Content

1. Introduction

Using the “SelTag” software the data on gene expression from the “notterman2001_set1.txt” file were analyzed. The data consist of measured expression levels for approximately 7400 human cDNAs and ESTs (Human 6500 GeneChip Set, Affymetrix) in colon adenocarcinomas from several patients (Notterman *et al*., 2001). Expression levels were also measured in normal colon tissues. Totally the table contains the expression measurements for 18 tumor and 18 normal counterpart colon tissues. These data are divided into “Tumor” and “Normal” groups. Data in columns of expression matrix were normalized in such a way that average value was equal to 50. The aims of analysis were the following:

1. Identification of genes, highly expressed in tumorous tissues and lowly expressed in normal ones, which are supposed to be related to cells tumorigenicity;
2. Identification of genes, highly expressed in normal tissues and lowly expressed in tumorous ones;
3. Identification of genes with similar expression profiles;
4. Clustering genes by their expression profiles;
5. Analysis of principal components for genes expression matrix.

2. Application start and data loading.

2.1. On application start the “Login” dialog appears (fig. 2.1). Choose the “Anonymous” entry mode and press “OK”.

Note. “Anonymous” entry is purposed for use with demo data only.

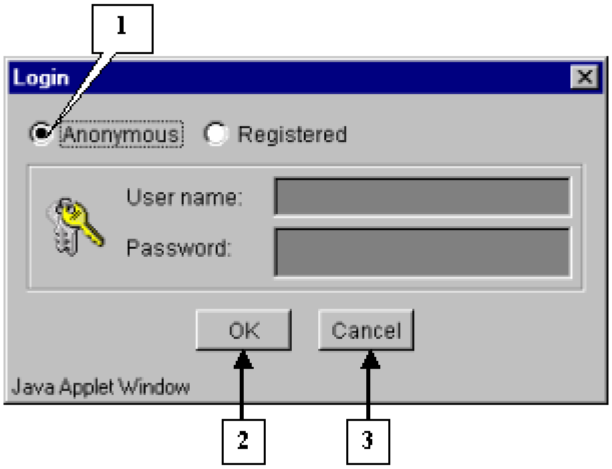


Figure 2.1.

1. Anonymous (non registered user) mode. **2.** Confirmation button. **3.** Cancel button.

2.2. Once the login procedure is over, the application main window appears. To load data from file, select the “File>Open data” command from the main menu (fig. 2.2).

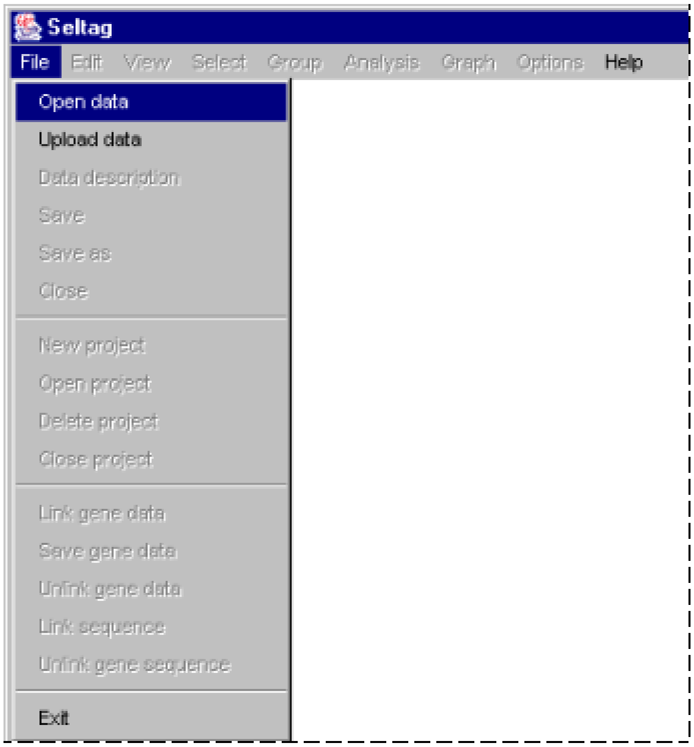
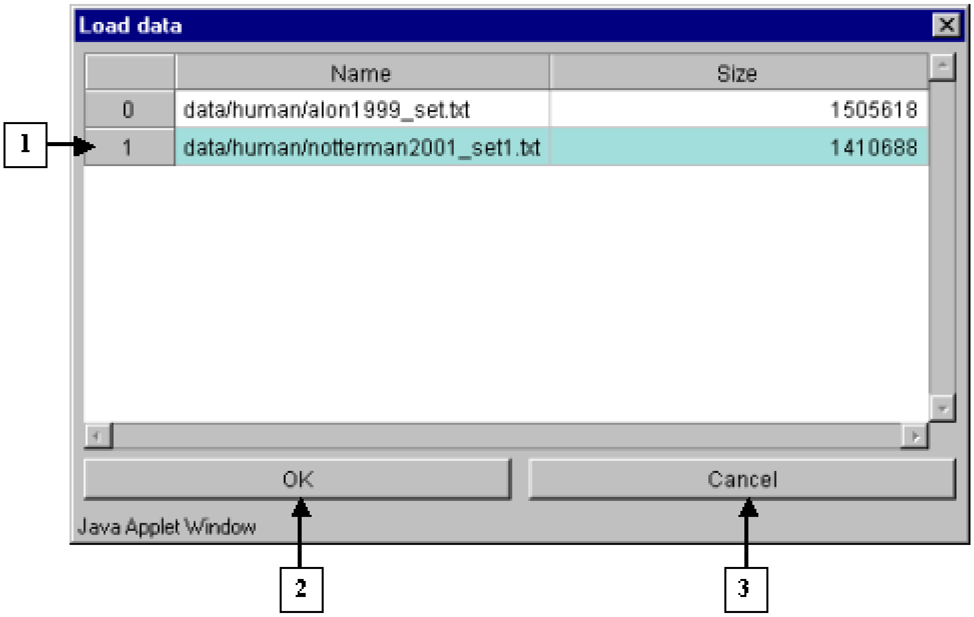


Figure 2.2.

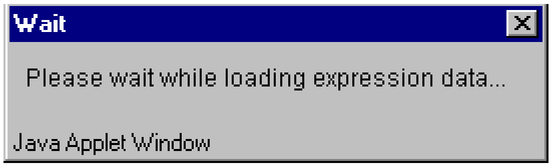
2.3. It will cause the “Load data” window, containing the names of files with table data and information on their size (fig. 2.3.1), to appear. Choose the appropriate filename and press “OK”. The “Wait” message window will appear indicating the data loading process, and once the process is over, it will disappear (fig. 2.3.2)



	Name	Size
0	data/human/alon1999_set.txt	1505618
1	data/human/notterman2001_set1.txt	1410688

Figure 2.3.1.

1. Selected file with expression data. **2.** Confirmation button. **3.** Cancel button.

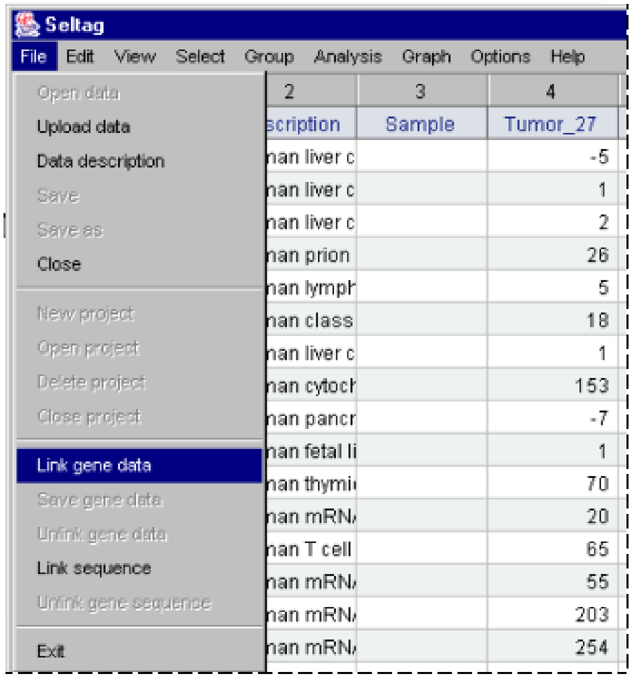


Please wait while loading expression data...

Figure 2.3.2.

