IJCTA, 9(23), 2016, pp. 125-134

Identification of Co Expressed Genes in Sinorhizobium Meliloti 1021 through Hierarchical Clustering

Shivani Chandra* Alka Grover Abha Kumari*** and Sampat Nehra******

Abstract : Clustering of gene expression microarray data has been proven a powerful tool for understanding gene function and gene regulation. Co expression of genes is indicated by their clustering into same groups. It is assumed that genes with similar expression patterns tend to have similar functions. In present research we used microarray data for *Sinorhizobium meliloti 1021* and used hierarchical clustering to study the gene expression pattern. The results of study clearly indicated a number of co-expressed gene clusters. The membership of a gene in a given cluster indicates the probable function of a gene. Based on cluster analysis functions of three genes of unknown function were assigned a probable function. The phylogenetic analysis also suggested that the gene clusters shared similar functions. Thus it can be inferred that hierarchical clustering may be used to identify functions of unknown genes.

Keywords : Gene expression, co expression, hierarchical clustering.

1. INTRODUCTION

DNA microarray technology has proved to be an elementary tool in studying expression of genes. A natural basis for organizing this information is to cluster genes with similar patterns of expression. Clustering is usually employed in microarray experiments to mark set of genes that share similar expressions[1].Identification of genes that exhibit similar expression patterns is one of the most important step in gene expression analysis. Every cluster is then analyzed separately.Each group is then associated with a specific biological function or biological process. Hierarchal clustering is the most widely used method for clustering gene expression profiles for the discovery of co regulated and functionally related genes. Clustering of genes using gene expression microarray data has been known for its application in functional annotation, classification of tissues, identification of regulatory motifs etc. Therefore, it can be inferred that clustering suggests the functional relationships between groups of genes. It may also help in identifying promoter sequence elements that are shared among genes. The main objective of this research was to determine the co expression of genes in *Sinorhizobium meliloti and to predict the function of some unknown genes based on clusters of co expressed genes. Sinorhizobium meliloti is one of the best known gram-negative nitrogen-fixing soil bacterium.* The genome size is 6.6 Mbwhich *contains three replicons: a 3.65 megabase chromosome and two mega*plasmids, *pSymA (1.35 Mb) and pSymB (1.68 Mb).*[3,4]

* Amity Institute of Biotechnology Amity University Uttar Pradesh, Noida, India Email: schandra4@amity.edu

** Amity Institute of Biotechnology Amity University Uttar Pradesh, Noida, India Email: agrover@amity.edu

*** Amity Institute of Biotechnology Amity University Uttar Pradesh, Noida, India Email: akumari@amity.edu

**** Birla Institute of Scientific Research Statue Circle, Jaipur, India Email: nehrasampat@gmail.com

2. EXPERIMENT AND RESULT

2.1. Materials and Method

Raw data of the transcription profiling experiments of *Sinorhizobium meliloti strain 1021* has been downloaded from ArrayExpress Home (http://www.ebi.ac.uk/microarray-as/aer/#ae). The data was collected from "Transcription profiling time series of *sinorhizobium meliloti* in response to an osmotic upshift elicited by salt or sucrose" experiment. This raw data was present in 28 excel file sets, expression data of each file was obtained at different time and different compound concentration.Sm6KOligo microarray raw data was downloaded from ArrayExpress Home (http://www.ebi.ac.uk/microarray-as/aer/#ae-main[0]). Hierarchical clustering was done using Gene Cluster software *(*http://rana.lbl.gov/EisenSoftware.htm) and the data was viewed graphically using Treeviewprogram. Clustal W was used for multiple sequence alignment and phylogenetic tree analysis.

2.2. Methodology

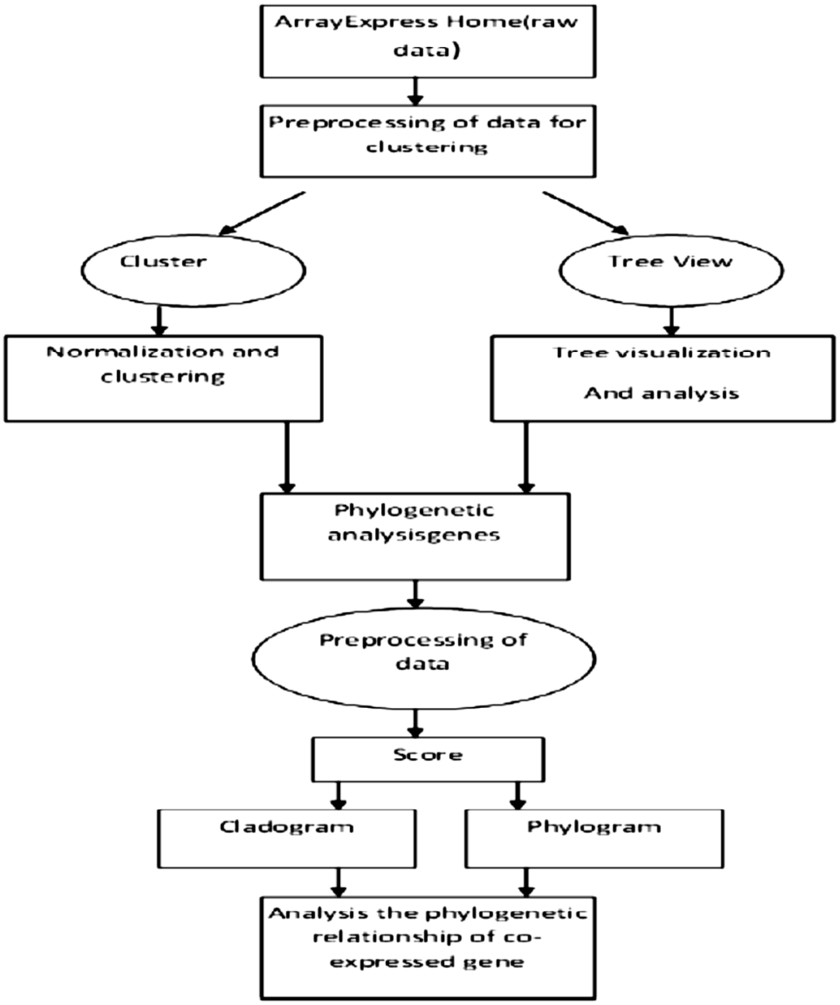

ArrayExpress Home(raw data)
Preprocessing of data for clustering
Cluster
Tree View
Normalization and clustering
Tree visualization And analysis
Phylogenetic analysisgenes
Preprocessing of data
Score
Cladogram
Phylogram
Analysis the phylogenetic relationship of co-expressed gene

Fig. 1.

2.3. Results and Discussion

Clustering of genes

The results from hierarchical clustering gave a total of six clusters (Fig 1.) Out of these six clusters three clusters were chosen for further analysis. The selections of these clusters were based on the presence of unknown genes. The description of the selected clusters is as follows :

Cluster-1(Fig 2.) : There were 46 genes in this cluster. Out of 46 genes, 28 genes code for flagellar hook proteins which are required in cell motility and rest 18 genes code for transferase proteins. Genes in this cluster were as follows :

SMc01167, SMc03017, SMc03023, SMc00638, SMc01597, SMc01939, SMa1754, SMc03030, SMc01239, SMc01922, SMc02562, SMc03051, SMc03048, SMa1750, SMc03049, SMc01500, SMc01947, Smc01930, SMc01929, SMc02121, SMc01923, SMc02120, SMc00245, SMc02047, SMc02048, SMc03024, SMc01513, SMc01919, SMa1851, SMa0745, SMc00913, SMc00912, SMa0744, SMc03787, SMc02122, SMc01912, SMc025 01, SMa2339, SMc02124, SMc01054, SMc01514, SMc02726, SMc03882, SMc04111, SMc03797, SMa2400.

