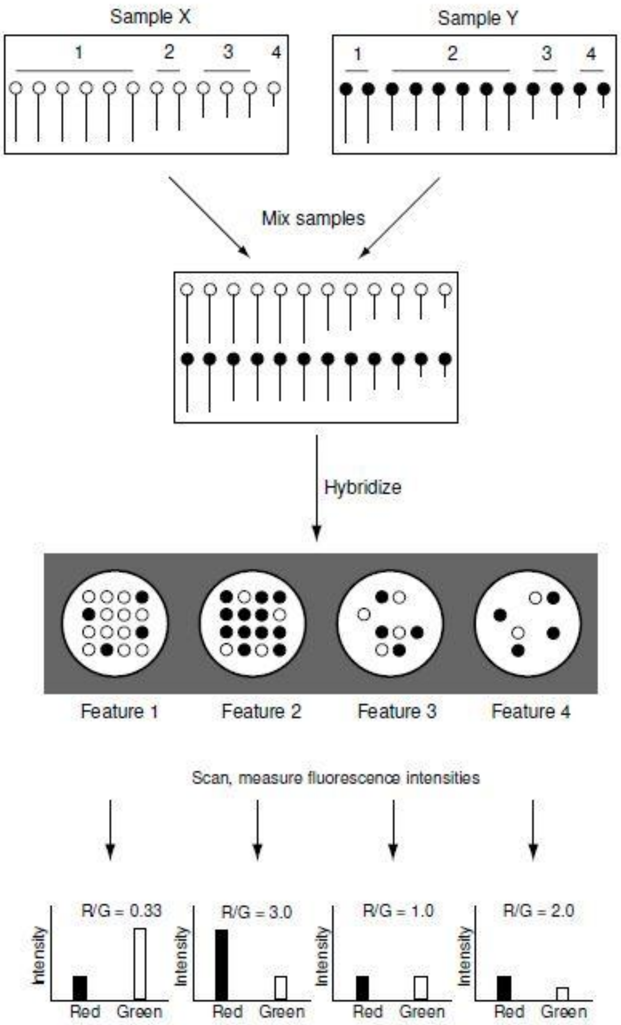
Chapter 11 Class Notes –
Measuring Expression of Genome Information

11.1. The Biological Problem: "even though all genes are generally found in all cells, not all genes are expressed at any one time, nor are all genes expressed in every cell," and genes may be differentially expressed over the cell cycle.

11.3. Principles and Practice of Microarray Analysis: spotted or oligonucleotide microarrays: area is approximately 1-2 cm^2

11.3.1. Basics of Nucleic Acids Used for Microarrays: from text: a probe is a particular DNA sequence corresponding (complementary) to an mRNA whose abundance, presence or absence within a sample is being evaluated; the target is the complex mixture of nucleic acid species being tested for the presence or absence of sequences related to the probe sequence… In microarray experiments, the probes are immobilized in a grid of positions on a substrate (usually glass)… Each gridded probe sample is a feature, which is indexed by its position within the array. The target depends upon the organism, the tissue, and the physiological conditions of the tissue at the time that the RNA is extracted; the target molecules may be radioactively labelled, but now it is far more common for them to be labelled with fluorescent dyes Cy3 or Cy5. The specific interaction between probe and target species is based upon DNA hybridization (the two strands need not come from the original duplex molecule). When labelled target molecules hybridize to a particular

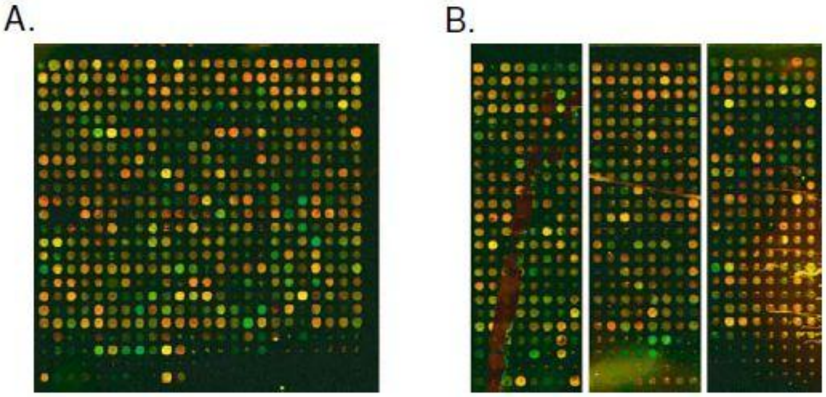
feature, the fluorescent label on the target species makes the feature capable of fluorescence when it is excited by light of an appropriate wavelength. The amplitude of the signal is proportional to the amount of hybridized target species. If the species is rare, the signal is correspondingly faint.



11.3.2. Making and Using Spotted Microarrays: Design is as above. Often we wish to understand how gene expression differs for two different conditions (e.g., malignant versus normal cells – above X and Y). The amount of fluorescence intensity corresponding to each condition at each microarray

feature can therefore be detected. If cDNA derived from RNA expressed during condition X is labelled with Cy3 and cDNA derived from condition Y is labelled with Cy5, and excess of green over red fluorescence at any feature (gene) indicates that the gene corresponding to that spot is more highly expressed under condition X than under condition Y.

Problems can arise as below (B). In A, studying *dsd*^{D} mutant *D. melanogaster*, red spots correspond to genes whose expression is higher in the mutant, and green spots correspond to genes with reduced expression in the mutant. Panel B shows some problems that can appear in such experiments (see p.302).



11.4. Analysis of Microarray Data: "we analyze either absolute amounts of transcripts in cells or, more commonly, the ratios of these amounts under two different experimental conditions." We need to first process the data to obtain a gene expression matrix (GEM): most commonly, n rows for each gene or feature and m columns one for each condition or time point; the content of each element of the GEM is either a fluorescence intensity or a ratio of 2 fluorescence intensities.

11.4.1. Normalization: important – there usually is a dye bias that needs to be corrected in a process called normalization; "in fact, Cy5 intensities are systematically lower than Cy3 intensities when equivalent amounts of sample are present." If linearly related, then $R = kG, k < 1, so\ \log_2 \frac{R}{G} = \log_2 k$. There are two approaches: (1) global normalization (only sometimes valid) or (2) intensity-dependent normalization (more complicated usually but worth the effort) as below. (1) is often invalid since the dye bias in $\log_2 \frac{R}{G}$ is not constant but varies with intensity; thus, we produce a MA plot by first finding $M = \log_2 \frac{R}{G} = \log_2 R - \log_2 G$ and ($\log_2$ of geometric mean intensity) $A = \frac{1}{2}\log_2 RG = \frac{1}{2}(\log_2 R + \log_2 G)$.