2.4. Table with selected data will be represented in the application main window.

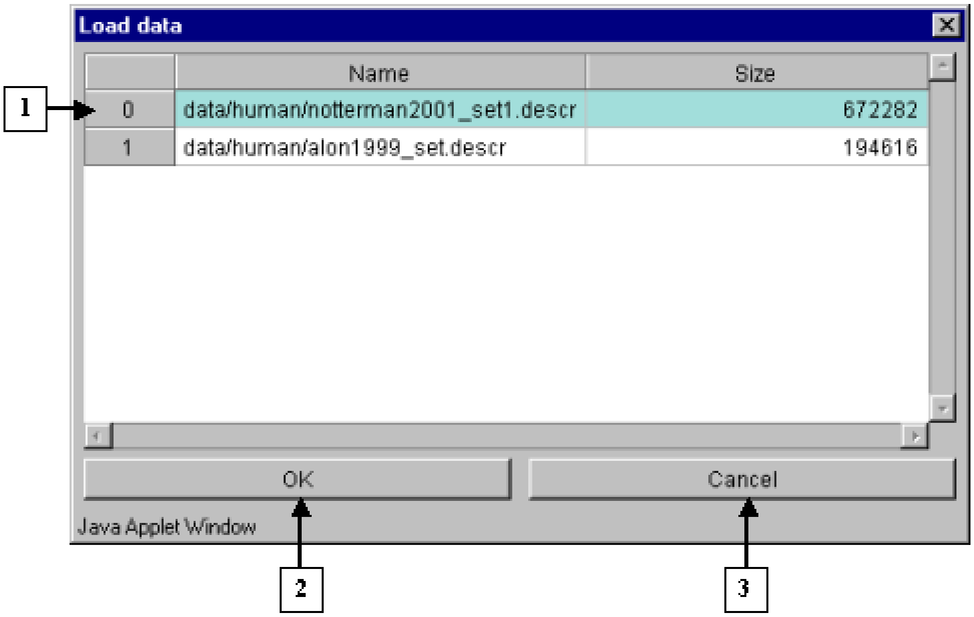
2.5. To load a file with genes description, select the “File>Link gene data” command from the main menu (fig. 2.5).



2	3	4
scription	Sample	Tumor_27
nan liver c		-5
nan liver c		1
nan liver c		2
nan prion		26
nan lymph		5
nan class		18
nan liver c		1
nan cytoch		153
nan pancr		-7
nan fetal li		1
nan thymi		70
nan mRN		20
nan T cell		65
nan mRN		55
nan mRN		203
nan mRN		254

Figure 2.5.

2.6. The “Load data” dialog with list of files and information on their size (fig. 2.6) will appear. Choose the file of interest and press “OK”.

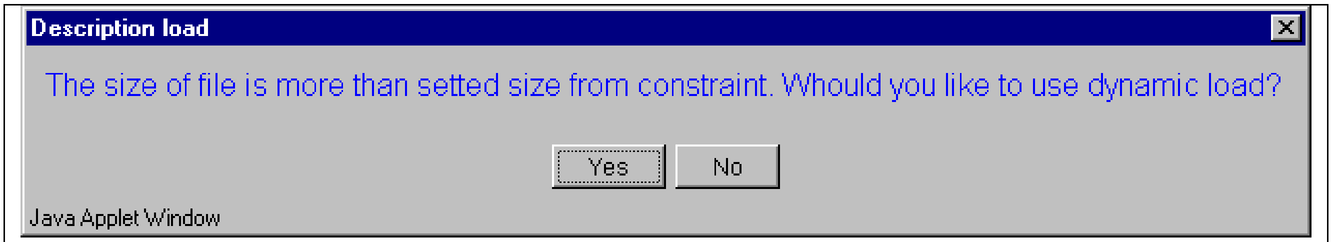


	Name	Size
0	data/human/notterman2001_set1.descr	672282
1	data/human/alon1999_set.descr	194616

Figure 2.6.

1. Selected file with genes description. **2.** Confirmation button. **3.** Cancel button.

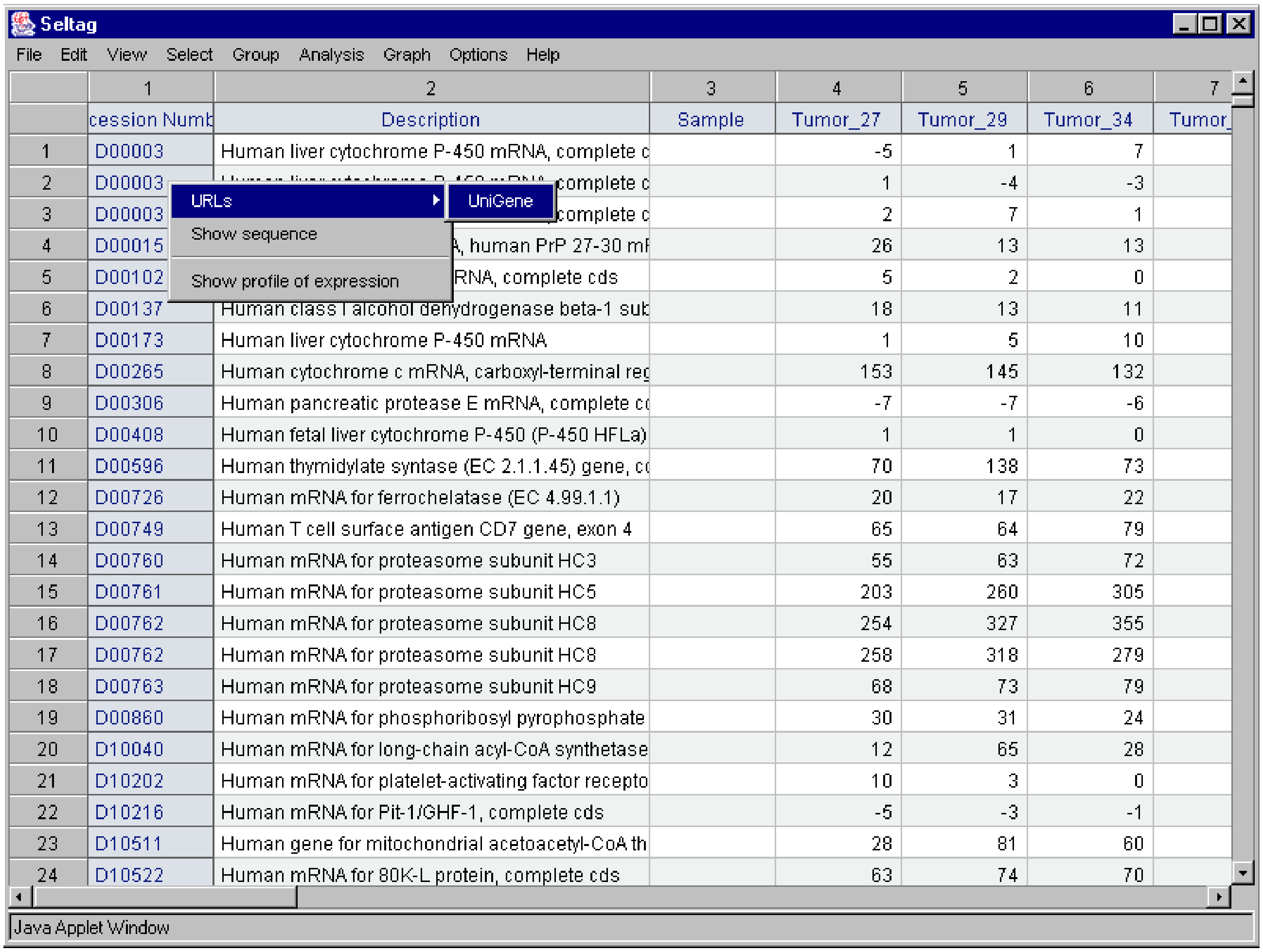
2.7. The “Description load” dialog with invitation to use dynamic file loading mode (fig. 2.7) will appear. Press the “Yes” button.



The size of file is more than setted size from constraint. Whould you like to use dynamic load?

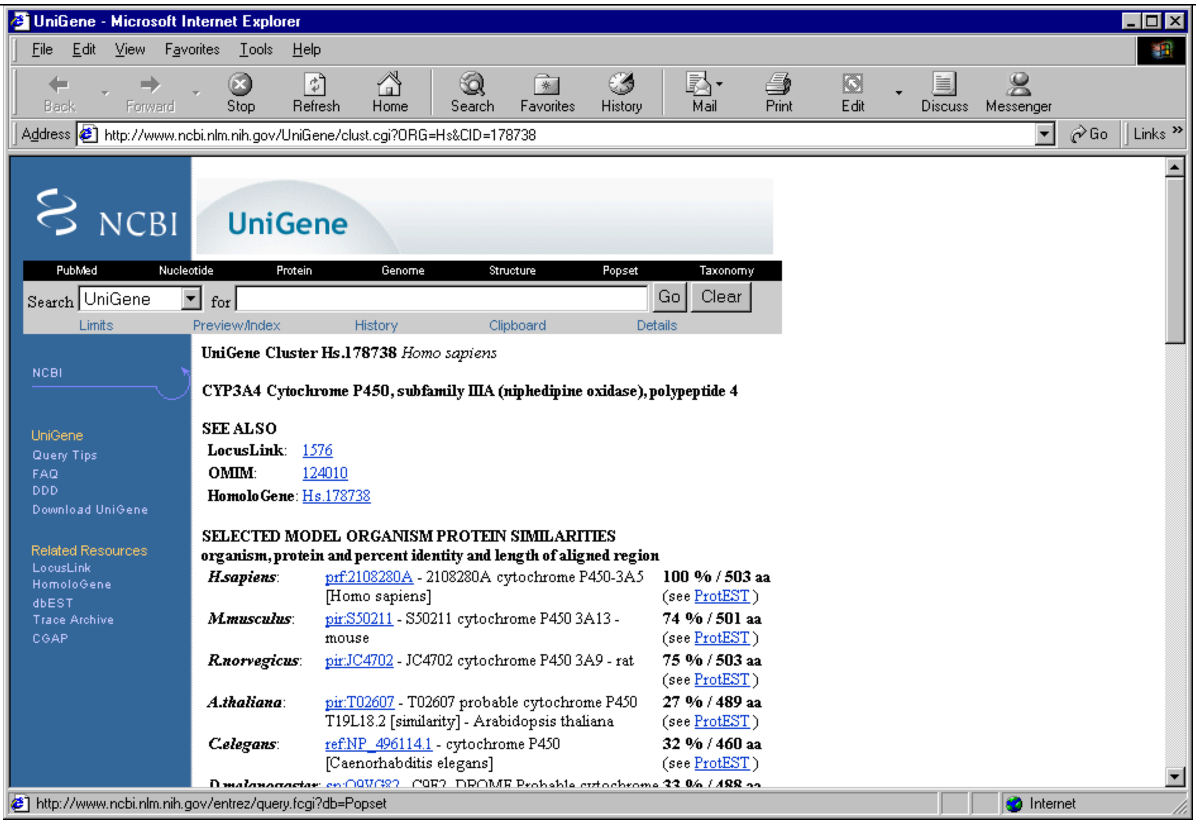
Figure 2.7.

