Discriminant analysis to evaluate clustering of gene expression data

Marco A. Méndez[a,1], Christian Hödar[a,1], Chris Vulpe[b], Mauricio González[a], Verónica Cambiazo[a,*]

[a] *Laboratorio de Bioinformática y Expresión Génica, INTA, Universidad de Chile, Macul 5540, Macul, Santiago, Chile*
[b] *Nutrition and Toxicology, 119 Morgan Hall, University of California, Berkeley, CA, USA*

Received 5 March 2002; revised 16 May 2002; accepted 16 May 2002

First published online 3 June 2002

Edited by Julio Celis

Abstract In this work we present a procedure that combines classical statistical methods to assess the confidence of gene clusters identified by hierarchical clustering of expression data. This approach was applied to a publicly released *Drosophila* metamorphosis data set [White et al., Science 286 (1999) 2179–2184]. We have been able to produce reliable classifications of gene groups and genes within the groups by applying unsupervised (cluster analysis), dimension reduction (principal component analysis) and supervised methods (linear discriminant analysis) in a sequential form. This procedure provides a means to select relevant information from microarray data, reducing the number of genes and clusters that require further biological analysis.

Key words: Microarray; Gene expression data; Cluster analysis; Principal component analysis; Discriminant analysis; *Drosophila*

1. Introduction

Advances in microarray technology have enabled us to measure the simultaneous expression of thousands of genes under multiple experimental conditions [2–4]. A main step in the analysis of gene expression data is the detection of gene groups with similar expression patterns. Several approaches to the computational analysis of gene expression data attempt functional classification of genes using clustering algorithms [2,4]. The visual presentation of hierarchical clustering enables easy recognition of groups of genes (clusters) that may be related in terms of their biological functions [2,5,6]. These clustering algorithms, however, do not measure whether gene correlations are higher than would be expected by chance, and thus provide no information on the statistical confidence of a particular cluster.

In the present report we propose a general procedure, that permits evaluation of both the statistical support for gene clusters described by a hierarchical clustering method and the correct classification of genes within the groups. Our procedure is based on two multivariate techniques, principal component analysis (PCA) and linear discriminant analysis (LDA). Following a hierarchical clustering of the expression data, groups of genes were defined from the branching pattern of the cluster tree. Using gene expression data values from raw data as input for PCA, we recovered new matrices with loading values, and applied LDA to determine the statistical support of the selected clusters. This allowed us to identify genes that remain in the clusters as correctly classified, after LDA. This procedure provides a means to extract information with statistical confidence from microarray data, reducing the complexity of large data sets, and therefore the number of genes and clusters requiring further biological analyses. We applied this approach to a published gene expression data set [1].

2. Materials and methods

2.1. Biological data

The data set used in this work was generated by White et al. [1] and it is publicly available at http://quantgen.med.yale.edu. The data contain the ratios of red/green (R/G) gene expression values for 4500 EST clones along with control cDNAs, corresponding to ecdysone-regulated genes. The data were collected at six time-points spanning two pulses of ecdysone: the late larval ecdysone pulse that arises 6 h prior puparium formation (PF), peaks at puparation and rapidly declines and the prepupal pulse that peaks 10 h after PF. Time points were examined relative to PF: ≥ 18 h and 4 h before PF; and 3, 6, 9 and 12 h after PF. The analysis revealed that 534 genes exhibited three-fold or more differential expression during early metamorphosis.

2.2. PCA

PCA is a multivariate statistical tool that simplifies complex data sets [7]. It has been previously applied to gene expression data obtained from microarray experiments [8–10]. In general, PCA changes the original variables into new independent and uncorrelated variables called principal components that explain the observed variability. The first components explain the majority of variability and concentrate the maximal amount of information from the experiment. For each component, it is possible to find one eigenvalue with an associated variance value (explained variance). The n eigenvalues and their corresponding eigenvectors originate from the $n \times n$ covariance matrix obtained from the original data. In our procedure, we considered the eigenvalue of each gene as a variable, generating a 'principal gene analysis', which shows the gene expression behavior that best explains the observed experimental response [9]. After run PCA, new matrices with loading expression values for each transformed gene were obtained.

