Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 3): pp. 3420-3431
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.03.357

Gene Pre-Processing Methodology In Cancer Identification System

Dr. A. SURESH KUMAR, Professor, CSE, Graphic Era Deemed to be University, Dehradun.

Rohit Pandey, School of Computing, Graphic Era Hill University, Dehradun, Uttarakhand, rohitpandey1705it@gmail.com

Abstract

A Novel pre-processing of gene expression data is proposed, whichemphasizes the filtering and normalization steps since these steps determine the set of probes used in the subsequent analyses. There are two parameters that are set during filtering. In the beginning, for every review, a cut-off is established for the recognition p-value, and a sample is assessed as existing if its detection p-value is smaller than the cut-off. In addition, an established limit is established. For a reviewto be incorporated into the dataset, it must contain a certain number of samples. Information can then be normalized after filtering. Several methods are compared,and for the dataset, the normalized information on the original scale produces themost stable results.

I. Introduction

The resulting gene expression pattern datasets are massive tables with numerous columns, one for each predicted experimental condition, and hundreds of rows for each gene or clone included in the DNA array. Typically, a spreadsheet is used to preprocess gene expression patterns, however due to memory constraints, even simple computations might be difficult on a small workstation.

Due to the magnitude of the dataset, visual analysis is not practical, making it challenging to estimate the amount of missing values, establish whether any gene has an excessive number of misplaced values, or search for duplicate genes. Additionally, carrying out these actions on a web server enables you to make use of the server's features and maintain your programme current. A consistent interface for processing the input eliminates potential issues caused by a particular file format and produces a clear, standard file for long-lasting examination with various tools. Based on analysis, they inference can be made that as the information is updated, and with marginally different choices made during the preprocessing stages, the subsets of genes are chosen for the different versions of the dataset vary. To fully comprehend the significance of the preprocessing stages, one needs to observe what choices are being made and how each influences the resulting gene set. A dataset comprises gene expression data for blood cells, a negative control data matrix, and background information on every sample.

In the background information, identifiers of each case and control are givenin the background information. There are three stages for cleaning information:

- Removing probes for genes that require gene expression data.
- Remove case-control pairs when one or both of the pairs are outliers in terms of quality.
- Perform background correction using negative control probes.
- Filter out probes that are not perceptible or adequately present and translate from probes to genes.

It is important to keep this in mind when investigating the gene expression matrix. A microarray-based gene expression measurement would be considerably more valuable if there were consistency and restrictions on specific microarray platforms for specific types of measurements, as well as cross-platform association and normalization.

II. Materials and Methods

A Novel pre-processing methodology is applied on the gene expression data toget the cleaned dataset for the next step is proposed here. To obtain filtered data, the work employs steps to carry out the cleaning process. There are four preprocessing functions: (1) Transformation, (2) Missing Value Removal, (3) Filtering Data, and (4) Proposed Filtering Algorithms. In Figure 1, the pre-processing functions are described. This figure illustrates how the pre-processing layer works. This layerincludes removing null values and unwanted values from the dataset and filtering the gene expression data by removing over-expressed and under-expressed values.