2.8. In contextual menu of the application main table (contextual menu can be called out by mouse right click) the “URLs>UniGene” command (fig. 2.8.1) will become active. This command opens you default web browser and, using the appropriate link, loads a gene card from UniGene database (fig. 2.8.2).



	1	2	3	4	5	6	7
	cession Numb	Description	Sample	Tumor_27	Tumor_29	Tumor_34	Tumor_
1	D00003	Human liver cytochrome P-450 mRNA, complete c		-5	1	7	
2	D00003	complete c		1	-4	-3	
3	D00003	complete c		2	7	1	
4	D00015	A, human PrP 27-30 mR		26	13	13	
5	D00102	RNA, complete cds		5	2	0	
6	D00137	Human class I alcohol dehydrogenase beta-1 sub		18	13	11	
7	D00173	Human liver cytochrome P-450 mRNA		1	5	10	
8	D00265	Human cytochrome c mRNA, carboxyl-terminal reg		153	145	132	
9	D00306	Human pancreatic protease E mRNA, complete cd		-7	-7	-6	
10	D00408	Human fetal liver cytochrome P-450 (P-450 HFLa)		1	1	0	
11	D00596	Human thymidylate syntase (EC 2.1.1.45) gene, cc		70	138	73	
12	D00726	Human mRNA for ferrochelatase (EC 4.99.1.1)		20	17	22	
13	D00749	Human T cell surface antigen CD7 gene, exon 4		65	64	79	
14	D00760	Human mRNA for proteasome subunit HC3		55	63	72	
15	D00761	Human mRNA for proteasome subunit HC5		203	260	305	
16	D00762	Human mRNA for proteasome subunit HC8		254	327	355	
17	D00762	Human mRNA for proteasome subunit HC8		258	318	279	
18	D00763	Human mRNA for proteasome subunit HC9		68	73	79	
19	D00860	Human mRNA for phosphoribosyl pyrophosphate		30	31	24	
20	D10040	Human mRNA for long-chain acyl-CoA synthetase		12	65	28	
21	D10202	Human mRNA for platelet-activating factor recepto		10	3	0	
22	D10216	Human mRNA for Pit-1/GHF-1, complete cds		-5	-3	-1	
23	D10511	Human gene for mitochondrial acetoacetyl-CoA th		28	81	60	
24	D10522	Human mRNA for 80K-L protein, complete cds		63	74	70	

Figure 2.8.1.


UniGene - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Messenger
Address http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=178738
Go Links
NCBI UniGene
PubMed Nucleotide Protein Genome Structure Popset Taxonomy
Search UniGene for Go Clear
Limits Preview/Index History Clipboard Details
NCBI
UniGene
Query Tips
FAQ
DDD
Download UniGene
Related Resources
LocusLink
HomoloGene
dbEST
Trace Archive
CGAP
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Popset
Internet

UniGene Cluster Hs.178738 *Homo sapiens*

CYP3A4 Cytochrome P450, subfamily IIIA (niphedipine oxidase), polypeptide 4

SEE ALSO

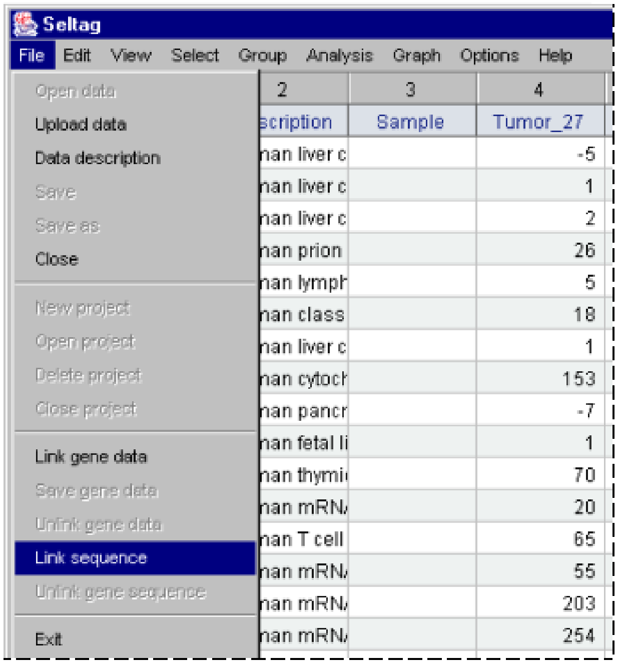
LocusLink:	1576
OMIM:	124010
HomoloGene:	Hs.178738

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H.sapiens:	prf:2108280A - 2108280A cytochrome P450-3A5 [Homo sapiens]	**100 % / 503 aa** (see ProtEST)
M.musculus:	pir:S50211 - S50211 cytochrome P450 3A13 - mouse	**74 % / 501 aa** (see ProtEST)
R.norvegicus:	pir:JC4702 - JC4702 cytochrome P450 3A9 - rat	**75 % / 503 aa** (see ProtEST)
A.thaliana:	pir:T02607 - T02607 probable cytochrome P450 T19L18.2 [similarity] - Arabidopsis thaliana	**27 % / 489 aa** (see ProtEST)
C.elegans:	ref:NP_496114.1 - cytochrome P450 [Caenorhabditis elegans]	**32 % / 460 aa** (see ProtEST)
D.melanogaster:	sp:Q9VG82 - C9F2_DROME Probable cytochrome	**33 % / 488 aa**

Figure 2.8.2.

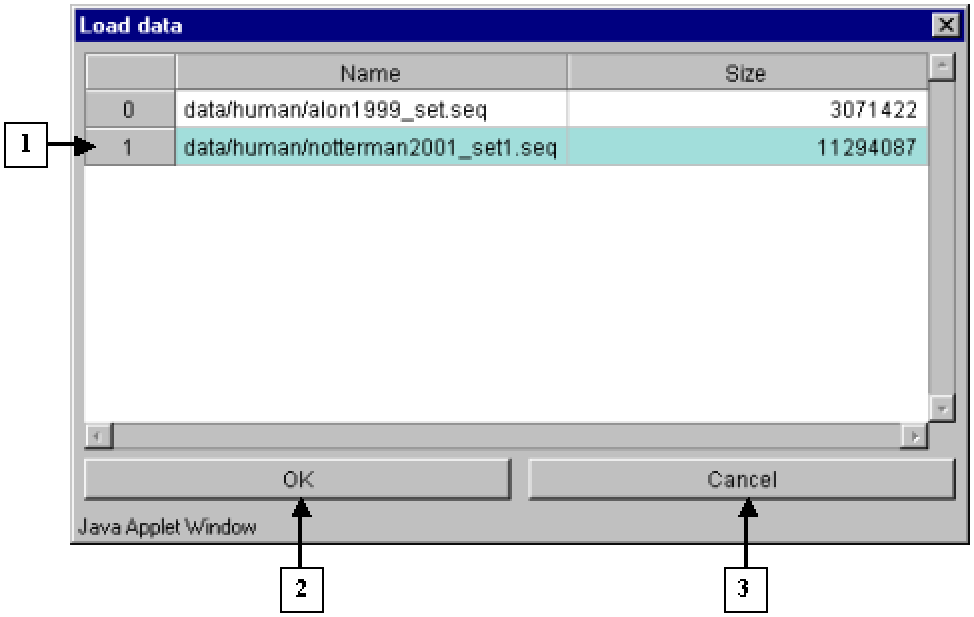
2.9. To load a file with genes nucleotide sequences, select the “File>Link sequence” command from the main menu (fig. 2.9).



2	3	4
scription	Sample	Tumor_27
nan liver c		-5
nan liver c		1
nan liver c		2
nan prion		26
nan lymph		5
nan class		18
nan liver c		1
nan cytoch		153
nan pancr		-7
nan fetal li		1
nan thymi		70
nan mRN/		20
nan T cell		65
nan mRN/		55
nan mRN/		203
nan mRN/		254

Figure 2.9.