2.3. LDA

This analysis uses a classification function (f_i) to calculate scores of each variable in the different groups, following the general formula:

$$f_i = \sum_j w_{ij} x_j + \text{constant};\ i = 1, ..., n;\ j = 1, ..., n$$

where i corresponds to groups, w_{ij} is the weight for the jth variable, in

*Corresponding author. Fax: (56)-2-221 4030.
E-mail address: vcambiaz@uec.inta.uchile.cl (V. Cambiazo).

[1] These authors contributed equally to this work.

0014-5793/02/$22.00
PII: S0014-5793(02)02873-9

the computation of the classification score for the ith group, and x_j is the observed value for the corresponding jth variable. We used LDA, instead of other discriminant techniques, such as classification trees, since LDA performs better than other when log transformed data are used [11]. This technique enables us to identify relationships between qualitative variables or classes (in our case, clusters of genes) and quantitative predictor variables (in our case, eigenvectors) [7]. Since we know the clusters we built a linear discriminant function to estimate the significance of gene classification within each one. Cross-validation and jackknife procedures were used to produce unbiased estimates [12]. Using both procedures we obtained practically the same number of correct classifications (data not shown).

2.4. Strategy of analysis

The original R/G values ($n = 534$) of gene expression from White et al. [1] were log transformed and subjected to PCA to obtain a covariance matrix that is available at http://www.inta.cl/genexpression/new/index.htm. PCA was applied to both the entire set of gene expression values and to the values corresponding to genes in groups 4 and 8. The matrix containing the eigenvalues for each gene expression value in the first three components and the groups recovered by ClusterAnalysis were used as input for LDA. These input data met the assumption underlying LDA, they had a normal multivariate distribution [13], and the covariance matrices for every class were equally distributed, since not-significant differences were found after applying the Sen and Puri test [14].

3. Results and discussion

We have analyzed the published data of White et al. [1], containing expression ratios for 534 genes from *Drosophila*. We used ClusterAnalysis software [2] to recover the tree shown in Fig. 1A, which was identical to that previously obtained [1]. We used the average-linkage method and Pearson correlation distances to perform this clustering analysis. Software implementation of the algorithm can be obtained from http://rana.lbl.gov/EisenSoftware.htm. Groups were selected by visual inspection of the dendrogram. Nine groups (1–9) of genes were located at a similar branching level, near the middle region. In addition, for groups 4 and 8, eight and seven subgroups were defined near the terminal region of the dendrogram (Fig. 1A).

3.1. Identification of five clusters with high percentage of correctly classified genes

PCA analysis of the entire gene expression data showed that the first three components contained 91.9% of the variability (first component 55.5%, second component 27.2% and third component 9.2%), indicating that we can summarize the

