NONPARAMETRIC MULTIVARIATE COMPARISON OF HCS DATA

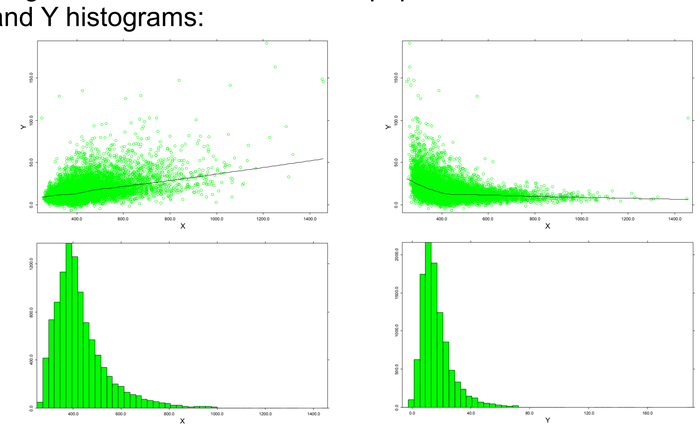
The Chi-Square Works, Inc. (http://chi-square-works.com)

Abstract

High content screening (HCS) data sets are multivariate in nature. Histograms and the Kolmogorov-Smirnov test are among the methods most frequently used to analyze such data sets. However, these 2 methods are univariate in nature because they are based on the marginal distribution of a single variable. When comparing 2 samples of measurements of variables, for each of the variables to have the same marginal distribution does not necessarily mean these 2 samples are from the same population. All the variables have to be considered jointly to effectively use HCS data for any two-sample tests. Using HCS data from dose-response experiments, this poster demonstrates a few nonparametric multivariate methods based on minimal spanning tree techniques. The analysis results can be presented graphically via a "focusing-and-linking" approach to better uncover and describe any differences.

Introduction

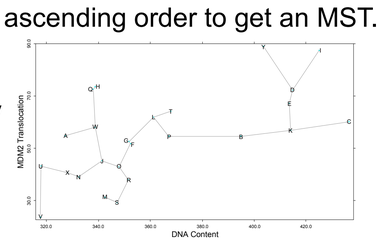
- HCS data are inherently multivariate: Hundreds to thousands of cells in each well of microplates are imaged in multiple fluorescent channels; tens or hundreds parameters are reported for each cell.
- Histograms and Kolmogorov-Smirnov (KS) tests are frequently used to compare HCS (and flow cytometry) data.
- These methods are based on the marginal distribution of a SINGLE variable ONLY and do not take relationships between variables into account. Quite likely important information is not revealed as a result.
- When comparing 2 samples of multivariate data, similar-looking histograms (hence, nonsignificant KS statistics) for each of the variables do not necessarily imply the same population. The following data come from 2 different populations but have the same X and Y histograms:



- We should examine the JOINT distributions of HCS variables both ANALYTICALLY and GRAPHICALLY. These can be achieved with advanced statistical techniques such as those based on minimal spanning trees.

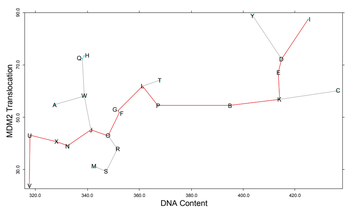
Minimal Spanning Trees

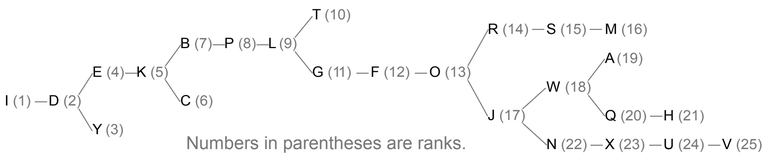
- A graph consists of a set of nodes and a set of node pairs called edges.
- A path between 2 prescribed nodes is a sequence of nodes with the prescribed nodes as first and last elements, all other nodes distinct. a path of m nodes has m - 1 edges.
- A cycle is a path beginning and ending with the same node.
- A spanning tree is a graph that connects all the nodes together with acyclic paths.
- Assign a weight to each edge of a spanning tree. A minimal spanning tree (MST) is a spanning tree for which the sum of edge weights is a minimum.
- The Euclidean distance between 2 nodes is commonly used as the weight for the edge defined by the 2 nodes.
- HCS context: N cells identified by an HCS reader with p measurements taken on each cell ⟶ N nodes in a p-dimensional space (R^p).
- 1-D example: Sort the numbers in ascending order to get an MST.
- 2-D example: 25 cells with 2 variables each. The location of each cell in this 2-D space is marked by an alphabet. There are 24 edges in the depicted MST.



