PatGeneClus: Supplementary Materials

1. Working of Dclique: An Example:

In our approach we consider the genes as nodes of a graph. If the similarity between a pair of genes is more than a user given threshold value, (represented by simTH) then we assume an edge between the nodes represented by the two genes. We assume that our similarity matrix is as given in Fig. 1 and say the simTH=60. Then, the graph will look like the graph given in Fig 1. In a graph a clique is a sub graph in which all vertices are connected to all other vertices. Hence from our graph if we extract a clique, we will get a subset of the given the genes which are similar to each other more than simTH. Our clustering algorithm is based on this concept. To extract these cliques from the graph under consideration, we consider *mark* as a Boolean attribute which if true indicates that a vertex wont be considered for further processing. Initially for all vertices this attribute is false. Here we consider g1 as the first gene. g_1 is not marked. Then we iterate over the rest of the genes to find another gene g_x such that g_1, g_x edge is not there in any cluster. Say g_x is g_2 in the graph as shown in Fig. 2 (a). We start from g_2. Genes g_1 and g_2 are not there in any cluster. So, we form a new cluster and add g_1 and g_2 to it as shown in Fig. 2 (b). Then for the rest of the unmarked genes we check if they can be added to this cluster or not. A gene is added to this cluster if it is connected to the rest of the genes present in the cluster. Since g_3 is connected to both g_1 and g_2 , we add it to the cluster as shown in Fig 2 (c). Similarly g_4 is added. Genes g_5 and g_6 wont be added since they do not meet the requirement. Since no other genes are there, we get the first cluster as g_1, g_2, g_3, g_4 depicted in Fig. 3. We repeat the process. We again consider g_1 as the first gene. g1 is not marked.Then we iterate over the rest of the genes to find another gene g_x such that g_1, g_x edge is not there in any cluster. We start from g_2. g_1, g_2 is there in cluster 1. We iterate to g_3. g_1, g_3 is also there in cluster1. Similarly g_1, g_4 is also there in cluster1. This iteration is shown in Fig. 4 (a), (b) and (c). For gene g_5, we see that g_1, g_5 is not there in any cluster. Hence, we create a new cluster and add g_1 and g_5 to it as shown in Fig. 5(a). Next, we iterate over the list of genes once again to add genes to this cluster.Genes g_2 and g_3 cant be added because they do not meet the requirement of a clique *i.e.,* there are no edges between g_5, g_2 and g_3, g_5. Gene g_4 satisfies the requirement and therefore we add it to the current cluster as given in Fig. 5(b). Gene g_6 is then again not added to the current cluster as it doesnot satisfy the requirement. We iterate over all genes. Finally, we add this current cluster to the list of generated clusters.

Thus we have, $C_1 = g_1, g_2, g_3, g_4$ and $C_2 = g_1, g_4, g_5$

We again repeat the process. Consider g_1 as the first gene. g_1 is not marked. We then iterate over the rest of the genes to find another gene g_x such that g_1, g_x edge is not there in any cluster. We start from g_2. g_1, g_2 is already there in a previous cluster. Similarly, g_3, g_4 and g_5. Since no new cluster can be found starting with g_1, it is marked and shown in red in Fig. 5(c). Then we consider g_2 as the first gene and try to find g_x, such that g_2, g_x is not there in any of the previous cluster. It can be seen that for g_2, g_3 and g_4 also no g_x can be found. Hence they are also marked as shown in Fig. 6(a). Then we consider g_5. For g_5, when we try to find g_x, we skip g_1, g_2, g_3, g_4 as they are marked. For gene g_6, edge g_5, g_6 is not there in any cluster. So we create a new cluster and add g_5, g_6 to it. Since no more genes are there we get another cluster $C_3 = g_5, g_6$ and depicted in Fig. 6(b). Then the process is repeated.

That is, we again consider g_1 as the first gene. Genes g_1, g_2, g_3, g_4 are marked and hence we skip them. Gene g_5 is not marked. For g_5 we try to find g_x such that edge g_5, g_x is not there in any cluster. No g_x can now be found. Therefore, mark g_5. Then for g_6 also no cluster can be formed so we mark it. All genes are now marked as shown in Fig. 6(c) and hence we stop the algorithm.

Practical implementation of this approach has the following problem: The number of cliques present may be very large. In such a case we may end up with a set of almost similar cliques. For example the example graph has two almost similar cliques. To prevent generating such similar cliques, we use another threshold value, $relaxTH$. In our approach if all nodes in a sub graph is connected to $relaxTH\%$ of other nodes, we consider it as a clique. To further eliminate the effect of generating almost similar cluster, we do a merging step once a cluster is generated.

	g1	g2	g3	g4	g5	g6
g1	100	60	70	65	73	35
g2		100	60	80	25	33
g3			100	70	10	10
g4				100	66	03
g5					100	80
g6						100

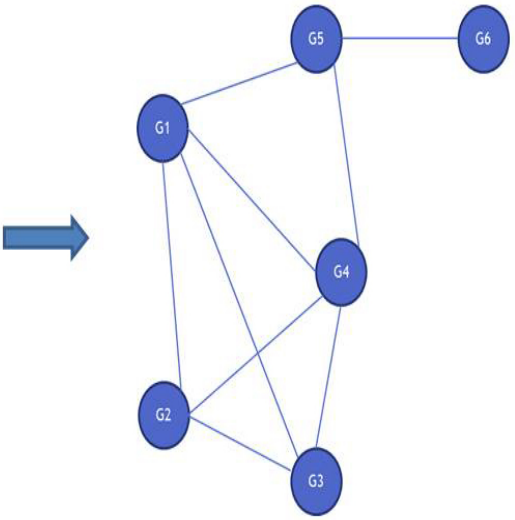


Figure 1: The example similarity matrix and its corresponding graph

2. PWCTM: Similarity to Distance Conversion

This section presents the proofs of different metric properties of PWCTM measure and establishes the various advantages of it. We can convert our similarity measure into a distance measure as $(d(g, g') = 1 - s(g, g'))$. Now, we shall show that $(d(g, g')$ gives a distance function for the genes. i.e., $(d(g, g')$ should satisfy the properties of a distance metric. Next, we show that our measure follow the distance properties.

Property 1. *: Non-negativity*
To satisfy non-negativity property, PWCTM distance (d) of two genes should be always greater than zero, $d(g, g') \geq 0$

Proof. We know that $0 \leq s(g, g') \leq 1$. Therefore, $0 \leq d(g, g') \leq 1$ □

Property 2. *: Identity*
To satisfy identity property, the distance between identical genes g, g' *should be equal to zero.* $d(g, g') = 0$ *iff* $g = g'$

Proof. $s(g, g') = 1$ iff $g = g'$. Therefore, $d(g, g') = 1 - 1 = 0$ and hence the proof. □

Property 3. *: Symmetricity*
To satisfy the symmetricity property, for any two genes g_1 *and* g_2, $d(g_1, g_2)$ *should be equal to* $d(g_2, g_1)$ *, i.e.,* $d(g_1, g_2) = d(g_2, g_1)$.

Proof.

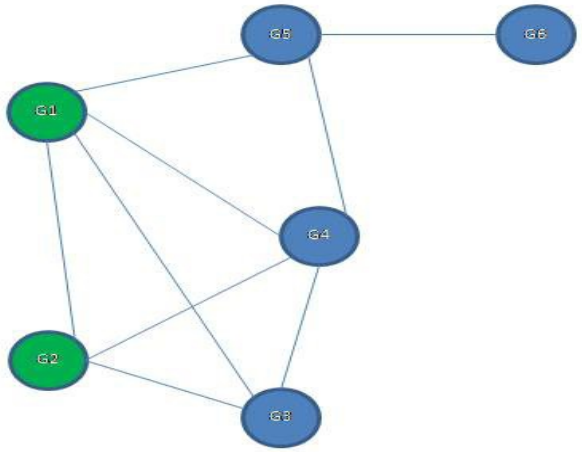
$$d(g_1, g_2) = 1 - s(g_1, g_2) = 1 - \frac{\sum_{k=1}^{q}(w_k \times \chi_k^{g_1,g_2})}{\sum_{k=1}^{q} w_k} \tag{1}$$

$$d(g_2, g_1) = 1 - s(g_2, g_1) = 1 - \frac{\sum_{k=1}^{q}(w_k \times \chi_k^{g_2,g_1})}{\sum_{k=1}^{q} w_k} \tag{2}$$

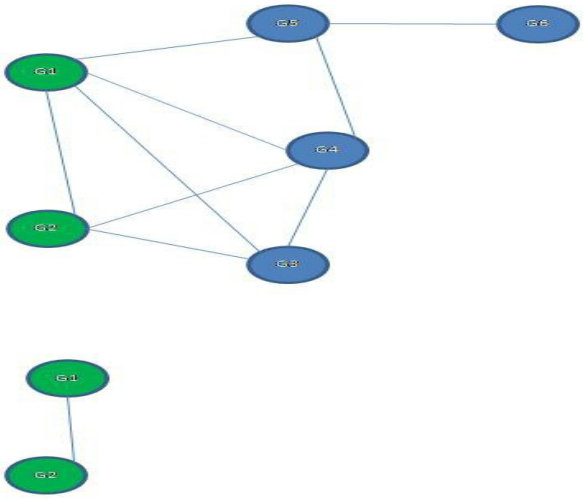
where, w_k is weight at k^{th} condition and q is the total number of possible pairs of conditions as explained in section 15. We know,

$$\chi_k^{g_1,g_2} = \chi_k^{g_2,g_1} = \begin{cases} 1 & \text{if } CT_{g_1}(k) = CT_{g_2'}(k), k = 1, \cdots, q \\ 0 & \text{otherwise.} \end{cases}$$