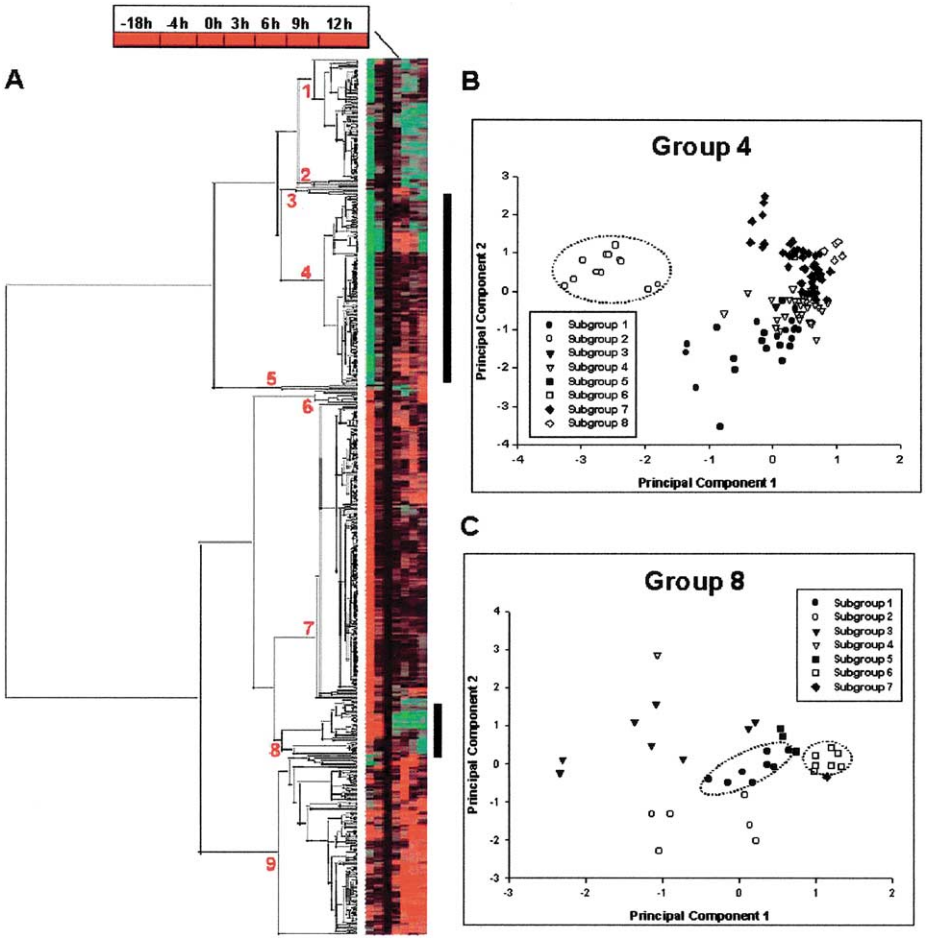

Fig. 1. Hierarchical clustering was applied to expression data from a set of 534 genes measured across six time-points during early *Drosophila* metamorphosis (A). Each time-point compared to PF is indicated above each column. Gene expression patterns are shown in the rows. Low or high gene expression levels are indicated by green or red colors. Nine groups of genes recovered by the clustering method are indicated in the dendrogram. The bars to the right indicate groups 4 and 8. Genes contained at group 4 (B) and at group 8 (C) were plotted with respect to the first and second principal components. Ellipses represent subgroups with 100% of classification values.

Table 1
Discriminant classification matrices[a]

A										
Groups	1	2	3	4	5	6	7	8	9	%[b]
1	**68**	5	0	0	1	0	0	0	0	92
2	0	**4**	0	0	0	0	0	0	0	100
3	0	0	**4**	1	0	0	0	0	1	80
4	25	1	15	**69**	6	0	0	0	0	59
5	0	0	0	0	**3**	0	0	0	0	100
6	0	0	0	0	4	**3**	1	0	0	38
7	0	17	0	0	1	0	**123**	39	0	68
8	0	10	3	0	0	0	9	**12**	0	35
9	0	0	30	9	25	13	4	0	**29**	26
B										
Groups	Subgroups	1	2	3	4	5	6	7	8	%[b]
4	1	**17**	0	6	0	0	0	0	0	74
4	2	0	**12**	0	0	0	0	0	0	100
4	3	0	0	**2**	0	0	0	0	0	100
4	4	0	0	1	**17**	11	0	0	0	59
4	5	0	0	0	0	**3**	0	0	0	100
4	6	0	0	0	0	1	**1**	1	0	33
4	7	0	17	0	0	6	11	**21**	1	54
4	8	0	0	0	0	0	0	1	**4**	80
8	1	**8**	0	0	0	0	0	0		100
8	2	1	**5**	0	0	0	0	0		83
8	3	1	0	**5**	0	2	0	0		63
8	4	0	0	1	**0**	0	0	0		0
8	5	0	0	0	0	**2**	1	0		67
8	6	0	0	0	0	0	**7**	0		100
8	7	1	0	0	0	0	0	**0**		0

The number of genes correctly assigned to each group or subgroup is indicated.
[a]Groups of genes recovered from the tree shown in Fig. 1A or subgroups contained within groups 4 and 8 (Fig. 1B,C) were used as qualitative variables to build a linear discriminant function. The eigenvalues for each gene expression value in the first three components of PCA were used as predictor variables.
[b]Percentages of correctly classified genes.

data in just three gene expression features that explain most of the total variability observed. As has been mentioned by [8] the use of PCA enables one to make a classification with biologically meaningful characteristics. The advantages of this approach to analyze the data have been previously described by [8–10]. However, PCA is a graphical technique to visualize the gene expression data, it does not give us information on the existence of different groups with an associated probability. In our approach we apply a linear discriminant function to obtain this information and to resolve the proper classification of each gene within the groups. Our results of a LDA using the gene loading matrix obtained from PCA and the groups recovered by ClusterAnalysis revealed that the nine groups were statistically different (Wilks' $\lambda = 0.3671$; $F = 26.089$; df = 24, 1517; $P < 0.0001$). The discriminant classification matrix showed that only five of the nine groups contained $> 60\%$ of genes with correct classification. Only two had 100% correctly classified genes (Table 1A). If we consider genes that were correctly assigned to one of the five groups with $> 60\%$ classification value, then the original 534 genes under analysis can be reduced to 202. A practical consequence of this method is an increase in confidence of the genes selected for future studies and a reduction in the number of genes to be considered.