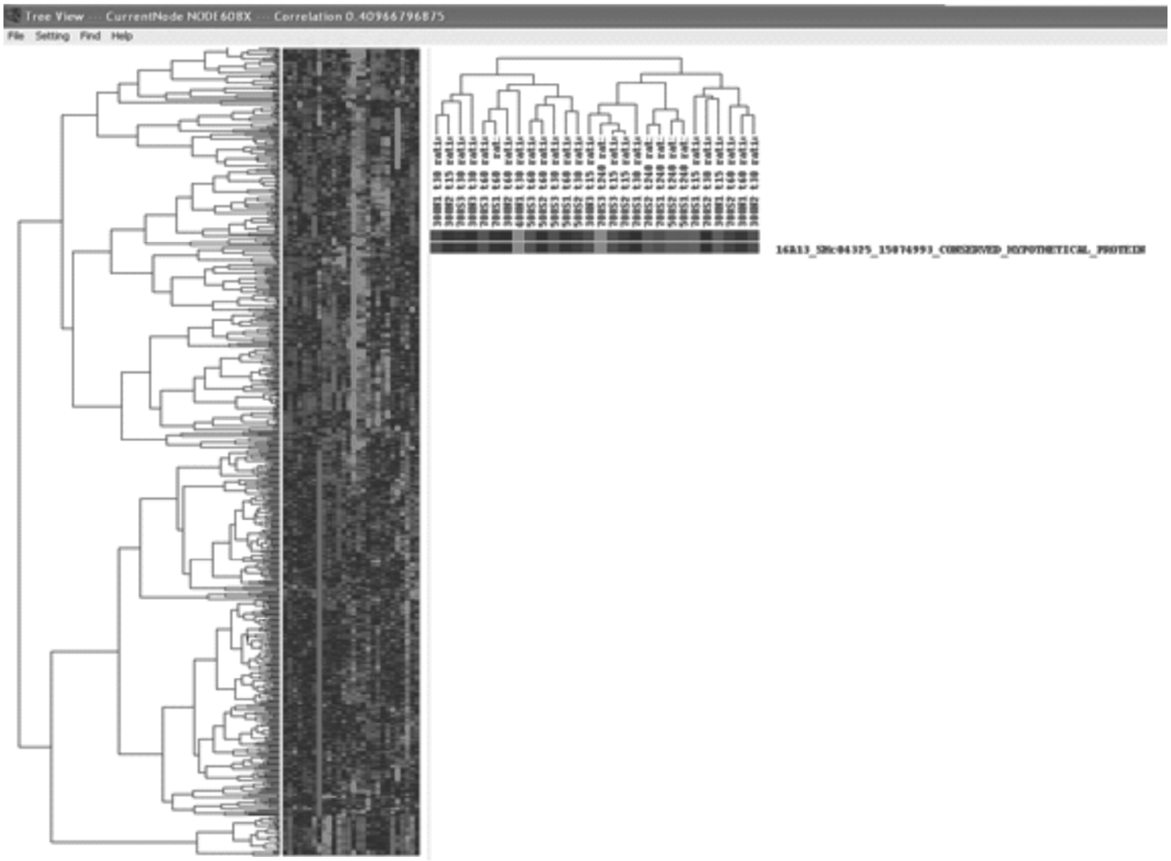
Cluster-2(Fig 3.) : There were 99 genes in this cluster. Out of 99 genes, 43 genes code for transmembrane proteins which are required in cellular processes and signaling. 38 genes code for reductase proteins and 18 genes code for dehydratase proteins. Genes in this cluster were as follows :

SMc01340, SMb21154, SMc04292, SMc01360, SMc02791, SMc03205, SMb21255, SMc02319, SMc04118, SMc00820, SMc01216, SMb20216, SMb21148, SMc01717, SMc01882, SMc03242, SMc00730, SMb21097, SMc03021, SM c00231, SMc00730, SMc00702, SMc02083, SMc02107, SMa1745, SMa1857, SMc00992, SMc01016, SMc00998, SMc00611, SMb20992, SMb21018, SMa0301, SMb20158, SMb21055, SMc01630, SMb20182, SMa1103, SMa1155, SMb20110, SMa1122, SMa1846, SMa0214, SMb21411, SMc03145, SMb20968, SMa0903, SMa0067, SMc00562, SMb20950, SMb21327, SMb20029, SMb20005, SMa1103, SMc01171, SMa0257, SMc03942, SMc01654, SMc03843, SMb20500, SMa1704, SMc01040, SMc00974, SMa0079, SMa0168, SMc00676, SMb20968, SMb20158, SMb21669, SMa1718, SMc01955, SMc00968, SMb20926, SMb20116, SMb20098, SMc00106, SMb21468, SMa0288, SMa1957, SMa1233, SMb21532, SMa0322, SMb21532, SMa1191, SMb20757, SMb21222, SMa1294, SMa2175, SMb20534, SMa0157, SMa2383, SMb20592, SMa1948, SMa0113, SMb20092, SMb20403, SMc02906, SMb20458, SMb20084.

Cluster-3 (Fig 4.) : There were 136 genes in this cluster. Out of 136 genes, 95 genes code for transporter proteins which are required in cellular processes and signaling. 14 genes code for transcriptional regulation, 16 genes code for synthetase proteins, and 11 genes code for isomerase proteins. Genes in this cluster were as follows :

