Modelling of an offline and online software for normalization of microarray data of gene expression by Perl, Bioperl and PerlTk and Perl-CGI

Gaurav Kumar Srivastava[1], Dr. Santosh Kumar[2], Dr. Himanshu Pandey[3]

•

[1]Research Scholar, Maharishi University of Information technology, Lucknow
[2]Assosiate Professor, Maharishi University of Information technology, Lucknow
[3]Assistant Professor, BBDNIIT, Lucknow

Abstract

B-Chip Reverence is an online database which isfreely accessible for microarray redundancy removal & normalizationand various data analysis techniques are applied on the data. This software accurately handle the massive amount of data.The growing use of DNA microarrays in biomedical research has led to the proliferation of analysis tools. These software programs address different aspects of analysis (e.g. normalization and clustering within and across individual arrays) as well as extended analysis methods (e.g. clustering, annotation and mining of multiple datasets).After studying all the terms and problems related to Microarray technique, we tried to make an-open and user friendly software to deal with all the problems and to run all the steps of this technique, so that we used Perl & Perl-cgi. perl-cgi stands for *Common Gateway Interface*, is a standard programming interface between Web servers and external programs. perl-cgi executes external programs on the webserver.

I. Introduction

A microarray database is a repository containing microarray gene expression data. The key uses of a microarray database are to store the measurement data, manage a searchable index, and make the data available to other applications for analysis and interpretation. The concept behind this, a microarray is a pattern of ssDNA probes which are immobilized on surface of chip or a slide. The probe sequences are designed and placed on an array in regular patter of spots. The chip or slide is usually made of glass or nylon and is manufactured using technologies developed for silicon computer chips. Each microarray chip is arranged as a checkerboard of 105 or 106 spots or features, each spot containing millions of copies of a unique DNA probe (often 25 nt long).Microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. Even in replicated experiment, some variations are commonly observed. Normalization is the process of removing some sources of variation which affect the measured gene expression levels. In gene expression microarray data analysis, selecting a small number of discriminative genes from thousands of genes is an important problem for accurate classification of diseases or phenotypes.The ability of microarray chip to capture the expressional level of thousands of genes in one snapshot becomes a major attraction for biologists. By performing parallel microarray experiments under different conditions, biologists seek useful information of the underlying biological process that lies in the hundreds of thousands of data points obtained. The first step of task is classifies data through a single step partition. In such task,cluster the genes into biological meaningful groups according to their pattern of expression, based on the assumption that expressional similarity of genes implies their functional similarity. The clustering methods which are used for this include conventional clustering methods (such as k-means clustering, and self-organizing maps).k-means clustering is a simple and a divisive approach. In this method, data are partitioned into k-clusters, which are prespecified at the outset. Self-Organizing Maps is pattern recognition algorithm employs neural networks and based on the machine-learning method.

B chip reverence database dealing the removing microarray redenduncy and normalization, for that using bioperl (kmean and SOM executed by cpan module) dreamweaver8 (for designing web pages on website) photoshop (for logo designing) PERL CGI (for programs to interface with information servers such as HTTP (web servers.)

II. Aim & Object

- To remove duplicates, repetitive and blank genes from our raw data. After removing redundancy, normalize the datasets using the PERL-CGI.
- To make user friendly, open and easily accessible interactive CGI interface for database and various tools for analysis of clustering (k-mean and SOM) for microarray using PERL-CGI algorithms.

III. Materials & Methods

- **Perl & Perl-cgi**

Perl is a programming language developed by Larry Wall, designed for text processing. ThoughPerl is not officially an acronym but many times it is used as it stands for *Practical Extraction and Report Language*. It runs on platforms like Windows, Mac OS, and UNIX.
Perl CGI is the Common Gateway Interface, a standard for programs to interface with information servers such as HTTP (web) servers. CGI allows the HTTP server to run an executable program or script in response to a user request, and generate output on the fly. This allows web developers to create dynamic and interactive web pages. Perl is a very common language for CGI programming as it is largely platform independent and the language's features make it very easy to write powerful applications.

- **Bio Perl**

BioPerl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications. It has played an integral role in the Human Genome Project.BioPerl is an active open source software project supported by the Open Bioinformatics Foundation.

- **Self-Organizing Maps (SOM) and K-Means Clustering (KMC)**

As a machine-learning method, a SOM belongs to the category of neural networks. It provides a technique to visualize the HD input data on an output map of neurons. The map is often presented in a 2D grid of neurons.KMC is a simple and widely used partitioning method for data analysis. It's helpfulness in discovering group of co-expressed genes has been demonstrated.

- **Dreamweaver8 and WampServer**

Dreamweaver8 allows to create professional web pages and also quickly add objects and functionality to pages without having to program the HTMLcode manually.WampServer is a windows web development environment for Apachey, MySQL, PHP databases. It's also virtual server for windows platform, allows it user to manage Website and its components.