3.2. Subgroups within cluster with low percentage of correctly classified genes

In order to extract additional information from the data, groups with $< 60\%$ correct classification that contained a high number of genes were re-examined. For this, the matrices of eigenvalues for genes in groups 4 and 8 and the corresponding subgroups recovered by ClusterAnalysis were used as input for LDA. Group 4 contains the subgroup of control genes named L71 and seven others; and group 8 contains control genes SGS (salivary gland secretion) and six additional subgroups. Both L71 and SGS subgroups were reported by [1] as containing sets of genes that cluster together and show expression patterns of ecdysone-regulated genes. In the case of group 4, PCA analysis revealed that the first three components explained 89.2% of the variability (first component 59.0%, second component 20.6% and third component 9.6%). When we applied LDA to this data set we found that the subgroups were statistically different (Wilks' $\lambda = 0.026$; $F = 37.371$; df = 21, 304; $P < 0.0001$). The discriminant classification matrix showed that five of the eight subgroups presented $> 70\%$ correctly classified genes ($n = 38$) (Table 1B). Two subgroups (3 and 5) with 100% correct classification were detected in addition to the L71, yet they contain only one and two genes, respectively. In agreement with the coordinated expression of the L71 genes during late larval ecdysone pulse [15], when the variance pattern is graphically represented, they fall into a well-defined cluster (Fig. 1B).

PCA analysis of genes in group 8 showed that the first three components explained 91.2% of the variability (first component 62.6%, second component 21.2% and third component 7.4%). LDA revealed that subgroups were statistically different (Wilks' $\lambda = 0.011$; $F = 15.344$; df = 18, 71; $P < 0.0001$). In this case three subgroups with $> 70\%$ correct classification were detected ($n = 20$ genes), among them two subgroups (1 and 6) had 100% correctly classified genes ($n = 15$). The subgroup containing control genes SGS showed that only 63% correct classifications (Table 1B) since gene SGS4 was not

Table 2
Genes contained in groups 4 and 8[a]

Group	Sub-group	GenBank	Name	Description	Biological process
4	1	K00670	actin 42A	structural cytoskeletal protein	cytoskeleton organization
4	1	AE003608	–	putative ankyrin	membrane–cytoskeleton linker protein
4	1	AF277390	dystroglycan-like	dystrophin complex component	cytoskeletal anchoring protein
4	1	U19909	corkscrew	tyrosine phosphatase	EGF/Torso receptors signaling pathways
4	1	M55099	ecdysone-inducible gene E2	uncharacterized secreted protein	imaginal disk morphogenesis
4	1	AF137269	innexin 2	component of the gap junction	cell–cell communication
4	1	AF168467	smell impaired 35A	serine/threonine kinase	olfaction
4	1	–	ecdysone-induced gene 87F	unknown	response to ecdysone
4	7	AF106932	plexin A	transmembrane receptor	axon guidance
4	7	X53837	neurotactin	cell adhesion, transmembrane protein	axon guidance
4	7	AF040989	roundabout	transmembrane receptor	axon guidance, midline recognition
4	7	AF038842	midline fasciclin	cell adhesion, transmembrane protein	axonogenesis
4	7	AF275903	echinoid	cell adhesion, transmembrane protein	EGF receptor signaling pathways
4	7	AA220496	p120ctn	adherens junction component	intercellular adhesion, cell migration
4	7	AF197345	prominin-like	putative transmembrane glycoprotein	generation of membrane protrusions
4	7	Y17922	unc-13	diacylglycerol binding protein	neurotransmitter release
4	7	U76378	LK6	serine/threonine kinase	microtubule binding and organization
4	7	NM_079414	reaper	caspase-dependent apoptosis activator	induction of apoptosis
4	7	X56689	protein on ecdysone puffs	hnRNA binding protein	mRNA processing and stability
4	7	AA202479	–	putative Na/PO_4 cotransporter	unknown
4	7	AI945337	–	RNA binding protein	unknown
4	7	AE003628	–	transcriptional repressor	unknown
4	7	AE003629	–	transcription factor	unknown
4	8	AF104357	Nedd2-like caspase	caspase-2	apoptosis
4	8	U25686	Eip93F	transcription factor	apoptosis and autophagy
4	8	AF119332	brain tumor	translational repressor	brain/imaginal disk regulation of growth
4	8	AF211192	sulfated	*N*-acetylglucosamine-6-sulfatase	pattern specification
8	1	AB43874	frost	putative secreted protein	cold resistance
8	1	AF311747	peroxiredoxin 5037	thioredoxin peroxidase	antioxidant
8	1	AE003740	–	putative glutamate-cysteine ligase	glutathione synthesis
8	1	U51047	α-esterase-5	carboxylesterase	unknown
8	2	AE003564	–	putative serine-type endopeptidase	unknown
8	2	M34147	vermillion	tryptophan 2,3-dioxygenase	tryptophan metabolism
8	2	AE003695	Cyp9f2	cytochrome P450	metabolic detoxification
8	6	AI947049	Sec61β	protein transporter translocon component	
8	6	AF181658	SrpRβ	signal recognition particle receptor	
8	6	AF160923	SsRβ	signal sequence receptor	targeting of secreted and membrane
8	6	AI945207	–	putative translocon-associated protein γ	proteins to endoplasmic reticulum
8	6	AL109630	–	putative TRAM protein	
8	6	AF160889	–	putative signal peptidase	