Multivariate Generalization of the KS Test

- Traditional KS tests sort the pooled univariate observations in ascending order and compute the test statistic based on the ranks of observations in the sorted list.
- Need to order multivariate observations in such a way that a strong relationship between the absolute difference in ranks between pairs of observations and their distance in the observation space is maintained.
- MST's tend to connect points that are close and can be used to rank multivariate observations:
  1. Pick one end of a path that has the maximum number of edges. For example, the I node in the red path.
  2. Root the MST at the selected node in Step 1.



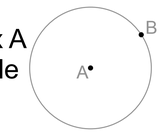


Numbers in parentheses are ranks.

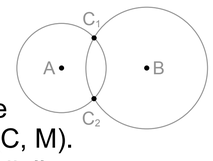
  3. Recursively, visit the root first and then visit its subtrees in ascending order of their heights. The height of a rooted tree is the maximum number of edges between the root and any node in the tree. For example, the height of the subtree rooted at node J is 4. The first visited node has rank 1, the second 2, ..., etc.
- Apply the standard univariate KS test to the resulting ranks.
- Multivariate P-P plots can be constructed with these ranks to explore the nature of the difference between the 2 samples.

MST Planing

- Basic idea:
  - Given 2 points, A and B, in a high-dimensional space (R^p). To map A and B onto a 2-D plane, fix A anywhere first and B can be any point on the circle centered at A with a radius equal to d_p(A, B), the distance between A and B in R^p.



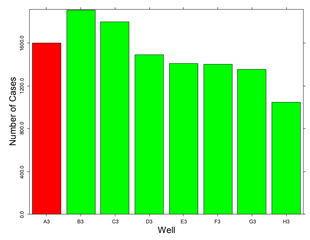
  - Given 3 point, A, B, and C, in R^p. C can be mapped to either C_1 or C_2. If there exists a point, M, already mapped, this ambiguity can be resolved by picking the one minimizing the absolute difference between d_2(C, M) and d_p(C, M).



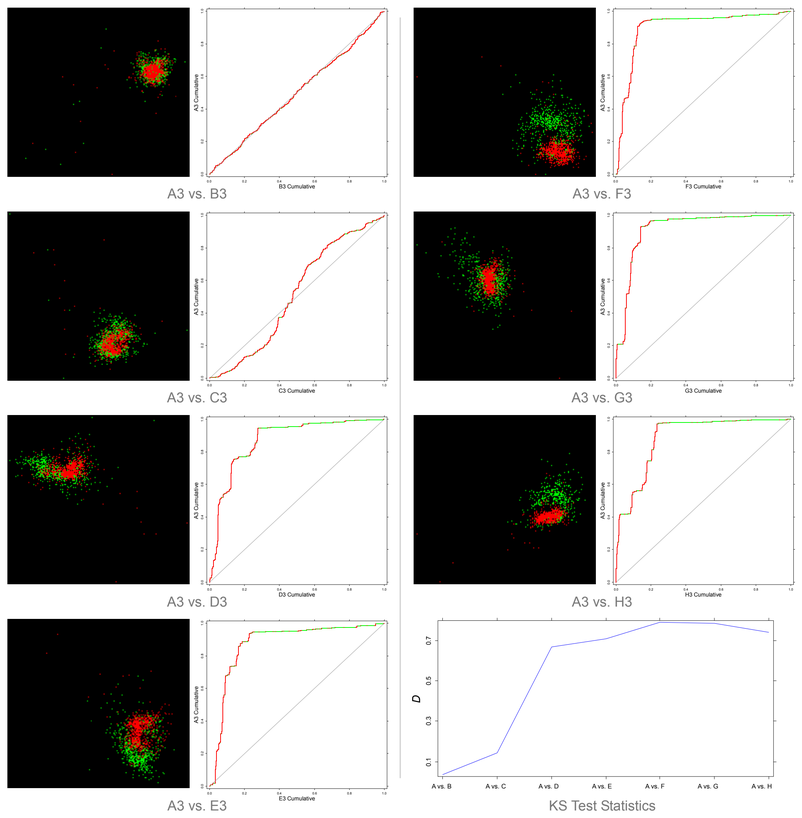
  - Starting with the MST center and mapping radially outward with increasing depth under the constraint that all MST edge lengths and distances from each node to its sister node farthest from their parent node are preserved.
- The resulting configuration of points in a 2-D plane is intended to reflect the interpoint distance relationships in the original R^p.