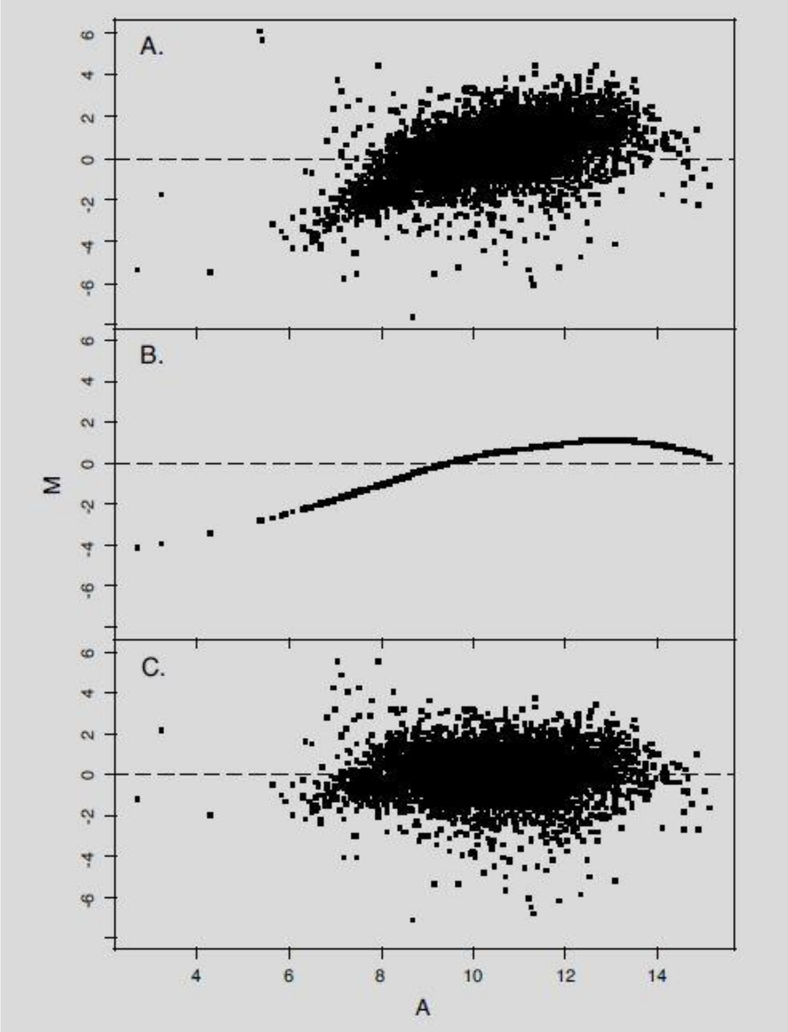
Computational Example 11.1: *dsx*^{D} mutant *Drosophila*, with 11 usable blocks and with A and M in columns 5 and 6; since $\bar{M} = 0.2903$ we get the estimate $\hat{k} = 2^{0.2903} = 1.2229$.

```
micro2=read.csv("c:\\CE11p1a.csv",sep=",",header = TRUE)
mean(micro2[,6])
[1]  0.2903073
2^mean(micro2[,6])
[1]  1.222901
```

This is used in global normalization (surprisingly R intensity is on average 22.3% higher than the G intensity), but we prefer to do the intensity-dependent normalization:

```
par(pin=c(4,2),mfrow=c(3,1))
plot(micro2[,5],micro2[,6],pch=".",xlab="A",ylab="M")
```

Plot below shows an upward (nonlinear?) trend:



Next, fit the trend curve in B (above) via LOESS regression, remove the trend and keep the residuals (above in C)

```
MA.ls<-loess(micro2[,6]~micro2[,5])
#Note argument order: dependent variable is listed first.
tmp<-predict(MA.ls,micro2[,5])
# tmp is a vector of predicted values
# predict is a standard R function
plot(micro2[,5],tmp,pch=".",xlab="A",ylab="M",ylim=c(-8,6))
#Plot with same scale used in first panel.
```

```
MA.norm<-micro2
MA.norm[,6]<-MA.norm[,6]-tmp
#Subtracting predicted value
mean(MA.norm[,6])
[1] -0.003620709
plot(MA.norm[,5],MA.norm[,6],pch=".",xlab="A",ylab="M",
ylim=c(-8,6))
```

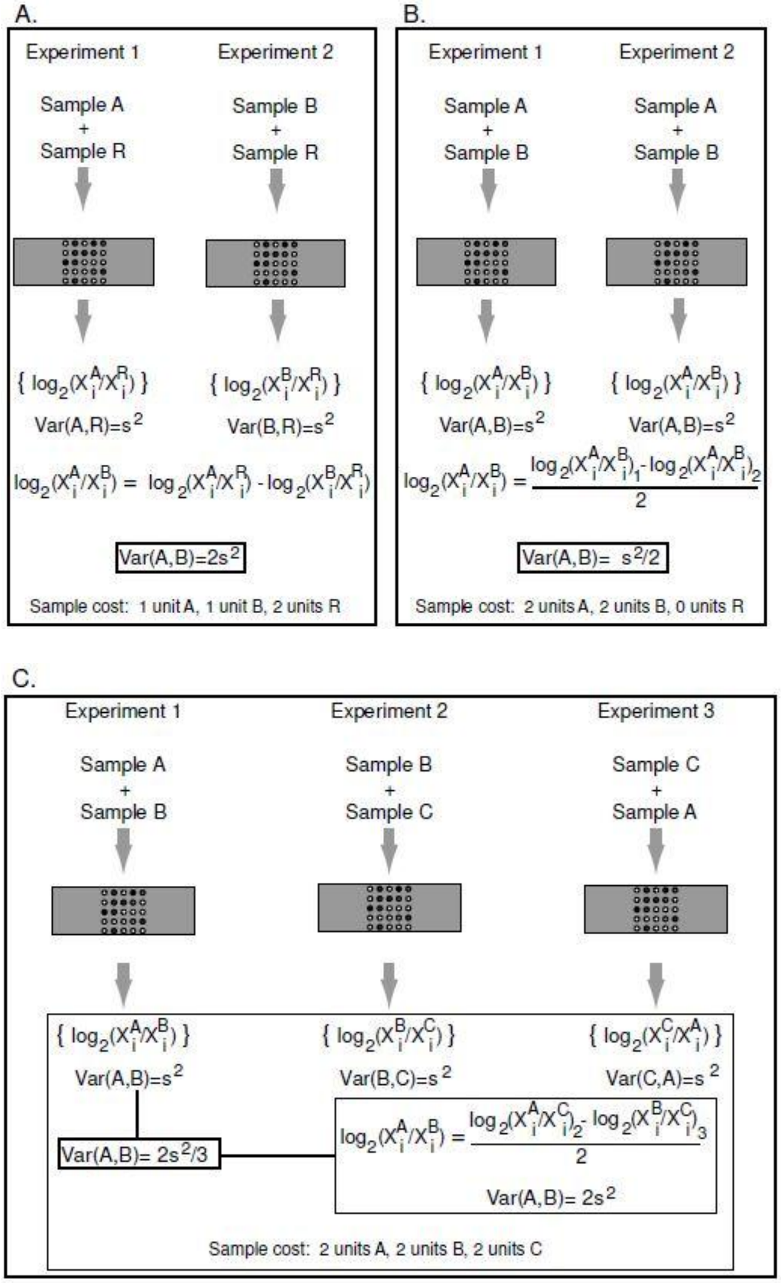
Now we have normalized the data (ready for statistical analysis) so that the mean of normalized $\log_2 \frac{R}{G}$ is near zero.