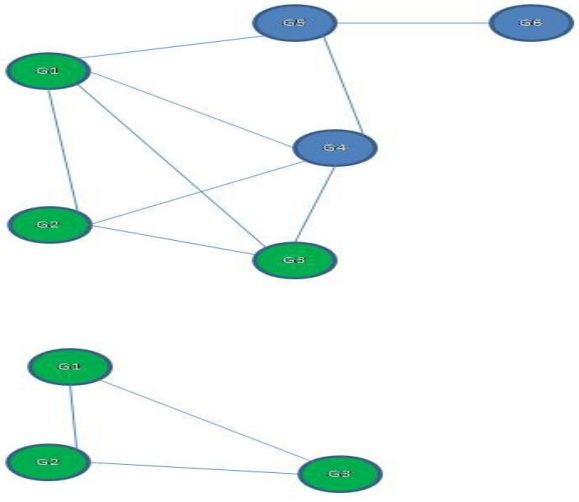
Then, from equation 19 and 2, we get, $d(g_1, g_2) = d(g_2, g_1)$ (Denominator of both equations are same) and hence the proof. □



(a)



(b)



(c)

Figure 2: (a)Genes g_1 and g_2 under consideration. (b) Genes g_1 and g_2 added to cluster as shown by the graph below. (c) Gene g_3 added to cluster.

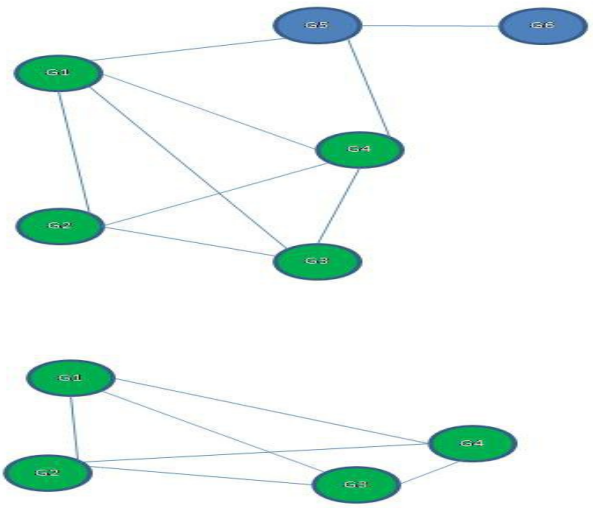


Figure 3: The first cluster

Property 4. *: Subadditivity or Triangle Inequality*
To satisfy triangular inequality property, for any three genes g_1, g_2 and g_3, the following condition should hold: $d(g_1, g_2) \leq d(g_1, g_3) + d(g_3, g_2)$

Proof. : Let,

$$\chi_k^{i,j} = \begin{cases} 1 & \text{if } CT_{g_i}(k) = CT_{g_j}(k), k = 1, \cdots, q \\ 0 & \text{otherwise.} \end{cases}$$

We shall prove that,

$$1 + \chi_k^{1,2} \geq \chi_k^{1,3} + \chi_k^{3,2} \; for \; k = 1, \cdots, q \tag{3}$$

if $CT_{g_1}(k) \neq CT_{g_2}(k), \chi_k^{1,2} = 0$ and either $CT_{g_3}(k) \neq CT_{g_1}(k)$ or $CT_{g_3}(k) \neq CT_{g_2}(k)$.

Hence, either $\chi_k^{1,3} = 0$ or $\chi_k^{2,3} = 0$. Thus equation 3 is satisfied. Otherwise, if $CT_{g_1}(k) = CT_{g_2}(k), \chi_k^{1,2} = 1$.

Therefore, $1 + 1 = 2 \geq \chi_k^{1,3} + \chi_k^{2,3}$, i.e., equation 3 is satisfied in this case also equation 3 is satisfied in all cases.

Multiplying equation 3 by w_k and summing over k we have,

$$\sum_{k=1}^{q} w_k + \sum_{k=1}^{q} w_k \times \chi_k^{1,2} \geq \sum_{k=1}^{q} w_k \times \chi_k^{1,3} + \sum_{k=1}^{q} w_k \times \chi_k^{3,2} \tag{4}$$

But, $\sum_{k=1}^{q} w_k \times \chi_k^{1,2} = \sum_{k=1}^{q} w_k \times s(g_1, g_2)$, $\sum_{k=1}^{q} w_k \times \chi_k^{1,3} = \sum_{k=1}^{q} w_k \times s(g_1, g_3)$ and $\sum_{k=1}^{q} w_k \times \chi_k^{3,2} = \sum_{k=1}^{q} w_k \times s(g_3, g_2)$.

Therefore, from equation 4 we have,

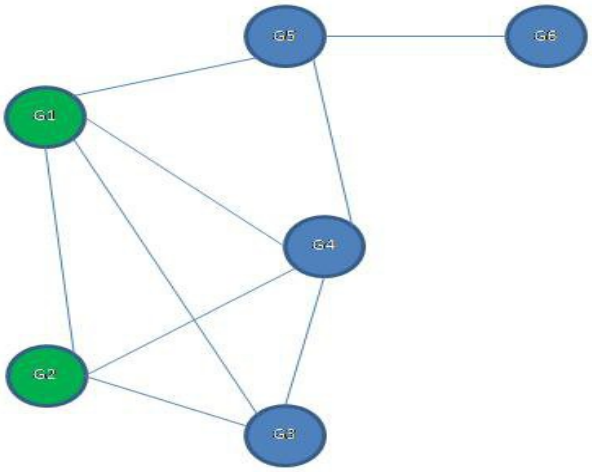
$$\sum_{k=1}^{q} w_k(1 + s(g_1, g_2)) \geq (\sum_{k=1}^{q} w_k)(s(g_1, g_3) + s(g_3, g_2))$$

Hence,

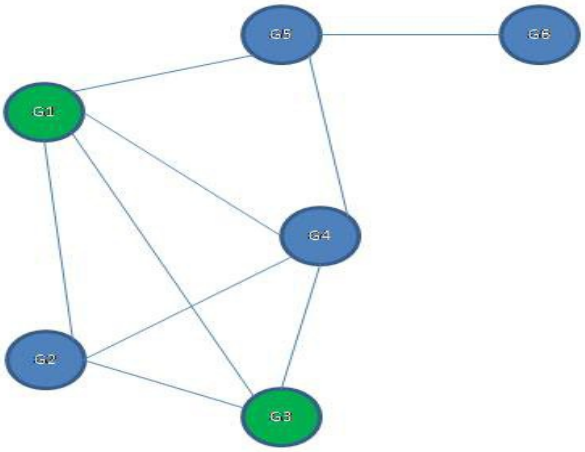
$$1 - s(g_1, g_2)) \leq (1 - s(g_1, g_3)) + (1 - s(g_3, g_2))$$

Therefore, $d(g_1, g_2) \leq d(g_1, g_3) + d(g_3, g_2)$ and hence the triangle inequality is proved. □

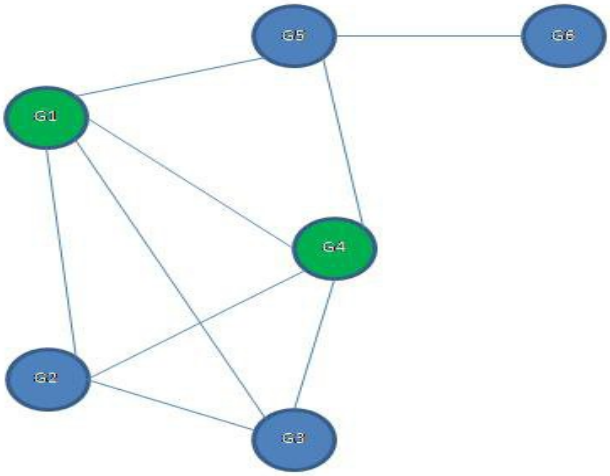
Thus, $d(g, g')$ defines a distance function and the similarity function $s(g, g') = 1 - d(g, g')$ is based on this underlying distance function.



(a)

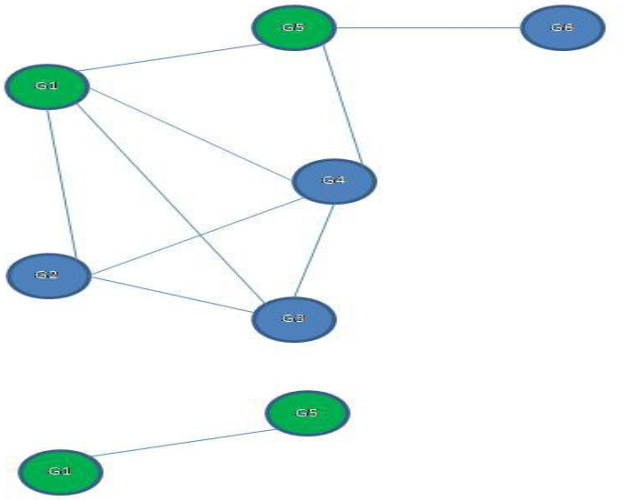


(b)