Example 1: Etoposide Dose Response of U-2 OS Cells

- Comparing the effects of etoposide on U-2 OS cells.
- Cellular targets monitored: DNA, pRb, and p53.
- No etoposide in well A3. Concentrations of etoposide increase with a common ratio of 3 from well B3 to well H3.



- The joint distribution of 8 variables from each of the 7 "green" wells is compared with that from the red well (A3) to test for any concentration effect.
- The 8 variables:
  - DNA stain intensity, nuclear area
  - pRb & p53: cytoplasmic intensity, nuclear intensity, and cytoplasmic area.
- Results:

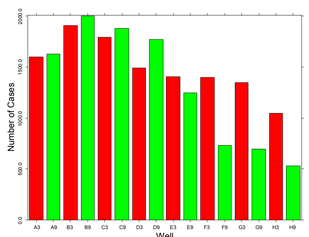


  - MST planing plots are those with a black background. There is a multivariate P-P plot to the right of each MST planing plot.
  - Cells from well A3 are red; cell from other wells are green. Note that red cells and green cells gradually separate out in R^8 as etoposide concentrations increase.
  - The P-P plot in the A3-vs.-B3 panel is typical of nearly identical samples.
  - The P-P plot in the A3-vs.-C3 panel is typical of samples that differ in scale. The MST planing plot also hints a scale difference.
  - The rest 5 P-P plots all suggest strong location differences, which are easy to see in corresponding MST planing plots.
  - P-values of the multivariate KS tests are 0.16 for the A3-vs.-B3 comparison and 0 for the rest.
  - Etoposide has no significant effect in low concentration (well B3). Its effect probably peaks out at the concentration level in well F3, as indicated in the profile plot of the KS test statistics. The same conclusion can probably be reached by examining the MST planing plot, too.

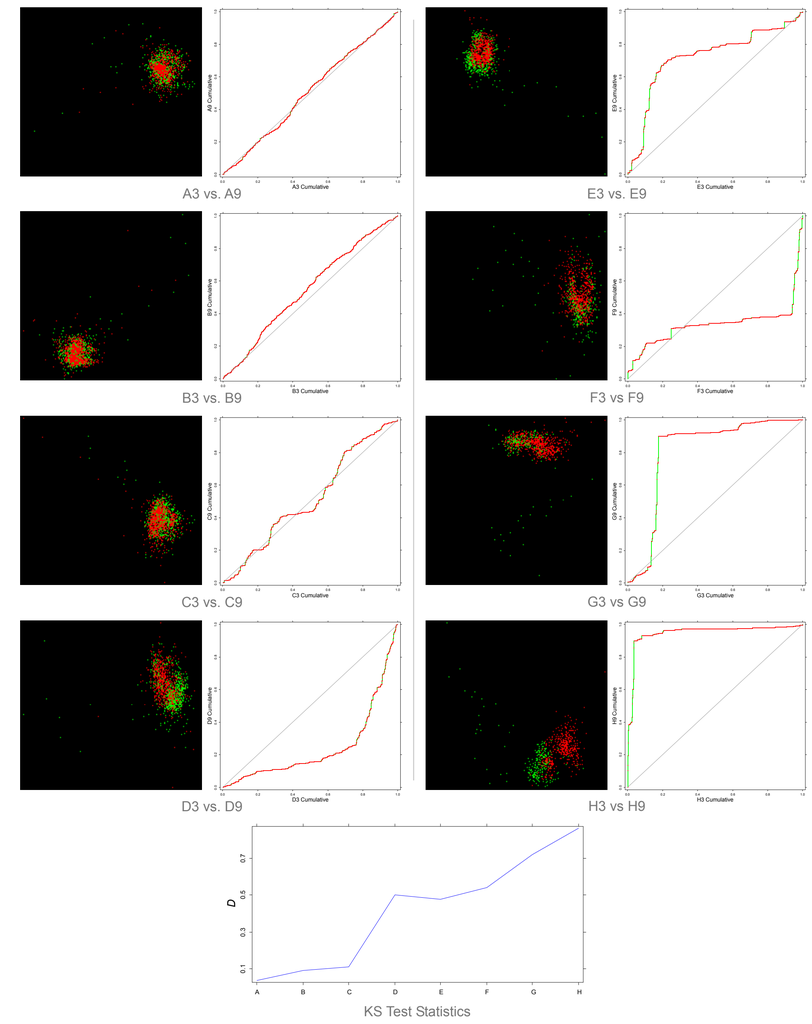
Example 2: Comparing the Effects of Etoposide and Vinblastin on U-2 OS Cells