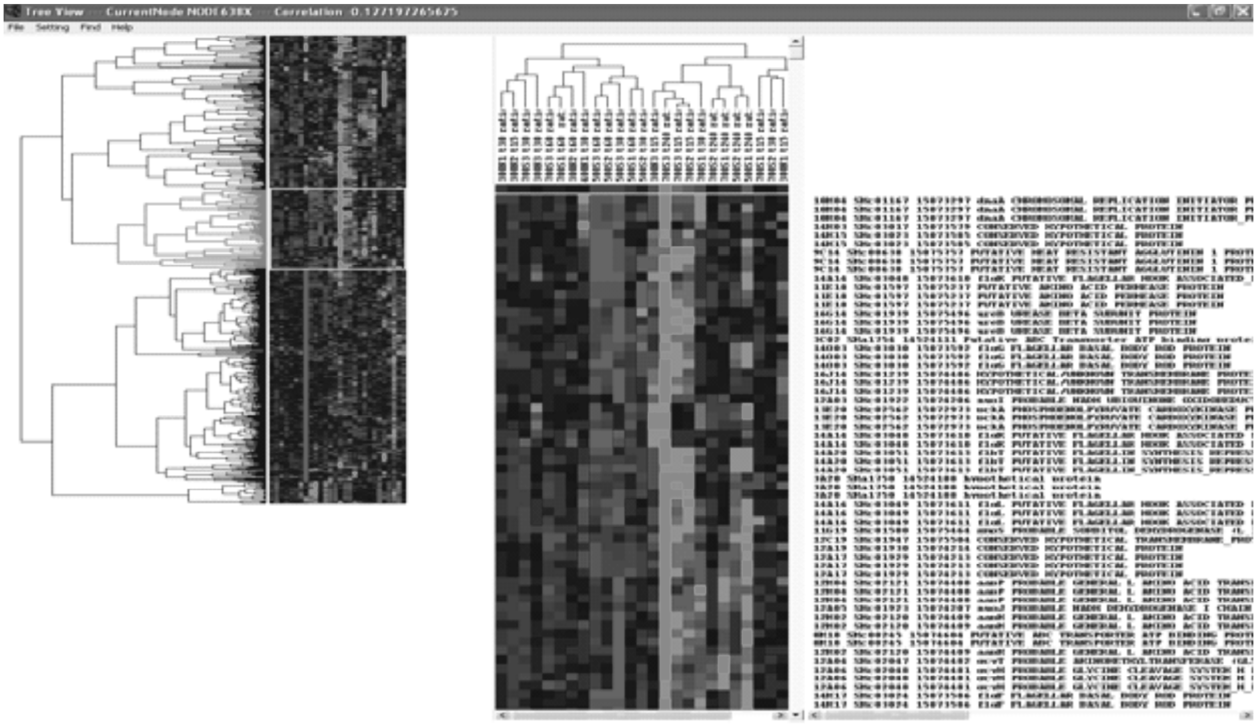
SMa1513, SMc03163, SMc03242, SMc03187, SMc02736, SMb20269, SMc00336, SMc02736, SMc00406, SMb20829, SMc04142, SMc02346, SMa1427, SMb20611, SMc02353, SMc01282, SMc00879, SMc03286, SMb20374, SMb20829, SMa2211, SMc02880, SMb20836, SMb21173, SMc02773, SMb20182, SMb20035, SMb20269, SMa1476, SMc03286, SMc04136, SMb20836, SMa1363, SMb20288, SMb21173, SMc00356, SMc02372, SMb20508, SMa0689, SMa0383, SMc02325, SMa2099, SMa1283, SMc04196, SMa0653, SMa2087, SMa2325, SMa0217, SMa2389, SMc01222, SMc03287, SMa1872, SMa1373, SMb20216, SMb21057, SMb20611, SMb21250, SMc03287, SMa1146, SMb21487, SMa0551, SMa1191, SMa1723, SMb21564, SMa1331, SMb20692, SMc00291, SMc01662, SMb20699, SMa0702, SMa0699, SMb21664, SMa1245, SMc01829, SMc00483, SMb20325, SMc02606, SMa1933, SMa0380, SMa1523, SMa1882, SMa0436, SMa1414, SMa0751, SMa0527, SMc00269, SMa0599, SMa0776, SMc00387, SMc01761, SMc00974, SMc01624, SMc02089, SMa1283, SMc03868, SMa1666, SMb20122, SMa0247, SMa0346, SMb20980, SMa0596, SMc03003, SMc02731, SMb21018, SMc02077, SMc01147, SMb20506, SMc01932, SMa2097, SMa2097, SMc00588, SMc03976, SMa1562, SMc00026, SMc01490, SMc03894, SMb20452, SMc01046, SMc00007, SMc02065, SMa2031, SMb21204, SMc01412, SMc03095, SMc01785, SMc02235, SMa0359, SMc03961, SMa1817, SMc00521, SMc03061, SMa0637, SMc03211, SMc00852, SMa0485, SMb20183, SMb20667, SMa0083, SMb20905, SMb21030, SMc00937, SMc00001, SMa1461, SMc00476, SMc00297.

Clustered display of data


Tree View --- CurrentNode NODE608X --- Correlation 0.40966796875
File
Setting
Find
Help

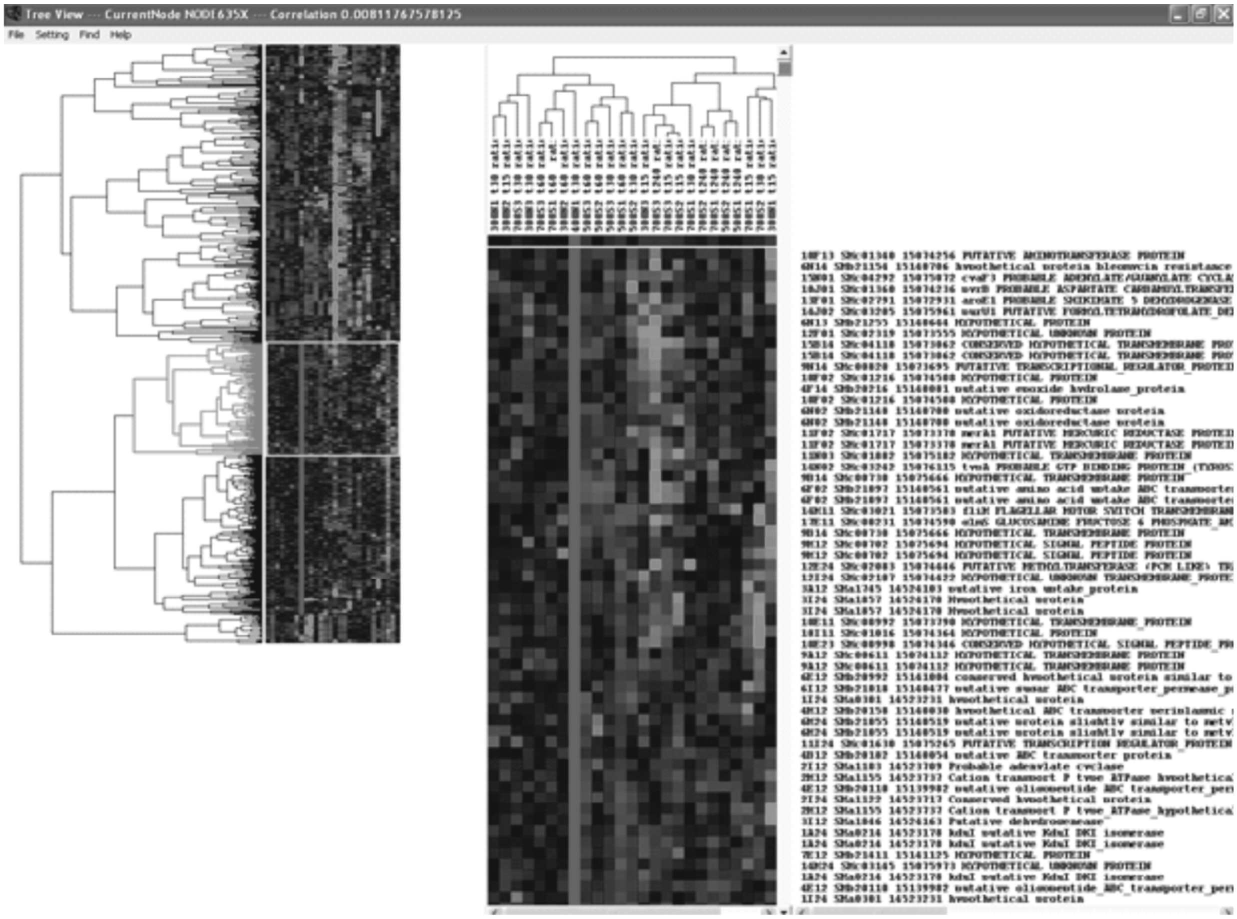
16A13_SMc04325_15074993_CONSERVED_HYPOTHETICAL_PROTEIN

Fig. 1. Groups of coexpressed genes representing diverse expression patterns. Genes with log mean ratios of zero are colored with black, increasingly positive log mean ratios with reds of increasing intensity and increasingly negative log mean ratios with greens of increasing intensity.


