area	0
Perimeter	0
Extension	0
Dispersion	0
Elongation	0
Orientation	0
Compact Factor	0
Border Dist/Nucleus Dist	0
Closest FA Dist	0
Short Axis	0
Skewness	0
Int Uniformity	0
Int Entropy	0
H_mean	0
H_std	0
V_mean	0
V_std	0
D_mean	0
D_std	0

Features	Step 2
Closest FA Dist	100
Int Entropy	100
Area	96
Int Std	73
Compact Factor	44
Int Uniformity	34
Smoothness	14
H_entropy	8
Border Dist/Nucleus Dist	7
Perimeter	6
Long Axis	6
Short Axis	5
D_std	5
Skewness	2
Extension	0
Dispersion	0
Elongation	0
Orientation	0
H_mean	0
H_std	0
V_mean	0
V_std	0
V_entropy	0
D_mean	0
D_entropy	0

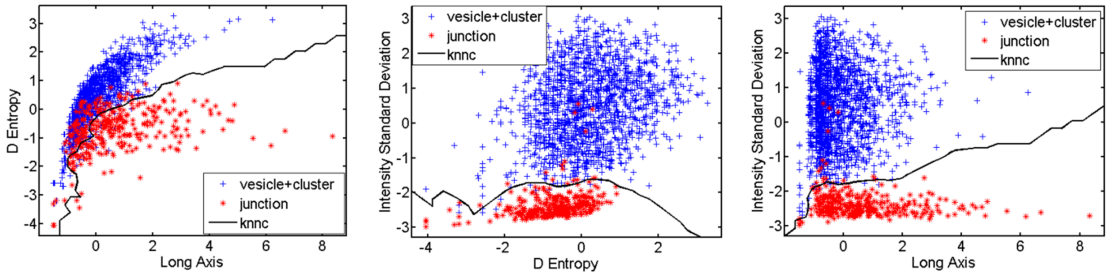


Figure 3.8. Step 1 Scattered Plot.

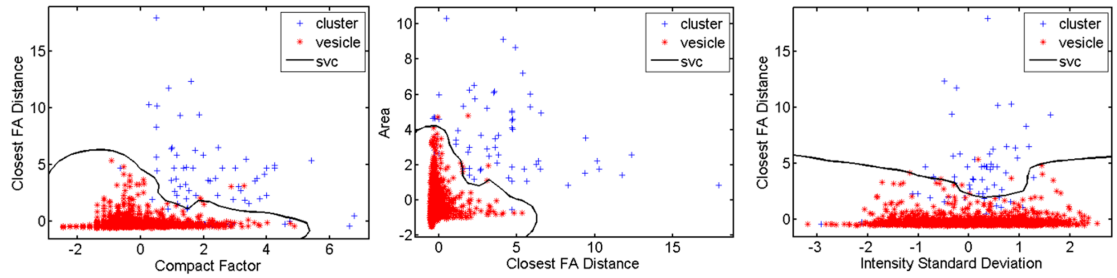
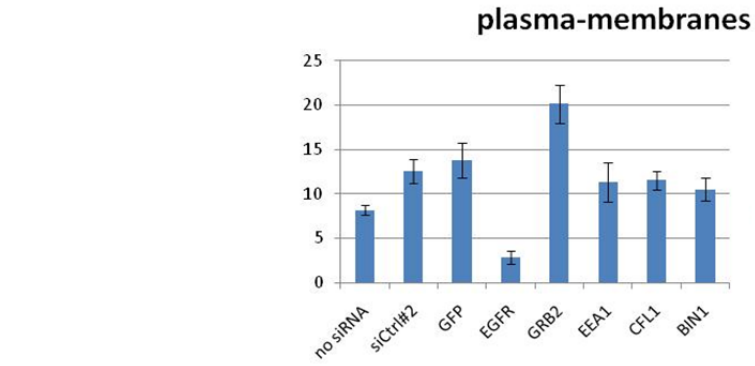


Figure 3.9. Step 2 Scattered Plot.