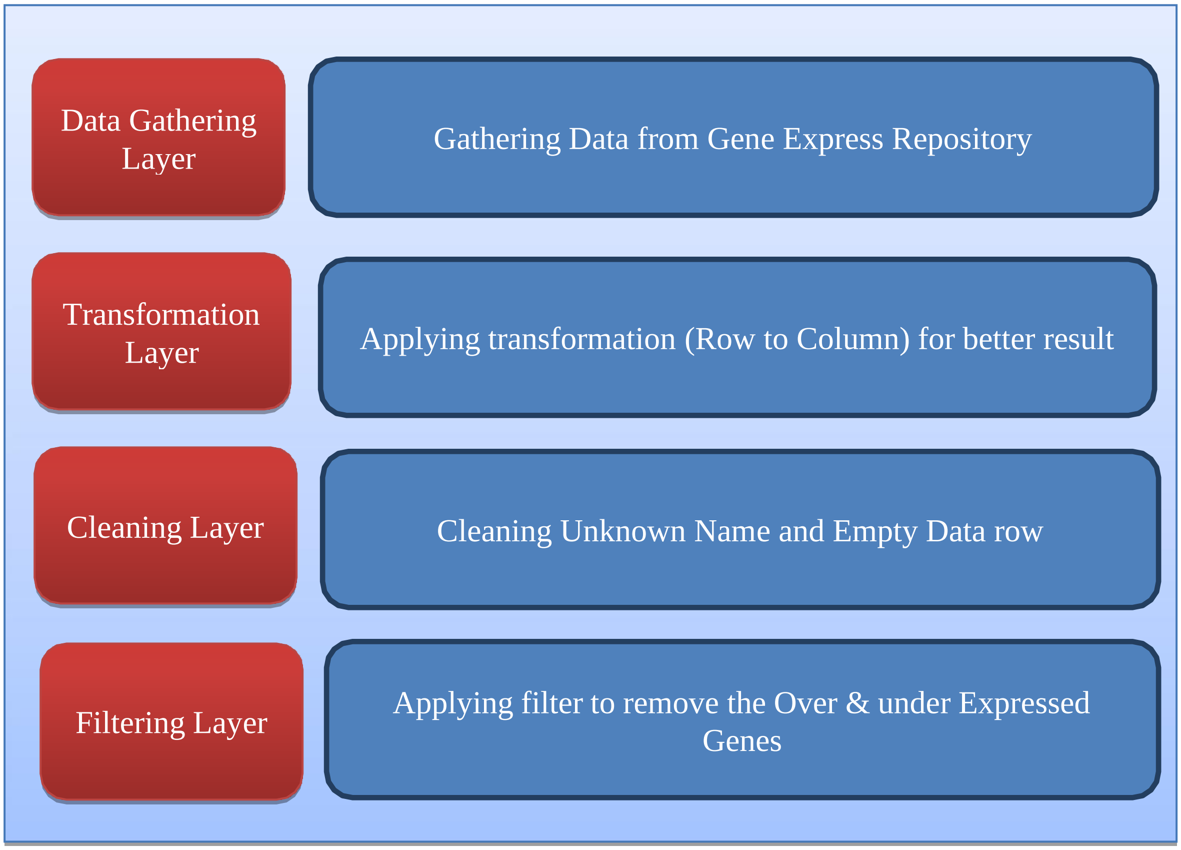
Data Gathering Layer
Gathering Data from Gene Express Repository
Transformation Layer
Applying transformation (Row to Column) for better result
Cleaning Layer
Cleaning Unknown Name and Empty Data row
Filtering Layer
Applying filter to remove the Over & under Expressed Genes

Figure 1: Flow of Pre-Processing Methodology

2.1. Gene Expression Data

Gene expression data provides thousands of genes involved in a biological process. These gene expression data are important in identifying diseases. Five different cancer datasets were obtained in this study to identify diseases. For this framework, lung cancer, cervical cancer, prostate cancer, tongue cancer, and breast cancer are included in the dataset. These datasets are transformed and applied to the pre-processing stages. Table 1 provides a description of the lung cancer dataset.

Detail of the Dataset	Informations / Values
Name of the Dataset	Cigarette smoking effect on lung adenocarcinoma (GDS3257)

Size of the Data	22283 X 109
Total Features (Rows)	22283
Total Samples + ID References + Identifier	107 + 1 + 1
Cancer	58
Normal	49

Table 1: Information about Dataset (For Ex.: Lung Cancer)

The table contains the description regarding lung cancer. The lung cancer dataset is obtained with the size of 22283 X 109 contains total samples of 107 consists of 58 affected data and 49 is normal.