Tree View --- CurrentNode NODE638X --- Correlation 0.127197265625
File
Setting
Find
Help

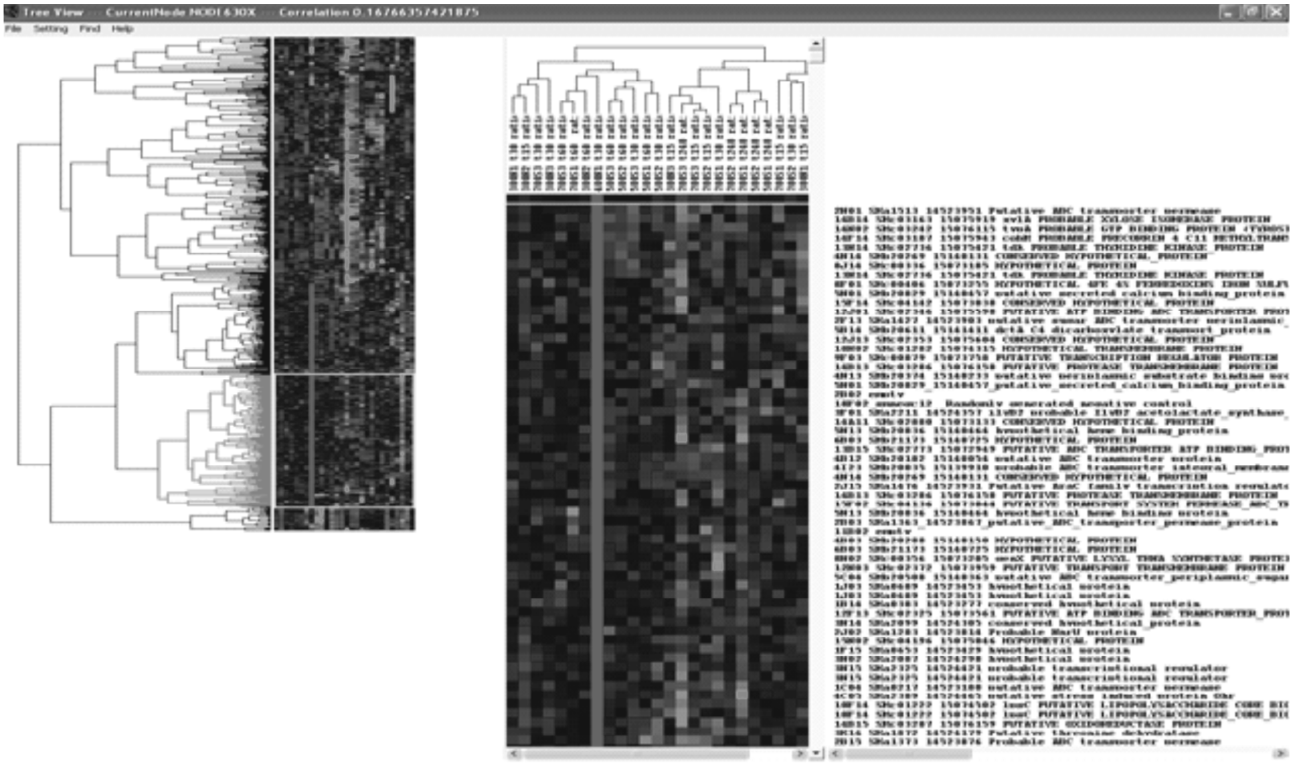
```
10H04 SMc01167 15073297 dnaA CHROMOSOMAL REPLICATION INITIATOR P
10H04 SMc01167 15073297 dnaA CHROMOSOMAL REPLICATION INITIATOR P
10H04 SMc01167 15073297 dnaA CHROMOSOMAL REPLICATION INITIATOR_P
14H03 SMc03017 15073579 CONSERVED HYPOTHETICAL PROTEIN
14H15 SMc03023 15073585 CONSERVED HYPOTHETICAL PROTEIN
14H15 SMc03023 15073585 CONSERVED HYPOTHETICAL PROTEIN
9C14 SMc00638 15075757 PUTATIVE HEAT RESISTANT AGGLUTININ 1 PROT
9C14 SMc00638 15075757 PUTATIVE HEAT RESISTANT AGGLUTININ 1 PROT
9C14 SMc00638 15075757 PUTATIVE HEAT RESISTANT AGGLUTININ 1 PROT
14A14 SMc03048 15073610 flgK PUTATIVE FLAGELLAR HOOK ASSOCIATED_
11E18 SMc01597 15075217 PUTATIVE AMINO ACID PERMEASE PROTEIN
11E18 SMc01597 15075217 PUTATIVE AMINO ACID PERMEASE PROTEIN
11E18 SMc01597 15075217 PUTATIVE AMINO ACID PERMEASE PROTEIN
16G14 SMc01939 15075496 ureB UREASE BETA SUBUNIT PROTEIN
16G14 SMc01939 15075496 ureB UREASE BETA SUBUNIT PROTEIN
16G14 SMc01939 15075496 ureB UREASE BETA SUBUNIT PROTEIN
3C07 SMa1756 14524131 Putative ABC Transporter ATP binding prote
14D03 SMc03030 15073592 flgG FLAGELLAR BASAL BODY ROD PROTEIN
14D03 SMc03030 15073592 flgG FLAGELLAR BASAL BODY ROD PROTEIN
14D03 SMc03030 15073592 flgG FLAGELLAR BASAL BODY ROD PROTEIN
16J14 SMc01239 15074486 HYPOTHETICAL/UNKNOWN TRANSMEMBRANE PROTE
16J14 SMc01239 15074486 HYPOTHETICAL/UNKNOWN TRANSMEMBRANE PROTE
16J14 SMc01239 15074486 HYPOTHETICAL/UNKNOWN TRANSMEMBRANE PROTE
12A03 SMc01922 15074206 nuoI PROBABLE NADH UBIQUINONE OXIDOREDUCT
13E20 SMc02562 15072973 pckA PHOSPHOENOLPYRUVATE CARBOXYKINASE P
13E20 SMc02562 15072973 pckA PHOSPHOENOLPYRUVATE CARBOXYKINASE P
13E20 SMc02562 15072973 pckA PHOSPHOENOLPYRUVATE CARBOXYKINASE P
14A14 SMc03048 15073610 flgK PUTATIVE FLAGELLAR HOOK ASSOCIATED
14A14 SMc03048 15073610 flgK PUTATIVE FLAGELLAR HOOK ASSOCIATED
14A20 SMc03051 15073613 flbT PUTATIVE FLAGELLIN SYNTHESIS REPRES
14A20 SMc03051 15073613 flbT PUTATIVE FLAGELLIN SYNTHESIS REPRES
14A20 SMc03051 15073613 flbT PUTATIVE FLAGELLIN_SYNTHESIS_REPRES
3A20 SMa1750 14524100 hypothetical protein
3A20 SMa1750 14524100 hypothetical protein
3A20 SMa1750 14524100 hypothetical protein
14A16 SMc03049 15073611 flgL PUTATIVE FLAGELLAR HOOK ASSOCIATED
14A16 SMc03049 15073611 flgL PUTATIVE FLAGELLAR HOOK ASSOCIATED
14A16 SMc03049 15073611 flgL PUTATIVE FLAGELLAR HOOK ASSOCIATED
11G19 SMc01500 15075444 smoS PROBABLE SORBITOL DEHYDROGENASE (L
12C19 SMc01947 15075504 CONSERVED HYPOTHETICAL TRANSMEMBRANE_PRO
12A19 SMc01930 15074214 CONSERVED HYPOTHETICAL PROTEIN
12A17 SMc01929 15074213 CONSERVED HYPOTHETICAL PROTEIN
12A17 SMc01929 15074213 CONSERVED HYPOTHETICAL PROTEIN
12A17 SMc01929 15074213 CONSERVED HYPOTHETICAL PROTEIN
12H04 SMc02121 15074408 aapP PROBABLE GENERAL L AMINO ACID TRANS
12H04 SMc02121 15074408 aapP PROBABLE GENERAL L AMINO ACID TRANS
12H04 SMc02121 15074408 aapP PROBABLE GENERAL L AMINO ACID TRANS
12A05 SMc01923 15074207 nuoJ PROBABLE NADH DEHYDROGENASE I CHAIN
12H02 SMc02120 15074409 aapM PROBABLE GENERAL L AMINO ACID TRANS
12H02 SMc02120 15074409 aapM PROBABLE GENERAL L AMINO ACID TRANS
8H18 SMc00245 15074604 PUTATIVE ABC TRANSPORTER ATP BINDING PROT
8H18 SMc00245 15074604 PUTATIVE ABC TRANSPORTER ATP BINDING PROT
12H02 SMc02120 15074409 aapM PROBABLE GENERAL L AMINO ACID TRANS
12A06 SMc02047 15074482 gcvT PROBABLE AMINOMETHYLTRANSFERASE (GL
12A06 SMc02048 15074481 gcvH PROBABLE GLYCINE CLEAVAGE SYSTEM H
12A06 SMc02048 15074481 gcvH PROBABLE GLYCINE CLEAVAGE SYSTEM H
12A06 SMc02048 15074481 gcvH PROBABLE GLYCINE CLEAVAGE SYSTEM H_
14H17 SMc03024 15073586 flgF FLAGELLAR BASAL BODY ROD PROTEIN
14H17 SMc03024 15073586 flgF FLAGELLAR BASAL BODY ROD PROTEIN
```

Fig. 2. Selected Cluster-1 in pink and name of genes belongs to this cluster.