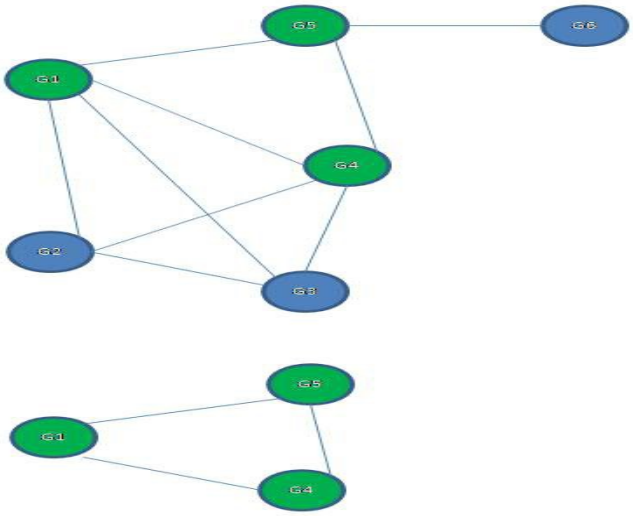


(c)

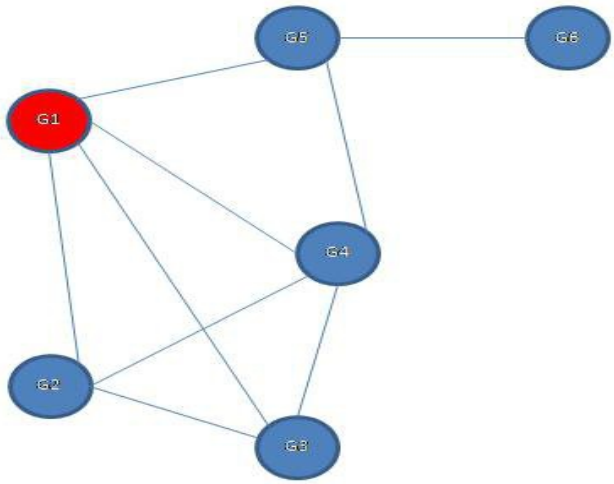
Figure 4: (a)Genes g_1 and g_2 under consideration. (b) Genes g_1 and g_3 under consideration. (c) Genes g_1 and g_4 under consideration. None is added to cluster.



(a)

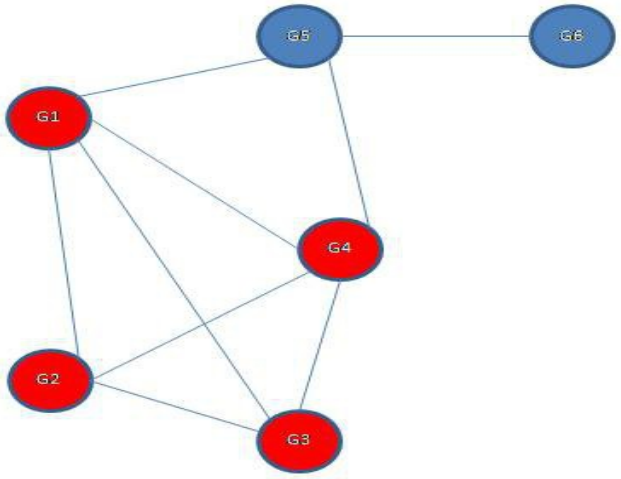


(b)

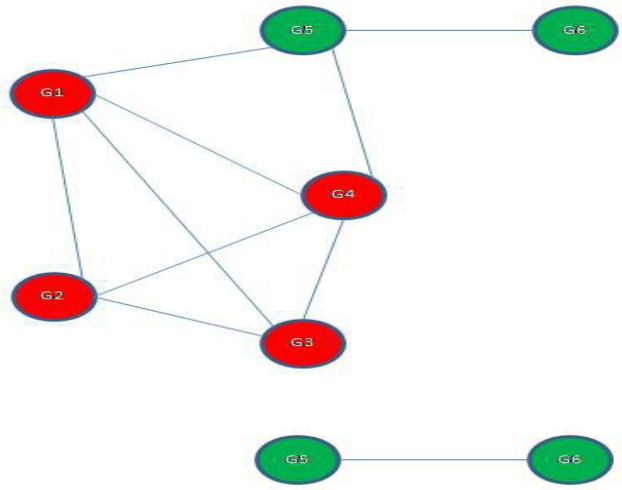


(c)

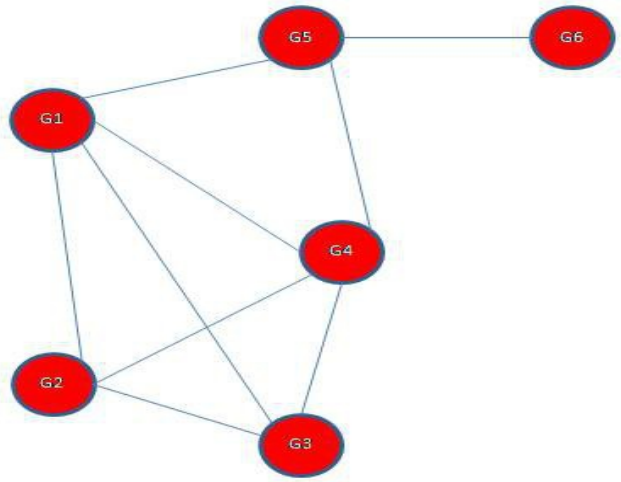
Figure 5: (a)New cluster with g_1 and g_5. (b) Gene g_4 added to cluster. (c) Gene g_1 is marked and will not be considered further.



(a)



(b)



(c)

Figure 6: (a)Genes g_2, g_3 and g_4 are marked and will not be considered further. (b) Cluster C_3 with genes g_5, g_6 is formed. (c) All genes are now marked

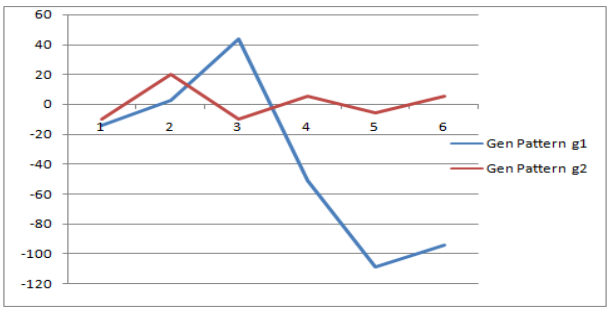


Figure 7: The original patterns of gene g_1, g_2.

3. PWCTM is not affected by normalization of data

According to [1], "Normalization allows comparing different experiments based on the same application independent from the scale of the features". In many applications, normalization is used as a pre-processing step. Various normalization procedures are available such as the min/max normalization, the z-transformation, the log transformation, Student's t-statistic, and the rank transformation [1, 2]. Our similarity measure, PWCTM, is unaffected by normalization performed on the data. To prove this we have taken two normalization procedures: min-max normalization and the z-transformation.

3.1. Min-Max normalization

Min-Max normalization is the process of taking data measured in any engineering units (for example: miles per hour, degrees C, etc) and transforming it to a value between 0.0 and 1.0. The minimum (lowest) value is set to 0.0 and the maximum (highest) value is set to 1.0. Thus using this normalization it becomes easy way to compare values measured in different scales or different units of measure. The min-max normalization is obtained as given in equation 5. A gene sequence g_i is a sequence $(g_{i1}, g_{i2}, \cdots, g_{in})$ of length n, where n is the number of conditions and g_{ij} is the expression value of gene g_i at condition j. The min-max normalized expression value g_{ij} is given by,

$$N_{ij} = \frac{(g_{ij} - Min_g)}{(Max_g - Min_g)} \tag{5}$$

where, Min_g and Max_g are the minimum and maximum value of the whole gene expression dataset under consideration. To show the effect of min-max normalization on PWCTM similarity measure, we consider a pair of genes g_1 and g_2 as illustrated in Fig. 7 with expression values as given in Table 1 and find the PWCTM similarity value on the original data. Then, we will compute the PWCTM similarity value on the normalized data. If both the values are same, we can say that our measure is not affected by normalization. The changing tendency (CT) of genes g_1, g_2 alongwith their $\chi_k^{g_1,g_2}$ and corresponding weights (Wts) for each of the possible pairs of conditions, (q), is shown in Table 2. For this example, $q = n \times (n-1)/2 = 6 \times 5/2 = 15$.

Table 1: Original Data

g_1	-14	3	44	-51	-109	-94
g_2	-10	20	-10	5	-6	5

From Table 2, we calculate the Total weight as 8.69 and score between the two gene pairs as,

$$\begin{aligned} score(g_1, g_2) &= 1*1\ 0*0.5\ 0*0.33\ 0*0.25\ 0*0.2\ 0*1\ 1*0.5 \\ &\quad 1*0.33\ 1*0.25\ 0*1\ 0*0.5\ 0*0.33\ 1*1 \\ &\quad 0*0.5\ 1*1 \\ &= 4.08. \end{aligned}$$

Table 2: CT, $\chi_k^{g_1,g_2}$ and weight (Wts) Values for the 15 pairs of conditions on the Original Data of Table 1

	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
CT of g_1	U	U	D	D	D	U	D	D	D	D	D	D	D	D	U
CT of g_2	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
$\chi_k^{g_1,g_2}$	1	0	0	0	0	0	1	1	1	0	0	0	1	0	1
Wts	1	0.5	0.33	0.25	0.2	1	0.5	0.33	0.25	1	0.5	0.33	1	0.5	1

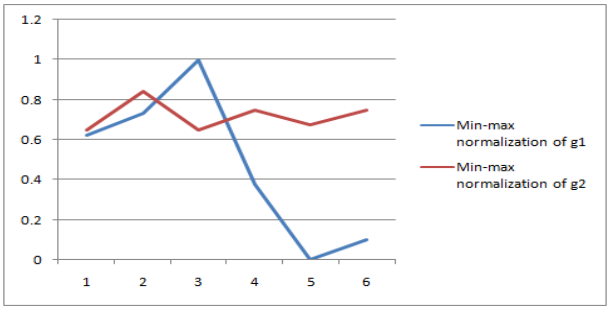


Figure 8: The Min-max normalized patterns of gene g_1, g_2.

