**Systems Biology – HW1** Due: Hand in printout before class 2016/11/03

學號:404410030

姓名:鄭光宇

1. (35 pts) **USING EXCEL TO DO THIS PROBLEM!** You are given to a microarray raw data file called “problem1.gpr” (please download it from the E-course web site).
2. (30 pts) First, any median intensity less than 100 after subtracting the background is deleted. Then, normalize the data using total intensity (or global) normalization method. What are the 5 most induced and repressed genes (please exclude those control spots and write down the gene names and IDs) and their log2(Cy5/Cy3)? Please upload your excel file to線上作業中的HW1 in the E-course web site)

|  |  |  |
| --- | --- | --- |
| Name | ID | log2(Cy5/Cy3) |
| IMAGE:1009353 | g1847166 | hAE0093 | 6.080980061 |
| IMAGE:1129745 | g2466805 | hAD9943 | 6.061628408 |
| MIG | Hs.77367 | IMAGE:503617 | AA131406 | hAA2205 | 5.860661206 |
| MIG | Hs.77367 | IMAGE:503617 | AA131406 | hAA2205 | 5.772793723 |
| SELL | Hs.82848 | IMAGE:149910 | H00756 | hAA0103 | 5.682549811 |

1. (5 pts) What are the assumptions when you apply global normalization to these raw data?  
   當log2(Cy5)對log2(Cy3)的分布趨勢呈現線性關係的時候，也就是當log2(Cy5)對log2(Cy3)散布圖的點較集中在一直線上時。
2. (25 pts) Using MIDAS to normalized problem1.gpr microarray data. (1) You use ExpressConverter to convert problem1.gpr to problem1.mev. Another problem1.ann (including information on the ID and name of each gene) will be generate at the same time.
3. (12 pts ) Use MIDAS to first filter out any IA (Cy3) or IB (Cy5) less than 10000, and then normalize problem1.mev by total intensity method. What are the 5 most induced and repressed genes (please exclude those control spots and write down the gene names and IDs) and their log2(Cy5/Cy3). (4) Compare the above result with that in Problem 1(a).

|  |  |  |
| --- | --- | --- |
| Name | ID | log2(Cy5/Cy3) |
| IGFBP1 | Hs.102122 | IMAGE:668685 | AA2\* | hAA0323 | 5.239190896 |
| GPC3 | Hs.119651 | IMAGE:878564 | AA775\* | hAB7211 | 5.116605617 |
| IGFBP3 | Hs.77326 | IMAGE:898218 | AA59\* | hAA1898 | 4.440168582 |
| TGFB2 | Hs.169300 | IMAGE:666218 | AA23\* | hAA0044 | 4.27381772 |
| TGFB2 | Hs.169300 | IMAGE:666218 | AA23\* | hAA0044 | 4.207740103 |

結果與1(a). 有很大的差距，也許是因為1(a). 的過程並沒有要求去掉Cy5和Cy3過高的極端值，而2(a).的MIDAS有，導致1(a).前幾列的log2(Cy5/Cy3)過高。

1. (13 pts) This time, use LOWESS to normalize the problem1.mev. What are the 5 most induced and repressed genes (please exclude those control spots and write down the gene names and IDs) and their log2(Cy5/Cy3). (4) Compare the above result with that in Problem 2(a).