Tree View --- CurrentNode NODE635X --- Correlation 0.00811767578125
File Setting Find Help

```
10F13 SMc01340 15074256 PUTATIVE AMINOTRANSFERASE PROTEIN
6N14 SMb21156 15140706 hypothetical protein bleomycin resistance
15B01 SMc04292 15075072 cyaF3 PROBABLE ADENYLATE/GUANYLATE CYCLA
10J01 SMc01360 15074236 pyrB PROBABLE ASPARTATE CARBAMOYLTRANSFE
13F01 SMc02791 15072931 aroE1 PROBABLE SHIKIMATE 5 DEHYDROGENASE
14J02 SMc03205 15075961 purU1 PUTATIVE FORMYLTETRAHYDROFOLATE_DE
6N13 SMb21255 15140644 HYPOTHETICAL PROTEIN
12F01 SMc02319 15073555 HYPOTHETICAL UNKNOWN PROTEIN
15B14 SMc04110 15073062 CONSERVED HYPOTHETICAL TRANSMEMBRANE PRO
15B14 SMc04110 15073062 CONSERVED HYPOTHETICAL TRANSMEMBRANE PRO
9N14 SMc00020 15073695 PUTATIVE TRANSCRIPTIONAL_REGULATOR_PROTEI
10F02 SMc01216 15074508 HYPOTHETICAL PROTEIN
4F14 SMb20216 15140081 putative epoxide hydrolase_protein
10F02 SMc01216 15074508 HYPOTHETICAL PROTEIN
6N02 SMb21140 15140700 putative oxidoreductase protein
6N02 SMb21140 15140700 putative oxidoreductase protein
11F02 SMc01717 15073370 merA1 PUTATIVE MERCURIC REDUCTASE PROTEI
11F02 SMc01717 15073370 merA1 PUTATIVE MERCURIC REDUCTASE PROTEI
11N03 SMc01082 15075182 HYPOTHETICAL TRANSMEMBRANE PROTEIN
14N02 SMc03242 15076115 typA PROBABLE GTP BINDING PROTEIN (TYROS
9D14 SMc00730 15075666 HYPOTHETICAL TRANSMEMBRANE PROTEIN
6F02 SMb21097 15140561 putative amino acid uptake ABC transporte
6F02 SMb21097 15140561 putative amino acid uptake ABC transporte
14M11 SMc03021 15073503 fliM FLAGELLAR MOTOR SWITCH TRANSMEMBRAN
12E11 SMc00231 15074590 glmS GLUCOSAMINE FRUCTOSE 6 PHOSPHATE_AM
9D14 SMc00730 15075666 HYPOTHETICAL TRANSMEMBRANE PROTEIN
9M12 SMc00702 15075694 HYPOTHETICAL SIGNAL PEPTIDE PROTEIN
9M12 SMc00702 15075694 HYPOTHETICAL SIGNAL PEPTIDE PROTEIN
12E24 SMc02083 15074446 PUTATIVE METHYLTRANSFERASE (PCM LIKE) TR
12I24 SMc02107 15074422 HYPOTHETICAL UNKNOWN TRANSMEMBRANE_PROTE
3A12 SMa1745 14524103 putative iron uptake_protein
3I24 SMa1857 14524170 Hypothetical protein
3I24 SMa1857 14524170 Hypothetical protein
10E11 SMc00992 15073790 HYPOTHETICAL TRANSMEMBRANE_PROTEIN
10I11 SMc01016 15074364 HYPOTHETICAL PROTEIN
10E23 SMc00998 15074346 CONSERVED HYPOTHETICAL SIGNAL PEPTIDE_PR
9A12 SMc00611 15074112 HYPOTHETICAL TRANSMEMBRANE PROTEIN
9A12 SMc00611 15074112 HYPOTHETICAL TRANSMEMBRANE PROTEIN
6E12 SMb20992 15141004 conserved hypothetical protein similar to
6I12 SMb21018 15140477 putative sugar ABC transporter_permease_p
1I24 SMa0301 14523231 hypothetical protein
4M12 SMb20150 15140030 hypothetical ABC transporter periplasmic
6M24 SMb21055 15140519 putative protein slightly similar to mety
6M24 SMb21055 15140519 putative protein slightly similar to mety
11I24 SMc01630 15075265 PUTATIVE TRANSCRIPTION REGULATOR_PROTEIN
4D12 SMb20182 15140054 putative ABC transporter protein
2I12 SMa1103 14523709 Probable adenylate cyclase
2M12 SMa1155 14523737 Cation transport P type ATPase hypothetica
4E12 SMb20110 15139982 putative oligopeptide ABC transporter_per
2I24 SMa1122 14523717 Conserved hypothetical protein
2M12 SMa1155 14523737 Cation transport P type_ATPase_hypothetica
3I12 SMa1846 14524163 Putative dehydrogenase
1A24 SMa0214 14523170 kduI putative KduI DKI isomerase
1A24 SMa0214 14523170 kduI putative KduI DKI_isomerase
7E12 SMb21411 15141125 HYPOTHETICAL PROTEIN
14M24 SMc03165 15075973 HYPOTHETICAL UNKNOWN PROTEIN
1A24 SMa0214 14523170 kduI putative KduI DKI isomerase
4E12 SMb20110 15139982 putative oligopeptide_ABC_transporter_per
1I24 SMa0301 14523231 hypothetical protein
```

Fig. 3. Selected Cluster-2 in pink and name of genes belongs to this cluster.