IV. Techniques/Databases Used

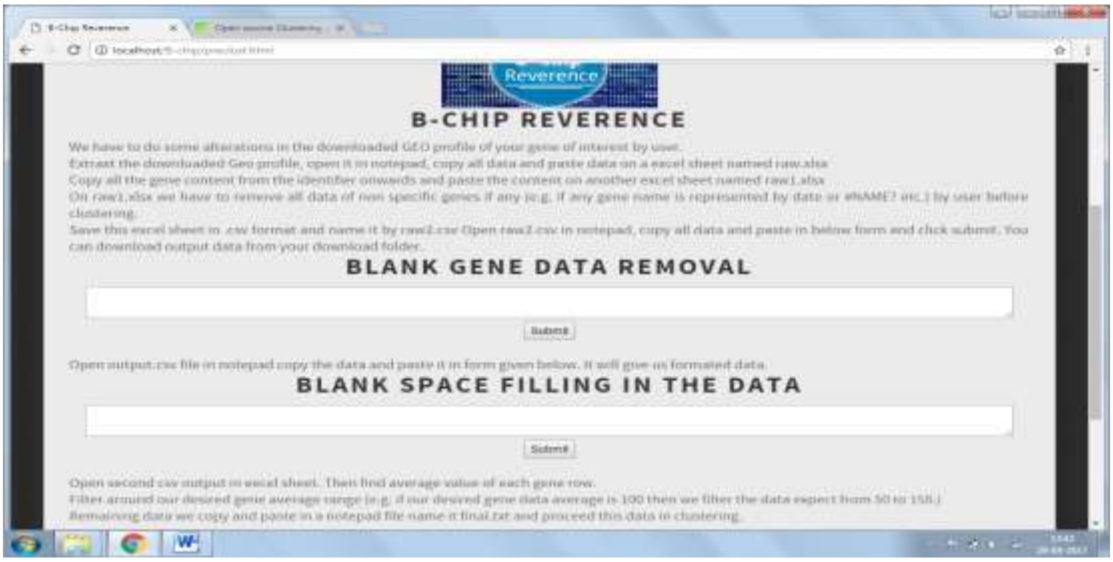
- Westudied information about genes through the method of Gene Ontology by Gene Cards, next we did Microarray data retrieval from NCBI Geo profiles of SMAD7. Data Normalization & Redundancy Removal of Gene Expression do with the help of Microsoft office excel.
- Next we studied Perl elementary and their different algorithms and logics in Perl includes:-
  a) **Regular Expression**
  A regular expression is a string of characters that defines the pattern or patterns you are viewing. The syntax of regular expressions in Perl is very similar to what you will find within other regular expression.There are three regular expression operators within Perl.
    - Substitute Regular Expression - s///
    - Transliterate Regular Expression - tr///
    - Match Regular Expression - m//

  b) **File Handling**
  A filehandle is a named internal Perl structure that associates a physical file with a name. All filehandles are capable of read/write access, so you can read from and update any file or device associated with a filehandle. However, when you associate a filehandle, you can specify the mode in which the filehandle is opened. Three basic file handles are - **STDIN**, **STDOUT**, and **STDERR,** which represent standard input, standard output and standard error devices respectively.

  c) **Sub-Routines**
  A Perl subroutine or function is a group of statements that together performs a task. You can divide up your code into separate subroutines. How you divide up your code among different subroutines is up to you, but logically the division usually is so each function performs a specific task.

  d) **Parsing**
  parsing is the process of analyzing an input sequence (read from a file or a keyboard, for example) in order to determine its grammatical structure with respect to a given formal grammar. It is formally named syntax analysis. A parser is a computer program that carries out this task.
- And Normalization & Redundancy removal (to make data sorted and fine/removal of repeated genes and blank genes).
- Next we did the major/main part of the Microarray technique i.e. clustering: the clustering problem of microarray data only as an analysis to find genes that behave similarly over the experimental conditions. The first generation of clustering techniques includes hierarchical clustering, K-Means clustering and Self-Organizing Maps, we used K-means clustering (KMC): helpful in discovering group of co-expressed genes,& Self-organization map (SOM): provides a technique to visualize the HD input data on an output map of neurons.
- Beside all the steps which we did, we also studied CGI, HTML, Perl toolkit packages etc. in which we used Perl-cgi to develop the online web server B-Chip reverence software for users to perform Microarray Technique computationally.
- We also used BioPerl, we correlated K-Means Clustering and Self-Organizing Maps (SOM) algorithms to perform clustering with Bio Perl. Bio Perl is a collection of Perl modules and it facilitates the development of Perl scripts for bioinformaticsapplications.

V. Result



B-CHIP REVERENCE

We have to do some alterations in the downloaded GEO profile of your gene of interest by user.
Extract the downloaded Geo profile, open it in notepad, copy all data and paste data on a excel sheet named raw.xlsx
Copy all the gene content from the identifier onwards and paste the content on another excel sheet named raw1.xlsx
On raw1.xlsx we have to remove all data of non specific genes if any (e.g. if any gene name is represented by date or #NAME? etc.) by user before clustering.
Save this excel sheet in .csv format and name it by raw2.csv Open raw2.csv in notepad, copy all data and paste in below form and click submit. You can download output data from your download folder.