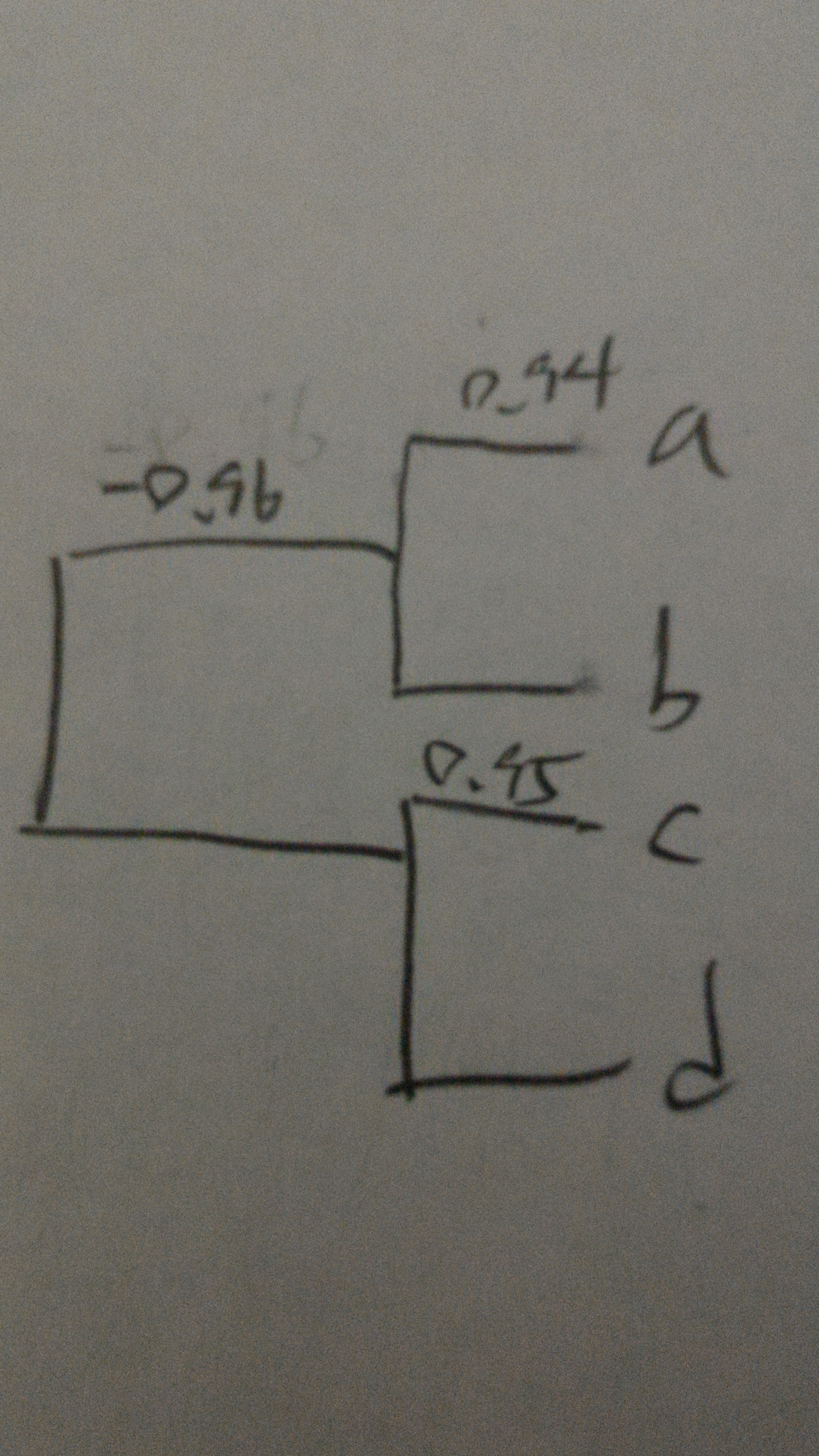
|  |  |  |
| --- | --- | --- |
| Name | ID | log2(Cy5/Cy3) |
| IGFBP1 | Hs.102122 | IMAGE:668685 | AA2\* | hAA0323 | 5.239190896 |
| GPC3 | Hs.119651 | IMAGE:878564 | AA775\* | hAB7211 | 5.116605617 |
| IGFBP3 | Hs.77326 | IMAGE:898218 | AA59\* | hAA1898 | 4.440168582 |
| TGFB2 | Hs.169300 | IMAGE:666218 | AA23\* | hAA0044 | 4.27381772 |
| TGFB2 | Hs.169300 | IMAGE:666218 | AA23\* | hAA0044 | 4.207740103 |

