pvcaPreprocessor Documentation

June 30, 2011

Description The tool that converts gene expression data into a format usable by pvcaCore.

Author Adam B. Norberg (ISB); anorberg@systemsbiology.org

1 Summary

While the PVCA algorithm itself, as a mathematical algorithm based in linear algebra, is appropriate to implement in R, the task of loading and filtering the input file is not. R's documentation explicitly states that R is bad at this, and its data structures simply don't support the efficient record manipulation necessary to handle divergent table formats in any reasonable amount of time or code.

This is a Java tool that takes gene expression or methylation data in the formats that exist in level_3 files in the DCC, filters the number of features down to a level that can be processed reasonably, and outputs the remaining records, transposed, in a single output file format optimized for the task at hand.

Features with the greatest post-normalization median absolute deviation *of the variance* are retained. This prefers a wide "splay" of data and thus selects features that are most informative as to batch effects. Experimentation suggests that as little as 12% of the features tends to provide PVCA results nearly identical to the results for the entire data set, so 15% is chosen as a default.

If computational power is at a fixed premium, `pvcaPreprocessor` can be configured to output a fixed number of records instead of a proportion

of the input data. Specifying a `keeprate` greater than 1 will cause it to be interpreted as a quantity rather than a fraction.

2 Parameters

Input File A tab-delimited gene expression or methylation data file, with exactly one or two header rows or columns. Features must be along the rows, with TCGA samples along the columns. The top column header must be the TCGA sample barcode. Methylation data has a great deal of special code handling it and must be considered highly fragile. Due to a lack of formal data standards, parsing is on a best-effort basis.

Keep Rate Either a floating-point value between 0 and 1, representing the proportion of features to maintain in the data set, or an integer 2 or greater, representing the exact number of features to maintain. The specified fraction or number of features is selected based on which features show the greatest median absolute deviation from the mean.

Output File name of the output. Should generally be specified in a fixed manner when this is used in a pipeline; the recommended extension is "`.PTD`", short for "PVCA tabular data". It is a TSV format, but it is not designed to be useful to other tools. (That said, it may well be.)

3 Output Files

(specified output file) A tab-delimited representation of the data, retaining only the fraction or count of features specified by the keep rate. The data will have been transposed, with TCGA barcodes down the first column and feature names forming the first row.

4 Platform Dependencies

Java Compiled against Java 6 or later. If strictly necessary, it might be possible to recompile this to Java 5; versions of Java before that would require extensive rewrites.