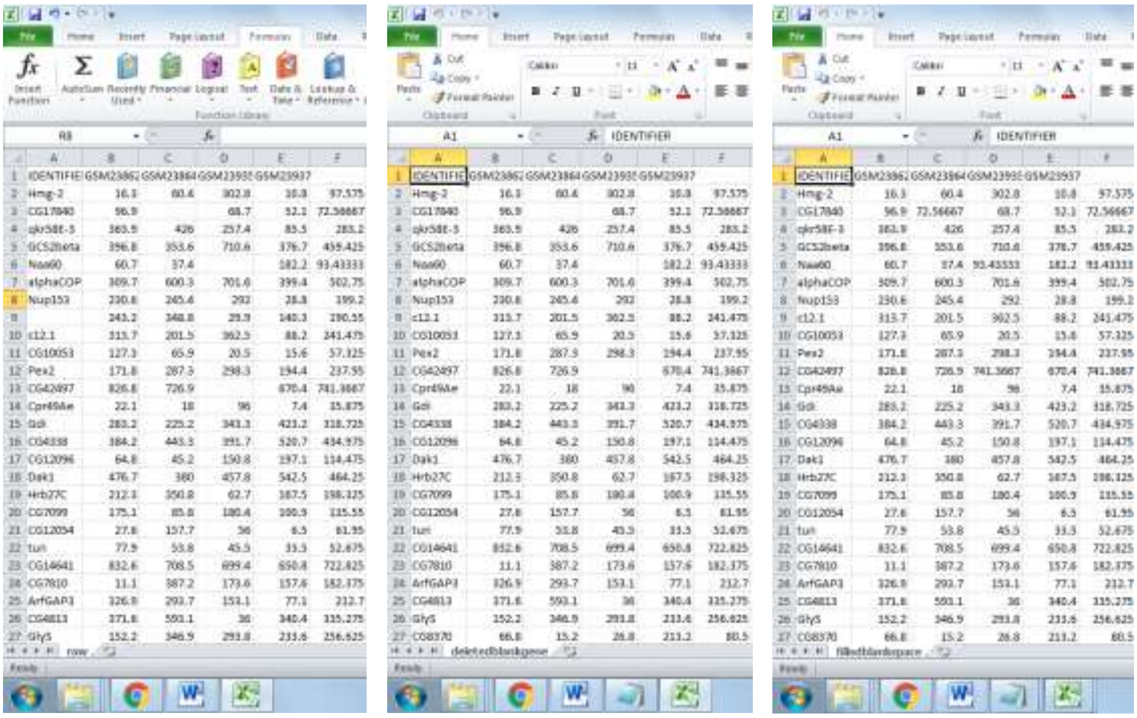
BLANK GENE DATA REMOVAL

Open output.csv file in notepad copy the data and paste it in form given below. It will give us formated data.

BLANK SPACE FILLING IN THE DATA

Open second csv output in excel sheet. Then find average value of each gene row.
Filter around our desired gene average range (e.g. if our desired gene data average is 100 then we filter the data expect from 50 to 150.)
Remaining data we copy and paste in a notepad file name it final.txt and proceed this data in clustering.

Fig 1: Display page of Pre-Clustering



A8

	A	B	C	D	E	F
1	IDENTIFIE	GSM2386	GSM23864	GSM2393	GSM23937	
2	Hmg-2	16.3	60.4	302.8	10.8	97.575
3	CG17840	96.9		68.7	52.1	72.56667
4	qkr58E-3	363.9	426	257.4	85.5	283.2
5	GCS2beta	396.8	353.6	710.6	376.7	459.425
6	Naa60	60.7	37.4		182.2	93.43333
7	alphaCOP	309.7	600.3	701.6	399.4	502.75
8	Nup153	230.6	245.4	292	28.8	199.2
9		243.2	348.8	29.9	140.3	190.55
10	c12.1	313.7	201.5	362.5	88.2	241.475
11	CG10053	127.3	65.9	20.5	15.6	57.325
12	Pex2	171.8	287.3	298.3	194.4	237.95
13	CG42497	826.8	726.9		670.4	741.3667
14	Cpr49Ae	22.1	18	96	7.4	35.875
15	Gdi	283.2	225.2	343.3	423.2	318.725
16	CG4338	384.2	443.3	391.7	520.7	434.975
17	CG12096	64.8	45.2	150.8	197.1	114.475
18	Dak1	476.7	380	457.8	542.5	464.25
19	Hrb27C	212.3	350.8	62.7	167.5	198.325
20	CG7099	175.1	85.8	180.4	100.9	135.55
21	CG12054	27.8	157.7	56	6.3	61.95
22	tun	77.9	53.8	45.5	33.5	52.675
23	CG14641	832.6	708.5	699.4	650.8	722.825
24	CG7810	11.1	387.2	173.6	157.6	182.375
25	ArfGAP3	326.9	293.7	153.1	77.1	212.7
26	CG4813	371.6	593.1	36	340.4	335.275
27	GlyS	152.2	346.9	293.8	233.6	256.625