結果與2(a). 的結果相同。

1. (40 pts) The problem involves 4 genes whose expression is measured over 5 hours after drug treatment. The microarray normalized data in the form of Cy5/Cy3 ratios are listed as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Gene** | **0 hours** | **1 hours** | **2 hours** | **3 hours** | **4 hours** | **5 hours** |
| a | 1 | 8 | 12 | 16 | 12 | 8 |
| b | 1 | 3 | 4 | 4 | 3 | 2 |
| c | 1 | 0.125 | 0.0833 | 0.0625 | 0.0833 | 0.125 |
| d | 1 | 0.5 | 0.33 | 0.25 | 0.33 | 0.5 |

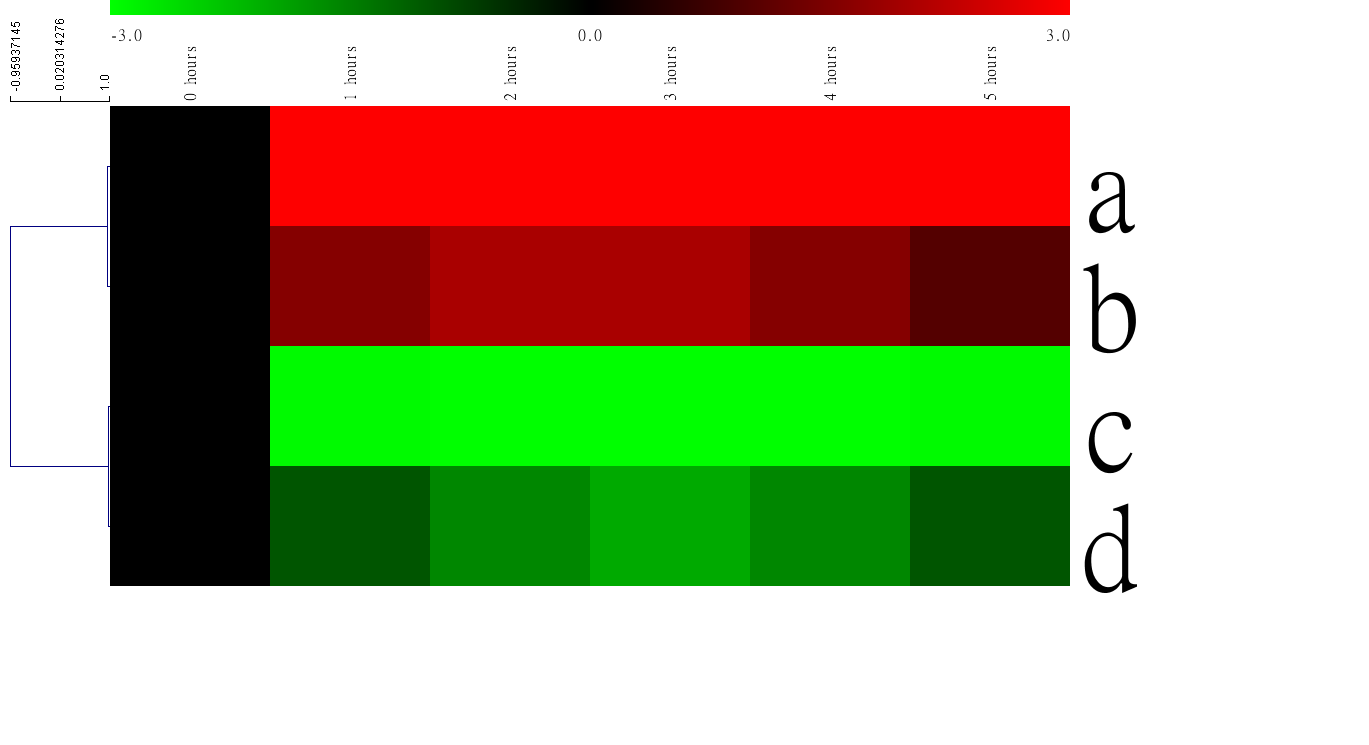
1. (10 pts) Please perform hierarchical clustering analysis (using Pearson Correlation) by hand step by step. Calculation steps include (1) log2 transformation, (2) construction of Pearson correlation matrix for each gene pair, and (3) gene tree construction. Finally, plot your cluster tree below.