- Cellular targets monitored: DNA, pRb, and p53.
- Plate layout:

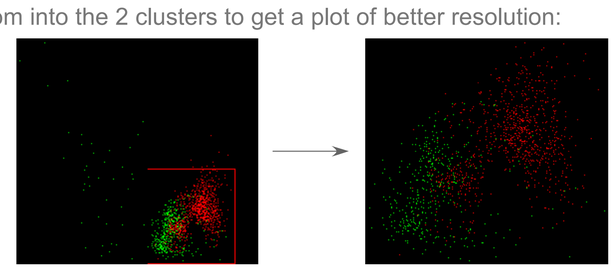
	3	9
A	[etoposide] = 0	[vinblastin] = 0
B	[etoposide] = c	[vinblastin] = c
C	[etoposide] = c·3	[vinblastin] = c·3
D	[etoposide] = c·3²	[vinblastin] = c·3²
E	[etoposide] = c·3³	[vinblastin] = c·3³
F	[etoposide] = c·3⁴	[vinblastin] = c·3⁴
G	[etoposide] = c·3⁵	[vinblastin] = c·3⁵
H	[etoposide] = c·3⁶	[vinblastin] = c·3⁵



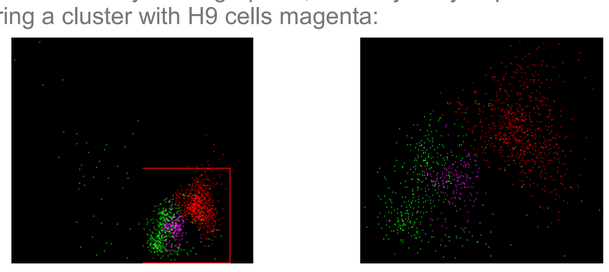
- The joint distributions of the same set of 8 variables as those in Example 1 are compared for each rows of wells.
- Results:



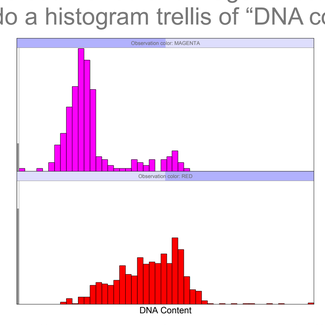
  - P-values: 0.22 for the A3-vs.-A9 comparison, 2.4×10^{-7} for the B3-vs.-B9 comparison, and zero for the rest.
  - As expected, no difference between the control wells A3 and A9.
  - There is a sharp increase in *D*, the KS test statistic, between row C and row D.
  - Within the range of concentrations in this experiment, the general trend is for differences between drug effects to increase as drug concentrations increase. However, there are 2 ranges of concentrations within which differences remain roughly constant.
- With our unique focusing-and-linking graphics implementation, information is just a few clicks away. For example, there are 2 clusters in the above H3-vs.-H9 panel. One cluster contains almost exclusively H3 cell; the other one contains both H3 and H9 cells. How do H3 cells in these 2 clusters differ?
  - Zoom into the 2 clusters to get a plot of better resolution:



  - With the aid of dynamic graphics, it is very easy to paint those H3 cells sharing a cluster with H9 cells magenta:



  - Just one click, take out the red and magenta cells from either of the above plots and do a histogram trellis of "DNA content":



Dramatically different DNA profiles; indicative of effects on cell cycle progression

Summary

- HCS data are inherently multivariate.
- Analyzing multivariate data using methods univariate in nature (histograms, the KS test) runs the risk of missing important content of high-content screening data sets.
- Nonparametric methods are required to properly decipher HCS data sets.
- The minimal spanning tree is a versatile tool:
  - It can generalize the traditional univariate KS test to handle multivariate data. For screening, the multivariate KS test provides an objective way (p-values or D's) to compare 2 HCS samples; no more need to squint at a bunch of heat maps.
  - In addition to the KS test, the Wald-Wolfowitz runs test can also be generalized by MST.
  - Data points in R^p can also be mapped onto a 2-D space by MST planing for visualization.
  - MST planing and the multivariate P-P plot can help describe how 2 samples differ from each other.
- All data analysis and plots in this poster were done with Panmo, a dynamic graphics system for exploring HCS data.