raw

A1 IDENTIFIER

	A	B	C	D	E	F
1	IDENTIFIE	GSM2386	GSM23864	GSM2393	GSM23937	
2	Hmg-2	16.3	60.4	302.8	10.8	97.575
3	CG17840	96.9		68.7	52.1	72.56667
4	qkr58E-3	363.9	426	257.4	85.5	283.2
5	GCS2beta	396.8	353.6	710.6	376.7	459.425
6	Naa60	60.7	37.4		182.2	93.43333
7	alphaCOP	309.7	600.3	701.6	399.4	502.75
8	Nup153	230.6	245.4	292	28.8	199.2
9	c12.1	313.7	201.5	362.5	88.2	241.475
10	CG10053	127.3	65.9	20.5	15.6	57.325
11	Pex2	171.8	287.3	298.3	194.4	237.95
12	CG42497	826.8	726.9		670.4	741.3667
13	Cpr49Ae	22.1	18	96	7.4	35.875
14	Gdi	283.2	225.2	343.3	423.2	318.725
15	CG4338	384.2	443.3	391.7	520.7	434.975
16	CG12096	64.8	45.2	150.8	197.1	114.475
17	Dak1	476.7	380	457.8	542.5	464.25
18	Hrb27C	212.3	350.8	62.7	167.5	198.325
19	CG7099	175.1	85.8	180.4	100.9	135.55
20	CG12054	27.8	157.7	56	6.3	61.95
21	tun	77.9	53.8	45.5	33.5	52.675
22	CG14641	832.6	708.5	699.4	650.8	722.825
23	CG7810	11.1	387.2	173.6	157.6	182.375
24	ArfGAP3	326.9	293.7	153.1	77.1	212.7
25	CG4813	371.6	593.1	36	340.4	335.275
26	GlyS	152.2	346.9	293.8	233.6	256.625
27	CG8370	66.8	15.2	26.8	213.2	80.5

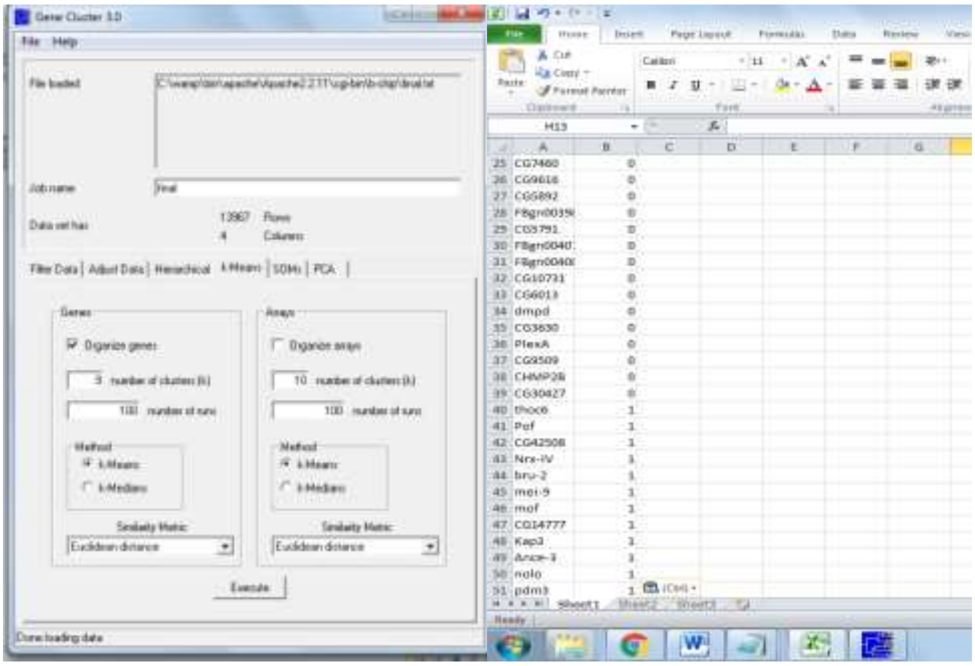
deletedblankgene

A1 IDENTIFIER

	A	B	C	D	E	F
1	IDENTIFIE	GSM2386	GSM23864	GSM2393	GSM23937	
2	Hmg-2	16.3	60.4	302.8	10.8	97.575
3	CG17840	96.9	72.56667	68.7	52.1	72.56667
4	qkr58E-3	363.9	426	257.4	85.5	283.2
5	GCS2beta	396.8	353.6	710.6	376.7	459.425
6	Naa60	60.7	37.4	93.43333	182.2	93.43333
7	alphaCOP	309.7	600.3	701.6	399.4	502.75
8	Nup153	230.6	245.4	292	28.8	199.2
9	c12.1	313.7	201.5	362.5	88.2	241.475
10	CG10053	127.3	65.9	20.5	15.6	57.325
11	Pex2	171.8	287.3	298.3	194.4	237.95
12	CG42497	826.8	726.9	741.3667	670.4	741.3667
13	Cpr49Ae	22.1	18	96	7.4	35.875
14	Gdi	283.2	225.2	343.3	423.2	318.725
15	CG4338	384.2	443.3	391.7	520.7	434.975
16	CG12096	64.8	45.2	150.8	197.1	114.475
17	Dak1	476.7	380	457.8	542.5	464.25
18	Hrb27C	212.3	350.8	62.7	167.5	198.325
19	CG7099	175.1	85.8	180.4	100.9	135.55
20	CG12054	27.8	157.7	56	6.3	61.95
21	tun	77.9	53.8	45.5	33.5	52.675
22	CG14641	832.6	708.5	699.4	650.8	722.825
23	CG7810	11.1	387.2	173.6	157.6	182.375
24	ArfGAP3	326.9	293.7	153.1	77.1	212.7
25	CG4813	371.6	593.1	36	340.4	335.275
26	GlyS	152.2	346.9	293.8	233.6	256.625
27	CG8370	66.8	15.2	26.8	213.2	80.5