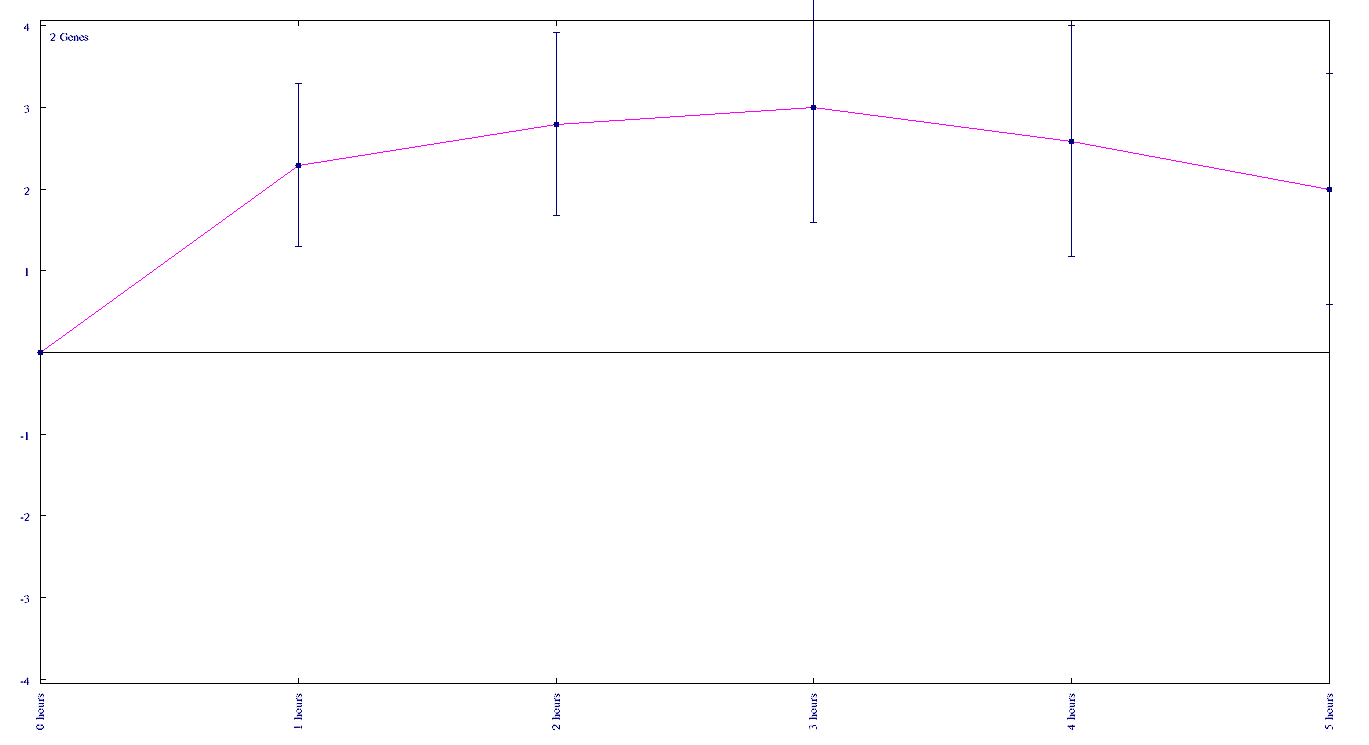
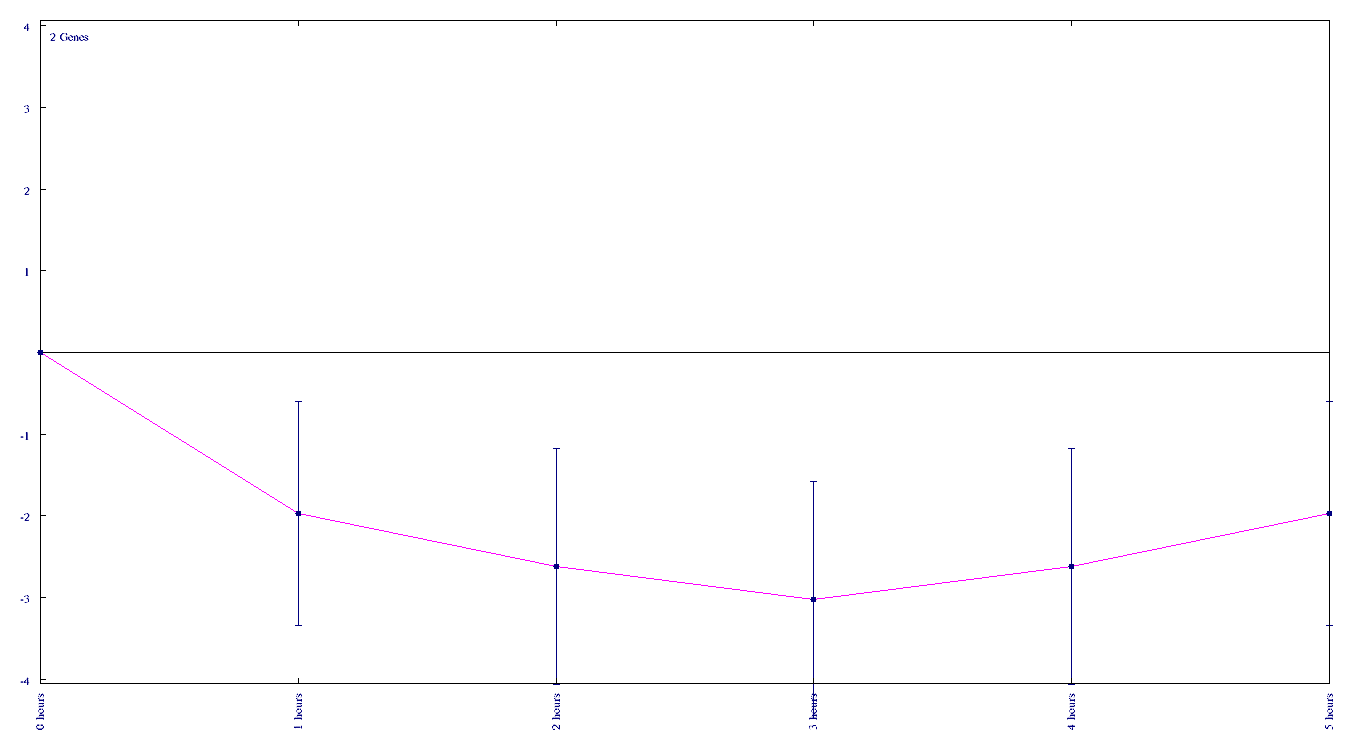
1. (10 pts) Please perform hierarchical clustering analysis (using distant) by hand step by step. Calculation steps include (1) log2 transformation, (2) construction of Pearson correlation matrix for each gene pair, (3) conversion of Pearson correlation to distant matrix, and (4) gene tree construction. Finally, plot your cluster tree below.



1. (10 pts) Please perform hierarchical clustering analysis using MeV with parameters: (1) gene tree only, (2) correlation coefficient and (3) average linkage cluster. Paste your cluster tree and color bar plot below.



1. (5 pts) (continue from c) If we want to get two main clusters (or nodes), please plot average gene profile (vs. time) graph for each cluster.  
   (Node cd.)  
   (Node ab.)



1. (5 pts) Compare these clustering results among (a)-(c).  
   三者的Clustering結果是相同的，ab一個cluster、cd一個cluster、abcd是一整個cluster，MeV顯示出來的距離也與結果(b)類似，ab與cd兩node之間距離較大，a到b、c到d之間的距離較小。而從結果(a)較看不出距離關係。