2.10. The “Load data” dialog with files and information on their size (fig. 2.10) will appear. Select the file of interest and press “OK”.



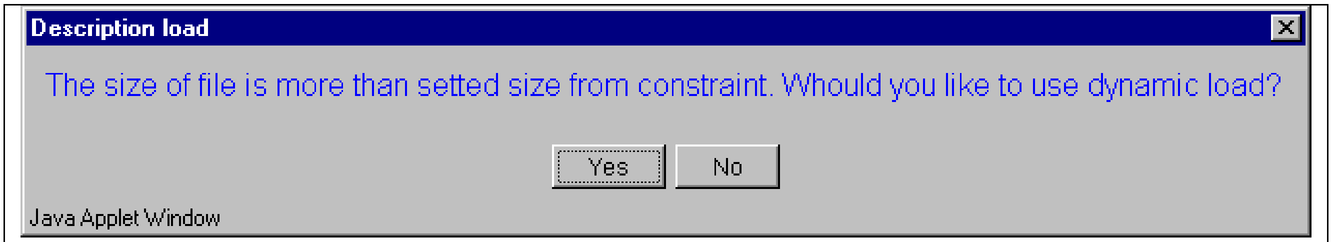
Load data

	Name	Size
0	data/human/alon1999_set.seq	3071422
1	data/human/notterman2001_set1.seq	11294087

Figure 2.10.

1. Selected file. **2.** Confirmation button. **3.** Cancel button.

2.11. The “Description load” dialog with invitation to use dynamic file loading mode (fig. 2.11) will appear. Press the “Yes” button

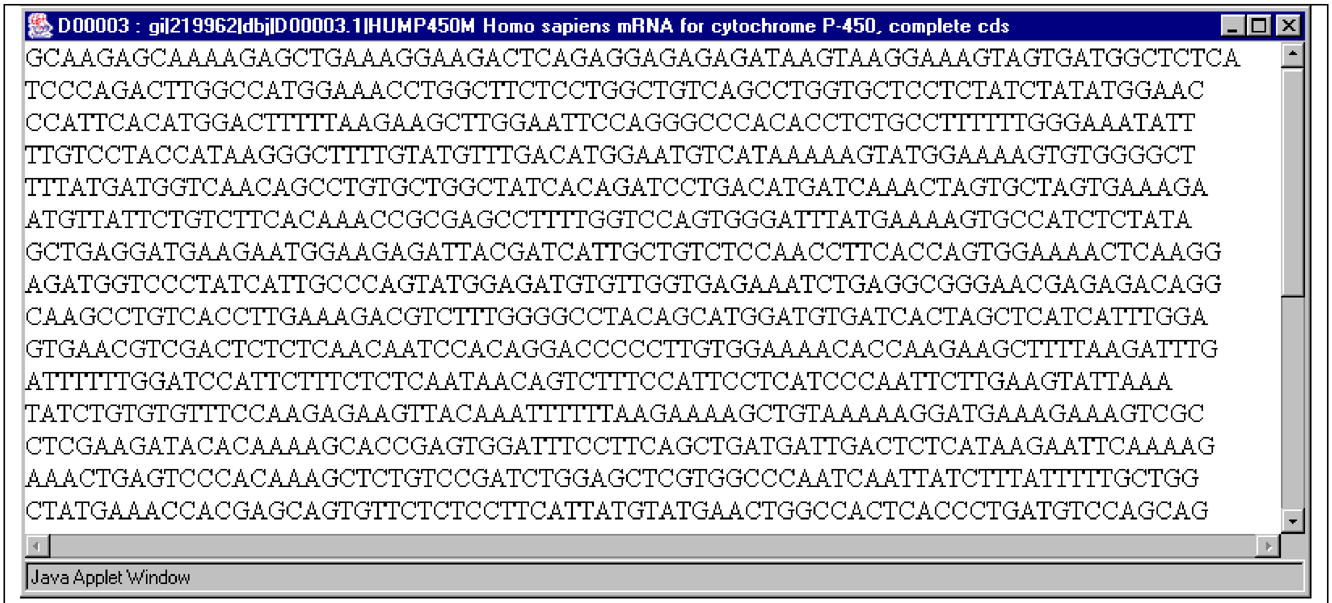


Description load

The size of file is more than setted size from constraint. Whould you like to use dynamic load?

Figure 2.11.

2.12. In contextual menu of the application main table (contextual menu can be called out by mouse right click) the “Show sequence” command (fig. 2.8.1) will become active. This command opens a window with gene’s nucleotide sequence (fig. 2.12).



D00003 : gi|219962|dbj|D00003.1|HUMP450M Homo sapiens mRNA for cytochrome P-450, complete cds

```
GCAAGAGCAAAAGAGCTGAAAGGAAGACTCAGAGGAGAGAGATAAGTAAGGAAAGTAGTGATGGCTCTCA
TCCCAGACTTGGCCATGGAAACCTGGCTTCTCCTGGCTGTCAGCCTGGTGCTCCTCTATCTATATGGAAC
CCATTCACATGGACTTTTTAAGAAGCTTGGAATTCCAGGGCCCACACCTCTGCCTTTTTGGGAAATATT
TTGTCCTACCATAAGGGCTTTTGTATGTTTGACATGGAATGTCATAAAAAGTATGGAAAAGTGTGGGGCT
TTTATGATGGTCAACAGCCTGTGCTGGCTATCACAGATCCTGACATGATCAAACTAGTGCTAGTGAAAGA
ATGTTATTCTGTCTTCACAAACCGCGAGCCTTTTGGTCCAGTGGGATTTATGAAAAGTGCCATCTCTATA
GCTGAGGATGAAGAATGGAAGAGATTACGATCATTGCTGTCTCCAACCTTCACCAGTGGAAAACTCAAGG
AGATGGTCCCTATCATTGCCCAGTATGGAGATGTGTTGGTGAGAAATCTGAGGCGGGAACGAGAGACAGG
CAAGCCTGTCACCTTGAAAGACGTCTTTGGGGCCTACAGCATGGATGTGATCACTAGCTCATCATTTGGA
GTGAACGTCGACTCTCTCAACAATCCACAGGACCCCCTTGTGGAAAACACCAAGAAGCTTTTAAGATTTG
ATTTTTTGGATCCATTCTTTCTCTCAATAACAGTCTTTCCATTCCTCATCCCAATTCTTGAAGTATTAAA
TATCTGTGTGTTTCCAAGAGAAGTTACAAATTTTTTAAGAAAAGCTGTAAAAAGGATGAAAGAAAGTCGC
CTCGAAGATACACAAAAGCACCGAGTGGATTTCCTTCAGCTGATGATTGACTCTCATAAGAATTCAAAAG
AAACTGAGTCCCACAAAGCTCTGTCCGATCTGGAGCTCGTGGCCCAATCAATTATCTTTATTTTTGCTGG
CTATGAAACCACGAGCAGTGTTCTCTCCTTCATTATGTATGAACTGGCCACTCACCCTGATGTCCAGCAG
```

Java Applet Window

Figure 2.12.