Therefore, the PWCTM similarity between the two gene profiles on the original data is found as follows.

$$s(g_1, g_2) = 4.08/8.69 = 0.469505$$

The normalized values of g_1, g_2 after performing min-max normalization is given in Table 5 and their expression profiles are illustrated in Fig. 8. The changing tendency (CT) of genes g_1, g_2 alongwith their $\chi_k^{g_1,g_2}$ and corresponding

Table 3: Min-Max Normalized Data

g_1	0.620915033	0.732026144	1	0.379084967	0
g_2	0.647058824	0.843137255	0.647058824	0.745098039	0.673202614	0.745098039

weights (Wts) for each of the possible pairs of conditions, (q), is shown in Table 4. From this Table 4, we see that the values remain the same as obtained from the original data and given in Table 2. Thus, as obtained on the previous

Table 4: CT, $\chi_k^{g_1,g_2}$ and weight (Wts) Values of the 15 pairs of conditions

	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
CT of g_1	U	U	D	D	D	U	D	D	D	D	D	D	D	D	U
CT of g_2	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
$\chi_k^{g_1,g_2}$	1	0	0	0	0	0	1	1	1	0	0	0	1	0	1
Wts	1	0.5	0.33	0.25	0.2	1	0.5	0.33	0.25	1	0.5	0.33	1	0.5	1

occassion when PWCTM was applied on the original data, we obtain the same similarity value between the two gene

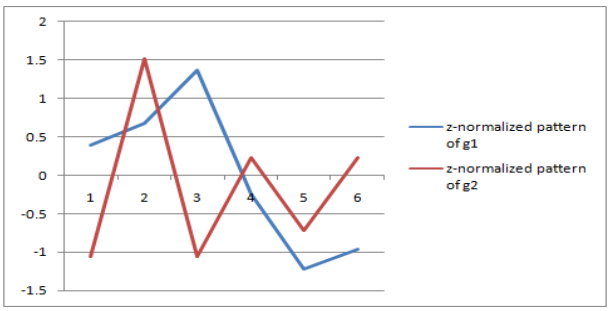


Figure 9: The z-normalized patterns of gene g_1, g_2.

pairs after min-max normalization.

$$\begin{aligned} score(g_1, g_2) &= 1*1\ 0*0.5\ 0*0.33\ 0*0.25\ 0*0.2\ 0*1\ 1*0.5 \\ &\quad 1*0.33\ 1*0.25\ 0*1\ 0*0.5\ 0*0.33\ 1*1 \\ &\quad 0*0.5\ 1*1 \\ &= 4.08. \end{aligned}$$

Therefore, PWCTM similarity between the two normalized gene profiles is also found to be same as given below.

$$s(g_1, g_2) = 4.08/8.69 = 0.469505$$

3.2. z-normalization

Another simple way of normalizing vectors is variance and mean normalization, also known as the z-transformation. The z-value of a given observation is the (signed) number of standard deviations an observation or data value is above/below the mean. A positive z-value will represent that the data value is above the mean, while a negative z-value represents the data value is below the mean. It is a dimensionless quantity and is obtained as given in equation 6. A gene sequence g_i is a sequence $(g_{i1}, g_{i2}, \cdots, g_{in})$ of length n, where n is the number of conditions and g_{ij} is the expression value of gene g_i at condition j. The z-value of an expression value g_{ij} is given by,

$$z_{ij} = \frac{g_{ij} - \mu_i}{\sigma_i} \tag{6}$$

where, μ_i and σ_i are the mean and standard deviation of gene g_i across the j conditions.

If we are just looking for profile similarity, that is the shape of the lines, normalization prior to distance calculation is appropriate and allows a simple distance measure (e.g. Euclidean) to be used. However, if the absolute value of the vector has some meaning, this will be lost after variance normalization.

Here, we will show that our PWCTM measure is not affected by z-normalization. The z-normalized data of the two gene pairs given in Table 1 is given in Table 5 and their expression profiles are shown in Fig. 9. As can be seen

Table 5: z-Normalized Data

g_1	0.387967193	0.67681868	1.3734605	-0.240709572	-1.22620288	-0.971333921
g_2	-1.054992005	1.511204764	-1.054992005	0.228106379	-0.712832436	0.228106379

from Table 6, the changing tendency (CT) of the z-normalized genes g_1, g_2 alongwith their $\chi_k^{g_1,g_2}$ and corresponding weights (Wts) for each of the possible pairs of conditions, (q), we find that they are the same as obtained from the original data and given in Table 2.

$$\begin{aligned} score(g_1, g_2) &= 1*1\ 0*0.5\ 0*0.33\ 0*0.25 \\ &\quad 0*0.2\ 0*1\ 1*0.5\ 1*0.33 \\ &\quad 1*0.25\ 0*1\ 0*0.5\ 0*0.33 \\ &\quad 1*1\ 0*0.5\ 1*1 \\ &= 4.08 \\ PWCTM\ value, s(g_1, g_2) &= 4.08/8.69 = 0.469505. \end{aligned}$$

Table 6: CT, $\chi_k^{g_1,g_2}$ and weight (Wts) Values of the 15 pairs of conditions

	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
CT of g_1	U	U	D	D	D	U	D	D	D	D	D	D	D	D	U
CT of g_2	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
$\chi_k^{g_1,g_2}$	1	0	0	0	0	0	1	1	1	0	0	0	1	0	1
Wts	1	0.5	0.33	0.25	0.2	1	0.5	0.33	0.25	1	0.5	0.33	1	0.5	1

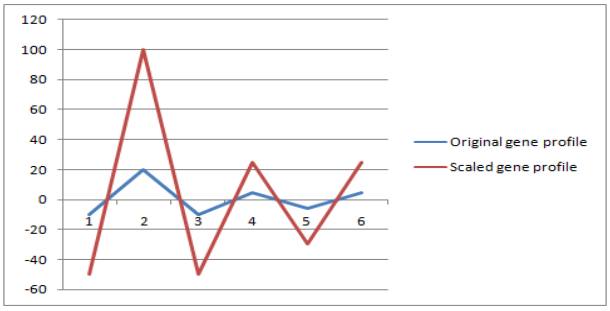


Figure 10: The original and the scaled patterns of gene g_2.

Thus, as obtained on the previous occassion when PWCTM was applied on the original data, we obtain the same similarity value between the two gene pairs after z-normalization. Therefore, we can now say that our PWCTM measure is unaffected by normalization.

4. PWCTM is not affected by transformation

Our proposed PWCTM measure is not affected by scaling, shifting and shift-scale transformations. To establish this fact, we take the gene sequence g_2 of Table 1. We then scale this gene profile by a factor of 5 to obtain the gene profile g_2', shift it by a factor of 50 to obtain the gene profile g_2'' and scale-shift the original by first scaling it by 5 and then shifting it by 50 to obtain g_2''' as shown in Table 7. The gene profiles are depicted in Fig. 10, 11 and 12. On obtaining the changing tendency of original profile g_2, scaled profile g_2', shifted profile g_2'' and scale-shifted pattern g_2''', we observe from Table 8 that their values are same and therefore their χ-values are also same. Computing PWCTM

Table 7: Original, Scaled and Shifted Data

g_2	-10	20	-10	5	-6	5
g_2'	-50	100	-50	25	-30	25
g_2''	40	70	40	55	44	55
g_2''	0	150	0	75	20	75

similarity measure between original profile g_2 and scaled pattern g_2', we observe that their PWCTM similarity value is 1 as given in equation 7. Therefore, both g_2 and g_2' have the same similarity value of 1 which proves that they are

Table 8: CT, $\chi_k^{g_1,g_2}$ and weight (Wts) Values of the 15 pairs of conditions

	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(2,3)	(2,4)	(2,5)	(2,6)	(3,4)	(3,5)	(3,6)	(4,5)	(4,6)	(5,6)
CT of g_2	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
CT of g_2'	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
CT of g_2''	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
CT of g_2'''	U	E	U	U	U	D	D	D	D	U	U	U	D	E	U
$\chi_k^{g_2,g_2'}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\chi_k^{g_2,g_2''}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$\chi_k^{g_2,g_2'''}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Wts	1	0.5	0.33	0.25	0.2	1	0.5	0.33	0.25	1	0.5	0.33	1	0.5	1

they follow the same pattern even after scaling transformation.

$$\begin{aligned} score(g_2, g_2') &= 1*1\ 1*0.5\ 1*0.33\ 1*0.25 \\ &\quad 1*0.2\ 1*1\ 1*0.5\ 1*0.33\ 1*0.25 \\ &\quad 1*1\ 0*0.5\ 1*0.33\ 1*1\ 1*0.5\ 1*1 \\ &= 8.69 \\ PWCTM\ value, s(g_2, g_2') &= 8.69/8.69 = 1. \end{aligned} \tag{7}$$

Also, we see from equation 8, the PWCTM similarity measure between original profile g_2 and shifted pattern g_2'' is 1. Thus, both g_2 and g_2'' have the same similarity value of 1 and follow the same pattern even after shifting transformation.

$$\begin{aligned} score(g_2, g_2'') &= 1*1\ 1*0.5\ 1*0.33\ 1*0.25 \\ &\quad 1*0.2\ 1*1\ 1*0.5\ 1*0.33 \\ &\quad 1*0.25\ 1*1\ 0*0.5\ 1*0.33 \\ &\quad 1*1\ 1*0.5\ 1*1 \\ &= 8.69 \\ PWCTM\ value, s(g_2, g_2'') &= 8.69/8.69 = 1. \end{aligned} \tag{8}$$

Similarly, from equation 9, the PWCTM similarity measure between original profile g_2 and scale-shifted pattern g_2''' is 1. Thus, both g_2 and g_2''' have the same similarity value of 1 and follow the same pattern even after scale-shift transformation.

$$\begin{aligned} score(g_2, g_2''') &= 1*1\ 1*0.5\ 1*0.33\ 1*0.25 \\ &\quad 1*0.2\ 1*1\ 1*0.5\ 1*0.33 \\ &\quad 1*0.25\ 1*1\ 0*0.5\ 1*0.33 \\ &\quad 1*1\ 1*0.5\ 1*1 \\ &= 8.69 \\ PWCTM\ value, s(g_2, g_2''') &= 8.69/8.69 = 1. \end{aligned} \tag{9}$$

Therefore, the PWCTM measure remains unchanged when scaling, shifting and scale-shift transformations are applied to the original pattern.

