BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 14 2012, pages 1945–1946
doi:10.1093/bioinformatics/bts288

Data and text mining Advance Access publication May 13, 2012

HCS-Analyzer: open source software for high-content screening data correction and analysis

Arnaud Ogier and Thierry Dorval*

Cellular Differentiation and Toxicity Prediction, Institut Pasteur Korea 696 Sampyeong-dong, Bundang-gu, Seongnam-si, Gyeonggi-do, 463-400 Korea

Associate Editor: Martin Bishop

ABSTRACT

Motivation: High-throughput screening is a powerful technology principally used by pharmaceutical industries allowing the identification of molecules of interest within large libraries. Originally target based, cellular assays provide a way to test compounds (or other biological material such as small interfering RNA) in a more physiologically realistic *in vitro* environment. High-content screening (HCS) platforms are now available at lower cost, giving the opportunity for universities or research institutes to access those technologies for research purposes. However, the amount of information extracted from each experiment is multiplexed and hence difficult to handle. In such context, there is an important need for an easy-to-use, but still powerful software able to manage multidimensional screening data by performing adapted quality control and classification. HCS-analyzer includes: a user-friendly interface specifically dedicated to HCS readouts, an automated approach to identify systematic errors potentially occurring during screening and a set of tools to classify, cluster and identify phenotypes of interest among large and multivariate data.

Availability: The application, the C# .Net source code, as well as detailed documentation, are freely available at the following URL: http://hcs-analyzer.ip-korea.org.

Contact: dorvalt@ip-korea.org

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on March 6, 2012; revised on May 8, 2012; accepted on May 9, 2012

1 INTRODUCTION

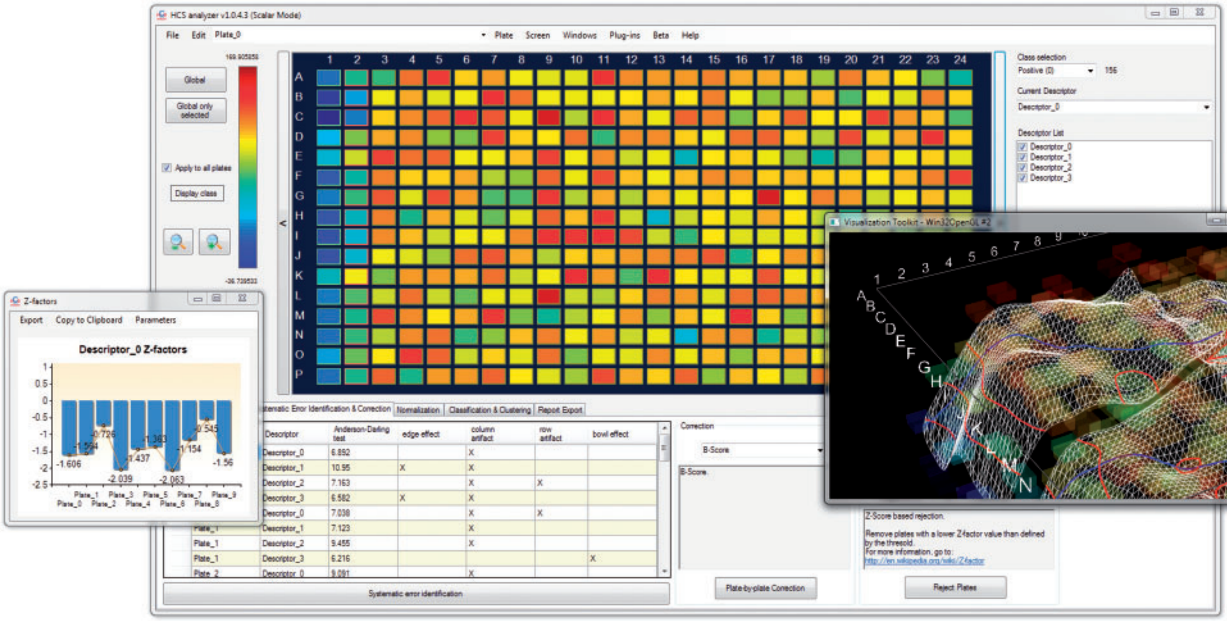
High-content screening (HCS) can be seen as an extension of high-throughput screening when multiplexing readout. This technology usually relies on complex biological systems such as cells. Combined with novel automated imaging platforms, this technology currently allows acquisition of a huge amount of experimental data. Followed by image analysis, this approach performs the extraction of high-dimensional signatures called phenotypes. Here we propose an open source front end interface dedicated to analyze high-dimensional data generated during HCS campaigns. We embedded Weka [Hall *et al.* (2009)], Alglib (http://www.alglib.net/) and Accord (http://accord-net.origo.ethz.ch) open source scientific libraries and created a link to the KEGG database soap server (http://www.kegg.jp/kegg). In a very easy and efficient way, the users can import screening datasets, visualize them, evaluate and correct potential biases (edge effect, dispensing effect, etc.), cluster and classify the data, all in a supervised or unsupervised mode. In the case of siRNA screening, gene and pathway information can be extracted, as well as pathways frequency. In addition, one can export the results and generate a report including tables, graphs and annotated images. Thus, our application is well suited for analysis of complex HCS data. This method can be used to further develop high-level plugins, taking advantage of statistical and mathematical toolboxes which can be shared within the community.