2.13. To get a description of loaded data, select the “File>Data description” command from the main menu (fig. 2.13)

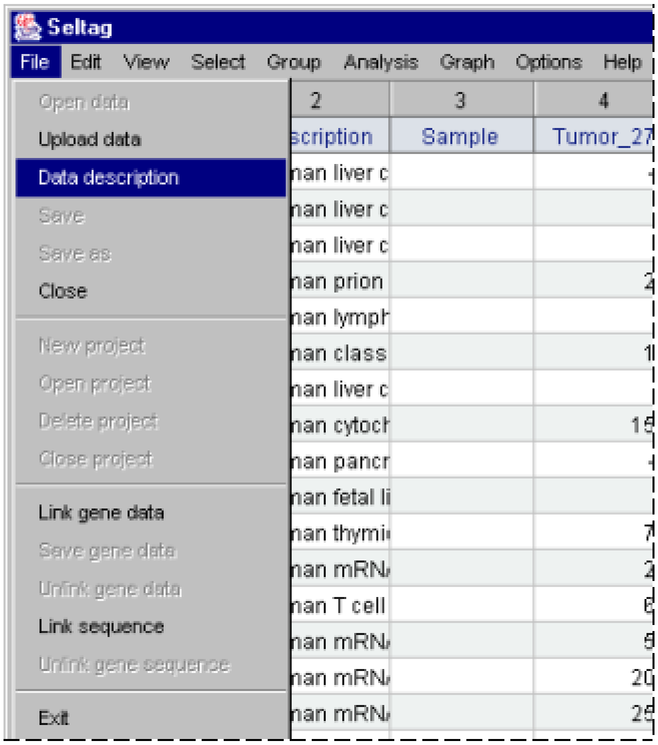
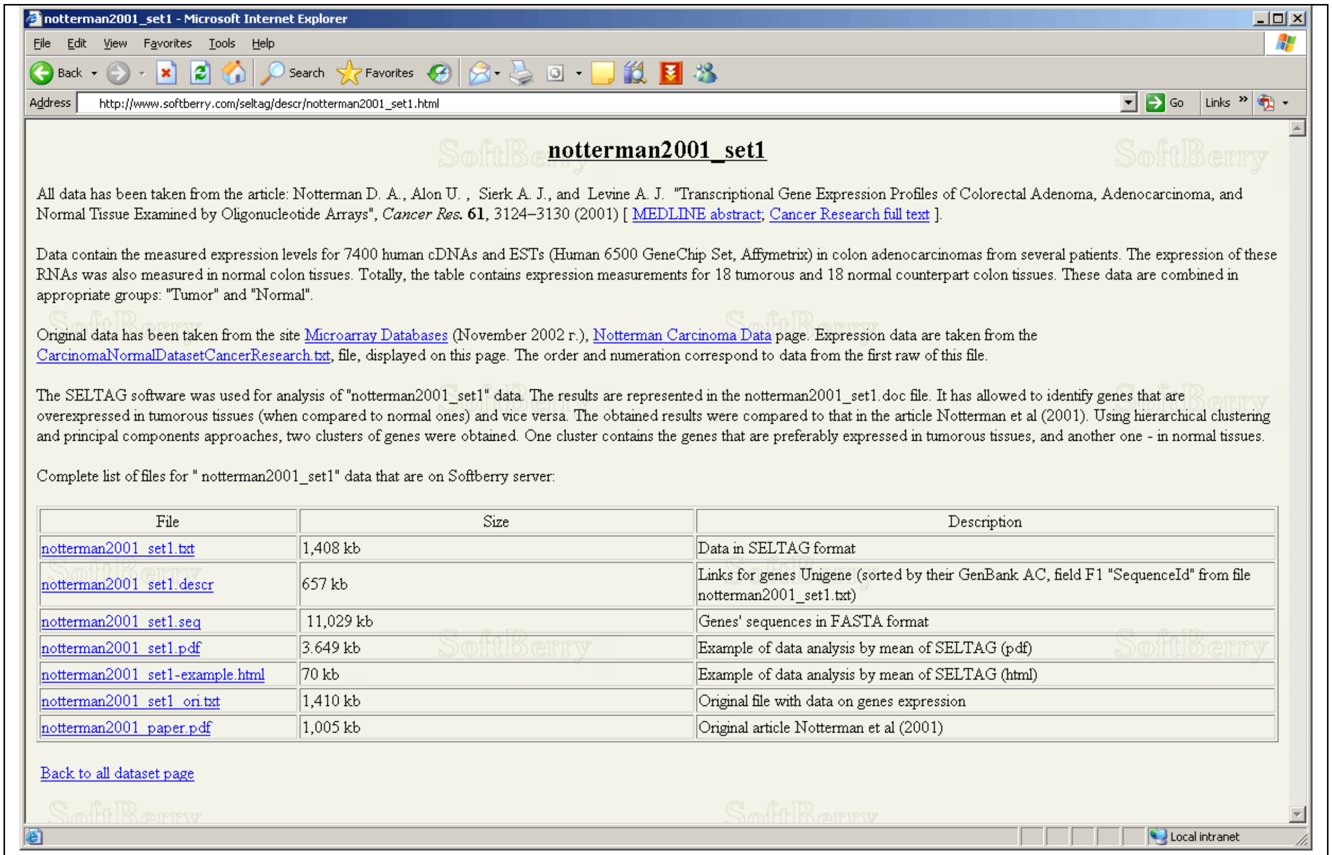


Figure 2.13.

2.14. It will open the document with description and list of files for data "notterman2001_set1", located on the Softberry server (fig. 2.14).


notterman2001_set1 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.softberry.com/seltag/descr/notterman2001_set1.html
Go Links
Local intranet

notterman2001_set1

All data has been taken from the article: Notterman D. A., Alon U. , Sierk A. J., and Levine A. J. "Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays", *Cancer Res.* **61**, 3124–3130 (2001) [MEDLINE abstract; Cancer Research full text].

Data contain the measured expression levels for 7400 human cDNAs and ESTs (Human 6500 GeneChip Set, Affymetrix) in colon adenocarcinomas from several patients. The expression of these RNAs was also measured in normal colon tissues. Totally, the table contains expression measurements for 18 tumorous and 18 normal counterpart colon tissues. These data are combined in appropriate groups: "Tumor" and "Normal".

Original data has been taken from the site Microarray Databases (November 2002 г.), Notterman Carcinoma Data page. Expression data are taken from the CarcinomaNormalDatasetCancerResearch.txt, file, displayed on this page. The order and numeration correspond to data from the first raw of this file.

The SELTAG software was used for analysis of "notterman2001_set1" data. The results are represented in the notterman2001_set1.doc file. It has allowed to identify genes that are overexpressed in tumorous tissues (when compared to normal ones) and vice versa. The obtained results were compared to that in the article Notterman et al (2001). Using hierarchical clustering and principal components approaches, two clusters of genes were obtained. One cluster contains the genes that are preferably expressed in tumorous tissues, and another one - in normal tissues.

Complete list of files for " notterman2001_set1" data that are on Softberry server:

File	Size	Description
notterman2001_set1.txt	1,408 kb	Data in SELTAG format
notterman2001_set1.descr	657 kb	Links for genes Unigene (sorted by their GenBank AC, field F1 "SequenceId" from file notterman2001_set1.txt)
notterman2001_set1.seq	11,029 kb	Genes' sequences in FASTA format
notterman2001_set1.pdf	3.649 kb	Example of data analysis by mean of SELTAG (pdf)
notterman2001_set1-example.html	70 kb	Example of data analysis by mean of SELTAG (html)
notterman2001_set1_ori.txt	1,410 kb	Original file with data on genes expression
notterman2001_paper.pdf	1,005 kb	Original article Notterman et al (2001)

Back to all dataset page

Figure 2.14.

3. Selecting genes with high expression level in tumor tissues.

In this chapter there is an example of how to select genes, which are expressed above average (more than 50) in, at least, 80% of tumorous tissues and, at the same time, below average (less than 50) in, at least, 80% of normal ones. The previously described dividing of experiments into two groups (in accordance with tissue types) – tumorous (G1) and normal (G2) – in the notterman2001_set1.txt file makes the solving of this task much easier.

To perform this task, the following steps are required:

3.1. Select the "Select>Select genes by query…" command from the main menu (fig. 3.1).

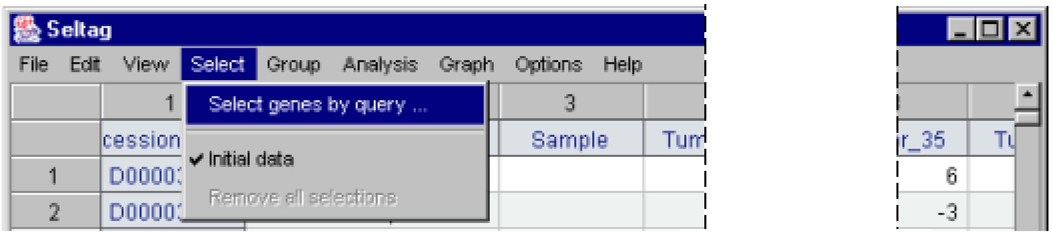
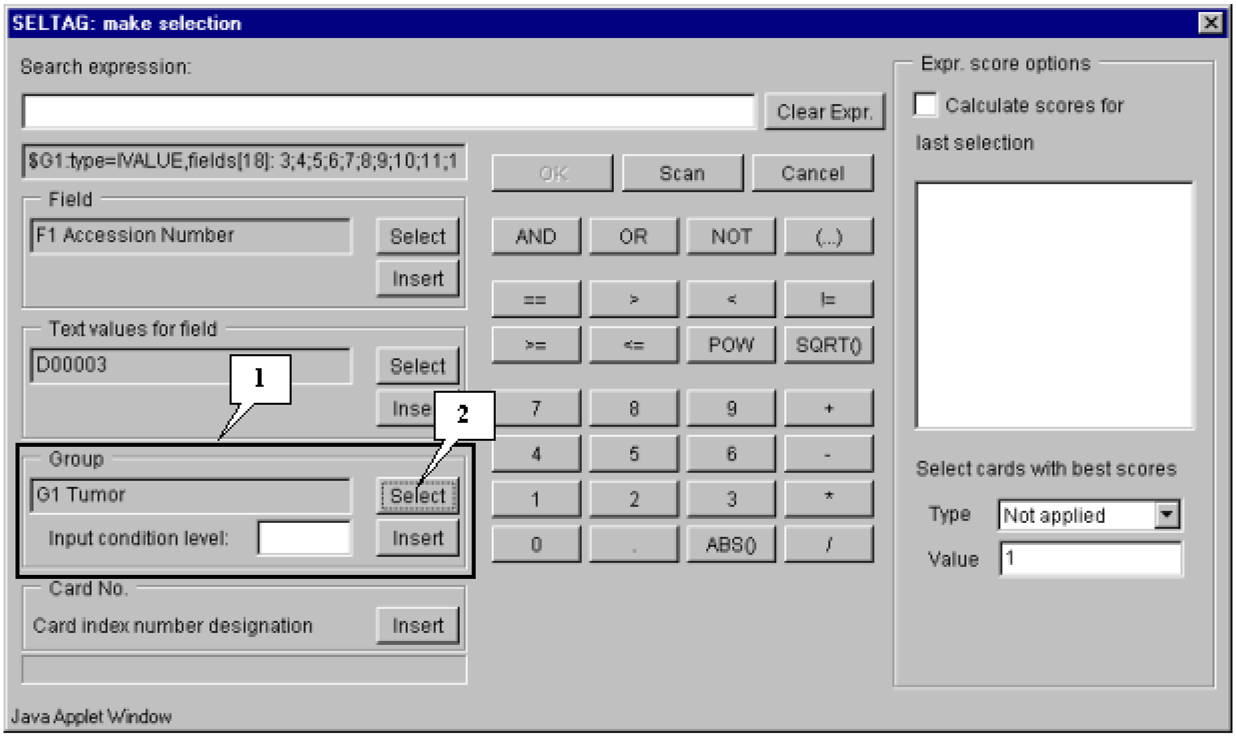


Figure 3.1.