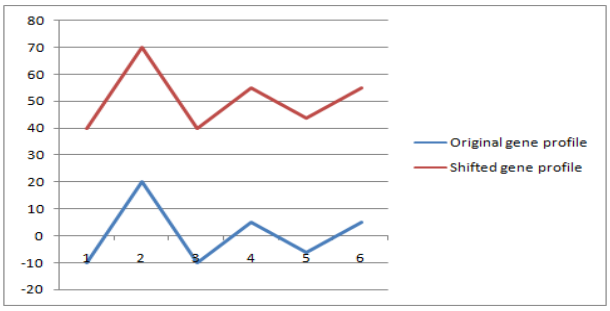


Figure 11: The original and the shifted patterns of gene g_2.

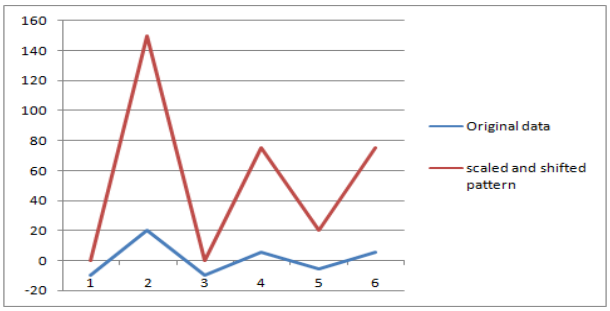


Figure 12: The original and the scale-shifted patterns of gene g_2.

Table 9: Datasets used for evaluating PWCTM

Serial No.	Dataset	No. of genes	No of con-di-tions	Source
1	Yeast Sporula-tion	474	7	http://cmgm.stanford .edu/pbrown/sporul ation/index.html
2	Subset of Yeast Cell Cycle [3]	384	17	http://faculty.washin gton.edu/kayee /cluster
3	Yeast Diauxic Shift [4]	6089	7	http://www.ncbi.nlm .nih.gov/geo/query
4	Asbestos treat-ment of human lung cancer cells [5]	37866	5	http://0-www.ncbi.nlm.nih. gov.opac.acc.msmc. edu/geo/query/acc. cgi?acc=GSE6013

5. Results

5.1. K-means Clustering Results

The result of the k-means clustering with k=4 (taking the best of 50 repetitions) using PWTCM and PCC are illustrated in Fig. 13 and Fig. 14 for Dataset 2 and Fig. 15 and Fig. 16 for Dataset 3. On applying k-means

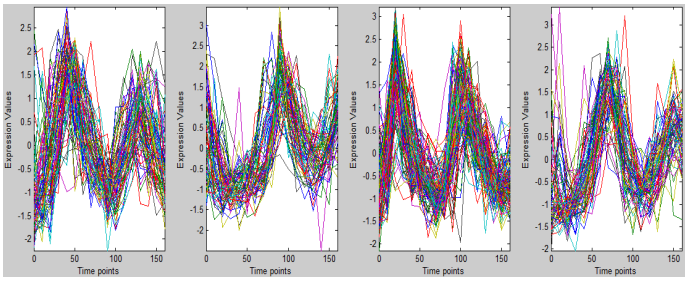


Figure 13: The clusters detected using k-means with PWCTM as the similarity measure for k=4 for Dataset 2. The vertical axes denote expression values of the genes and horizontal axes denote the 17 time points during the yeast cell cycle dataset.

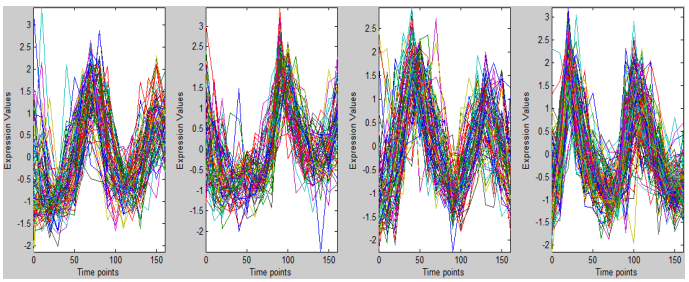


Figure 14: The clusters detected using k-means with Pearson correlation as the similarity measure for k=4 for Dataset 2. The vertical axes denote expression values of the genes and horizontal axes denote the 17 time points during the yeast cell cycle dataset.

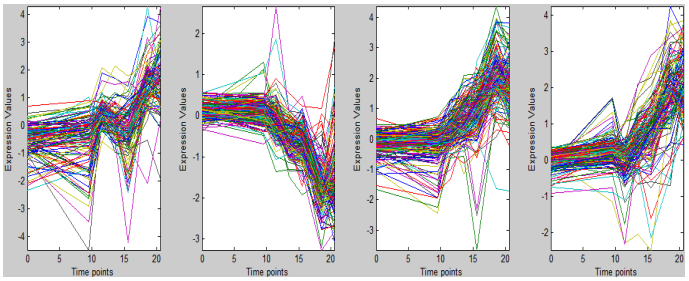


Figure 15: k-means clusters using PWCTM as the similarity measure for Dataset 3 at k=4. The vertical axes denote expression values of the genes and horizontal axes denote the 7 time points during the diauxic shift.

algorithm over the reduced Dataset 3 for k=16 using different proximity measures we obtain the clusters as given in Fig. 17 (a), 17 (b), 18(a), 18 (b), 18 (c).

The enriched functional categories for two of the clusters obtained by k-means using PWCTM and Pearson's correlation(PCC) on Dataset 2 are listed in Table 10. We report functional categories with p-values $< e-21$ for cluster 1 and p-value $< e-09$ for cluster 2 in order to restrict the size of the table. The values shown in Table 10 indicate that the genes categorized in the corresponding clusters through this algorithm are biologically significant in the respective clusters due to their low p-values. From the table, we see that PWCTM gave better result than Pearson correlation coefficient, most of the times. In cluster C1 we see that our PWCTM obtained a lower p-value of 3.633 e-31 for the

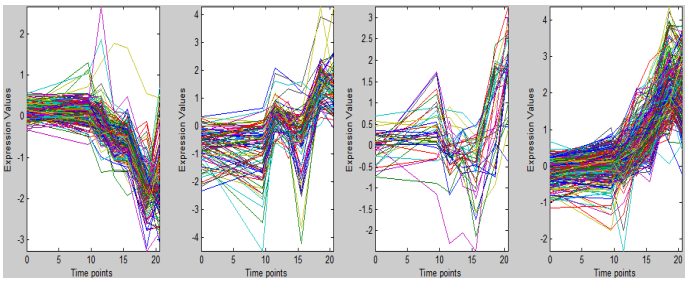
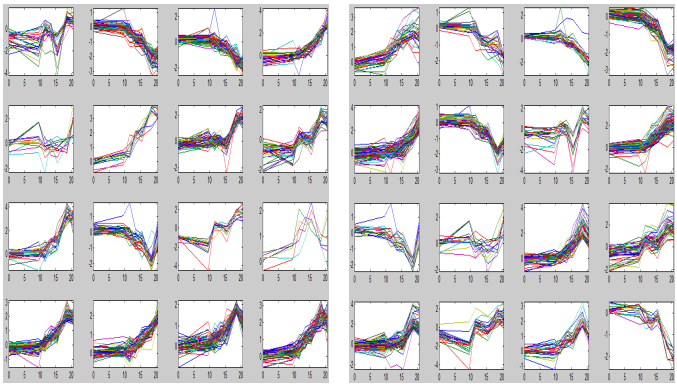


Figure 16: k-means clusters using PCC as the similarity measure for Dataset 3 at k=4.



(a) (b)

Figure 17: k-means clusters over reduced Dataset 3 using (a) Euclidean distance and (b) Pearson correlation as the proximity measure for k=16

highly enriched functional category of MCM complex while using PCC this same GO category had a p-value of 1.843 e-9 in C2. The lowest p-value obtained by k-means with PCC is 7.047 e-28 for GO category Chromosomal part.

The values shown in Table 10 indicates that the genes categorized in the corresponding clusters through the k-means algorithm with PWCTM as the similarity measure are biologically significant in the respective clusters due to their low p-values.

5.2. PatGeneClus Results

The result of PatGeneClus for Dataset 1 with the appropriate $simTH$, $relaxTH$, $mergeTH$, GO weight w_2 and the proximity measures used are illustrated in Fig. 19, 20 and 21. The result of PatGeneClus for Dataset 2 with the appropriate appropriate thresholds is illustrated in Fig. 22, 23 and 24. The result for Dataset 3 is given in Fig. 25 and 26. The result of PatGeneClus on this subset of Dataset 4 with the appropriate $simTH$, $relaxTH$, $mergeTH$, GO weight w_2 and the proximity measures used are illustrated in Fig. 27, 28 and 29.