ID_REF	IDENTIFIER	GSM254629	GSM254648	GSM254694	GSM254701	GSM254728
1007_s_at	DDR1	10.9885	10.6919	10.8978	11.7494	10.9028
1053_at	RFC2	6.82603	6.9096	6.80295	6.81802	6.83816
117_at	HSPA6	7.77559	7.68374	7.88498	7.9384	8.01043
121_at	PAX8	9.85506	10.1321	9.8411	9.90026	9.87185
1255_g_at	GUCA1A	4.82396	4.98489	4.87689	4.70904	4.78877
1294_at	UBA7	9.1043	8.99562	9.26155	9.68563	9.03172
1316_at	THRA	6.19335	6.31397	6.2566	6.1893	6.30964
1320_at	PTPN21	6.11913	6.02121	6.07582	6.00636	6.01719
1405_i_at	CCL5	7.7529	8.14121	7.44056	7.86048	8.23439
1431_at	CYP2E1	4.96849	5.13706	4.91496	4.87536	4.86365
1438_at	EPHB3	8.19361	7.76094	8.18528	8.0899	7.80732
1487_at	ESRRA	8.58764	8.34476	8.54456	8.83581	8.24958
1494_f_at	CYP2A6	7.65884	8.01993	7.83222	7.66733	7.83591
1598_g_at	GAS6	10.1701	11.0343	11.1643	11.2073	11.0185
160020_at	MMP14	9.08051	9.31352	9.40717	9.28509	9.68386
1729_at	TRADD	8.78659	9.39552	9.02206	8.94696	8.87095
1773_at	FNTB	7.28541	7.1028	7.22706	7.18666	7.19293
177_at	PLD1	5.90399	5.94174	5.98608	5.80192	5.9449
179_at	DTX2P1-UPK3BP1-PMS2P11	10.2857	10.1377	10.3647	10.2953	10.1259
1861_at	BAD	7.00868	6.79834	7.45689	6.99576	7.1817
200000_s_at	PRPF8	9.96122	10.1488	9.92106	10.1802	9.79175
200001_at	CAPNS1	11.9046	11.7204	11.6531	11.9463	11.9584
200002_at	RPL35	12.0549	12.2	11.7516	12.6799	12.046
200003_s_at	RPL28	12.4906	12.8648	13.2475	13.0116	13.1381
200004_at	EIF4G2	11.8279	11.7049	11.6893	11.4128	11.6058
200005_at	EIF3D	11.0025	10.0671	10.6479	10.4744	10.9013
200006_at	PARK7	11.4926	11.175	11.6577	11.8399	11.8031
200007_at	SRP14	11.8644	12.0154	11.3669	11.9931	12.1085

Table 2: Sample Dataset (For Ex.: Lung Cancer)

The sample dataset for lung cancer is shown in Table 2. The lung cancer dataset contains more than thousand gene expression data, but the table contains a sample set of records from the dataset. The table depicts the IDENTIFIER of the proteins and the reference id (ID_REF) of the proteins contained in the set and also contains gene data.

2.2. Missing Value Removal

Missing data can occur for a number of reasons, such as survey non-response or data entry errors. Even while it would appear that all missing data is the same, this is not the case. There are three broad categories in which missing data might be placed:

- Missing completely at random.
- Missing at random.
- Missing not at random.

No	Type of Missing Value	Description
1	Missing completely at random (MCAR)	When all observations are equally likely to be missing, data is missing completely at random.
2	Missing at random (MAR)	In the case of missing data at random (MAR), the probability of a missing data point is not related to the missing data, but to other observed data.
3	Missing not at random (MNAR)	The likelihood of a missing observation increases with its value when data is missing not at random (MNAR). Missing data values can be difficult to identify since they are unobserved. Data can be distorted as a result.

Table 3: Type of Missing Values

As mentioned above, the different types of missing values is depicted inTable 3. The types of missing values are presented with its description. Anyone or more than one missing values may occurs in the dataset.

2.3. Duplicate Value Detection

The objective of this stage is to identify and fuse duplicate features in Microarray datasets. A goal of this research is to examine whether fusing duplicate features can improve the predictive power of data while reducing training time. Due to the necessity of pairwise comparisons, it is not possible to manually evaluate a Microarray dataset for duplicate features. A small dataset with 100 features would require an expert to assess almost 5K pairs. Therefore, machine-based detection will be used to minimize the domain specialist's efforts. Duplicate values are detected based on the identifier in this work.

Detail of the Dataset	Information / Values
Total Data	**22283**
Missing Value	**51**
Duplicate Value	**6478**

Table 4: Missing and Duplication Values (For Ex.: Lung Cancer)

The details of the dataset which includes total number of data, missing values and duplicate values for lung canceris shown in Table 4. The table shows the total data '22283' included in Lung Cancer dataset and it contains '51' number of missing values and '6478' duplicate values.