3.2. The "Make selection" dialog will appear (fig. 3.2). For the first, select a target group to satisfy selection criteria. In the "Group" section press the "Select" button (fig. 3.2).



$G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;1

Field: F1 Accession Number

Text values for field: D00003

Group: G1 Tumor

Input condition level:

☐ Calculate scores for last selection

Type: Not applied

Value: 1

Figure 3.2.

1. The "Group" section. **2.** The "Select" button calls out the "Calculator options" dialog.

3.3. The "Calculator options" dialog with the complete list of groups in the table (fig. 3.3) will appear. In the list select a target group, for which the first part of constrains will be set. In our case, this is group «G1 Tumor». After selection press the "OK" button.

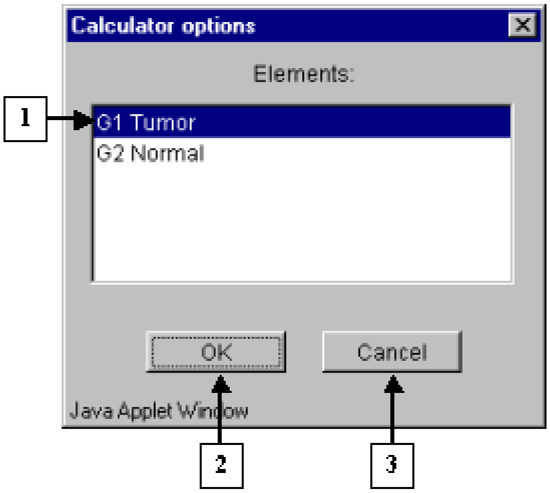


Figure 3.3.

1. Selected group. **2.** Confirmation button. **3.** Cancel button.

3.4. In the "Make selection" dialog the following changes will occur (fig. 3.4):

- In the text area, the number and the name of selected group will appear: G1 Tumor
- In the status bar, the information on this group will be shown: $G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;

3.5. In the field of additional constraints "Input condition level:" specify the threshold as a percentage share in a group:

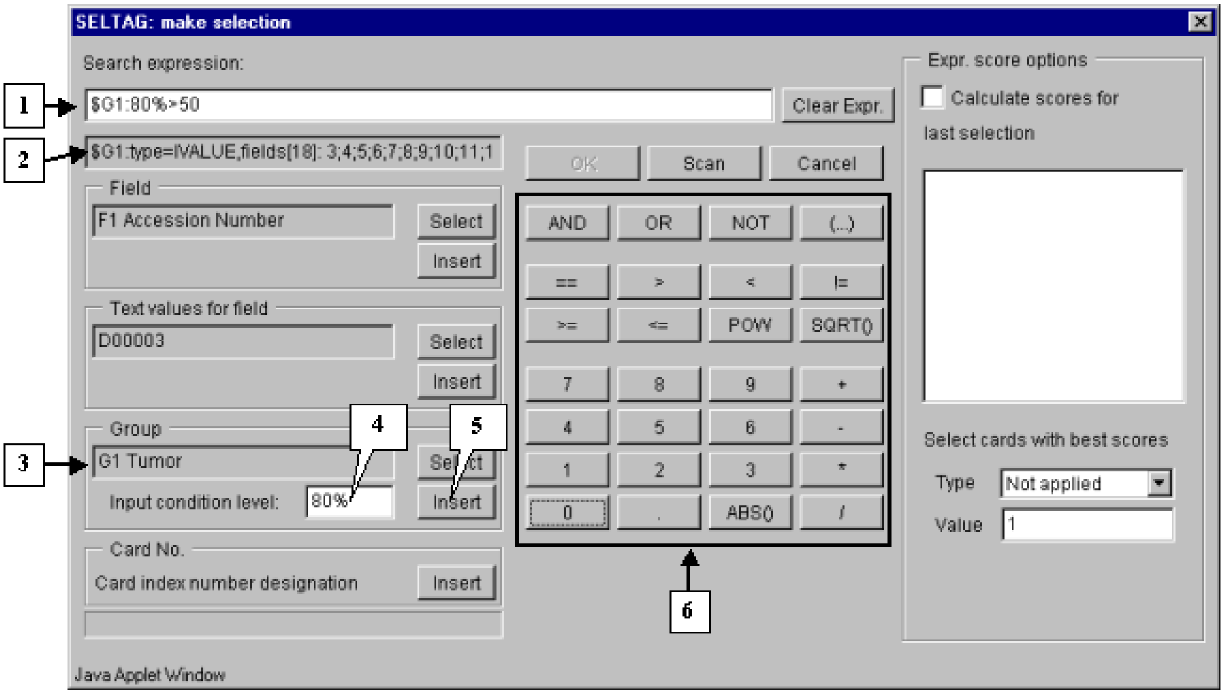
80%

3.6. The selected group ID and selection criteria should be inserted into expression line. To do this, press the "Insert" button. In the expression line the following will appear (fig. 3.4):

$G1:80%

3.7. Using buttons of query entering specify the first part of condition in the expression line (fig. 3.4):

$G1:80%>50



SELTAG: make selection

Search expression: $G1:80%>50

$G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;1

Field: F1 Accession Number

Text values for field: D00003

Group: G1 Tumor

Input condition level: 80%

Card No.: Card index number designation

Expr. score options

□ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

Figure 3.4.

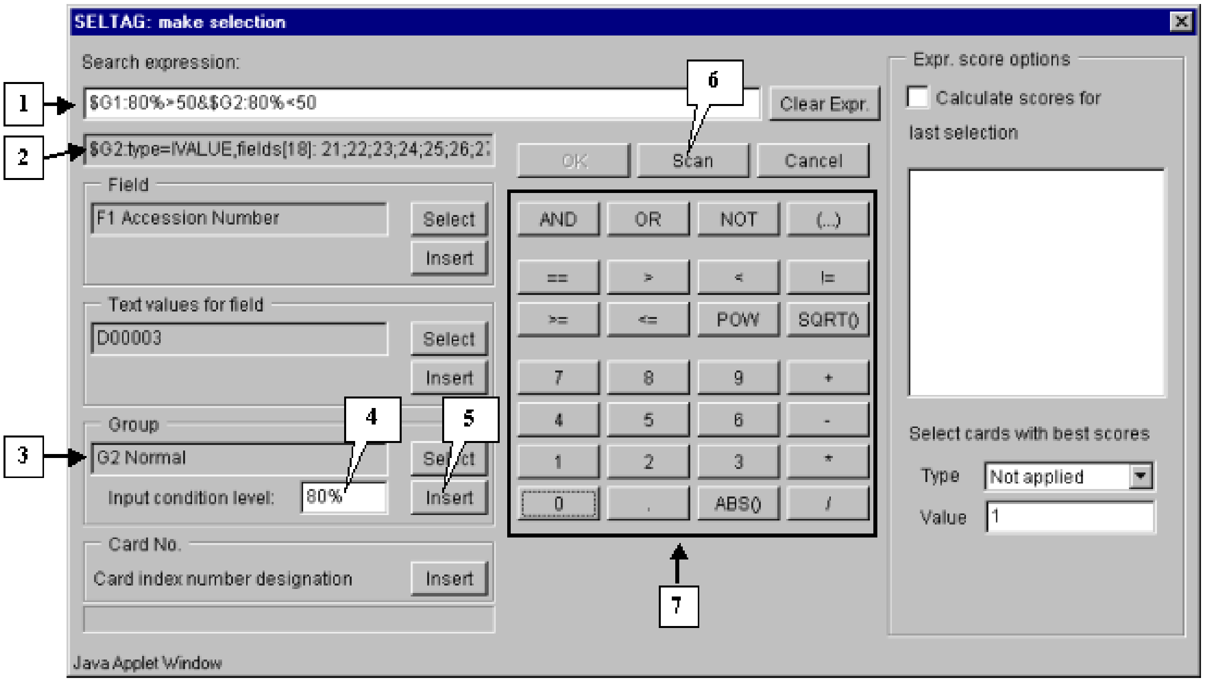
1. Expression line with selection criteria. **2.** Status bar with information on selected field. **3.** Number and ID of selected group. **4.** Field for additional constraints. **5.** The "Insert" button for insertion of the group ID with additional constraint into expression line. **6.** Query entering buttons.