11.4.2. Statistical Background: review of SD and two-sample pooled t-test is on pp.308-10, but note that "for microarray experiments, the number of replicates, n, will often be a small number much less than 10." "The null hypothesis would be that the expression level is no different in cancer cells compared with unaffected cells, and we would want to perform a hypothesis test to determine whether the observed mean value for X_j^t is significantly different from the mean value for X_j^c." $TS': t = \frac{\bar{X}_j^t - \bar{X}_j^c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$

Pooled t: $t = \frac{\bar{X}_j^t - \bar{X}_j^c}{s_p\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}, df = n_c + n_t - 2$

Note: since we're testing many, many genes, we can easily find a spurious result (false positive), so it's common in this multiple hypothesis testing situation to us a Bonferroni correction such as using $\alpha = \frac{\alpha_B}{N}, N = n \times m$; e.g., $\alpha_B = 0.05$

11.4.3. Experimental Design: wish to compare A and B; comparing with a reference R is often inefficient:


A.
Experiment 1
Experiment 2
Sample A + Sample R
Sample B + Sample R
$\{ \log_2(X_i^A/X_i^R) \}$
$\{ \log_2(X_i^B/X_i^R) \}$
Var(A,R)=s^2
Var(B,R)=s^2
$\log_2(X_i^A/X_i^B) = \log_2(X_i^A/X_i^R) - \log_2(X_i^B/X_i^R)$
Var(A,B)=2s^2
Sample cost: 1 unit A, 1 unit B, 2 units R
B.
Experiment 1
Experiment 2
Sample A + Sample B
Sample A + Sample B
$\{ \log_2(X_i^A/X_i^B) \}$
$\{ \log_2(X_i^A/X_i^B) \}$
Var(A,B)=s^2
Var(A,B)=s^2
$\log_2(X_i^A/X_i^B) = \frac{\log_2(X_i^A/X_i^B)_1 - \log_2(X_i^A/X_i^B)_2}{2}$
Var(A,B)= s^2/2
Sample cost: 2 units A, 2 units B, 0 units R
C.
Experiment 1
Experiment 2
Experiment 3
Sample A + Sample B
Sample B + Sample C
Sample C + Sample A
$\{ \log_2(X_i^A/X_i^B) \}$
$\{ \log_2(X_i^B/X_i^C) \}$
$\{ \log_2(X_i^C/X_i^A) \}$
Var(A,B)=s^2
Var(B,C)=s^2
Var(C,A)=s^2
Var(A,B)= 2s^2/3
$\log_2(X_i^A/X_i^B) = \frac{\log_2(X_i^A/X_i^C)_2 - \log_2(X_i^B/X_i^C)_3}{2}$
Var(A,B)= 2s^2
Sample cost: 2 units A, 2 units B, 2 units C

11.5. Data Interpretation: our goals are to:

- **Annotate anonymous genes based on their expression patterns over a number of conditions. If gene j shows a pattern similar to another/others, then we might** hypothesize that j functions in a similar manner or along a common pathway; called “guilt by association”

- Identify genes that are co-regulated (may function in the same biochemical pathway)
- Classify e.g. tumors based on gene expression patterns – could lead to genetic markers of clinically useful help

Here, “supervised” methods refer to whether information from outside the microarray experiment (e.g. biochemical info), and “unsupervised” methods only work with the multivariate data collected. As stated, usually n genes (rows) and m conditions; however “if the purpose of the experiment is tumor diagnosis, it may be of greater interest to consider m rows of conditions (e.g. m rows of tumor samples) measured over n genes (i.e. one column for each gene)…If the purpose of the experiment is classification of genes, then genes are the response variables and the conditions under which their expression was measured are the predictor variables. If the main interest is in classifying conditions or grouping similar conditions, then the conditions are the response variables and the genes are the predictor variables.” We’ll use clustering (genes) and/or data reduction (variables), as appropriate.

11.5.1. Clustering of Microarray Expression Data: in the typical (gene) study correlations and covariances are calculated between genes, and we form a correlation matrix ($\boldsymbol{R}$) or covariance matrix ($\boldsymbol{S}$); then, the distance matrix is

$$\boldsymbol{D} = \boldsymbol{J} - \boldsymbol{R}$$

(or equivalently, by $d_{ij} = 1 - r_{ij}$).