2.4. Transformation

It executes the transpose operation on the Cleaned Dataset. A row of data is converted into a column of features, and the columns are converted into rows. It is crucial for each of these processes to transform data in order to shape, standardize, and ensure consistency among different datasets**.**

IDENTIFIER	DDR1	RFC2	HSPA6	PAX8	GUCA1A	PLD1	DTX2P1-UPK3BP1-PMS2P11
GSM254629	10.9885	6.82603	7.77559	9.85506	4.82396	5.90399	10.2857

GSM254648	10.6919	6.9096	7.68374	10.1321	4.98489	5.94174	10.1377
GSM254694	10.8978	6.80295	7.88498	9.8411	4.87689	5.98608	10.3647
GSM254701	11.7494	6.81802	7.9384	9.90026	4.70904	5.80192	10.2953
GSM254728	10.9028	6.83816	8.01043	9.87185	4.78877	5.9449	10.1259

Table 5: Transformed Matrix (For Ex.: Lung Cancer)

The above Table 5 shows the transformed matrix for lung cancer. The table depicts that the dataset is transformed its row to its column that would obtain the filtered gene data.

2.5. Filtering Data

Microarray data filters are used to select a subset of the probes to be included or excluded in analysis. There are two types of filtering, non-specific and specific. First, removing genes with little or no variability is a general procedure. The second approach involves finding genes related to a particular phenotype of interest. Genes that are differentially expressed are identified by filtering. The following table 6 explains how filtering works.

No	Specific Filtering	Non-Specific Filtering
1	Filtering of the probes without regard to a classification or clustering objective	Based on the basis of missing values andthe variation
2	Analyst wants to remove those features which have no chanceof being predictive, regardless of the prediction problem	selection on the basis of level (more than k larger than A; ensuring that the basis for classification corresponds to a reasonable subgroup)

Table 6: Types of filtering

In this work, two approaches for selecting the number of components to expel are examined. One involves the "top three" rule. It is expected that three components with maximum values for the parameter of location, named high-level expressed genes, medium-level expressed genes, and low-level expressed genes, are informative, and this work recollects genes that coincide with these components and other genes are detached.

2.6. Gene Pre-Processor

The process of filtering is performed with this Gene Pre-Processor. The gene pre-processor involves in filtering the gene expression data that contains non-specific values. This would remove under expressed and over expressed data from the dataset. The following algorithm 1 describes the gene pre-processor that filters the data.

```
Algorithm GenePreProcessor()
{
Input: S-Size of the gene, X-Column Matrix, L-Length of the gene
Output: P-Preprocessed Data
{
    ReadGeneSet(D′) from Dataset D
    S=Size (D′) //read the size of the time domain for sample
    //represent x as column-
    vector4. X= X (D′)
    // compute length of the Data D′
    Remove Empty or Null values (X)
    Remove Duplication Value(X)
    Transform the Vector (X)
    L = Record Length(D′)

    //Applying Filtering
    For each i=1 to n from N // no. of genes in the dataset D′
    {
    For each j=1 to l from L
    {
    If (!isNull(n1))
    {
    ReadGene(n1)
    // Check for missing values
    If (missing>=n1)
    {
    If (Min (D′) <=n1 | Max (D′) >=n1) // checks for over expression genes
    {
    T=SelectGene(Top(n1)) // select genes to be top N of the highest maximum values
    }
    }
```

```
        }
        P= FilteredData(T)
        }
        //represent the filtered gene in the form of the original one
        Return P
    }
```

Algorithm 1: Gene Preprocessor

To filter the data, the cleaned gene dataset is obtained. Initially, the Gene Preprocessor computes the size of the dataset sample, then constructs the columnvector with the size and records of the dataset. Gene filtering takes into account the length of the dataset, missing values, duplicate values, and also the rows and columns.Gene data are fetched from the dataset to verify whether they are null or not. In the case that they are not null, the minimum to maximum genes of high correlation to target disease are arranged. The highest values of gene data are removed due to overexpression, and also the low values of genes are removed due to under expression,which would filter the dataset. The final filtered dataset is then used for gene selection.

2.7. Results and Discussion

This section illustrates the pre-processing, results, and discussion performed. The original dataset configuration and its pre-processed configuration with illustrations is given below. Based on the pre-processing steps, this illustration shows the results of filtered gene datasets.

Dataset	Original Size (in MB)	Processed Size (in MB)
CER_CAN	4.4	3.89
BRE_CAN	18.3	17.6
PRO_CAN	17.2	16.5
LUN_CAN	5.31	4.8
TON_CAN	11.6	10.8

Table 7: Performance of Gene preprocessor method with Data Size reduction

The five different cancer dataset includes cervical cancer, breast cancer, prostate cancer, lung cancer and tongue cancer is employed for pre-processing. The original and cleaned dataset size of every dataset is shown in Table 7.

