Biological model representation and analysis

Cao, L.

Chapter 7

Conclusion and Future Work

In this thesis, different ways to represent and analyse biological models from image dataset are discussed. These image datasets can be two dimensional (x,y) or three dimensional (x,y,z; x,y,t). We have used different feature sets and feature extraction methods. We focus on finding patterns for analysis in multi-dimensions both with spacial and temporal series. The images are captured in sufficient qualities to be able to find patterns in a population. This can be a high-throughput screening but also three dimensional modeling approaches are probed. In each chapter, the ground truth is used for validation of phenotype analysis and method evaluation. In temporal two dimensions, i.e. time series, we focus on a biological study on epidermal growth factor receptor (EGFR) signalling and receptor degradation. The defect in the receptor mechanism is considered to be closely related to the breast cancer progression. We have provided a solution to analyse high-throughput image datasets on the level of protein location in Chapter 2. Moreover in Chapter 3, in a EGFR endocytosis study, we have introduced a Hierarchical classification strategy to improve the categorization of three dynamic phenotypes in the EGFR endocytosis process. In Chapter 4 we have changed our working field to spacial 3D datasets and use the extracted features for reconstruction method evaluation. In Chapter 4 we have presented error estimation for four representative 3D surface reconstruction methods by comparing analytical features derived from a surface model, thereby evaluating the quality of the surface model. In Chapter 5, we are making use of the conclusions drawn in Chapter 4 and build a system for 3D surface representation and analysis from stack of

images; basically this consists of an optimization of the geometrical model. In Chapter 6, we further use the surface representation workflow derived in Chapter 5 for complex 3D model reconstruction and analysis. In addition, we introduce the L-system as a model for ground truth construction.

7.1 Image analysis and Pattern Recognition in High-throughput Screens

The concept of high-throughput screening is used to visualize various cell structures. One HTS experiment may produce up to half million images which is a quality for which it is not possible to be analysed without a clearly formulated plan of automation. Therefore, an automated analysis solution for HTS experiments is required by combining image analysis and pattern recognition. Chapter 2 uses a 1D episode to represent the biological model. These categorized episodes are described by features extracted from 2D images and further confirmed through classification using a ground truth model. As a result, an automated system is constructed to extract phenotype measurements for each object and characterize the objects into three characteristic episodes in the EGFR endocytosis process. We illustrate that the phenotype measurements from segmented images and categorization of phenotypic episodes can be done successfully using feature selection and classification. The best trained classifier has been used to classify three EGFR phenotypic episodes. Two case studies show the capability of our solution in identifying characteristic episodes and analysing a large scale siRNA screening.

7.2 Hierarchical classification strategy for EGFR phenotype extraction

From the good results accomplished in Chapter 2, we continued to improve the phenotype identification process so as to get an even higher classification score.

We had noticed that the way to quantify the prominent features from a segmented image is a crucial step in the identification process. In the data analysis part, the classification strategy is evenly important to make full use of all measurements. Our improvements are reported in Chapter 3. We have designed a scheme by employing a hierarchical classification strategy and adding wavelet-based texture measurements to further improve the recognition of phenotypic episodes of EGFR endocytosis. The hierarchical classification strategy is very capable in dealing with complex classification problems. For the case study in Chpater 3, we construct the classification process in a hierarchical way by separating three classes classification into two steps. Meanwhile, we can select prominent features for each step. As a result, this strategy makes full use of related features and improves the performance of the classifier. The motivation for us to use the wavelet-based texture measurements is to include extra prominent features in our set of observations and to lessen the impact of fluorescent intensity variation. After integrating all merits, the phenotype identification process shows a remarkable improvement and has been successfully used to find new regulators in the EGFR endocytosis process.

7.3 Analytical evaluation of point cloud surface reconstruction methods

Feature quantification can not only be used for classifier training, but can also be used for analytical evaluation. For 3D feature analysis we are interested in surface description for geometrical models derived from images. As far as we have concluded from the literature, the straightforward way to evaluate surface reconstruction methods is by convincing through visual inspection. Analytical evaluation seems to be missing for this purpose. In Chapter 4, we utilize 3D surface descriptors including surface distance, surface area and surface curvature to evaluate four representative surface reconstruction methods from a point cloud. Meanwhile, we validate 3D model through ground truth models. To that end we have introduced three analytical shapes: the unit sphere, the ellipsoid and the

ovoid to provide the ground truth values. From the results we have concluded that Poisson reconstruction method has the most stable performance in shape preservation and noise repression. We intend to utilize the results on biological 3D models to improve the quality of surface representation; because the former representation was not considered sufficiently for analysis.