filledblankspace

Fig 2: Pre-Clustering Results and analysis example



Gene Cluster 3.0

File loaded: C:\wamp\bin\apache\Apache2.2.11\cgi-bin\b-chip\final.txt

Job name: final

Data set has: 13967 Rows, 4 Columns

Filter Data | Adjust Data | Hierarchical | k-Means | SOMs | PCA

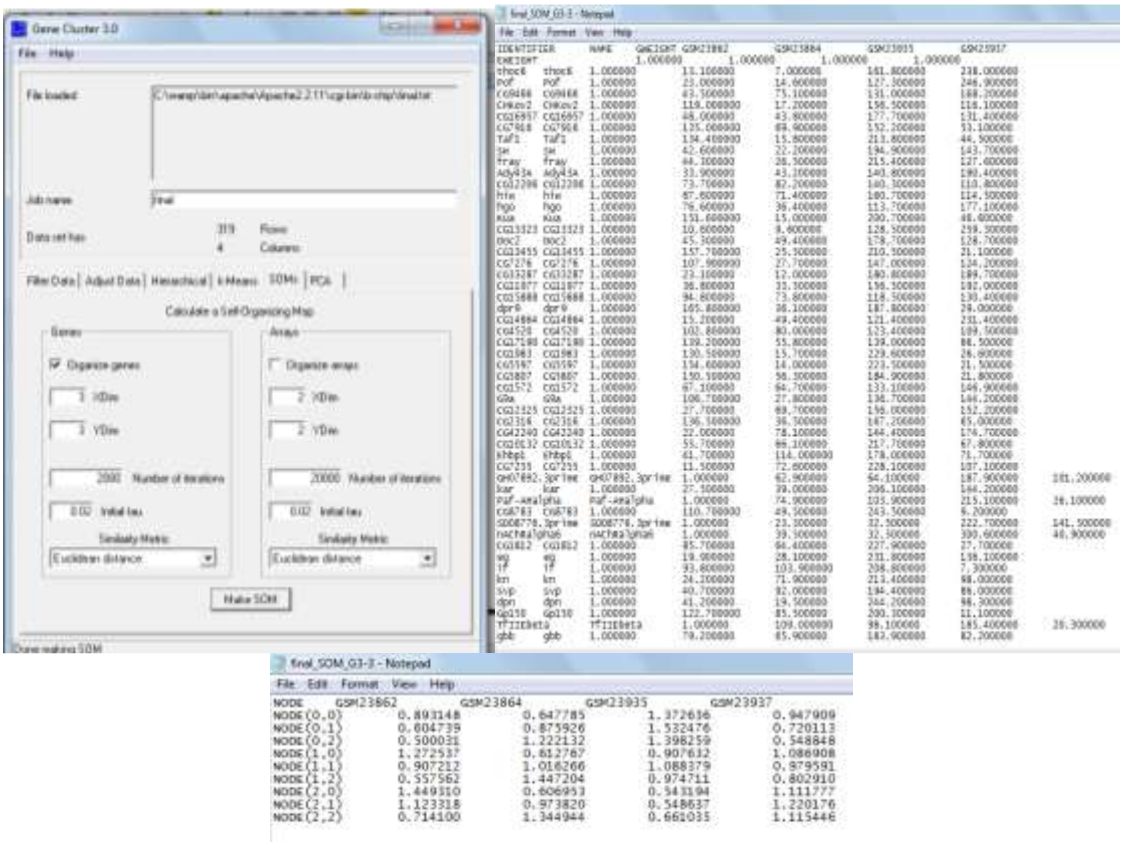
Genes: ☑ Organize genes; 3 number of clusters (k); 100 number of runs; Method: ● k-Means ○ k-Medians; Similarity Metric: Euclidean distance

Arrays: □ Organize arrays; 10 number of clusters (k); 100 number of runs; Method: ● k-Means ○ k-Medians; Similarity Metric: Euclidean distance

Done loading data

	A	B	C	D	E	F	G
25	CG7460	0					
26	CG9616	0					
27	CG5892	0					
28	FBgn00359	0					
29	CG5791	0					
30	FBgn0040	0					
31	FBgn0040	0					
32	CG10731	0					
33	CG6013	0					
34	dmpd	0					
35	CG3630	0					
36	PlexA	0					
37	CG9509	0					
38	CHMP2B	0					
39	CG30427	0					
40	thoc6	1					
41	Pof	1					
42	CG42508	1					
43	Nrx-IV	1					
44	bru-2	1					
45	mei-9	1					
46	mof	1					
47	CG14777	1					
48	Kap3	1					
49	Ance-3	1					
50	nolo	1					
51	pdm3	1					

Fig 3: KMeans-Clustering Results and analysis example



Gene Cluster 3.0

File loaded: C:\wamp\bin\apache\Apache2.2.11\cgi-bin\b-chip\final.txt

Job name: final

Data set has: 319 Rows, 4 Columns

Filter Data | Adjust Data | Hierarchical | k-Means | SOMs | PCA

Calculate a Self-Organizing Map

Genes: ☑ Organize genes; 3 XDim; 3 YDim; 2000 Number of iterations; 0.02 Initial tau; Similarity Metric: Euclidean distance

Arrays: □ Organize arrays; 2 XDim; 2 YDim; 20000 Number of iterations; 0.02 Initial tau; Similarity Metric: Euclidean distance