Computational Example 11.2: mRNA data for 12 yeast genes measured at 16 successive time points; we use K-means clustering to determine how many different expression patterns there are and to group together genes having similar patterns:

```
yeast.dat=read.csv("c:\\CE11p2.csv",sep=",",header=FALSE)
yeast<-cbind(df[,2:10],df[,12:18])

# standardize by row (gene)
syeast.dat<-as.matrix(yeast)
for(i in 1:12){
   syeast.dat[i,]<-(syeast.dat[i,]-mean(syeast.dat[i,]))/
   sqrt(var(syeast.dat[i,]))
}
row.names(syeast.dat)<-c("YBL023c","YBL072c","YBR202w",
"YDR258c","YEL032w","YER131w","YGL189C","YGR027C",
"YLL026w","YLR259C","YPL240C","YLR274W")
colnames(syeast.dat)<-c("t01","t02","t03","t04","t05","t06",
"t07","t08","t09","t10","t11","t12","t13","t14","t15","t16")
syeast.dat[1:4]
               t01          t02          t03          t04
YBL023c  -1.3202812   0.005524189  0.73029779  -0.7722816
YBL072c  -1.1625721  -0.617924997  0.83474443  -0.8399414
YBR202w  -1.1549769   0.009151956 -0.67907516  -1.0744397
YDR258c   3.4337065   1.136020908 -0.46415301  -0.3996771
YEL032w  -0.3421999   1.040063427 -0.27039405  -0.5396661
YER131w  -1.1608946  -0.868859161  0.45174234   0.8812062
YGL189C  -0.5141768  -0.994527569  1.25917287   1.5338940
```

```
YGR027C  -1.8475332   -0.448629053    0.42444587   -0.1956358
YLL026w   3.3412087    0.659218808   -0.47331522   -0.5791014
YLR259C   2.6423248    0.553840762    0.47351445   -1.1330117
YPL240C   3.3712044    1.212655197    0.25669716   -0.2691817
YLR274W  -0.3135798    0.366511474   -0.03127778   -1.1989817
```

```
# k=2 now do the K-MEANS
ckm2<-kmeans(syeast.dat,2,iter.max=10)
ckm2
K-means clustering with 2 clusters of sizes 8, 4

Clustering vector:
YBL023c YBL072c YBR202w YDR258c YEL032w YER131w
YGL189C YGR027C YLL026w YLR259C YPL240C YLR274W
      2       1       2       1       2       1       1       1       1       1       1       2

Within cluster sum of squares by cluster:
[1]  82.02579   10.34386
 (between_SS / total_SS =  37.0 %)

# k=3
ckm3<-kmeans(syeast.dat,3,iter.max=10)
ckm3
K-means clustering with 3 clusters of sizes 4, 4, 4

Clustering vector:
YBL023c   YBL072c   YBR202w   YDR258c   YEL032w   YER131w
YGL189C   YGR027C   YLL026w   YLR259C   YPL240C   YLR274W
      2       3       2       1       2       3       3       3       1       1       1       2
```

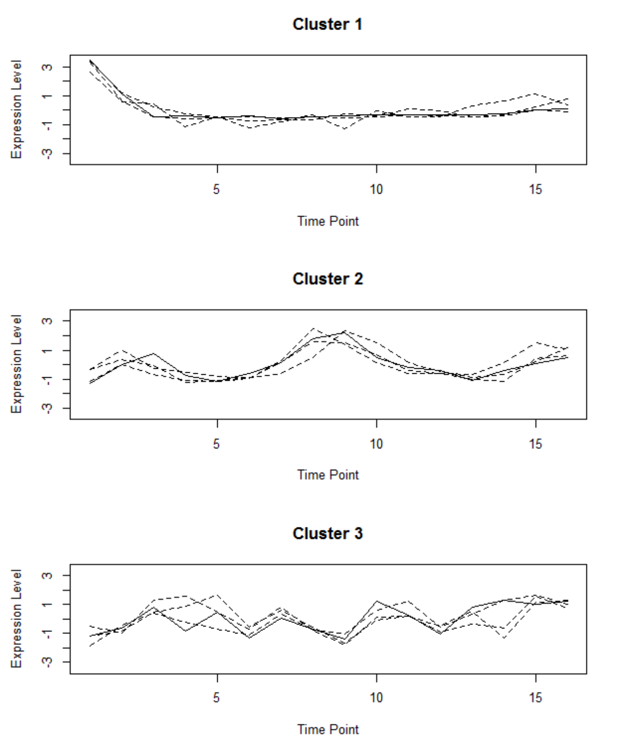
```
Within cluster sum of squares by cluster:
[1]  5.776459  10.343858  17.329197
 (between_SS / total_SS =  77.2 %)

# k=4
ckm4<-kmeans(syeast.dat,4,iter.max=10)
ckm4
K-means clustering with 4 clusters of sizes 2, 4, 2, 4

Clustering vector:
YBL023c YBL072c YBR202w YDR258c YEL032w YER131w
YGL189C YGR027C YLL026w YLR259C YPL240C YLR274W
    2       3       2       4       2       1       1       3       4       4       4       2

Within cluster sum of squares by cluster:
[1]  2.489429  10.343858   2.385605   5.776459
 (between_SS / total_SS =  85.7 %)

par(mfrow=c(3,1))
plot(syeast.dat[4,],ylim=c(-3.5,3.5),type="l",lty=1,
xlab="Time Point", ylab="Expression Level")
points(syeast.dat[9,],type="l",lty=2)
points(syeast.dat[10,],type="l",lty=2)
points(syeast.dat[11,],type="l",lty=2)
title("Cluster 1")
plot(syeast.dat[1,],ylim=c(-3.5,3.5),type="l",lty=1,
xlab="Time Point", ylab="Expression Level")
points(syeast.dat[3,],type="l",lty=2)
points(syeast.dat[5,],type="l",lty=2)
points(syeast.dat[12,],type="l",lty=2)
title("Cluster 2")
plot(syeast.dat[2,],ylim=c(-3.5,3.5),type="l",lty=1,
xlab="Time Point", ylab="Expression Level")
points(syeast.dat[6,],type="l",lty=2)
points(syeast.dat[7,],type="l",lty=2)
points(syeast.dat[8,],type="l",lty=2)
title("Cluster 3")
```

	Cluster 1	Cluster 2	Cluster 3
	4: YDR258C	1: YBL023c	2: YBL072C
	9: YLL026W	3: YBR202w	6: YER131W
	10: YLR259C	5: YEL032w	7: YGL189C
	11: YPL240C	12: YLR274W	8: YGR027C

11.5.2. Principal Components Analysis: before, we focused on rows (genes) – sometimes, we're interested in the columns (variables), and these too can be correlated. If column variables i and j are denoted x_i and x_j, then instead of working with these correlated variables, a data dimension

reduction is better done as follows: we convert the $\{x_i\}$ into new variables $\{y_i\}$ so that each y_i is a linear combination of the x_i (that is, $y_i = \sum_j a_{ij} x_j$) so that (1) the $\{y_i\}$ are uncorrelated and (2) $var(y_1) > var(y_2) \dots > var(y_m)$. This is called principal components analysis (PCA) and is equivalent to rotating the coordinate axes; a/the major goal of PCA is dimension reduction. Then: "the final step is to perform clustering of the genes based upon similarity in their expression patterns described by the set of variables y_i."

11.5.3. Confirmation of Results: "microarray experiments are often used to identify genes that should be examined further by extensive genetic, biochemical, or other "wet lab" approaches – which in turn are very costly and lengthy!