Tree View --- CurrentNode NODE630X --- Correlation 0.16766357421875
File Setting Find Help

```
2H01 SMa1513 14523951 Putative ABC transporter permease
14B14 SMc03163 15075919 xylA PROBABLE XYLOSE ISOMERASE PROTEIN
14N02 SMc03242 15076115 typA PROBABLE GTP BINDING PROTEIN (TYROSI
14F14 SMc03187 15075943 cobM PROBABLE PRECORRIN 4 C11 METHYLTRANS
13B14 SMc02736 15075421 tdk PROBABLE THYMIDINE KINASE_PROTEIN
4N14 SMb20269 15140131 CONSERVED HYPOTHETICAL_PROTEIN
8J14 SMc00336 15073105 HYPOTHETICAL PROTEIN
13B14 SMc02736 15075421 tdk PROBABLE THYMIDINE KINASE PROTEIN
8F01 SMc00406 15073255 HYPOTHETICAL 4FE 4S FERREDOXINS IRON SULF
5N01 SMb20829 15140657 putative secreted calcium binding_protein
15F14 SMc04142 15073030 CONSERVED HYPOTHETICAL PROTEIN
12J01 SMc02346 15075570 PUTATIVE ATP BINDING ABC TRANSPORTER PROT
2F13 SMa1427 14523903 putative sugar ABC transporter periplasmic_
5B14 SMb20611 15141411 dctA C4 dicarboxylate transport_protein
12J13 SMc02353 15075604 CONSERVED HYPOTHETICAL PROTEIN
14N07 SMc03202 15076315 HYPOTHETICAL TRANSMEMBRANE PROTEIN
9F03 SMc00879 15073750 PUTATIVE TRANSCRIPTION REGULATOR PROTEIN
14B13 SMc03206 15076158 PUTATIVE PROTEASE TRANSMEMBRANE PROTEIN
4N13 SMb20374 15140233 putative periplasmic substrate binding pro
5N01 SMb20829_15140657_putative_secreted_calcium_binding_protein
2B02 empty
10F02 emmroc12 Randomly generated negative control
3F01 SMa2211 14524357 ilvB2 probable IlvB2 acetolactate_synthase_
14A11 SMc02080 15073133 CONSERVED HYPOTHETICAL PROTEIN
5N13 SMb20836 15140664 hypothetical heme binding_protein
6B03 SMb21173 15140725 HYPOTHETICAL PROTEIN
13B15 SMc02773 15072949 PUTATIVE ABC TRANSPORTER ATP BINDING_PROT
4D12 SMb20182 15140054 putative ABC transporter protein
4I23 SMb20035 15139910 probable ABC transporter integral_membrane
4N14 SMb20269 15140131 CONSERVED HYPOTHETICAL PROTEIN
2J15 SMa1470 14523931 Putative AraC family transcription regulato
14B13 SMc03206 15076158 PUTATIVE PROTEASE TRANSMEMBRANE PROTEIN
15F02 SMc04136 15073044 PUTATIVE TRANSPORT SYSTEM PERMEASE_ABC_TR
5N13 SMb20836 15140664 hypothetical heme binding protein
2B03 SMa1363_14523867_putative_ABC_transporter_permease_protein
13B02 empty
4B03 SMb20700 15140150 HYPOTHETICAL PROTEIN
6B03 SMb21173 15140725 HYPOTHETICAL PROTEIN
8N02 SMc00356 15073205 genX PUTATIVE LYSYL TRNA SYNTHETASE PROTEI
12N03 SMc02372 15073959 PUTATIVE TRANSPORT TRANSMEMBRANE PROTEIN
5C04 SMb20500 15140363 putative ABC transporter_periplasmic_sugar
1J03 SMa0609 14523453 hypothetical protein
1J03 SMa0609 14523453 hypothetical protein
1B14 SMa0303 14523277 conserved hypothetical protein
12F13 SMc02325 15073561 PUTATIVE ATP BINDING ABC TRANSPORTER_PROT
3N14 SMa2099 14524305 conserved hypothetical_protein
2J02 SMa1283 14523814 Probable MarU protein
15N02 SMc04196 15075046 HYPOTHETICAL PROTEIN
1F15 SMa0653 14523429 hypothetical protein
3N02 SMa2087 14524298 hypothetical protein
3N15 SMa2325 14524421 probable transcriptional regulator
3N15 SMa2325 14524421 probable transcriptional regulator
1C04 SMa0217 14523180 putative ABC transporter permease
4C03 SMa2309 14524465 putative stress induced protein Ohr
10F14 SMc01222 15074502 lpsC PUTATIVE LIPOPOLYSACCHARIDE CORE BIO
10F14 SMc01222 15074502 lpsC PUTATIVE LIPOPOLYSACCHARIDE_CORE_BIO
14B15 SMc03207 15076159 PUTATIVE OXIDOREDUCTASE PROTEIN
3K14 SMa1872 14524179 Putative threonine dehydratase
2B15 SMa1373 14523876 Probable ABC transporter permease
```

Fig. 4. Selected Cluster-3 in pink and name of genes belongs to this cluster.

Analysis of Gene Expression pattern of co expressed gene

For gene expression analysis, expression values of genes clustered in a group were applied. The pattern showed that in given environmental conditions and time, the expressions of the genes changes according to the environment and time, and these fluctuations in expression patterns are similar in genes that are in same cluster.Figure 5 shows the expression of gene in given time and environmental condition, it was found that at some point it is very high for some gene. It shows that after clustering there is a chance of getting some false positive genes in cluster.

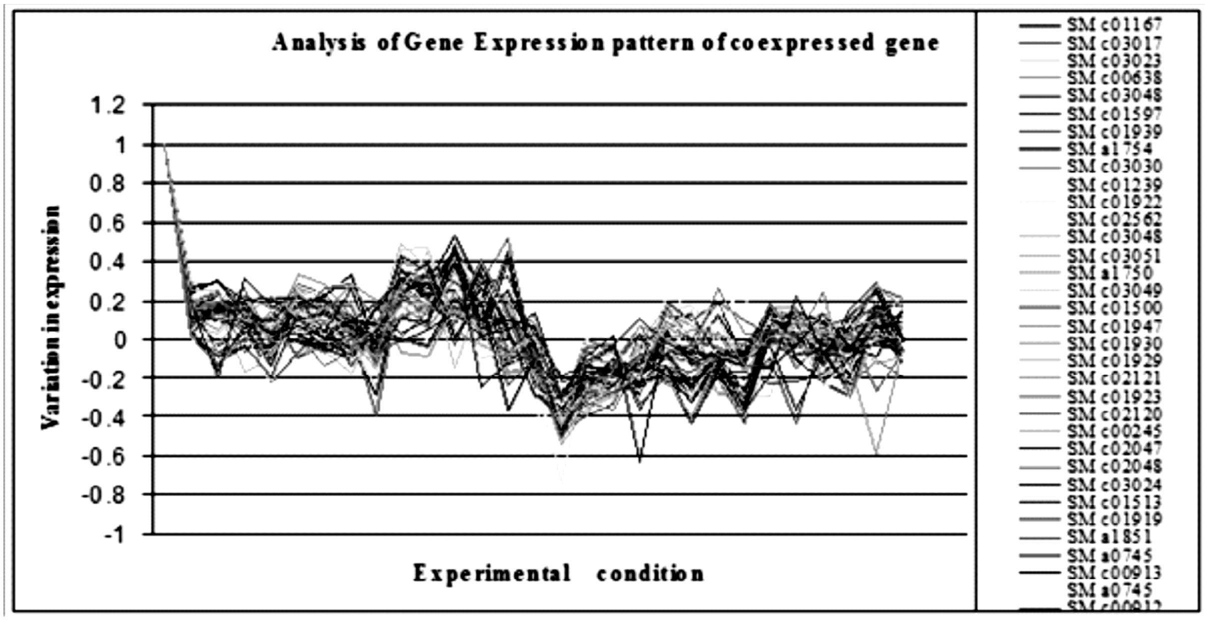


Fig. 5. The expression pattern of clustered genes in graph form, according to time and experimental (environmental) condition. Each gene is shown by different color line.

Phylogenetic Analysis of coexpressed genes

Phylogenetic analysis of the coexpressed genes, which are obtained from clustering method, was performed by CLUSTAL W. The phylogram was constructed which displayed that all the coexpressed genes are near about at the same distance which shows their amino acid sequence similarities. In phylogram there are four set of genes (*i*) genes gi|15073983 and gi|15073987 with distance value 0.10376 & 0.10494, (*ii*) genes gi|15965306 and gi|15964415 with distance value 0.02090 & 0.02561, (*iii*) genes gi|15073062 and gi|15963887 with distance value 0.00000, gene gi|15964378 also belongs the same parent group but different to these two genes with distance value 0.00153 and (*iv*) genes gi|15966751 and gi|16264618 with distance value 0.01705 & 0.01372 have nearly same sequence, same function and gene product, as compare to other genes (Phylogram not shown)

Analysis of function of some unknown genes

SMc04118 : The length of this gene is 584 bp (157283-157867bp). It is a product of conserved hypothetical transmembrane protein of length 194 amino acid. It's a "TadE" like protein. The members of this family are similar to a region of the protein product of the bacterial tadE locus. In various bacterial species, the tad locus is closely linked to flp-like genes, which encode proteins required for the production, cellular processes and signaling, intracellular trafficking, secretion, and vesicular transport. (**www.ncbi.nlm.nih.gov)**

SMb21204 : The length of this gene is 1112 bp (955955-957067bp). It codes for putative ABC transporter permease protein of length 370 amino acid. It's an ABC-2 type multidrug transporter which transports all of the small molecules in cell processes, signaling and defense mechanisms. (**www.ncbi.nlm.nih.gov)**

SMc03030: The length of this gene is 788 bp (726879-726091bp). It codes for flgG flagellar basal body rod protein FlgG of length 262 amino acid. This family of protein consists of a number of C-terminal domains which specific to flagellar basal-body rod and flagellar hook. It is ues in cellular processes, signaling and cell motility. (**www.ncbi.nlm.nih.gov)**

According to the assumption that coexpressed genes are related to same function, comparision of unknown genes with the known genes, which are present in same cluster, was done by gene expression pattern graph. By finding same expression pattern for these genes it can be assumed that these are related at their functional level and share same functional genomics.