3.8. Add the "&" symbol into expression line by pressing the "And" button in the query entering section.

3.9. Finally, specify the second part of selection criteria. It can be done by the same way, as described in 3.2-3.7, with the only exception: select the group G2 Normal and add a constraint for it:

$G2:80%<50

The final expression is shown on figure 3.5.



SELTAG: make selection

Search expression: $G1:80%>50&$G2:80%<50

$G2:type=IVALUE,fields[18]: 21;22;23;24;25;26;2

Field: F1 Accession Number

Text values for field: D00003

Group: G2 Normal

Input condition level: 80%

Card No.: Card index number designation

Expr. score options

□ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

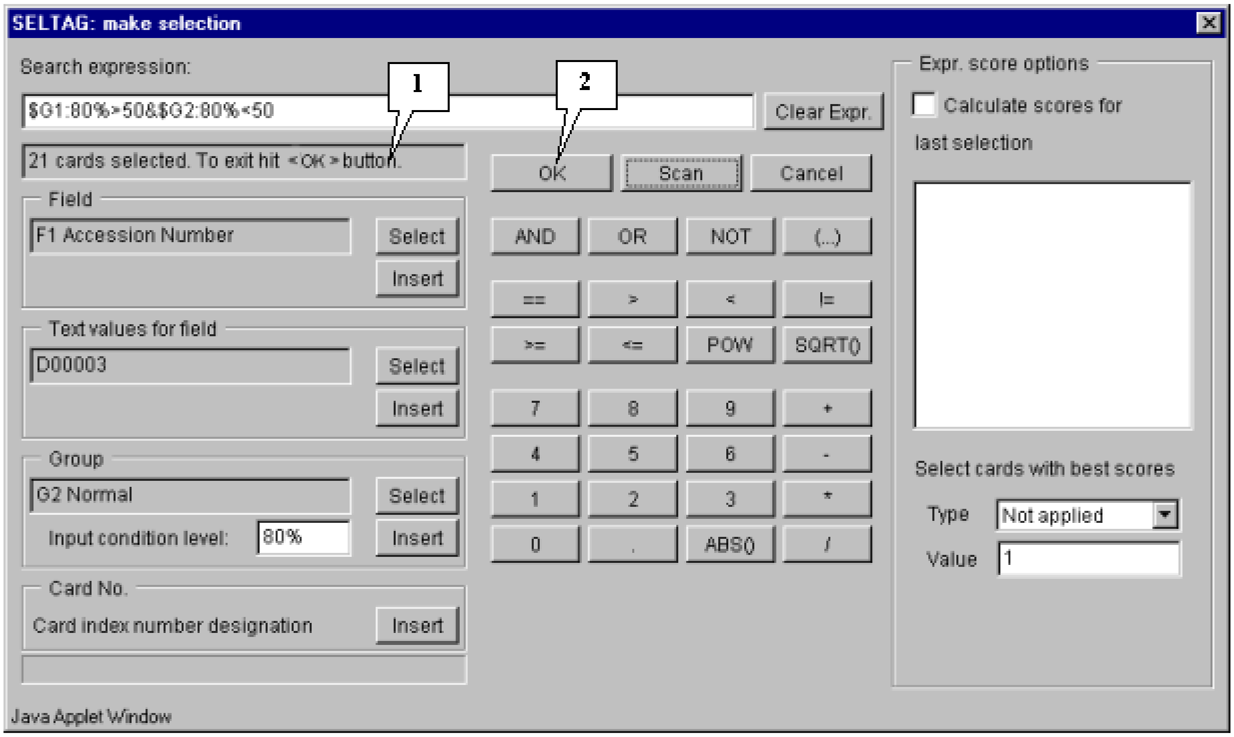
Java Applet Window

Figure 3.5.

1. Expression line with selection criteria. **2.** Status bar with information on selected field. **3.** Number and ID of selected group. **4.** Field for additional constraints. **5.** The “Insert” button for insertion of the group ID with additional constraint into expression line. **6.** Button for search start. **7.** Query entering buttons.

3.10. To start the search process press the “Scan” button (fig. 3.5). As a result, all genes, overexpressed in, at least, 80% of tumorous tissues, and downexpressed in 80% of normal ones, will be found and selected.

3.11. Once the selection is finished, information on the number of found genes will be represented in the status bar (fig. 3.6), and the “OK” button will become active.



SELTAG: make selection

Search expression: $G1:80%>50&$G2:80%<50

21 cards selected. To exit hit <OK> button.

Field: F1 Accession Number

Text values for field: D00003

Group: G2 Normal

Input condition level: 80%

Card No.: Card index number designation

Expr. score options: □ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

Figure 3.6.

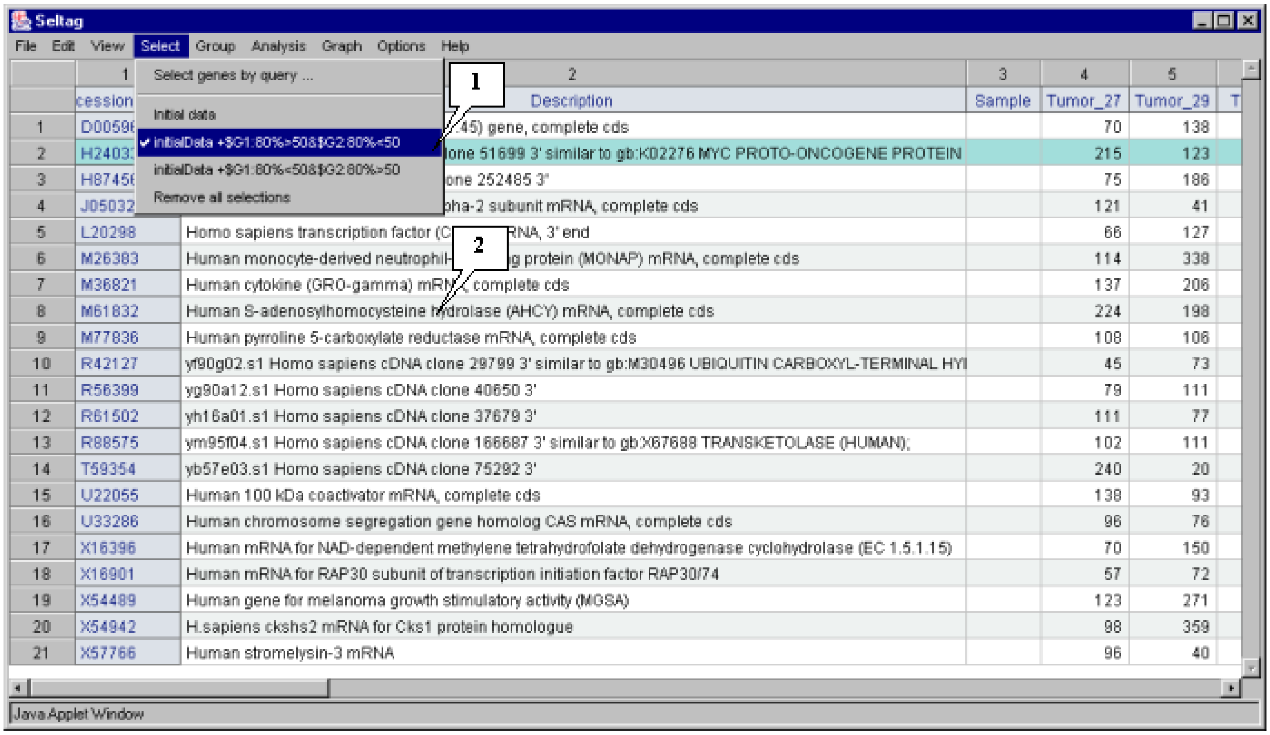
1. Information on the number of found genes. **2.** Button for accepting the search results.

3.9. Press the “OK” button.

3.10. In the application main window the table with selected genes (fig. 3.7) will be represented. In the «Select» section of the main window menu an additional item with the name corresponding to selected set of genes will appear. During the project run, the obtained sets of genes can be saved and remained available by simple switching between them. To remove the list of tables use the «Remove all selections» command.