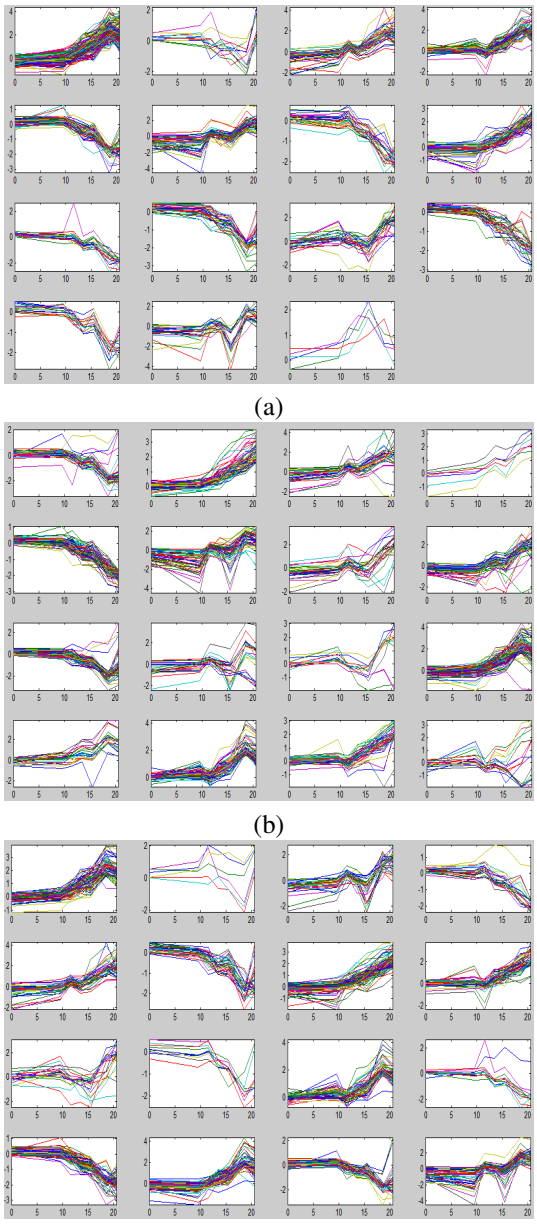
The p-values of two of the clusters obtained by PatGeneClus using three different proximity measures with threshold values $simTH = 75$, $relaxTH = 75$ and $mergeTH = 25$ $w_2 = 50$ is given in Table 11. From the table it can be seen that PWCTM obtained much better p-values in cluster C1 with the lowest p-value of 1.2E-41 for GO category ascospore wall assembly. This same category obtained a p-value of 1.30E-38 and 2.80E-37 with Euclidean distance and PCC respectively. PCC obtained the lowest p-value of 3.80E-38 for GO category sporulation which is comparable to the one obtained by PWCTM (2.20E-38) for the same category. Euclidean distance also obtained the lowest p-value for sporulation as 4.90E-39. Overall in C1 the result obtained for PWCTM is sufficiently better. In cluster C2, we see that Euclidean distance obtained much better p-values. However, PWCTM and PCC obtained comparably good result. The lowest value for all three measures were for GO category cytosolic ribosome with Euclidean distance obtaining the lowest p-value of 2.20E-65, followed by PCC with 4.60E-52 and PWCTM with 4.70E-51.

Table 10: P-values for k-mean clusters using PWCTM and PCC for Dataset 2

Clusters using PWCTM	P-value	GO number	GO category	Clusters using PCC	P-value	GO number	GO category
C1	3.63e-31	GO:0042555	MCM complex	C1	7.05e-28	GO:0044427	Chromosomal part
	8.22e-31	GO:0007049	Cell cycle		7.30e-28	GO:0022402	Cell cycle process
	3.09e-24	GO:0000084	S phase mitotic cycle		5.09e-23	GO:0007049	Cell cycle
	5.39e-24	GO:0051320	S phase		9.76e-23	GO:0044454	Nuclear chromosomal part
	2.59e-22	GO:0022403	Cell cycle phase		1.04e-20	GO:0006259	DNA metabolic process
C2	1.25e-12	GO:0022402	Cell cycle process	C2	4.37e-13	GO:0007049	Cell cycle
	5.60e-11	GO:0044427	Chromosomal part		7.51e-11	GO:0005935	Cellular bud neck
	5.54e-10	GO:0007049	Cell cycle		1.84e-09	GO:0042555	MCM complex
	5.54e-10	GO:0044454	Nuclear chromosomal part		3.25e-09	GO:0030427	Site of polarized growth
	9.63e-10	GO:0006259	DNA metabolic process		1.14e-08	GO:0022402	Cell cycle process
C3	1.13e-05	GO:0005935	Cellular bud neck	C3	7.4e-05	GO:0007076	Mitotic chromosome condensation

Table 11: P-value comparison of PatGeneClus clusters on Dataset 1

Euclidean distance	P-value	GO number	GO category	PCC	P-value	GO number	GO category	PWCTM	P-value	GO number	GO category
C1	4.90E-39	GO:0030435	sporulation	C1	3.80E-38	GO:0030435	sporulation	C1	1.20E-41	GO:0030476	ascospore wall assembly
	1.30E-38	GO:0030476	ascospore wall assembly		2.80E-37	GO:0030476	ascospore wall assembly		1.20E-41	GO:0042244	spore wall assembly
	1.30E-38	GO:0042244	spore wall assembly		2.80E-37	GO:0042244	spore wall assembly		2.20E-38	GO:0030435	sporulation
	1.60E-34	GO:0030154	cell differentiation		2.00E-33	GO:0030154	cell differentiation		1.30E-34	GO:0030154	cell differentiation
	1.60E-34	GO:0048869	cellular developmen tal process		2.00E-33	GO:0048869	cellular developmen tal process		1.30E-34	GO:0048869	cellular developmen tal process
	1.60E-33	GO:0030437	ascospore formation		3.30E-33	GO:0030437	ascospore formation		6.30E-33	GO:0030437	ascospore formation
	1.60E-33	GO:0034293	sexual sporulation		3.30E-33	GO:0034293	sexual sporulation		6.30E-33	GO:0034293	sexual sporulation
	9.00E-28	GO:0048610	reproductive cellular process		2.60E-27	GO:0048610	reproductive cellular process		1.10E-27	GO:0048610	reproductive cellular process
	5.80E-27	GO:0032502	developmental process		1.70E-25	GO:0032502	developmental process		1.90E-25	GO:0032502	developmental process
	1.40E-20	GO:0022414	reproductive process		5.90E-20	GO:0022414	reproductive process		5.30E-21	GO:0022414	reproductive process
	2.00E-19	GO:0007047	cell wall organization and biogenesis		9.50E-20	GO:0007047	cell wall organization and biogenesis		8.70E-20	GO:0007047	cell wall organization and biogenesis
	2.00E-19	GO:0045229	external encapsulating structure organization and biogenesis		9.50E-20	GO:0045229	external encapsulating structure organization and biogenesis		8.70E-20	GO:0045229	external encapsulating structure organization and biogenesis
C2	2.20E-65	GO:0022626	cytosolic ribosome	C2	4.60E-52	GO:0022626	cytosolic ribosome	C2	4.70E-51	GO:0022626	cytosolic ribosome
	9.50E-61	GO:0044445	cytosolic part		1.70E-47	GO:0044445	cytosolic part		1.70E-46	GO:0044445	cytosolic part
	9.90E-58	GO:0003735	structural constituent of ribosome		1.50E-44	GO:0003735	structural constituent of ribosome		1.50E-43	GO:0003735	structural constituent of ribosome
	9.90E-58	GO:0033279	ribosomal subunit		1.50E-44	GO:0033279	ribosomal subunit		1.50E-43	GO:0033279	ribosomal subunit
	7.50E-54	GO:0005840	ribosome		3.10E-44	GO:0043228	non-membrane-bounded organelle		3.10E-41	GO:0005840	ribosome
	4.40E-50	GO:0005198	structural molecule activity		3.10E-44	GO:0043232	intracellular non-membrane-bounded organelle		1.30E-40	GO:0043228	non-membrane-bounded organelle



(a)

(b)

(c)

Figure 18: The clusters detected using k-means over reduced Dataset 3 with (a) Spearman's rank correlation, (b) BioSim and (c) PWCTM as the similarity measure

6. ClustalW results of PatGeneClus

The result of ClustalW using PatGeneClus with PWCTM on Dataset 2 using thresholds $simTH = 70$, $relaxTH = 70$, $mergeTH = 25$ is given in Table 12. Table 12 and 13, reports the best alignment scores ≥ 90.0.

References

[1] A. Attig and P. Perner, "The problem of normalization and a normalized similarity measure by online data," *Transactions on Case-Based Reasoning*, vol. 4 (1), pp. 3–17, 2011.

