Biostatistics and Biometrics
Open Access Journal
ISSN: 2573-2633

Juniper PUBLISHERS key to the Researchers

Research Article
Volume 9 Issue 1 - January 2019
DOI: 10.19080/BBOAJ.2019.09.555752

Biostat Biometrics Open Acc J

Appendix 1

Description of Experimentally Derived Cancer Microarray Datasets

I. **Alon et al. [11]:** The gene expression data from Alon et al. [11] consists of 40 tumor and 22 normal colon tissue samples which were analyzed with an Affymetrix oligonucleotide Hum6000 array. The data are available from the "colonCA" package in Bioconductor. According to the "colonCA" package documentation, "Two thousand out of around 6,500 genes were selected based on the confidence in the measured expression levels (for details refer to publication). No further preprocessing (normalization, etc.) was done." Following the processing method of Boulesteix and Strobl [9], we removed nine duplicate genes which were "tiled a number of times on the chip, with different choices of feature sequences". The final dataset to be analyzed contains 1991 features.

II. **Golub et al. [12]:** In the study by Golub et al. [12], gene expression data on 6,817 features from 38 leukemia patient samples were obtained from Affymetrix high-density oligonucleotide arrays. Three step pre-processing was done as described by Dudoit et al.

a) Thresholding, floor of 100 and ceiling of 16,000;

b) Filtering, exclusion of genes with max/min < 5 or (max - min) < 500, where max and min refer to the maximum and minimum intensities for a particular gene across the 72 mRNA samples; and

c) Base 10 logarithmic transformation".

The final dataset to be analyzed is found in the "CMA" package in Bioconductor. This dataset has 3,051 genes from 27 acute lymphoblastic leukemia (ALL) cases and 11 acute myeloid leukemia (AML) cases.

III. **Khan et al. [13]:** The "khan" data in Bioconductor's CMA library contains cDNA microarray gene-expression data from 2,308 genes. These 2,308 genes were obtained by filtering the initial 6,567 genes "by requiring that a gene should have red intensity greater than 20 across all experiments". The Khan dataset contains 63 samples from four classes of small, round blue-cell tumors (SRBCTs). There were 23 samples from the Ewing family of tumors, 20 rhabdomyosarcoma samples, 12 non-Hodgkin lymphoma samples and 8 Burkitt lymphoma samples.

IV. **Singh et al. [14]:** For the Singh et al. 2002 study, Affymetrix (U95Av2 array) gene expression data on 12,600 genes were measured from 52 prostate cancer tumor samples and 50 normal prostate tissue samples. So as to render our analyses comparable to those of Boulesteix and Strobl [9], we analyzed data obtained from A-L Boulesteix which was said to be processed "as described in Singh et al. [14]". According to Singh et al. [14], "Raw expression values were normalized to the median array intensity and thresholds were set at 10 and 16,000 units." Singh et al. [14] describes the non-specific gene filtering used to obtain the final dataset analyzed in the present study which measured the expression levels of 5,908 genes.

V. **Sültmann et al. [15]:** The samples were hybridized to two-color cDNA arrays containing 4,224 genes. A reference sample obtained from pooling numerous renal cell carcinoma samples was labeled with the red dye on all arrays. The data were normalized using the "vsn" method (Huber et al. 2002) with the generalized log-ratio values quantifying the difference between the tumor samples and the reference sample. This data is found in the "kidpack" package in Bioconductor. The Sültmann et al. [15] dataset contains gene expression data on 74 renal cell carcinoma samples from three tumor classes (52 clear cell, 13 papillary and 9 chromophobe samples).

VI. **Alizadeh et al. [16]:** This study investigated the gene expression profiles of Diffuse Large B-Cell Lymphoma (DLBCL) subtypes. The Lymphochip microarray used in this study contained 17,856 cDNA clones. The pre-processing and normalization methods are described by Alizadeh et al. [16]. The dataset analyzed contained 4,026 genes. The data were downloaded from the website: http://stat.ethz.ch/~dettling/bagboost.html. This dataset contains 62 samples from three different stages (classes) of B-cell differentiation: 42 diffuse large B-cell lymphoma samples, 9 follicular lymphoma samples and 11 chronic lymphocytic leukemia samples.

VII. **Garber et al. [16]:** This is the "stanford" dataset from the "lungExpression" library in Bioconductor. The original cDNA microarray dataset from Garber et al. [16] contains 24,000 features to profile 67 lung tumors with eight histological patterns (classes). For the Parmigiani et al. comparison study, the original Garber et al. [16] cDNA intensities were log-transformed. The final Garber dataset to be analyzed contains 3,171 genes common to the Bhattacharjee et al. and Beer et al. studies. Five of the tumor classes had five or less observations. These five classes were excluded from the dataset to be analyzed since it would not be possible to use 5-fold CV for classes with fewer than five observations. The final dataset contained 53 samples from three tumor classes: 37 adenocarcinoma, 10 squamous cell and 6 lymph node tumor samples.

VIII. **Pomeroy et al. [18]:** As described by Pomeroy et al. [18], the Affymetrix gene expression data were filtered to exclude genes showing minimal variation across samples, and then each sample's data was normalized to have a mean expression of 0 with a variance of 1. The pre-processed data containing expression levels of 5,597 genes were downloaded from the website: http://stat.ethz.ch/~dettling/bagboost.html. The original dataset contained samples from five different kinds of central nervous system tumors but we restricted our analysis to four of the classes because one of the classes (human cerebella) only had four subjects, so analysis was not possible using 5-fold cross validation since there were fewer observations than folds. Thus, the dataset analyzed contained 38 samples from four classes: 10 medulloblastomas, 10 malignant gliomas, 10 atypical teratoid/rhabdoid tumors and 8 primitive neuro-ectodermal tumors.

IX. **Ramaswamy et al. [19]:** The original data set included 16,063 genes on the Affymetrix HU-6800 microarray platform. There were 64 primary adenocarcinoma tumor samples and 12 metastatic adenocarcinoma tumor samples (for both classes, the primary tumor came from one of a variety of sites, i.e., breast, prostate, lung, colon, uterus or ovary). The data were rescaled and filtered as described by Ramaswamy et al. [19] to yield a dataset with 9,868 genes. This dataset was downloaded from: http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html