Following are the graphs of comparison of gene expression patterns.

1. SMc04118 & SMc02319

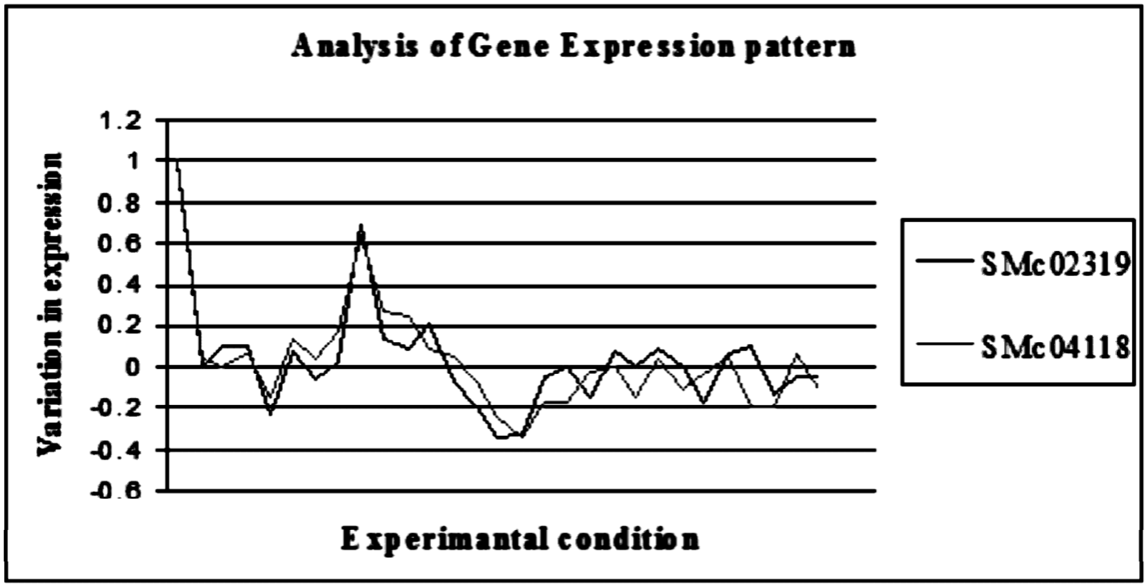


Fig. 6. Gene expression pattern graph, gene SMc02319 (unknown) show similarity in gene expression pattern with SMc04118 which code for conserved hypothetical transmembrane protein.

2. SMc03030 & SMc01239

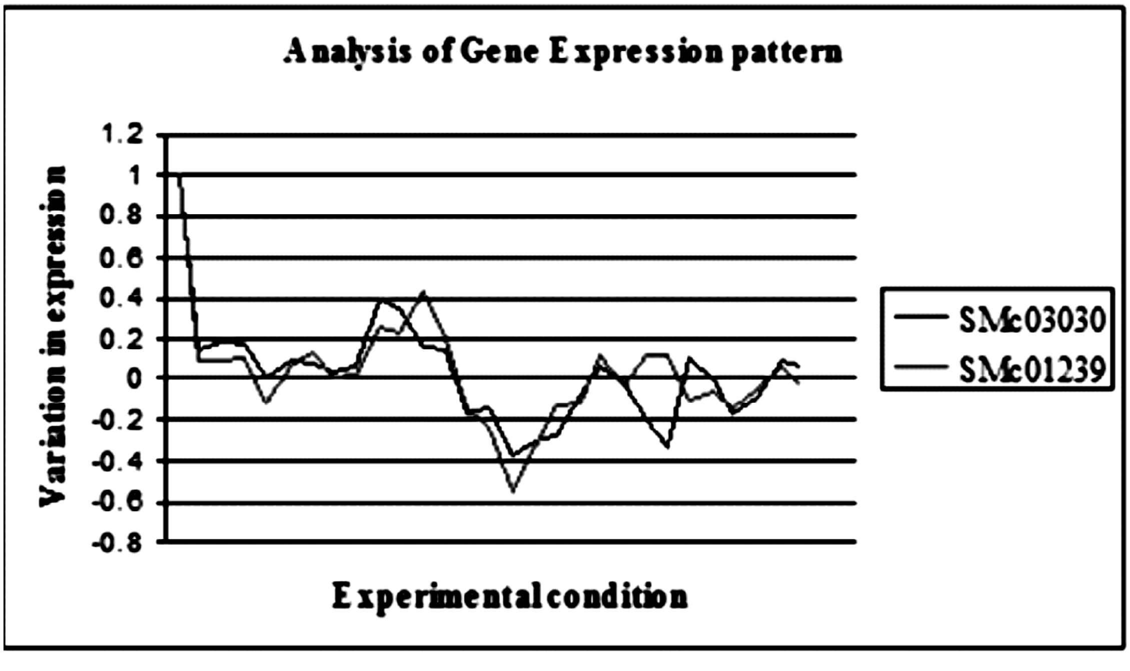


Fig. 7. Gene expression pattern graph, gene SMc1239 (unknown) show similarity in gene expression pattern with SMc03030 which code for flgG flagellar basal body rod protein FlgG.

3. SMB21204 & SMC03095

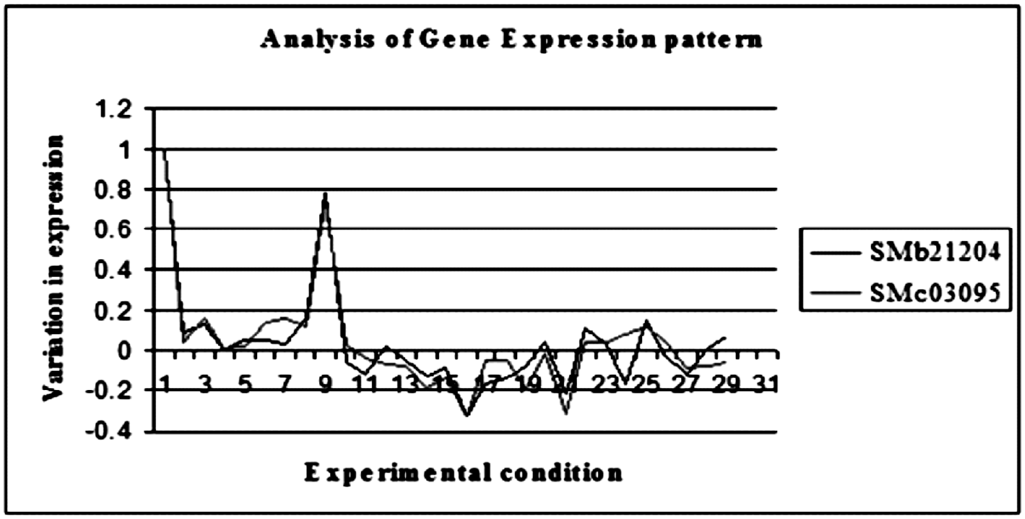


Fig. 8. Gene expression pattern graph, gene SMc03095 (unknown) show similarity in gene expression pattern with SMb21204 which code for putative ABC transporter permease protein.