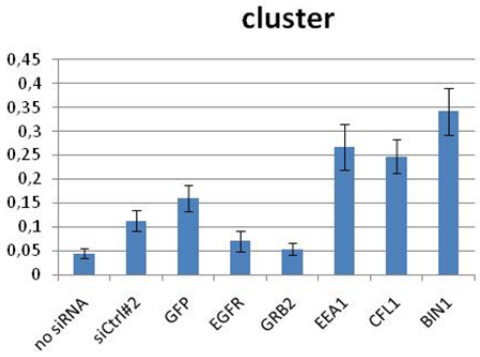
3.3 Results and discussion

3.3.1 EGFR endocytosis regulator identification

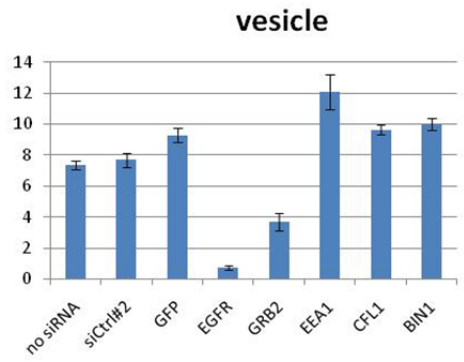
Our hierarchical classification strategy was used to validate siRNA-mediated knock-down of several known EGFR endocytosis regulators (e.g. siGrb2, siEEA1, siCFL). We selected 10 sample points per well from WT cells (not treated with siRNA), control siRNA treated cells (siCtrl#2 and siGFP), siEGFR treated cells and three target siRNAs. After image processing and data analysis, we calculated the number of objects belonging to each episode group per nucleus and compared the result. This is depicted in Figure 3.10. As expected, cells treated with siEGFR show a decreased level of all three classes since treatment of cells with siEGFR results in $> 90\%$ knock-down of EGFR. Cells incubated with siGrb2 show a drastic reduction in number of endosomes (vesicle and cluster) because siGrb2, as a known regulator of EGFR endocytosis, can significantly inhibit EGFR internalization [Jiang et al., 2003]. In general the increase in number of endosomes



(a)



(b)



(c)

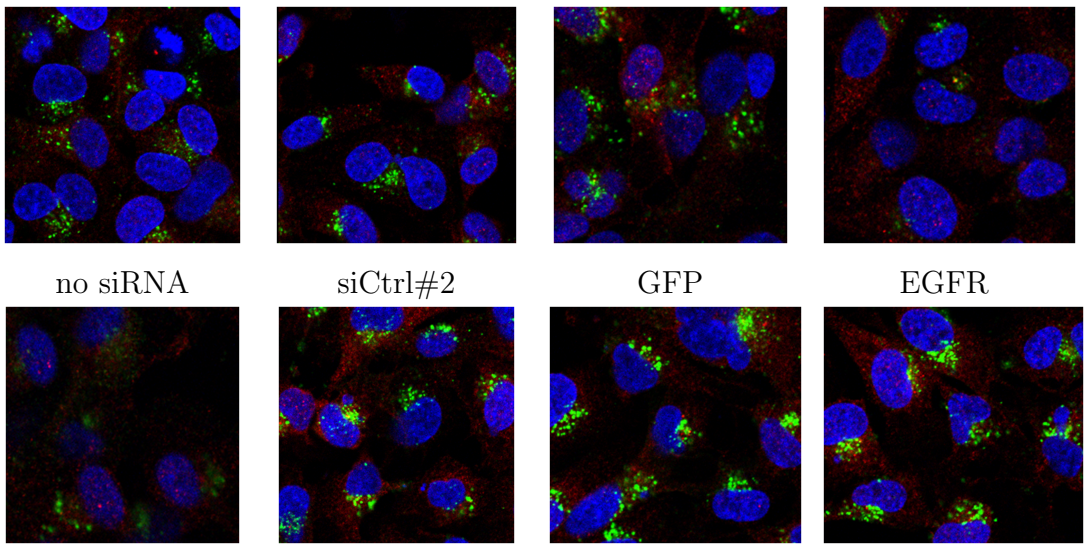


Figure 3.10. EGFR endocytosis regulator identification results. (a) Pixels of plasma-membrane, (b) Number of clusters, (c) Number of vesicles.