Done making SOM

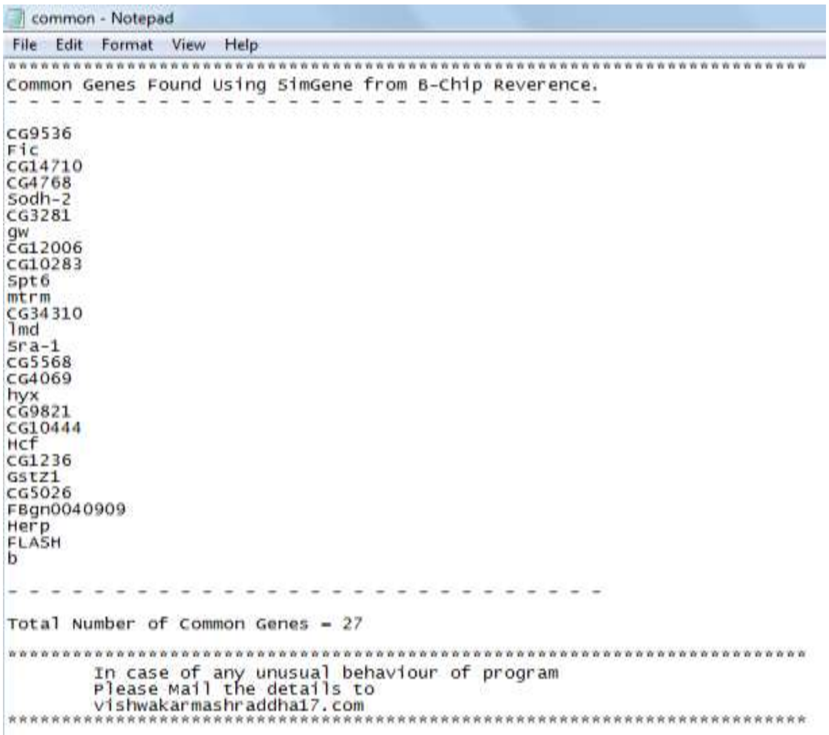
final_SOM_G3-3 - Notepad

```
IDENTIFIER        NAME      GWEIGHT GSM23862      GSM23864      GSM23935      GSM23937
EWEIGHT                     1.000000      1.000000      1.000000      1.000000
thoc6   thoc6   1.000000        13.100000     7.000000      161.800000    238.000000
Pof     Pof     1.000000        23.000000     14.600000     127.300000    246.900000
CG9468  CG9468  1.000000        43.500000     75.100000     131.000000    168.200000
CHKov2  CHKov2  1.000000        119.000000    17.200000     158.500000    118.100000
CG16957 CG16957 1.000000        48.000000     43.800000     177.700000    131.400000
CG7918  CG7918  1.000000        125.000000    89.900000     152.200000    53.100000
Taf1    Taf1    1.000000        134.400000    15.800000     213.800000    44.500000
se      se      1.000000        42.600000     22.200000     194.900000    143.700000
fray    fray    1.000000        44.300000     28.300000     215.400000    127.400000
Ady43A  Ady43A  1.000000        33.900000     43.200000     140.800000    190.400000
CG12206 CG12206 1.000000        73.700000     82.200000     140.300000    110.800000
htw     htw     1.000000        87.600000     71.400000     180.700000    114.500000
hgo     hgo     1.000000        76.600000     36.400000     113.700000    177.100000
kua     kua     1.000000        151.600000    15.000000     200.700000    48.600000
CG13323 CG13323 1.000000        10.600000     9.600000      128.500000    259.300000
Doc2    Doc2    1.000000        45.300000     49.400000     178.700000    128.700000
CG13455 CG13455 1.000000        157.700000    25.500000     210.500000    21.100000
CG7276  CG7276  1.000000        107.900000    27.700000     147.000000    124.200000
CG13287 CG13287 1.000000        23.100000     12.000000     180.800000    189.700000
CG11877 CG11877 1.000000        38.800000     33.300000     158.300000    182.000000
CG15688 CG15688 1.000000        94.800000     73.800000     118.500000    130.400000
dpr9    dpr9    1.000000        165.800000    38.100000     187.800000    29.000000
CG14884 CG14884 1.000000        15.200000     49.400000     121.400000    231.400000
CG4520  CG4520  1.000000        102.800000    80.000000     123.400000    109.500000
CG17198 CG17198 1.000000        139.200000    55.800000     139.000000    88.500000
CG1983  CG1983  1.000000        130.500000    15.700000     229.600000    26.600000
CG5597  CG5597  1.000000        134.600000    14.000000     223.500000    21.500000
CG3807  CG3807  1.000000        150.500000    58.300000     184.900000    21.800000
CG1572  CG1572  1.000000        67.100000     64.700000     133.100000    146.900000
G9a     G9a     1.000000        106.700000    27.800000     138.700000    166.200000
CG12325 CG12325 1.000000        27.700000     69.700000     158.000000    152.200000
CG2316  CG2316  1.000000        136.300000    36.300000     187.200000    65.000000
CG42240 CG42240 1.000000        22.000000     78.100000     144.400000    174.700000
CG10132 CG10132 1.000000        55.700000     66.100000     217.700000    67.800000
Khbp1   Khbp1   1.000000        41.700000     114.000000    178.000000    71.700000
CG7235  CG7235  1.000000        11.500000     72.600000     228.100000    107.100000
GH07892.3prime  GH07892.3prime  1.000000      62.900000     64.100000     187.900000    181.200000
kar     kar     1.000000        27.500000     39.000000     206.100000    144.200000
Paf-AHalpha     Paf-AHalpha     1.000000      74.900000     105.900000    215.100000    26.100000
CG8783  CG8783  1.000000        110.700000    49.500000     243.500000    6.200000
SD08778.3prime  SD08778.3prime  1.000000      23.300000     32.500000     222.700000    141.500000
nAChRalpha6     nAChRalpha6     1.000000      39.300000     32.300000     300.600000    40.900000
CG1812  CG1812  1.000000        85.700000     64.400000     227.900000    27.700000
wg      wg      1.000000        19.000000     28.100000     231.800000    158.100000
1f      1f      1.000000        93.800000     103.900000    208.800000    7.300000
kn      kn      1.000000        24.200000     71.900000     213.400000    98.000000
svp     svp     1.000000        40.700000     92.000000     194.400000    86.000000
dpn     dpn     1.000000        41.200000     19.500000     244.200000    98.300000
Gp150   Gp150   1.000000        122.700000    85.500000     200.300000    11.100000
TfIIEbeta       TfIIEbeta       1.000000      109.000000    98.100000     185.400000    26.300000
gbb     gbb     1.000000        79.200000     85.900000     182.900000    82.200000
```