11.6. Examples of Experimental Applications: a made-up eg:

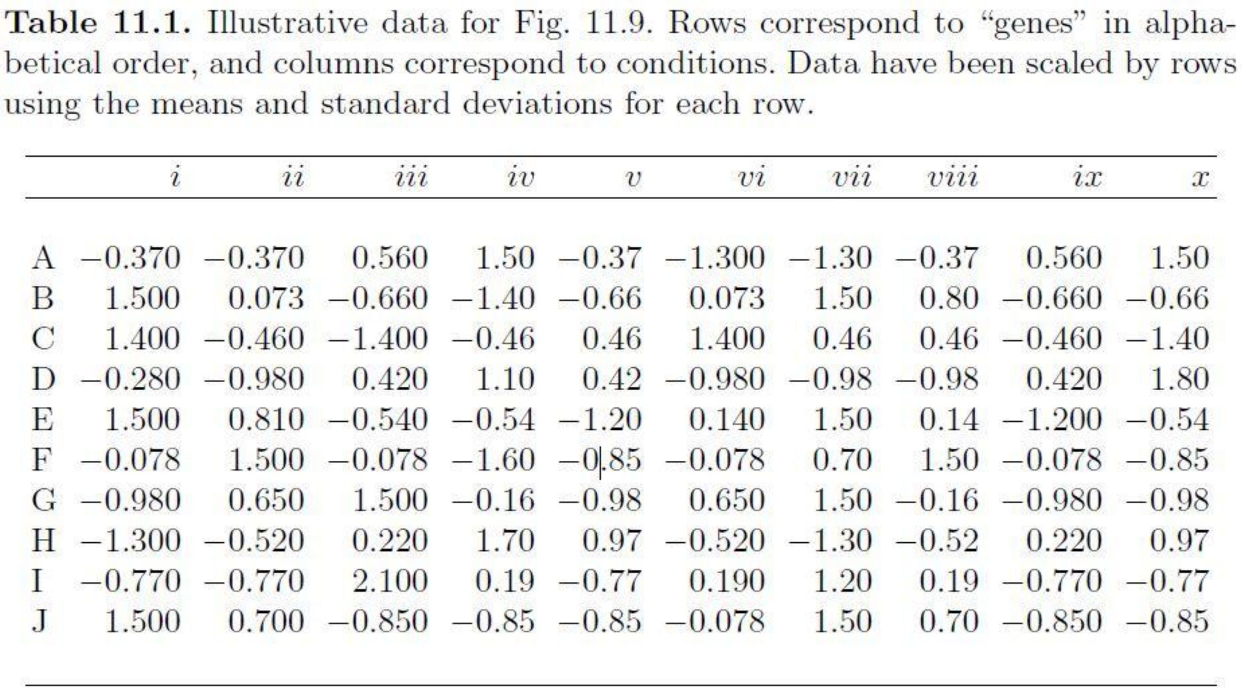
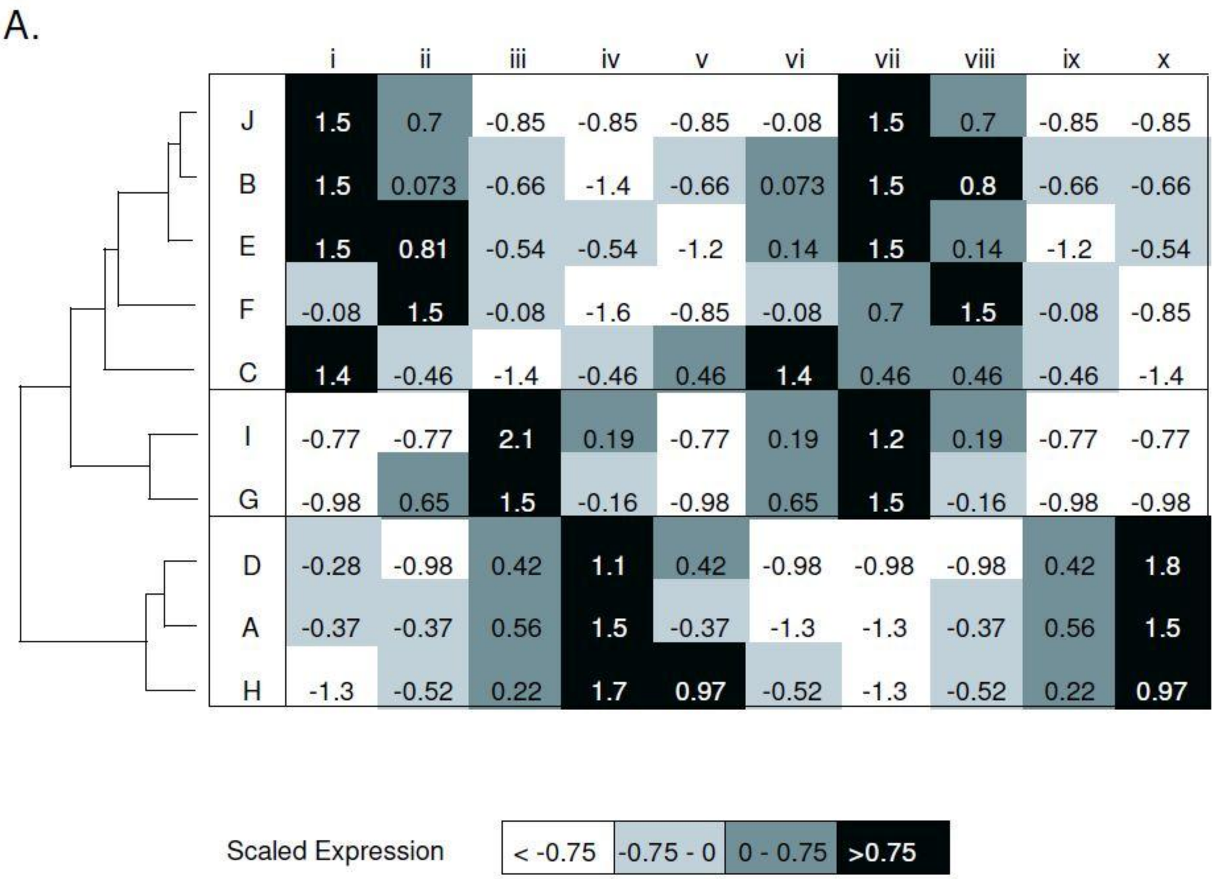


Table 11.1. Illustrative data for Fig. 11.9. Rows correspond to "genes" in alphabetical order, and columns correspond to conditions. Data have been scaled by rows using the means and standard deviations for each row.