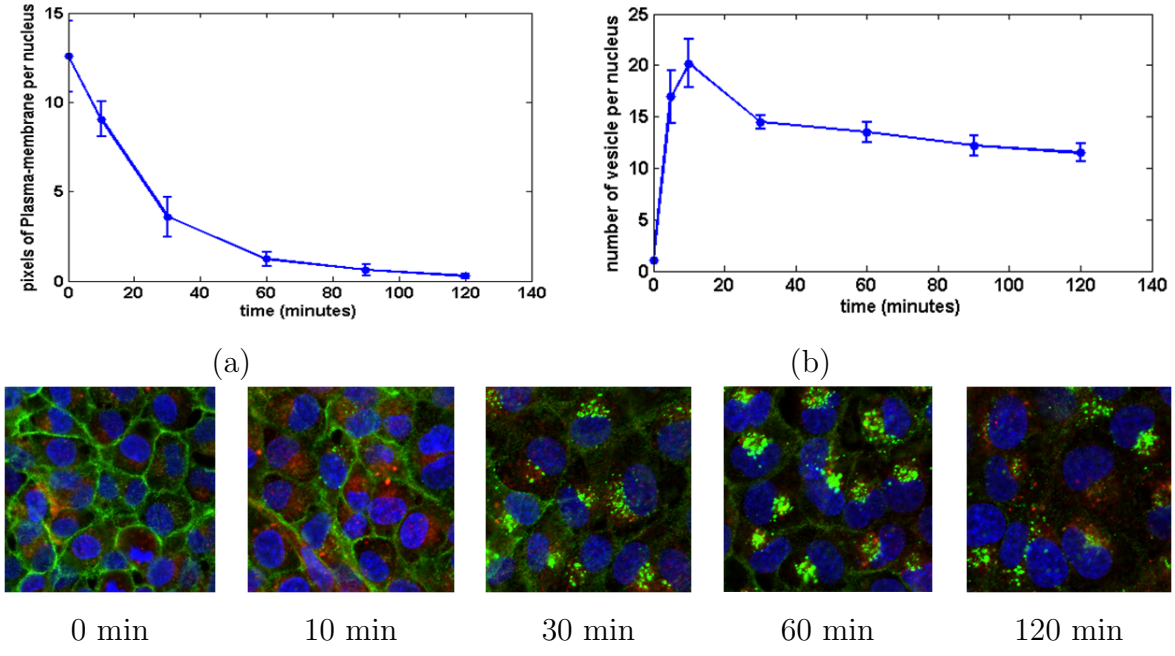


Figure 3.11. Dynamic stages of EGFR endocytosis. (a) Number of EGFR localized at Plasma-membrane (pixels/nucleus), (b) Number of EGFR as vesicle in early endosome (number/nucleus).

(vesicle and cluster) can be caused either by enhanced uptake of EGFR resulting in an enhanced EGFR endocytosis and EGFR degradation, or by delayed endocytosis and EGFR breakdown. For EEA1 (a member of the early endosomes), an increase in number of endosomes (vesicle and cluster) is due to delayed endocytosis [Leonard et al., 2008]. Cofilin (CFL) regulates the cytoskeleton and because of these changes in the actin cytoskeleton the endocytosis route of EGFR changes [Nishimura et al., 2006]. At present, there is not a lot of knowledge about BIN1 in EGFR endocytosis. Regarding the result, BIN1 depletion decreases EGFR at plasma-membrane, increases the number of vesicles and clusters, suggesting that it potentiates EGFR endocytosis and possibly signaling. These results demonstrate the robustness of our hierarchical classification scheme and the capability to predict new EGFR endocytosis regulators.

3.3.2 Dynamic EGFR endocytosis stage

In this case study, HBL100 cells were exposed to EGF (50 ng/ml) for indicated timepoints (Figure 11c). The number of EGFR localizations at the plasma-membrane and the vesicles was quantified. The amount of EGFR localized at the plasma-membrane, expressed as pixels of plasma-membrane per nucleus, decreases over time as shown in Figure 11a. This fits with the EGFR endocytosis process during which EGF exposure is causing a gradual EGFR re-distribution from the plasma-membrane into vesicles. Meanwhile, the number of vesicles per nucleus increases at early endosome caused by the EGFR internalization, then decreases at the late endosome stage when the vesicles form into a larger complex clusters and degrades at the end as illustrated in Figure 11b. These graphs indicate the trend of EGFR endocytosis process and are representative in illustrating the dynamics of EGFR endocytosis stages.

3.4 Conclusions

This chapter discusses an improved image and data analysis system for High-throughput screening including texture wavelet-based measurements and a hierarchical classification strategy. Having learned from earlier results [Cao et al., 2011] we have improved the phenotype description with the new texture features and improved the classification of the characteristic episodes with an alternative classification scheme. For the image analysis we use an innovative image segmentation algorithm combined with representative phenotype measurements which include relative texture features with wavelet transform to replace the absolute intensity feature so as to decrease the impact of fluorescent intensity variation. For the data analysis part, we change from a single-step multi-class classification solution into a two step Hierarchical classification strategy to categorize three dynamic phenotypes of EGFR endocytosis process. We include two feature normalization methods, two feature selection methods and five classifiers to find the best classification strategy. After evaluation of different combinations, we have chosen the combination of branch and bound feature selection with k-nearest neighbor classifier as for first step classification after normalization of variance.

As shown in Table 3.4, the combination of branch and bound feature selection with support vector machine classifier is chosen for the second step classification after having applied the same normalization method.
With the selected combination, the classifier shows a notable improvement in distinguishing plasma-membrane from the dataset. This improvement is due to three factors. First, it benefits from the hierarchical classification scheme which introduce multilevel classifiers to deal with two subset classes at a time. Second, we introduced the exact prior probability for the classifier training which improve the performance of the classification strategy significantly. Third, the relative texture measurements show their potential to describe the phenotype characteristics.
This explicit hierarchical classification solution can identify the characteristic episodes in the EGFR endocytosis process and is helpful to find new regulators in this crucial process relating to breast cancer progression. With all kinds of phenotype measurement and flexible classifier training strategy that we introduced in this work, it is easy to detect morphological changes of phenotype and extend our solution to cope with studies utilizing fluorescence microscopy in a siRNA based high-throughput screening (HTS).