final_SOM_G3-3 - Notepad

```
NODE      GSM23862      GSM23864      GSM23935      GSM23937
NODE(0,0)       0.893148      0.647785      1.372636      0.947909
NODE(0,1)       0.604739      0.875926      1.532476      0.720113
NODE(0,2)       0.500031      1.222132      1.398259      0.548848
NODE(1,0)       1.272537      0.612767      0.907632      1.086908
NODE(1,1)       0.907212      1.016266      1.088379      0.979591
NODE(1,2)       0.557562      1.447204      0.974711      0.802910
NODE(2,0)       1.449310      0.606953      0.543194      1.111777
NODE(2,1)       1.123318      0.973820      0.548637      1.220176
NODE(2,2)       0.714100      1.344944      0.661035      1.115446
```

Fig 4: SOM-Clustering Results and analysis example



```
************************************************************************
Common Genes Found Using SimGene from B-Chip Reverence.
- - - - - - - - - - - - - - - - - - - - - - - - - - - -

CG9536
Fic
CG14710
CG4768
Sodh-2
CG3281
gw
CG12006
CG10283
Spt6
mtrm
CG34310
lmd
Sra-1
CG5568
CG4069
hyx
CG9821
CG10444
Hcf
CG1236
GstZ1
CG5026
FBgn0040909
Herp
FLASH
b

- - - - - - - - - - - - - - - - - - - - - - - - - - - -

Total Number of Common Genes = 27

************************************************************************
        In case of any unusual behaviour of program
        Please Mail the details to
        vishwakarmashraddha17.com
************************************************************************
```

Fig 5: Common Gene of SOM and KMC Clustering

VI. Conclusion

B-Chip Reverence is a database, specially design for in silico analysis of microarray. In the field of Bioinformatics, B-chip Reverence fulfills the basic needs during the online microarray analysis.hence is one of a kind of its innovative database which provides web server for microarray redundancy removal & normalizationand various data analysis techniques are applied on the data.

VII. Summary

B-Chip Reverence is a online database which is freely accessible for microarray redundancy removal & normalization and various data analysis techniques are applied on the data. This software accurately handle the massive amount of data. The growing use of DNA microarrays in biomedical research has led to the proliferation of analysis tools. These software programs address different aspects of analysis (e.g. normalization and clustering within and across individual arrays) as well as extended analysis methods (e.g. clustering, annotation and mining of multiple datasets). After studying all the terms and problems related to Microarray technique, we tried to make an-open and user friendly software to deal with all the problems and to run all the steps of this technique, so that we used Perl & Perl-cgi. perl-cgi stands for *Common Gateway Interface*, is a standard programming interface between Web servers and external programs. perl-cgi executes external programs on the web server. We also used BioPerl, we correlated K-Means Clustering and Self-Organizing Maps (SOM) algorithms to perform clustering with Bio Perl. Bio Perl is a collection of Perl modules and it facilitates the development of Perl scripts for bioinformatics applications. And Dreamweaver8 used to create professional web pages and also quickly add objects and functionality to pages without having to program the HTML code manually.B-Chip Reverence is a database, specially design for in silico analysis of microarray. In the field of Bioinformatics, B-chip Reverence fulfills the basic needs during the online microarray analysis. hence is one of a kind of its innovative database which provides web server for microarray redundancy removal & normalization and various data analysis techniques are applied on the data.