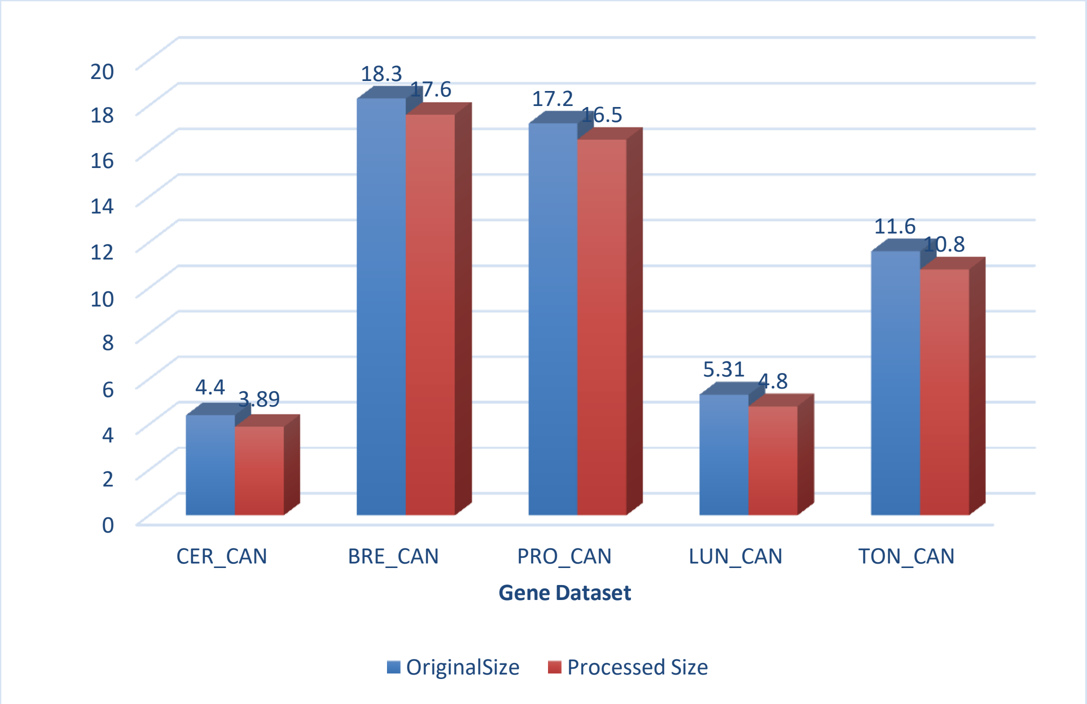

Figure 2: Original vs Pre-Processed Data Size

The illustration of the original and processed data size is shown in Figure 2. Thefigure contains five different cancer datasets and its data size evaluation.

3. Conclusion

This methodology covers the pre-processing of gene expression datasets, which filters the data and synthesizes it into an organized dataset for further analysis. The purpose of gene filtering is analyzed in this chapter. After that, the input datasets are used to perform the steps. Using five cancer datasets (CER_CAN, BRE_CAN, PRO_CAN, LUN_CAN, and TON_CAN), the missing and duplicate data in thedataset is identified and eliminated. In order to analyze overexpressed and under expressed data after the removal, the dataset is transformed. Data is filtered using Gene Preprocessor, which removes the under-expressed and over-expressed genes from the dataset. This will help with the development of optimized frame work for disease prediction model.

4. Reference

[1] Nguyen DV, Arpat AB, Wang N, Carroll RJ. DNA microarray experiments: biological and technological aspects. Biometrics 2002; 58: 701-17.

[2] Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. J Clin Oncol 2002; 20: 1932-41.

[3] Alizadeh AA, Ross DT, Perou CM, Rijn Mvd. Towards a novel classification of human malignancies based on gene expression patterns. J Pathol 2001; 195: 41-52.
[4] Nature-Genetics. The chipping forecast. Vol 21 Supplement 1999: 1-60
[5] Nature-Genetics. The chipping forecast II. Vol 32 Supplement 2002: 461-552.
[6] Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. Trends Genet 2003; 19: 649-59.
[7] Reimers M. Statistical analysis of microarray data. Addict Biol 2005; 10: 2335.
[8] Huber W, v.Heydebreck A, Vingron M. Analysis of microarray gene expression data., In: Balding DJ, Bishop M, Cannings C Eds, Handbook of Statistical Genetics. John Wiley & Sons, Chichester, 2003.