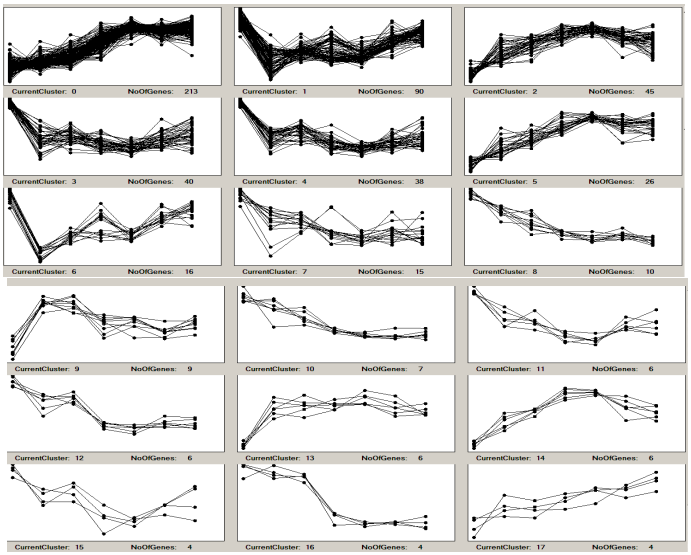


Figure 19: PatGeneClus clusters using Euclidean distance as the proximity measure for Dataset 1 with threshold values $simTH = 70$, $relaxTH = 70$ and $mergeTH = 30$ $w_2 = 40$.

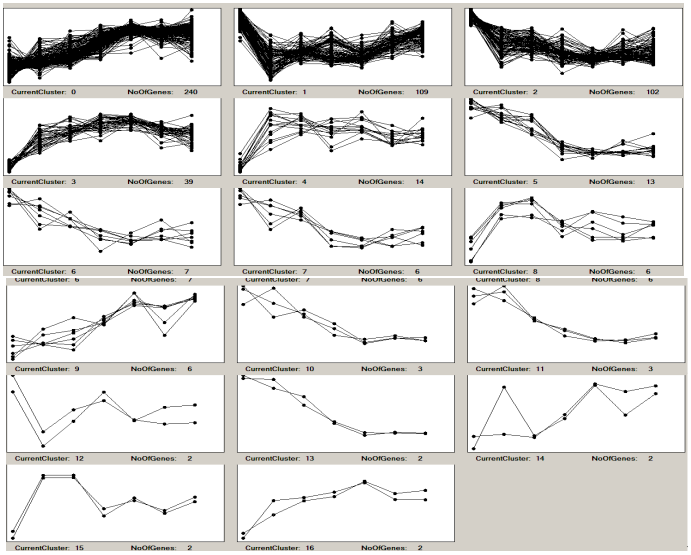


Figure 20: PatGeneClus clusters using Pearson's correlation as the similarity measure for Dataset 1 with threshold values $simTH = 70$, $relaxTH = 70$ and $mergeTH = 30$ $w_2 = 40$.

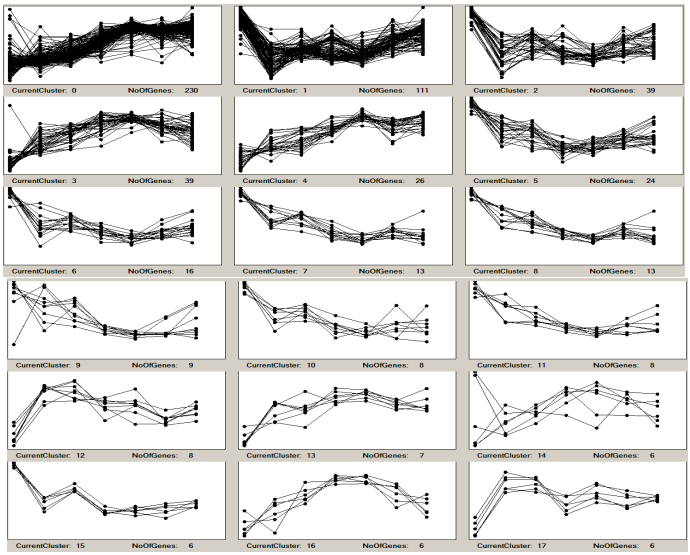


Figure 21: PatGeneClus clusters using PWCTM as the similarity measure for Dataset 1 with threshold values $simTH = 70$, $relaxTH = 70$ and $mergeTH = 30$ $w_2 = 40$.

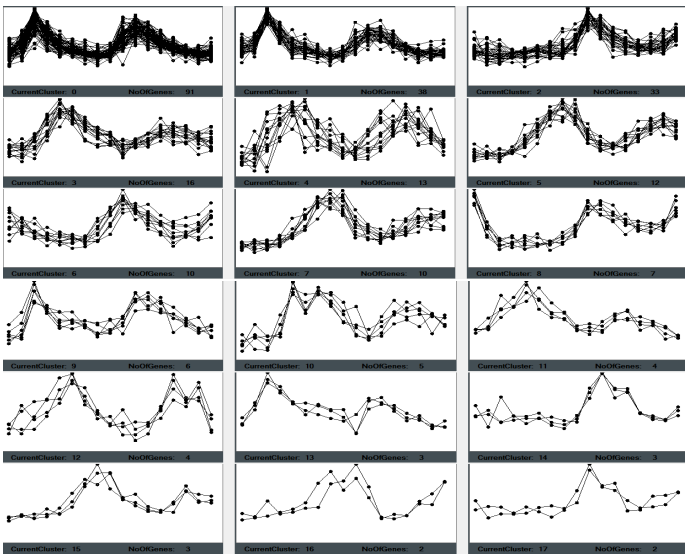


Figure 22: PatGeneClus clusters using Pearson's correlation measure for Dataset 2 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 20$ $w_2 = 40$.

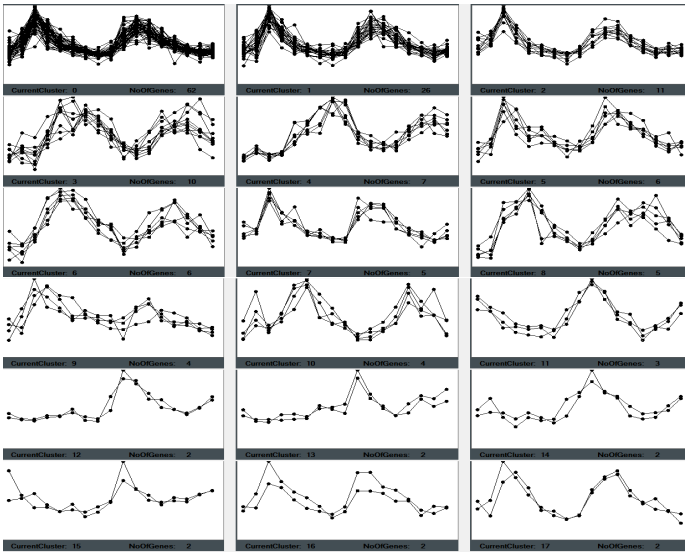


Figure 23: PatGeneClus clusters using PWCTM as the similarity measure for Dataset 2 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 20$ $w_2 = 40$.

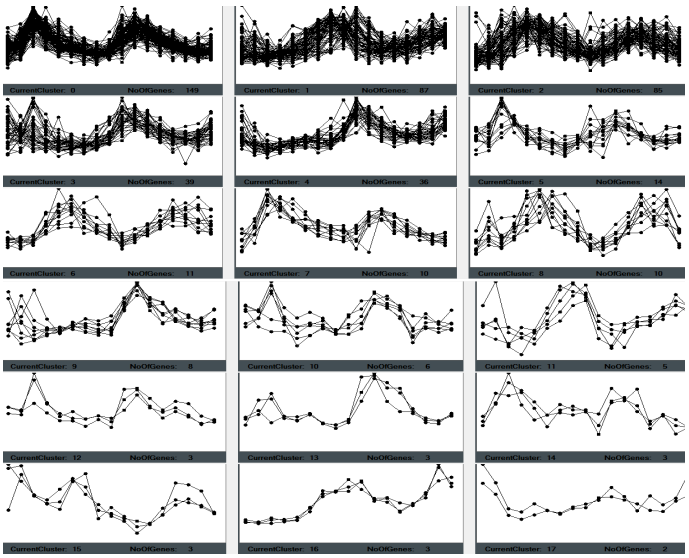


Figure 24: PatGeneClus clusters using PWCTM as the similarity measure for Dataset 2 with threshold values $simTH = 75$, $relaxTH = 75$ and $mergeTH = 20$ $w_2 = 40$.

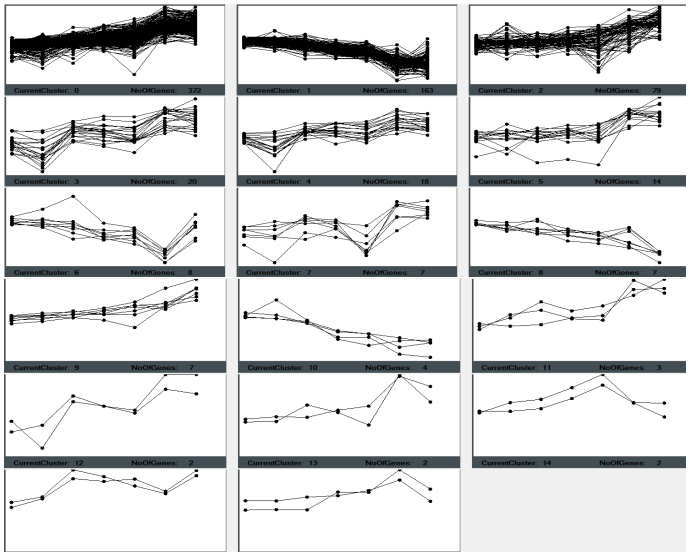


Figure 25: PatGeneClus clusters using Pearson's correlation as the similarity measure for Dataset 3 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 25$ $w_2 = 40$.

