High Content Screening Analysis in R

Insa Winzenborg*, Pierre Ilouga

Discovery Informatics and Statistics, Evotec, Hamburg
*Contact author: insa.winzenborg@evotec.com

Keywords: High Content Screening, Multivariate Analysis, Linear Discriminant Analysis

High content screening (HCS) is a drug discovery method which is used to identify chemical substances (e.g. small molecules) that change the cell phenotype in a desired manner for a certain disease indication. Automated microscopy systems allow running these high content assays in a high throughput (up to 5,000 - 10,000 substances per day). The images acquired by automated microscopy are analyzed using image processing software, which extracts biologically meaningful parameters like the number of cells, cell roundness, cell area or fluorescence intensity. Many dozens of parameters could be obtained as a result e.g. on cell morphology, intracellular structures, and activation of intracellular signaling cascades.

With growing complexity of assays, it often happens that none of the extracted parameters is capable of reliably discriminating between biologically inactive and active substances. In this case, it is essential to combine several parameters into a robust multi-parameter readout that is straightforward to compute and at the same time easy to interpret. The linear discriminant analysis (LDA) is used to address this objective. Linear combinations of selected parameters are derived and lead usually to much better separations of the active and inactive control groups than the best parameter alone. In order to derive the multi-parameter readout, test plates are analyzed that only contain known inactive and active substances, i.e. negative and positive controls. They are randomly divided into a training and a test group. Linear combinations of different numbers of parameters are calculated on the training group (via exhaustive search or forward selection) and finally evaluated based on the test group. This process is repeated several times (internal cross validation). The quality of the obtained dimensionless multi-parameter readout is assessed by means of the so-called (multivariate) Z' factor (Kuemmel et al. 2010), which is a quality measure that incorporates the absolute difference of means and variability of negative and positive control groups.

However, there may be challenges in practical applications of this method. As an HCS campaign usually runs over several weeks it is desirable, at least for consistency and interpretability purposes, to apply the derived linear combination to all data generated throughout the campaign, even if uncontrollable factors lead to some measurement variation over time. For this reason, it is crucial to derive a robust readout combination, i.e. a combination that results in good Z' factors and therefore in a good separation of inactive and active substances in independent validations on several plates in several days (external cross validation).

The data analysis procedure, i.e. how many and which parameters are needed to obtain the best separation and what improvement of the Z' factor is achieved, as well as the graphical representation that shows the importance of parameters in the respective combinations was implemented in *R*.

The talk will present this HCS analysis method and its application to a real screening scenario.

References

Kuemmel, A., Gubler, H., Gehin, P., Beibel, M., Gabriel, D. & Parker, C. (2010). Integration of Multiple Readouts into the Z' factor for Assay Quality Assessment. *J Biomol Screen*, 1, 95–101.