A methodology to analyse high-throughput screening data using regression

Szkalisity Ábel, I. évf. programtervező informatikus MSc

Témavezető: Dr. Horváth Péter

MTA SZBK Szintetikus és Rendszerbiológiai Egység

We live in the era of big data. Biological research is not an exception. Since the appearance of high-throughput screening microscopes biologists can acquire hundreds of thousands of images in a short time. The term high-throughput screening incorporates different components of this big data analysis starting from lab automation through cutting-edge microscopy to image analysis and machine learning software. For Computer Science people the latter fields of these pipelines are the most interesting. How can we analyse reliably extremely large amount of images and provide statistics that can support or reject biological hypotheses?

The most widespread approach to solve this problem is combining image analysis and machine learning, more precisely supervised classification. Although this is commonly used, annotation – which is crucial in supervised learning – often becomes troublesome because of its cost and the difficulty of labeling uncertain cases. In this work, we suggest to use regression instead of classification as a better platform to analyse biological processes especially those that does not result discrete cell types but are part of continuous process. The regression toolbox that is presented here as a submodule of the Advanced Cell Classifier software (developed by the Host group) enables field experts to virtually insert cells in a continuous space called the *regression plane*. Using *Gaussian Processes* which is based on Bayesian modelling, we introduce a novel active learning method and show that it decreases the required number of training samples to achieve a given level of prediction error. Using various regression techniques, the toolbox is also capable of predicting the location of any unknown sample on the regression plane and provide statistical output of cell populations to the field expert. Whilst the interpretation of supervised classification results is straightforward it is very difficult to understand the meaning of predictions in a continuous space. We provide a method to visually and quantitatively compare results of the regression analyses by using nonlinear dimensionality reduction methods.

Our methodology is evaluated on artificial and real biological samples. Synthetic results largely outperform random and human annotation-based ones. Results on real data show that the methodology is capable of retrieving information on drug treated cell cultures.