In this present study hierarchical clustering analysis has been done by using microarray gene expression data of *Sinorhizobium meliloti strain 1021*on the basis of gene expression pattern. The genes- SMc02319, SMc01239 and SMc03095, whose functions are unknown shows similarity in gene expression pattern with SMc04118, SMc03030 and SMb21204 genes respectively. After that the analysis of phylogenetic relationship with the help of Clustal W has been performed. The phylogenetic output figures shows amino acid sequences of SMc04118 and SMc0231,SMc03030 and SMc01239, SMb21204 and SMc03095 have very similar sequence as score table shows. So, it can be inferred that the gene pairs are closely related to each other at sequence level. In which SMc02319, SMc01239 and SMc03095 genes functions are unknown. Thus it can be assumed that functions of these genes are related to SMc04118, SMc03030 and SMb21204 genes.

Genome-wide expression studies in several organisms, suchDrosophila,[5,6,7] nematode,[8,9,10] mouse,[11,12,13] human,[14,15,16] and Arabidopsis,[17,18,19] have recently showed thatgenes with similar expression levels are non randomly distributed within genomes and tend to cluster within genomicneighbourhoods.[20] Gene clusters are known to be prominent features of bacterial chromosomes. One of the most striking features of prokaryotic gene clusters is that typically they are composed of functionally related genes. [21]

The results from this study confirm that conserved gene clusters accurately convey functional coupling between the genes present in them. Also, a pairwise measure of co expression of genes clusters contains functionally related genes. Several studies support this notion by noting that some genes known to have similar functions were grouped together.[1,22] These tight clusters provide powerful indications that co-clustered genes of currently unknown function are probably sharing the same functionality.

Phylogenetic analysis of the coexpressed genes, displayed that all the coexpressed genes are near about at the same distance which shows their amino acidsequence similarities. In phylogram there are four set of genes (*i*) genes gi|15073983 and gi|15073987 with distance value 0.10376 & 0.10494, (*ii*) genes gi|15965306 and gi|15964415 with distance value 0.02090 & 0.02561, (*iii*) genes gi|15073062 and gi|15963887 with distance value 0.00000, gene gi|15964378 also belongs the same parent group but different to these two genes with distance value 0.00153 and (*iv*) genes gi|15966751 and gi|16264618 with distance value 0.01705 & 0.01372 have nearly same sequence, same function and gene product, as compare to other genes. This suggests the biological relevance of the generated clusters, because if coexpression clusters are nonfunctional and/or purely coincidental, then phylogenetic conservation will not be observed.This also emphasize the importance of comparative genomics to elucidate evolutionary constraints imposed on clusters of coexpressed genes.

4. CONCLUSION

This present study examined the co-expressed genes pattern and their phylogenetic relation according to gene expression pattern diversity in *Sinorhizobium meliloti strain 1021.* The results from this study concluded that closely related (co-express) genes are in the same cluster. Here by the help of gene expression pattern plot, some unknown genes SMc02319, SMc01239 and SMc03095 show significant similarity with known genes SMc04118 SMc03030 and SMb21204 with low distance value in phylogenetic analysis thus it can be easily assumed that probably, unknown genes shared same functionality with known genes.

5. REFERENCES

1. Eisen M.B, Spellman P.T, Brown P.O, Botstein D. "Cluster analysis and display of genome-wide expression patterns". *PNAS*. 1998; 95 14): 863-14868.

2. Ray S.S, Bandyopadhyaya S, Pal S, "Gene ordering in partitive clustering using microarray expressions". *J. Biosci* . 2007; 32:5: 1019- 1025.

3. Guo X, Flores M, Mavingui P, Fuentes SI, Hernandez G, Davila G, Palacios R. " Natural genomic design in sinoshizobiummeliloti: Novel genomic architectures." *Gen Res*. 2003; 13 (8): 1810- 1817.

4. Turlough M.F, Stefan W, Wong, K, Jens B, Patrick C, Frank J, Vand A.P. "The complete sequence of the 1,683-kb pSymBmegaplasmid from the N2-fixing endosymbiont Sinorhizobium meliloti." *PNAS*. 2001; 98 :17: 9889-9894

5. Spellman P.T, Rubin G.M. "Evidence for large domains of similarly expressed genes in the Drosophila genome" *J. Biol*. 2002; 1: 5.

6. BoutanaevA.M ,Kalmykova A.I, Shevelyov Y.Y, Nurminsky D.I. " Large clusters of co-expressed genes in the Drosophila genome". *Nature*. 2002; 420: 666–669

7. Kalmykova A.I , Nurminsky D.I , Ryzhov D.V, Shevelyov Y.Y. "Regulated chromatin domain comprising c luster of co-expressed genes in Drosophila melanogaster," *Nucleic Acids Res*. 2005; 33:1435–1444

8. Lercher M.J, Blumenthal T, Hurst L.D. "Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes". *Genome Res*. 2003; 13: 238–243.

9. Roy P.J, Stuart J.M, Lund J, Kim S.K. "Chromosomal clustering of muscle expressed genes in *Caenorhabditis elegans*," *Nature* 2002; 418 :975–979.

10. Miller M.A, Cutter A.D, Yamamoto I, Ward S, Greenstein D. "Clustered organization of reproductive genes in the C. elegans genome". *Curr. Biol*. 2004; 14 :1284–1290

11. Williams E.J, Hurst L.D. "Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes". *J. Mol. Evol*. 2002; 54:511–518

12. Singer G.A, Lloyd A.T, Huminiecki L.B, Wolfe K.H. "Clusters of coexpressed genes in mammalian genomes are conserved by natural selection". *Mol. Biol. Evol*. 2005; 22:767–775.

13. Sémon, M.L. Duret,. "Evolutionary origin and maintenance of coexpressed gene clusters in mammals". *Mol. Biol. Evol.* 2006*;* 23: 1715–1723.

14. Purmann A, Toedling J, Schueler M, Carninc P, Lehrach H, Haayashizaki Y, Huber W, Sperling S. "Genomic organization of transcriptomes in mammals: coregulation and cofunctionality," *Genomics*. 2007; 89 : 580-587

15. Lercher M.J, Urrutia A.O. Hurst L.D, "Clustering of housekeeping genes provides a unified model of gene order in the human genome". *Nat. Genet*. 2002; 31: 180–183.

16. Caron *H*, van Schaik B, van der Mee M. "The human transcriptome map: clustering of highly expressed genes in chromosomal domains". *Science*. 2002; 291: 1289–1292

17. Williams E.J, Bowles D.J. "Coexpression of neighboring genes in the genome of Arabidopsis thaliana". *Genome Res.* 2004; 14: 1060–1067.

18. Ren X.Y, Fiers M.W, Stiekema W.J, Nap J.P. "Local coexpression domains of two to four genes in the genome of Arabidopsis". *Plant Physiol*. 2005; 138:923–934.

19. Zhan S, Horrocks J, Lukens L.N. "Islands of co-expressed neighbouring genes in Arabidopsis thaliana suggest higher-order chromosome domains." *Plant J*. 2006; 45: 347–357.

20. P Michalak, "Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes". *Genomics*. 2008; 91: 243–248

21. R Overbeek , M Fonstein, , M D'Souza, , GD Pusch, N Maltsev. "The use of gene clusters to infer functional coupling". *Proc. Natl. Acad. Sci.* 1999; 96: 2896–2901

22. Heyer, L.J. Kruglyak, S, Yooseph. S "Expression Data: Identification and Analysis of Coexpressed Genes." *Gen Res.* 1999; 9:1106–1115.