[a]Subgroups containing a significant number of genes with known or predicted functions are shown. Subgroups containing the control genes L71 and SGS are not listed. A complete list of genes in groups 4 and 8 can be found at http://www.inta.cl/genexpression/new/index.htm.

correctly classified into this subgroup, indicating that its expression pattern differs from the other SGS genes. This result is in agreement with previous reports on a non-coordinate expression of SGS genes. They comprise a family of genes expressed at high levels in the salivary glands of late third instar larvae in response to ecdysone. In contrast to the other SGS genes, SGS4 is turned on throughout *Drosophila* development and is not expressed exclusively in the larval salivary glands [16]. These results indicate that our protocol may be useful to detect, with statistical confidence, additional features of gene expression patterns that were overlooked by a hierarchical clustering method. A graphical representation of the first two components of PCA (Fig. 1C) shows the high degree of dispersion of SGS genes within the multivariate space, while genes at subgroups 1 and 6 fall into delimited clusters (Fig. 1C).

3.3. Discriminant analysis provides a tool to identify properly classified genes within a cluster

Discriminant analyses have been applied to various genome-wide gene expression studies in order to build statistical models to categorize new samples based on the gene expression of few selected genes [9,11,17–18]. Here, we proposed a different procedure to study microarray data by using traditional statistical methods in a sequential manner. This approach incorporates the gene classification identified by a clustering method (we based our analysis on hierarchical clustering but it can be applied to any other clustering method) with an eigenvector matrix of gene expression values to generate a discriminant classification of the cluster groups. This procedure provides statistical support for the selection of genes before beginning the biological analysis of gene expression microarray data. Different methods to estimate the

accuracy of the groups recovered by clustering techniques have been proposed [19]. For instance, Bootstrap has been used to assess statistical confidence of nodes in a tree, and this approach was used to find nodes with either high or low values of statistical support [13,20]. However, to our knowledge these approaches have not been used to assess the correct assignment of individual genes within the nodes.

3.4. Biological analysis of new subgroups identified by LDA within a cluster

We selected groups 4 and 8 for the analysis of their gene contents and distribution (a complete classification of the genes can be found at http://www.inta.cl/genexpression/new/index.htm). Following LDA, group 4 contained 69 correctly classified genes (Table 1B), 31 of these had a known function. Out of these 31 genes, eight correspond to ecdysone-regulated genes (ImpE2, L711-6, Hsp23, Actin 42A, Eip 63E). The other genes are mainly associated with tissue reorganization and differentiation, including: (1) members of the EGF receptor signaling pathway (corkscrew, echinoid); (2) genes involved in CNS remodeling (mub, plexin A, neurotactin, roundabout, unc-13, midline fasciclin), several of these genes cluster together at subgroup 7, along with unknown transcripts; (3) genes involved in programmed cell death (reaper, Dronc, Eip 93F); (4) genes encoding cytoskeletal regulators (stathmin, mig-2-like, dystroglycan, LK6), some of them are found in subgroup 1, along with the gene encoding actin 42A (Table 2).