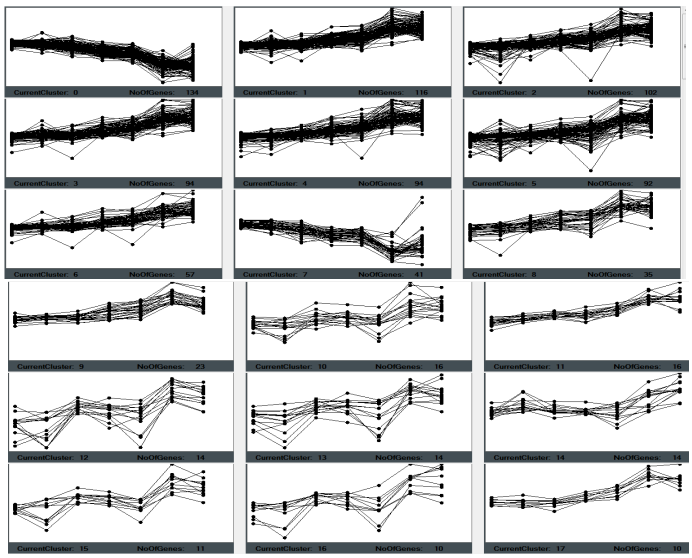


Figure 26: PatGeneClus clusters using PWCTM as the similarity measure for Dataset 3 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 25$ $w_2 = 40$.

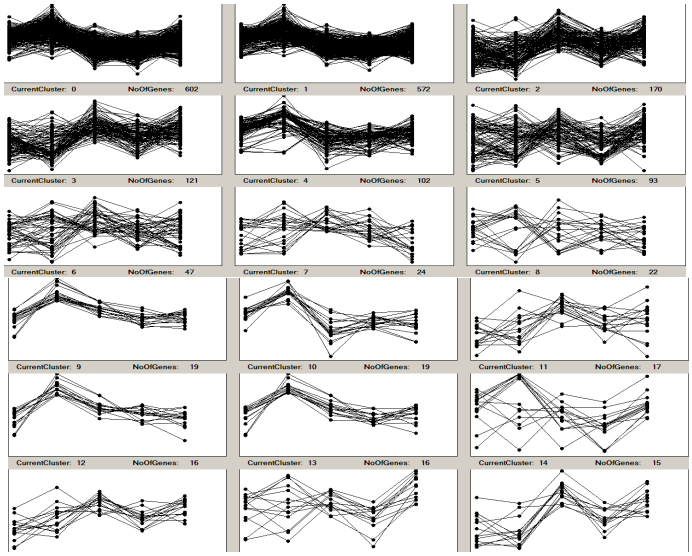


Figure 27: PatGeneClus clusters using Euclidean distance as the proximity measure for subset of Dataset 4 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 50$ $w_2 = 40$.

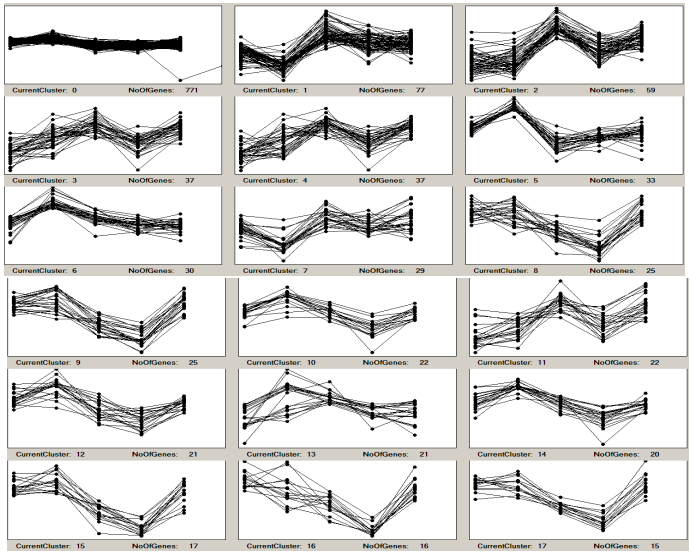


Figure 28: PatGeneClus clusters using Pearson's correlation as the similarity measure for subset of Dataset 4 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 30$ $w_2 = 40$.

Table 12: ClustalW Result of Dataset 2

SEQA	SEQB	SCORE
Cluster 1		
YER111C	YNL303W	91.09
YER111C	YBL002W	90.4
YNL046W	YLL021W	90.75
YNL046W	YBR275C	90.94
YLR050C	YKL101W	91.15
YLR050C	YBR275C	90.33
YMR076C	YJL173C	90.79
YMR076C	YLR236C	90.74
YMR076C	YNL303W	92.24
YFL008W	YJL173C	90.51
YFL008W	YNL303W	91.38
YJL074C	YNL303W	91.95
YJL074C	YBL002W	90.66
YNL233W	YNL303W	90.52
YDR507C	YNL303W	90.52
YNL102W	YJL173C	92.68
YNL102W	YBR252W	92.12
YNL102W	5 YNL303W	92.53
YNL102W	YBL002W	90.15
YNL102W	YKL066W	90.09
YBR278W	YNL262W	92.24
YNL262W	YOR074C	90.05
YNL262W	YBR252W	90.54
YNL262W	YPL267W	91.27
YNL262W	YCL022C	91.86
YNL262W	YHR110W	91.71
YNL262W	YBR089W	90.0
YNL262W	YDL018C	91.59
YNL262W	YDR013W	91.07
YNL262W	YNL303W	91.09
YNL262W	YBL003C	90.98
YNL262W	YBL002W	90.66
YNL262W	YAR008W	90.22
YNL262W	YKL066W	91.22
YJL173C	YLL021W	90.79
YJL173C	YKL101W	91.87
YJL173C	YCL061C	90.24
YJL173C	YBR275C	93.77
YER070W	YNL303W	90.52
YBR252W	YBR275C	91.22
YDR097C	YNL303W	93.68
YOL090W	YNL303W	90.52
YOL090W	YBL003C	90.48
YLR383W	YBL003C	90.23
YLL021W	YLR236C	92.9
YLL021W	YBR089W	90.0
YLL021W	YNL303W	91.67
YKL101W	YPL267W	90.0
YKL101W	YLR236C	91.05
YKL101W	YBR089W	90.0
YKL101W	YNL303W	91.09
YKL101W	YKL066W	90.32
YBR073W	YBL003C	90.48
YHR153C	YBR275C	91.62
YPL267W	YBR275C	91.9
YLR236C	YCL024W	90.12

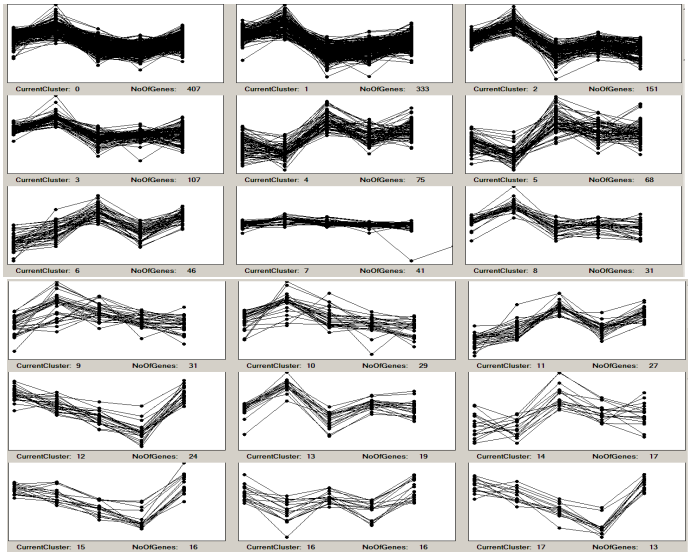


Figure 29: PatGeneClus clusters using PWCTM as the similarity measure for subset of Dataset 4 with threshold values $simTH = 80$, $relaxTH = 80$ and $mergeTH = 30$ $w_2 = 40$.

Table 13: ClustalW Result of Dataset 2 continued

SEQA	SEQB	SCORE
Cluster 2		
YBR200W	YGR183C	90.55
YBR202W	YLR395C	91.98
YBR202W	YGR183C	91.54
YJL157C	YGR183C	90.05
YLR395C	YGR281W	90.72
YGR183C	YPL058C	91.04
YGR183C	YGR281W	92.54
YMR256C	YPL058C	94.54
YGR281W	YMR254C	91.91
Cluster 5		
YNR016C	YJL173C	92.41
YNR016C	YBR252W	89.41
YNR016C	YKR083C	93.03
cluster 10		
YMR254C	YKL129C	91.26
YGR143W	YNL057W	90.39
YKL129C	YLR297W	91.54
cluster11		
YLR349W	YBR275C	90.53
YCL061C	YNL303W	91.09
YDR383C	YBR275C	90.1
YNL300W	YBR275C	90.94
YNL303W	YBR275C	92.82
YKL067W	YBR275C	90.26
YBR275C	YOR284W	90.16

[2] S. A., K. S., and F. L., "Nonparametric feature normalization for svm-based speaker verification," in *In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2008, pp. 1577 – 1580.
[3] R. J. Cho, M. Campbell, E. Winzeler, L. Steinmetz *et al.*, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, vol. 2(1), p. 6573, 1998.
[4] J. DeRisi and P. Iyer, V.R.and Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680–686, 1997.
[5] P. Nymark, P. Lindholm, M. Korpela, L. Lahti, S. Ruosaari, H. J. Kaski, S., S. Anttila, V. Kinnula, and S. Knuutila, "Gene expression profiles in asbestos-exposed epithelial andmesothelial lung cell lines," *BMC Genomics*, vol. 8:62, 2007.