2 FEATURES

HCS-Analyzer has been designed to follow a regular screening analysis pipeline. The software provides six distinct steps: load and display, data quality control, dimensionality reduction, data normalization, classification or clustering, and lastly, export results (Fig. 1). *Importing data* loads commonly used file formats such as comma-separated values .csv, .mtr [Makarenkov *et al.* (2006)], and regular .txt files (Supplementary Figs. S1–S3). The users can select different parameters they want to load and visualize. *Current Plate* displays the selected array of data. In this mode, the users can design the assay by selecting controls and defining classes (e.g. for a supervised classification). *Dimensionality Reduction* performs feature selection. Two approaches are available: supervised, where the reduced dimensions are those separating the most efficiently defined classes (depending on criteria); or unsupervised, where only one class is required. In the context of phenotypic screening, the feature selection represents not only a pre-process for classification but may also lead to a better understanding of the biological mechanism(s) of action involved. *Systematic Error Identification and Correction* achieves an automated quality evaluation on the active plates. A survey of existing methods can be found in Dragiev *et al.* (2011). We implemented here a novel approach to automatically identify systematic errors each plate is potentially subjected to. The algorithm is split into two distinct steps. The first one performs an automated uni-dimensional K-Means clustering based on each descriptor readout.

For this purpose, a signature related to the plate geometry is associated to each well previously clustered. Typically, this signature includes distance to the edges, distance to the center, column index and row index. In a normal case, those parameters should not be influenced by any of the clustering methods. Finally, a C4.5 classifier learning step is performed on those data. If successful, the tree is automatically analyzed to provide the user a comprehensive feedback identifying the systematic error. Two correction methods

*To whom correspondence should be addressed.



	Descriptor	Anderson-Darling test	edge effect	column artifact	row artifact	bowl effect
	Descriptor_0	6.892		X		
	Descriptor_1	10.95	X	X		
	Descriptor_2	7.163		X	X	
	Descriptor_3	6.582	X	X		
	Descriptor_0	7.038		X	X	
Plate_1	Descriptor_1	7.123		X		
Plate_1	Descriptor_2	9.455		X		
Plate_1	Descriptor_3	6.216				X
Plate_2	Descriptor_0	9.091		X		

Z-Score based rejection.

Remove plates with a lower Z-factor value than defined by the thresold.
For more information, go to:
http://en.wikipedia.org/wiki/Z-factor

Fig. 1. Schematic illustration of HCS-Analyzer features. HCS-Analyzer provides a user-friendly way for designing screening plates and for applying iteratively all the regular screening processes. For siRNA screening, the application is also linked to the KEGG database, allowing automated pathway analysis. A special effort has been made when designing the import and export controls to make these steps as flexible as possible. Graphs allow the assessment of the screening quality such as Z-factors evolution during the screening. A new algorithm has been developed to automatically identify systematic errors. After phenotype identification, the user can check the readout distribution using various scatter points or graphs

are proposed, the B-score [Brideau *et al.* (2003)] and the diffusion-based model [Carralot *et al.* (2012)]. Plates can also be rejected based on Z-score in one dimension.

Normalization normalizes the readouts of the different plates to reach a data consistency required for hit identification. This step is not mandatory with some classification approaches, typically when the learning step is performed plate by plate.

Classification and Clustering automatically classifies and identifies the phenotype(s) of interest in a multivariate way. The classification can be operated in a supervised (Support vector machine, Neural network, K-nearest neighbors, random forest) or unsupervised manner (K-Means, Expectation maximization, Hierarchical). The clustering and classification automatically updates the classes, and if a tree-based algorithm has been chosen, the graph can be displayed and exported.

Report Export generates a complete report of most of the performed operations as well as a description of the screening properties.

Finally, the application also includes a module to generate artificial realistic multidimensional screening data [Kwan and Birmingham (2010)], allowing the validation of additionally developed algorithms.

3 CONCLUSION

The stand-alone software presented here aims to simplify the process of correcting and analyzing multivariate screening datasets. The full C# source code is available as well as a plugin sample (Supplementary Fig. S4), allowing the user to adapt or to extend the currently provided version of the application. A tutorial as well as visual documentation can be downloaded at the same URL. Further development of the application is expected to adapt the software to evolving community requirements.

ACKNOWLEDGEMENTS

We gratefully acknowledge Ministry Of Education, Science And Technology (MEST) Korea, Gyonggido and KISTI. We wish to thank H. K. Moon for fruitful discussions and technical support, and Dr V. Makarenkov and Dr P. Dragiev for the B-score implementation.

Funding: This work was supported by the National Research foundation of Korea (NRF) grant funded by the Korea government (MEST) [2011-00244] and [2011-0019430 to T.D., in part], Gyeonggi-do and KISTI.

Conflict of Interest: none declared.

REFERENCES

Brideau,C. *et al.* (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.*, **8**, 634–647.

Carralot,J.-P. *et al.* (2012) A novel specific edge effect correction method for RNA interference screenings. *Bioinformatics*, **28**, 261–268.

Dragiev,P. *et al.* (2011) Systematic error detection in experimental high-throughput screening. *BMC Bioinformatics*, **12**, 25.

Hall,M. *et al.* (2009) The weka data mining software: An update; sigkdd explorations. *SIGKDD Explorations*, **11**, 10–18.

Kwan,P. and Birmingham,A. (2010) Noisemaker: simulated screens for statistical assessment. *Bioinformatics*, **26**, 2484–2485.

Makarenkov,V. *et al.* (2006) Hts-corrector: software for the statistical analysis and correction of experimental high-throughput screening data. *Bioinformatics*, **22**, 1408–1409.