	i	*ii*	*iii*	*iv*	*v*	*vi*	*vii*	*viii*	*ix*	*x*
A	−0.370	−0.370	0.560	1.50	−0.37	−1.300	−1.30	−0.37	0.560	1.50
B	1.500	0.073	−0.660	−1.40	−0.66	0.073	1.50	0.80	−0.660	−0.66
C	1.400	−0.460	−1.400	−0.46	0.46	1.400	0.46	0.46	−0.460	−1.40
D	−0.280	−0.980	0.420	1.10	0.42	−0.980	−0.98	−0.98	0.420	1.80
E	1.500	0.810	−0.540	−0.54	−1.20	0.140	1.50	0.14	−1.200	−0.54
F	−0.078	1.500	−0.078	−1.60	−0.85	−0.078	0.70	1.50	−0.078	−0.85
G	−0.980	0.650	1.500	−0.16	−0.98	0.650	1.50	−0.16	−0.980	−0.98
H	−1.300	−0.520	0.220	1.70	0.97	−0.520	−1.30	−0.52	0.220	0.97
I	−0.770	−0.770	2.100	0.19	−0.77	0.190	1.20	0.19	−0.770	−0.77
J	1.500	0.700	−0.850	−0.85	−0.85	−0.078	1.50	0.70	−0.850	−0.85

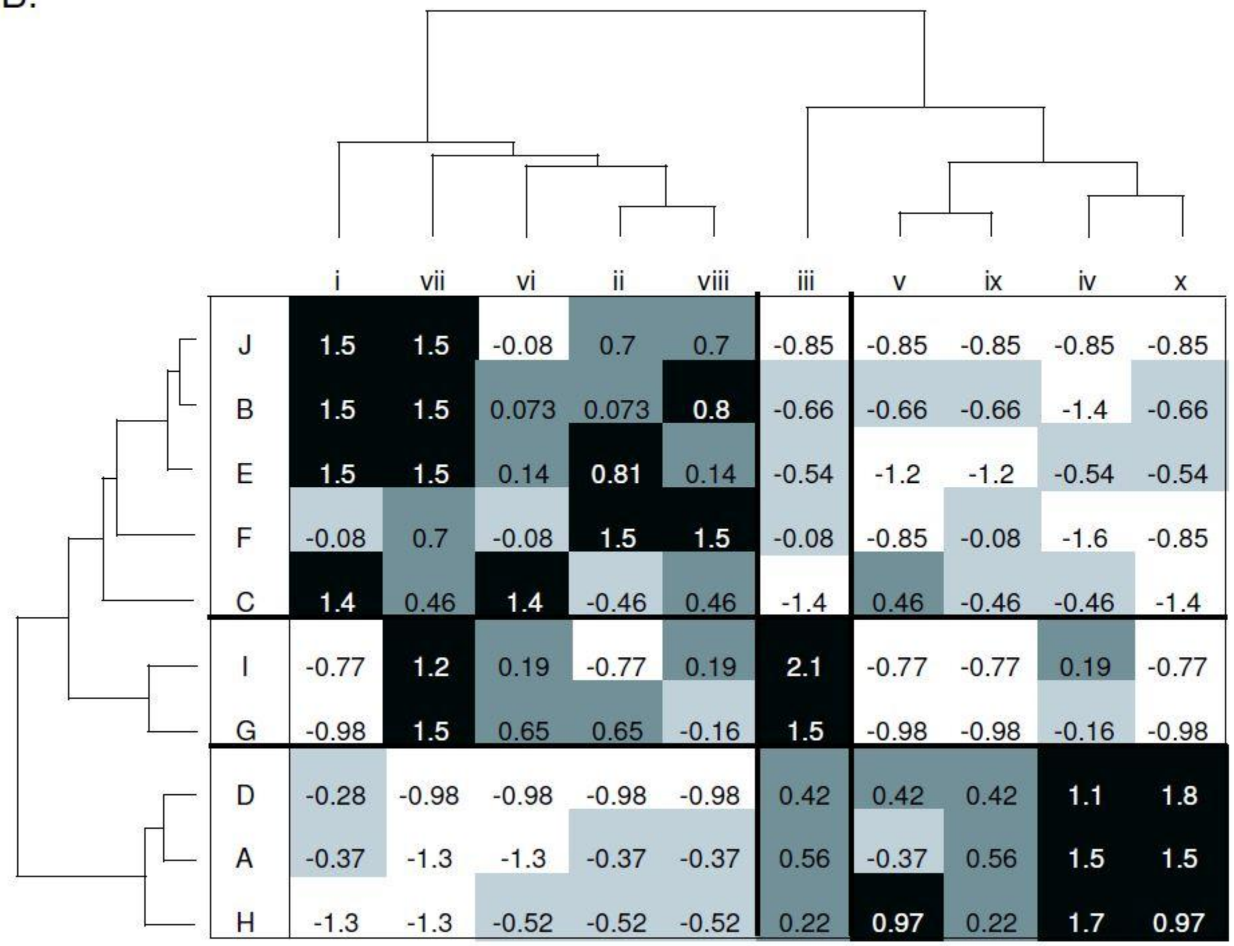
In the above, rows (letters) correspond to "genes" and columns (roman numerals) to "variables". In panel A, these data have been clustered by rows (using in R dist(,method= "euclidean") and hclust(,method="average") and then plclust). This defines three clusters of genes (which show similar expression patterns):



	i	ii	iii	iv	v	vi	vii	viii	ix	x
J	1.5	0.7	-0.85	-0.85	-0.85	-0.08	1.5	0.7	-0.85	-0.85
B	1.5	0.073	-0.66	-1.4	-0.66	0.073	1.5	0.8	-0.66	-0.66
E	1.5	0.81	-0.54	-0.54	-1.2	0.14	1.5	0.14	-1.2	-0.54
F	-0.08	1.5	-0.08	-1.6	-0.85	-0.08	0.7	1.5	-0.08	-0.85
C	1.4	-0.46	-1.4	-0.46	0.46	1.4	0.46	0.46	-0.46	-1.4
I	-0.77	-0.77	2.1	0.19	-0.77	0.19	1.2	0.19	-0.77	-0.77
G	-0.98	0.65	1.5	-0.16	-0.98	0.65	1.5	-0.16	-0.98	-0.98
D	-0.28	-0.98	0.42	1.1	0.42	-0.98	-0.98	-0.98	0.42	1.8
A	-0.37	-0.37	0.56	1.5	-0.37	-1.3	-1.3	-0.37	0.56	1.5
H	-1.3	-0.52	0.22	1.7	0.97	-0.52	-1.3	-0.52	0.22	0.97