Group 8 contains 20 correctly classified genes (Table 1B); out of the 11 genes with a known function, five are involved in stress response, such as frost (response to cold), prx5037 (antioxidant) and Cyp9f2 (detoxification). Three of them plus an unknown transcript are clustered at subgroup 1. Three genes are components of the SRP-dependent membrane-targeting complex (Sec62β, SrpRβ, SsRβ), they cluster together at subgroup 8, along with three uncharacterized genes with predicted functions on targeting proteins to the endoplasmic reticulum. The remaining three genes encode puparial glue proteins (SGS) (Table 2). Our exploration of groups 4 and 8 has used the features of known genes sharing expression patterns with unknown genes to provide clues on their function or common mechanism of gene regulation (Table 2). This is particularly the case for those genes that are clearly regulated by ecdysone and for the genes within subgroups 1 and 8 with 100% correct classification, since they can be rapidly assigned to distinguishable cellular functions.

Most of the time, the extraction of biological information from clusters of genes requires access to external information for every gene in a cluster that permits to recognize functional roles or to decipher networks of interactions. In addition, expression patterns of genes need to be verified by independent assays. Therefore, a great deal of effort is needed to confirm the information obtained from microarray data. The protocol described here enables filtering of the data by applying statistical criteria to clusters of genes and thus to identify those genes correctly assigned to the original groups defined by a clustering method. This procedure reduces the complexity of large sets of data and increases the accuracy in the selection of candidate genes susceptible to further biological analysis.

Acknowledgements: This work was supported by Fondecyt 1000852, 3000048 and 1010693, and by ICA/ILS-UC Berkeley Grant SA2962. We thank Alex Loguinov for critical discussions.

References

[1] White, K., Rifkin, S., Hurban, P. and Hogness, D. (1999) Science 286, 2179–2184.
[2] Eisen, M.B., Spellman, P., Brown, P. and Botstein, D. (1988) Proc. Natl. Acad. Sci. USA 95, 14863–14868.
[3] Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. and Davis, R.W. (1998) Trends Biotechnol. 16, 301–306.
[4] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. and Golub, T. (1999) Proc. Natl. Acad. Sci. USA 96, 2907–2912.
[5] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999) Proc. Natl. Acad. Sci. USA 96, 6745–6750.
[6] Getz, G., Levine, E. and Domany, E. (2000) Proc. Natl. Acad. Sci. USA 97, 12079–12084.
[7] Kachigan, S. (1986) Statistical Analysis: An Interdisciplinary Introduction to Univariate and Multivariate Methods, Radius Press, New York.
[8] Crescenzi, M. and Giuliani, A. (2001) FEBS Lett. 507, 114–118.
[9] Raychaudhuri, S., Stuart, J. and Altman, R. (2000) Pacific Symp. Biocomput. 5, 452–463.
[10] Yeung, K.Y. and Ruzzo, W.L. (2001) Bioinformatics 17, 763–774.
[11] Dudoit, S., Fridlyand, J. and Speed, T. (2002) J. Am. Statist. Assoc. 97 (457), 77–87.
[12] Hair, J., Anderson, R., Tatham, R. and Black, W. (1992) Multivariate Data Analysis with Readings, Macmillan, London.
[13] Zhang, K. and Zhao, H. (2000) Funct. Integr. Genomics 1, 156–173.
[14] Sen, P.K. and Puri, M.L. (1968) Sankhya 30, 1–22.
[15] Wright, L.G., Chen, T., Thummel, C.S. and Guild, G.M. (1996) J. Mol. Biol. 255, 387–400.
[16] Barnett, S.W., Flynn, K., Webster, M.K. and Beckendorf, S.K. (1990) Dev. Biol. 140, 362–373.
[17] Spanakis, E. and Brouty-Boyé, D. (1997) Int. J. Cancer 71, 402–409.
[18] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Science 286, 531–537.
[19] Yeung, K.Y., Haynor, D.R. and Ruzzo, W.L. (2001) Bioinformatics 17, 309–318.
[20] Sokal, R.S. and Rohlf, F.J. (1995) Biometry, 3rd edn., W.H. Freeman and Company, New York.