References

1. Vroh Bi I, McMullen MD, Sanchez-Villeda H, Schroeder S, Gardiner J, Polacco M, Soderlund C, Wing R, Fang Z, Coe EH., Jr Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci. 2006;46:12–21. doi: 10.2135/cropsci2004.0706. [Cross Ref]
2. Wright SI, Vroh Bi I, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. The effects of artificial selection on the maize genome. Science. 2005;308:1310–1314. doi: 10.1126/science.1107891. [PubMed] [Cross Ref]
3. Schwartz AS, Pachter L. Multiple alignment by sequence annealing. Bioinformatics. 2006;23:e24–e29. doi: 10.1093/bioinformatics/btl311. [PubMed] [Cross Ref]
4. Ewing B, Green P. Basecalling of automated sequencer traces using *phred*. II. Error probabilities. Genome Res. 1998;8:186–194. [PubMed]
5. Ewing B, Hillier L, Wendl M, Green P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 1998;8:175–185. [PubMed]
6. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res. 1994;22:4673–4680. doi: 10.1093/nar/22.22.4673. [PMC free article][PubMed] [Cross Ref]
7. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res. 2004;32:1792–1797. doi: 10.1093/nar/gkh340. [PMC free article] [PubMed] [Cross Ref]
8. Page RDM. TREEVIEW: An application to display phylogenetic trees on personal computers. Comp Appl Biosci. 1996;12:357–358. [PubMed]
9. Rozas J, Sánchez-DelBarrio SJ, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics. 2003;19:2496–2497. doi: 10.1093/bioinformatics/btg359. [PubMed] [Cross Ref]
10. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999;41:95–98.
11. Clamp M, Cuff J, Searle SM, Barton JG. The Jalview Java alignment editor. Bioinformatics. 2004;20:426–427. doi: 10.1093/bioinformatics/btg430. [PubMed] [Cross Ref]
12. Pible O, Imbert G, Pellequer J-L. INTERALIGN: interactive alignment editor for distantly related protein sequences. Bioinformatics. 2005;21:3166–3167. doi: 10.1093/bioinformatics/bti474. [PubMed] [Cross Ref]
13. Yamasaki M, Tenaillon MI, Vroh Bi I, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD. A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell. 2005;17:2859–2872. doi: 10.1105/tpc.105.037242. [PMC free article] [PubMed] [Cross Ref]
14. Canaran P, Stein L, Ware D. Look-Align: An interactive web-based multiple sequence alignment viewer with polymorphism analysis support. Bioinformatics. 2006;22:885–886. doi: 10.1093/bioinformatics/btl028. [PubMed] [Cross Ref]
15. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki M, Szilagyi A, Kihara D, (2003) TOUCHSTONE: A unified approach to protein structure prediction, *Proteins* (53 Suppl 6):469–479.
16. Stark A., Sunyaev S., Russell R.B. (2003) A model for statistical significance of local similarities in structure. J. Mol. Biol. 2003;326:1307–1316.
17. Sonnhammer, E., Eddy, S., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28, 405–420.
18. Tatusov, R., Galperin, M., Natale, D., and Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28, 33–36.

19. Pandey. H, Darbari. M and Singh. V.K, "Coalescence of Evolutionary Multi-Objective Decision making approach and Genetic Programming for Selection of Software Quality Parameter", International Journal of Applied Information System (IJAIS), Foundation of Computer Science, New York, USA, Volume 7, No. 11, PP. ISSN: 2249-0868, Nov. 2014.
20. Bansal. S and Pandey. H, "Develop Framework for selecting best Software Development Methodology", International Journal of Scientific and Engineering Research, Volume 5, Issue 4, PP. 1067-1070, ISSN: 2229-5518, Apr. 2014.
21. Srivastava. M and Pandey. H, "A Literature Review of E- Learning Model Based on Semantic Web Technology", International Journal of Scientific and Engineering Research" Volume 5, Issue 10, PP. 174-178, ISSN: 2229-5518, Oct. 2014.
22. Pandey. H and Singh. V.K, "A New NFA Reduction Algorithm for State Minimization Problem", International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA, Volume 8, No.3, PP. 27-30, ISSN: 2249-0868, Feb. 2015.
23. Pandey. H and Singh. V.K, "LR Rotation rule for creating Minimal NFA", International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA, Volume 8, No.6, PP. 1-4, ISSN: 2249-0868, Apr. 2015.
24. Pandey. H and Darbari. M, "Estimation of Software Quality Parameters Using Combination of Quality Function Deployment and Messy Genetic Algorithm", Grenze Scientific Society, Associate publisher: **McGraw-Hill Education**, ISBN: 978-93-392-2169-0. Feb. 2015.
25. Darbari. M, Srivastava. G and Pandey. H, "New Assumption of Cognitive Model for Information Foraging on Web", International Journal of Advances in Engineering & Technology, Volume 8, Issue 2, PP. 163-169, ISSN: 22311963, Apr. 2015.
26. Pandey. H and Singh. V.K, "A Fuzzy Logic based Recommender System for E-Learning System with Multi-Agent Framework", International Journal of Computer Applications, Foundation of Computer Science FCS, New York, USA, Volume 122, No.17, PP. 18-21, ISSN: 0975 – 8887, July. 2015.
27. Rai. V and Pandey. H, "Estimation of Maintainability in Object Oriented Design Phase: State of the art", International Journal of Scientific and Engineering Research" Volume 6, Issue 9, PP. 25-35, ISSN: 2229-5518, Sept. 2015.
28. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y et. al.: E-Cell: software environment for whole-cell simulation. Bioinformatics 1999, 15:72-84.
29. Von Dassow G, Meir E, Munro EM, Odell GM: The segment polarity network is a robust developmental module. Nature 2000, 406:188-92.
30. Wilson CA, Kreychman J, Gerstein M (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores, *J MolBiol*297(1):233–249.
31. Wu, C., Huang, H., Yeh, L., and Barker, W. (2003).Protein family classification and functional annotation.CompBiolChem 27, 37–47.