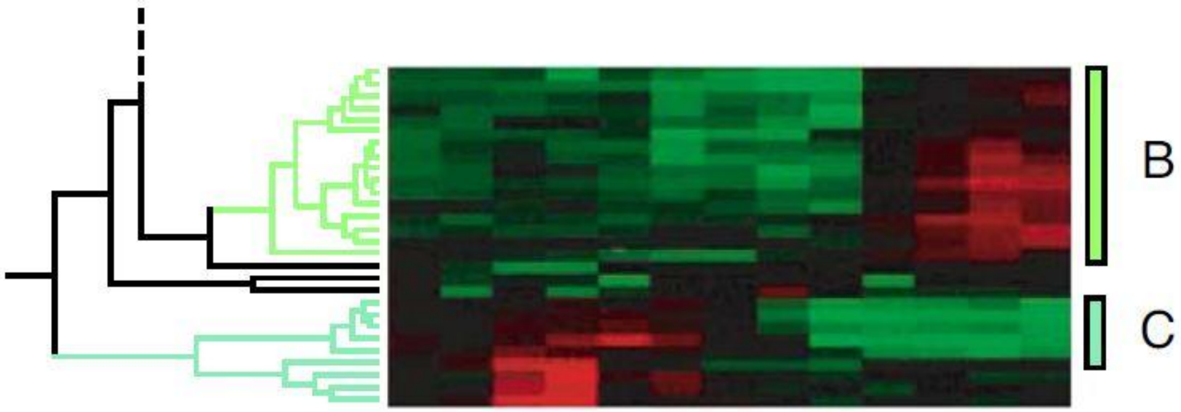
Next, since we're also interested in patterns among the variables, we transpose the above, and apply the same technique to the variables – we wish to identify groups of "conditions" whose genes show similar expression patterns. This results in panel B below.

B.

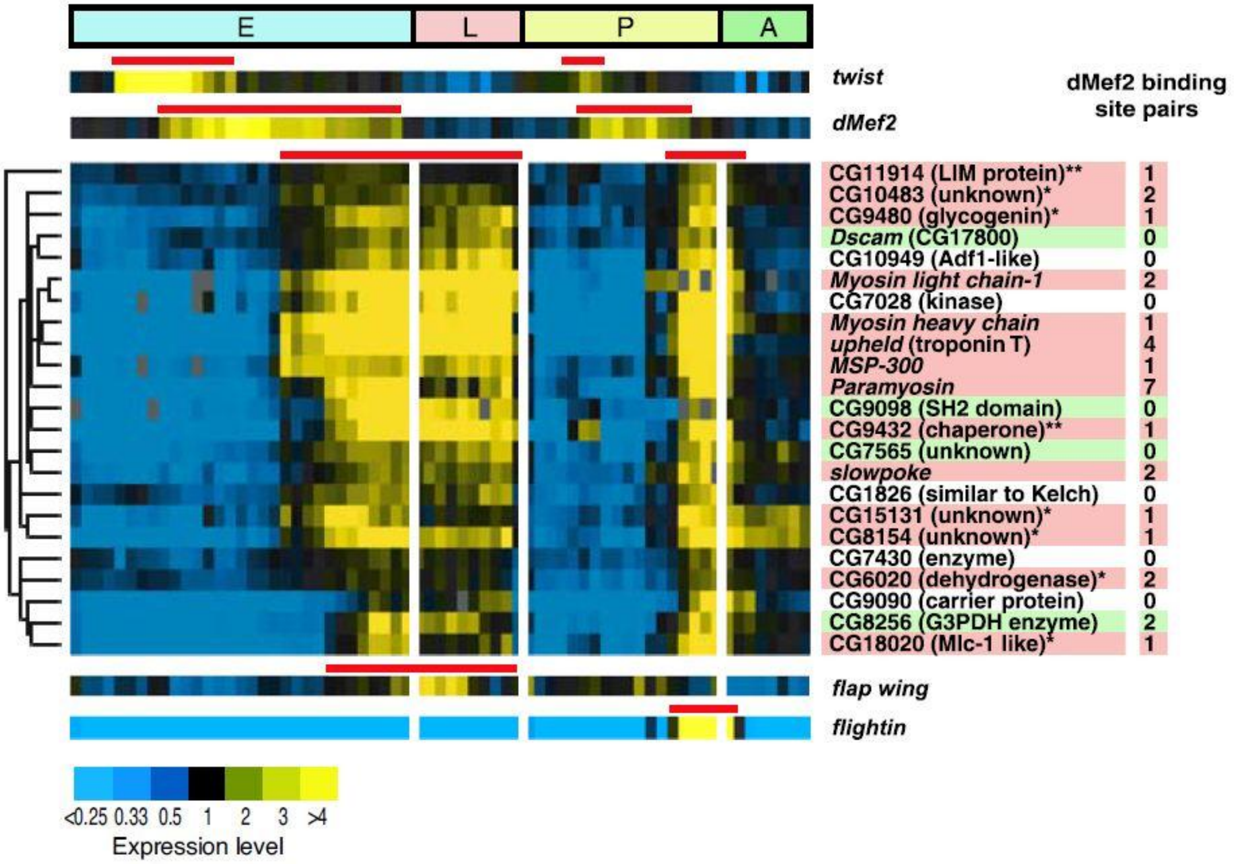


	i	vii	vi	ii	viii	iii	v	ix	iv	x
J	1.5	1.5	-0.08	0.7	0.7	-0.85	-0.85	-0.85	-0.85	-0.85
B	1.5	1.5	0.073	0.073	0.8	-0.66	-0.66	-0.66	-1.4	-0.66
E	1.5	1.5	0.14	0.81	0.14	-0.54	-1.2	-1.2	-0.54	-0.54
F	-0.08	0.7	-0.08	1.5	1.5	-0.08	-0.85	-0.08	-1.6	-0.85
C	1.4	0.46	1.4	-0.46	0.46	-1.4	0.46	-0.46	-0.46	-1.4
I	-0.77	1.2	0.19	-0.77	0.19	2.1	-0.77	-0.77	0.19	-0.77
G	-0.98	1.5	0.65	0.65	-0.16	1.5	-0.98	-0.98	-0.16	-0.98
D	-0.28	-0.98	-0.98	-0.98	-0.98	0.42	0.42	0.42	1.1	1.8
A	-0.37	-1.3	-1.3	-0.37	-0.37	0.56	-0.37	0.56	1.5	1.5
H	-1.3	-1.3	-0.52	-0.52	-0.52	0.22	0.97	0.22	1.7	0.97