7.4 3D model representation for phenotype analysis

For a 3D stack of images, model is represented in a geometrical model. We have optimized the 3D reconstruction method based on the knowledge derived from Chapter 4. First, we have reconstructed a 3D model from the stack of 2D contour images. Subsequently, we have been able to extract important features from the surface model. In Chapter 5, we have introduced such a system for 3D representation and analysis from a stack of images. We have augmented the model by resampling using a contour interpolation method. Our workflow has been validated with real image stack data. In this manner a stack of images is successfully converted into a uniformly distributed point cloud. We have applied the conclusion derived from Chapter 4 and have utilized the Poisson reconstruction to construct the surface model from the point cloud. We have chosen different shape features to adapt to different biology studies. We have used the sphericity shape factor in a study on development of zebrafish. This feature is a relative and compact descriptor. This case study validated the capability of our 3D representation and analysis system.

7.5 Nature inspired phenotype analysis with 3D model representation

In Chapter 6, we have intentionally been dealing with a much more complex 3D shape: the developing rodent mammary gland. The mammary gland is a branched structure and we have used centerline extraction method to establish the

topology of the branch structure and calculate the corresponding measurements from this simplified topological structure. Further we have introduced a ground truth model by using the L-system formation. We construct an L-system model for mouse mammary gland by which we can simulate the different results under different conditions. In this Chapter, we have validated the robustness of the 3D representation method and have demonstrated the potential of L-system as a ground truth model where no other is available but a good formalism can provide the necessary insight. The L-system resulted to show the effect of endocrine disruption.

7.6 Conclusion and future work

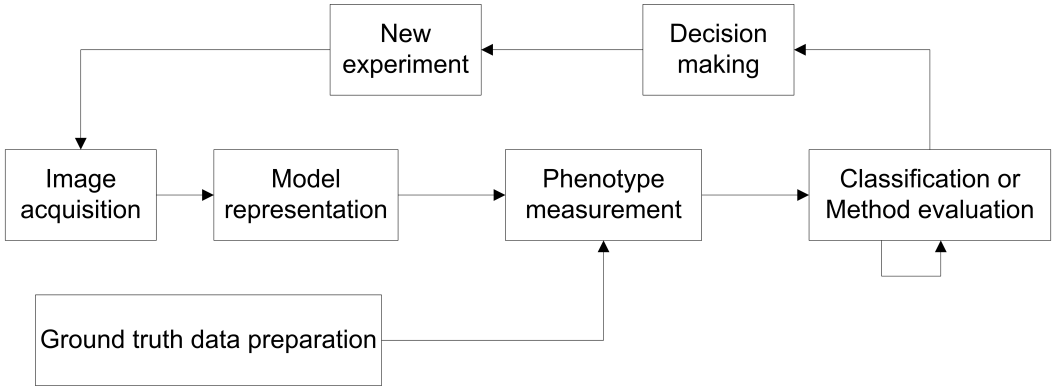

Figure 7.1. General workflow.

This thesis presents a number of studies in biological image dataset analysis both in 2D and 3D space. The general workflow is shown in Figure 7.1. In this thesis the recurring themes are pivotal to image analysis of large volumes of data. The studies are grouped on the themes of feature selection and ground truth in datasets. These studies enlighten us the general routing in coping with different datasets in N-Dimensional space and extracting related and interesting information from the original image dataset. They also have lead us to explore the specific ways for each image dataset how to pass this route. With all these experiences, we summarize the important points. First is the phenotype measurement. The measurements should meet the biological description and should be prominent to distinguish between different categories. We do not focus on the quality of

features but the quality of a few features. Therefore, the feature sets and feature extraction methods differ with different biological dataset. We should always be conscious to find the measureable and objective features during our future work. Ground truth data (training data) preparation is crucial for supervised classifier training and method evaluation. The amount of the ground truth data (training data) should be big enough and it should be relative to the complexity of the "true" function. According to our experience, for each category or each parameter, 50 ground truth samples is the lower bound for the classification validation and method evaluation. The reality is that ground truth data are always sparse. Therefore the question to set a proper proportion from limited training data should be a part of future work. 3D image stack representation and analysis system could also be improved in many aspects such as reconstruction method, contour interpolation method as well as centerline extraction method. Here we have presented a good start but this can be further elaborated in our future work.