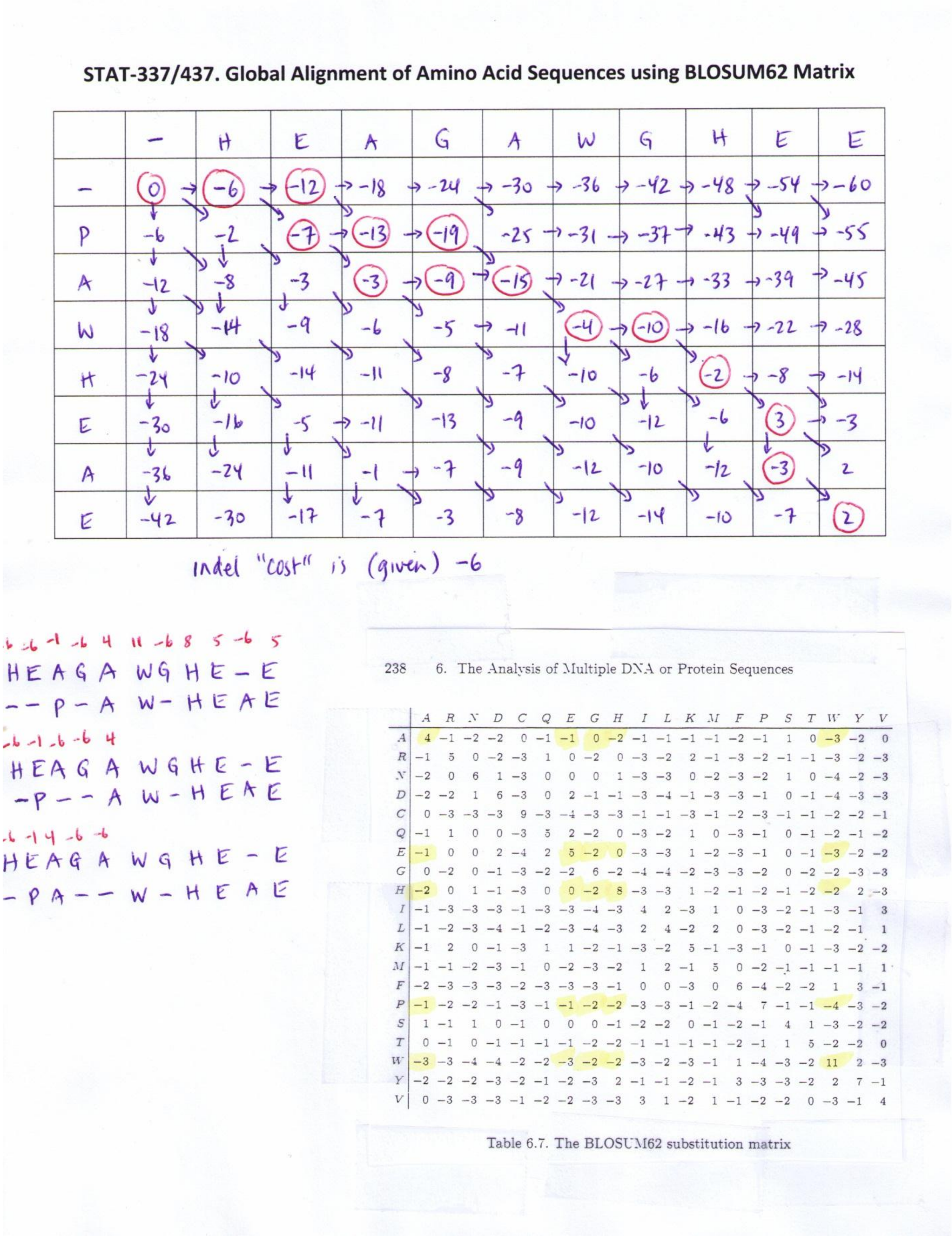
11.6.1. Gene Expression in Human Fibroblasts: from text: animal cells growing in tissue culture ordinarily require growth factors – here, by adding serum to the growth medium. Here, human fibroblast cells isolated from foreskins were cultured in a medium lacking serum, and after 48 hours serum was restored – 8600 genes were assessed over time (over 24 hours), and genes having like expression patterns were clustered to identify the stages of the cell cycle during which their expression was elevated or reduced. Here, cluster B represented genes involved in the cell cycle, and cluster C consists of genes involved in the immediate-early response. (Recall: red is up-regulated, and green is down-regulated.) Notice that at first, cluster B genes is down-regulated and then up-regulated, and the opposite occurs for cluster C.



11.6.2. Gene Expression During Drosphila Development: development of Drosphila progresses from embryo (E) through larval (L), pupal (P), and adult (A) stages, and understanding the genes involved in each stage is paramount. Here, 4028 genes were measured over the entire Drosophila life cycle. For the gene clusters: members of the resulting clusters of genes were often functionally related. For example, one cluster was enriched for genes active in terminally differentiated muscle.



Miscellany: Aligning Amino Acid Sequences



STAT-337/437. Global Alignment of Amino Acid Sequences using BLOSUM62 Matrix

	–	H	E	A	G	A	W	G	H	E	E
–	0	-6	-12	-18	-24	-30	-36	-42	-48	-54	-60
P	-6	-2	-7	-13	-19	-25	-31	-37	-43	-49	-55
A	-12	-8	-3	-3	-9	-15	-21	-27	-33	-39	-45
W	-18	-14	-9	-6	-5	-11	-4	-10	-16	-22	-28
H	-24	-10	-14	-11	-8	-7	-10	-6	-2	-8	-14
E	-30	-16	-5	-11	-13	-9	-10	-12	-6	3	-3
A	-36	-24	-11	-1	-7	-9	-12	-10	-12	-3	2
E	-42	-30	-17	-7	-3	-8	-12	-14	-10	-7	2

indel "cost" is (given) -6

-6 -6 -1 -6 4 11 -6 8 5 -6 5
HEAGA WGHE-E
--P-A W-HEAE

-6 -1 -6 -6 4
HEAGA WGHE-E
-P--A W-HEAE

-6 -1 4 -6 -6
HEAGA WGHE-E
-PA-- W-HEAE

238 6. The Analysis of Multiple DNA or Protein Sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	1	−2	−1	−2	−1	−2	−2	2	−3
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−3	−2	2	7	−1
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Table 6.7. The BLOSUM62 substitution matrix