Table obtained in this example includes 21 genes. Some of selected genes are also identified as transcripts that overexpressed in tumorous tissues (when compared to their normal counterparts) in the article by Notterman et al [1] (see table 1 in this article). These genes are:

X54489 Human gene for MGSA
U22055 Human 100 kDA coactivator mRNA, complete cds
M61832 Human S-adenosylhomocysteine hydrolase (AHCY) mRNA, complete cds
M36821 Human cytokine (GRO-g) mRNA, complete cds
U33286 Human chromosome segregation gene homolog CAS mRNA, complete cds
X54942 H. sapiens CKSHS2 mRNA for CKS1 protein homologue


Seltag
File Edit View Select Group Analysis Graph Options Help
Select genes by query ...
Initial data
✓ initialData +$G1:80%>50&$G2:80%<50
initialData +$G1:80%<50&$G2:80%>50
Remove all selections
1
2
Java Applet Window

	1	2	3	4	5	
	cession	Description	Sample	Tumor_27	Tumor_29	T
1	D0059	:45) gene, complete cds		70	138	
2	H2403	lone 51699 3' similar to gb:K02276 MYC PROTO-ONCOGENE PROTEIN		215	123	
3	H8745	lone 252485 3'		75	186	
4	J05032	pha-2 subunit mRNA, complete cds		121	41	
5	L20298	Homo sapiens transcription factor (C RNA, 3' end		66	127	
6	M26383	Human monocyte-derived neutrophil- g protein (MONAP) mRNA, complete cds		114	338	
7	M36821	Human cytokine (GRO-gamma) mRN complete cds		137	206	
8	M61832	Human S-adenosylhomocysteine hydrolase (AHCY) mRNA, complete cds		224	198	
9	M77836	Human pyrroline 5-carboxylate reductase mRNA, complete cds		108	106	
10	R42127	yf90g02.s1 Homo sapiens cDNA clone 29799 3' similar to gb:M30496 UBIQUITIN CARBOXYL-TERMINAL HYI		45	73	
11	R56399	yg90a12.s1 Homo sapiens cDNA clone 40650 3'		79	111	
12	R61502	yh16a01.s1 Homo sapiens cDNA clone 37679 3'		111	77	
13	R88575	ym95f04.s1 Homo sapiens cDNA clone 166687 3' similar to gb:X67688 TRANSKETOLASE (HUMAN);		102	111	
14	T59354	yb57e03.s1 Homo sapiens cDNA clone 75292 3'		240	20	
15	U22055	Human 100 kDa coactivator mRNA, complete cds		138	93	
16	U33286	Human chromosome segregation gene homolog CAS mRNA, complete cds		96	76	
17	X16396	Human mRNA for NAD-dependent methylene tetrahydrofolate dehydrogenase cyclohydrolase (EC 1.5.1.15)		70	150	
18	X16901	Human mRNA for RAP30 subunit of transcription initiation factor RAP30/74		57	72	
19	X54489	Human gene for melanoma growth stimulatory activity (MGSA)		123	271	
20	X54942	H.sapiens ckshs2 mRNA for Cks1 protein homologue		98	359	
21	X57766	Human stromelysin-3 mRNA		96	40	

Figure 3.7.

1. Additional menu item that corresponds to obtained genes set. **2.** Table with selected genes.

3.11. To visualize expression profiles for selected genes, use the "Graph>Specified with current selection" command from the main menu (fig. 3.8).

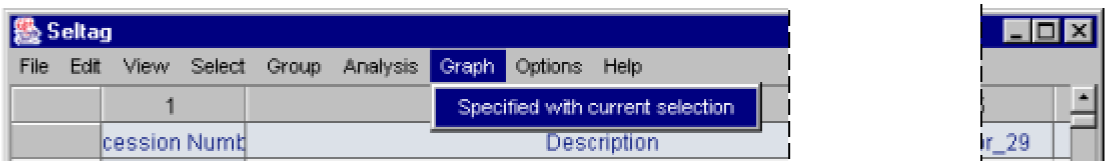
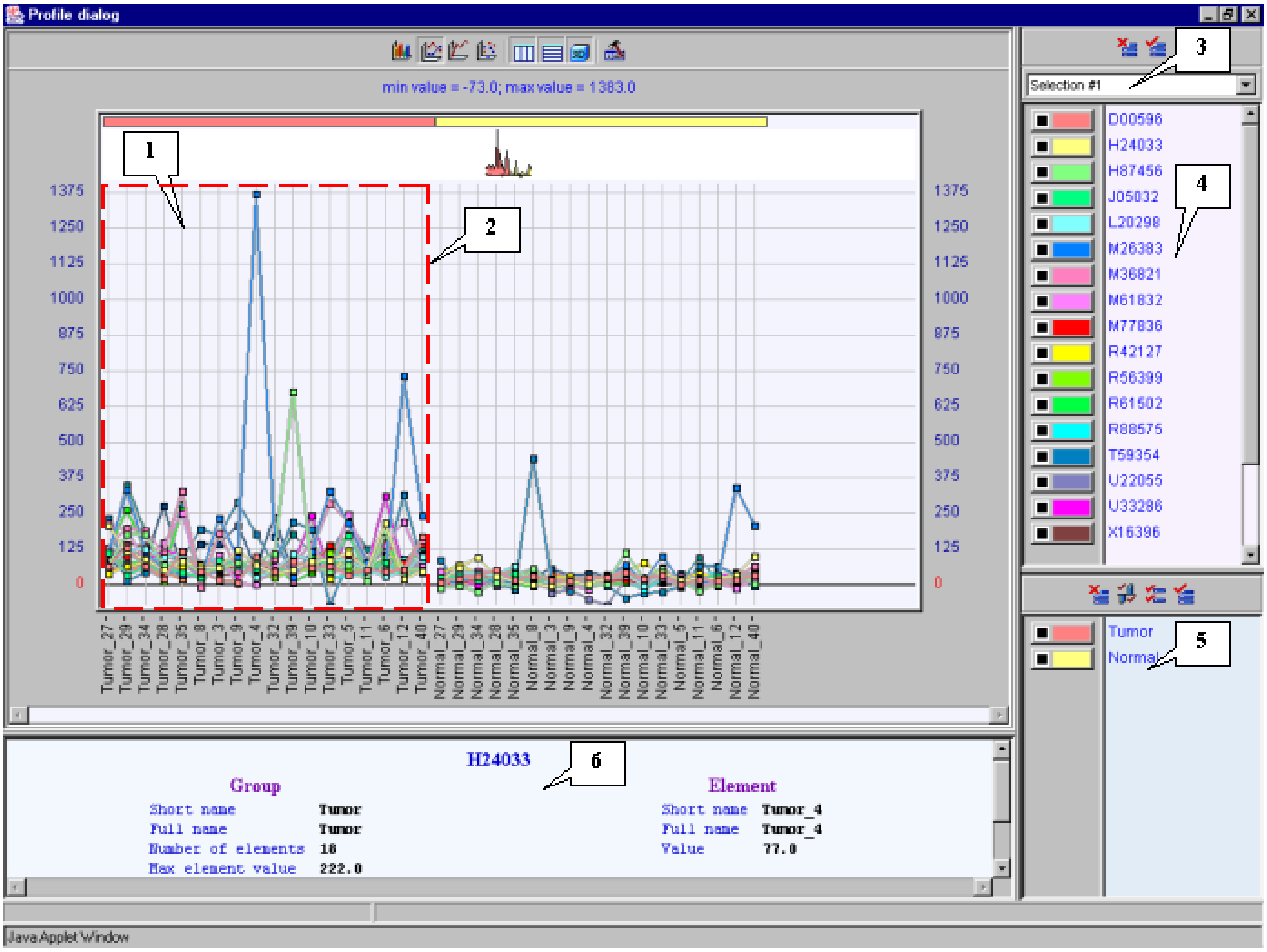

Seltag
File Edit View Select Group Analysis Graph Options Help
Specified with current selection
1
cession Numb
Description
r_29

Figure 3.8.

3.12. The “Profile dialog” window with expression profiles for selected genes (fig. 3.9) will appear. It is notable that expression profiles of all represented genes are higher for tumorous tissues (profiles inside the red rectangle) than for normal ones (profiles outside the red rectangle).



H24033

Group

Short name Tumor

Full name Tumor

Number of elements 18

Max element value 222.0

Element

Short name Tumor_4

Full name Tumor_4

Value 77.0

Figure 3.9.

1. Area of expression profiles visualization. **2.** Expression values for tumorous tissues. **3.** List of genes sets. **4.** List of genes. **5.** List of tissues groups. **6.** Information on gene, profile of which is pointed by mouse.

4. Selecting genes with high expression level in normal tissues.

In this chapter there is an example of how to select genes, which are expressed above average (more than 50) in, at least, 80% of normal tissues and, at the same time, below average (less than 50) in, at least, 80% of tumorous ones.

To perform this task, the following steps are required:

4.1. Select the “Select>Select genes by query…” command from the main menu (fig. 4.1).

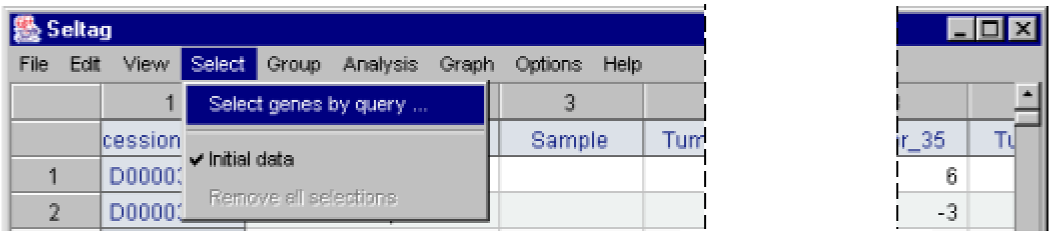
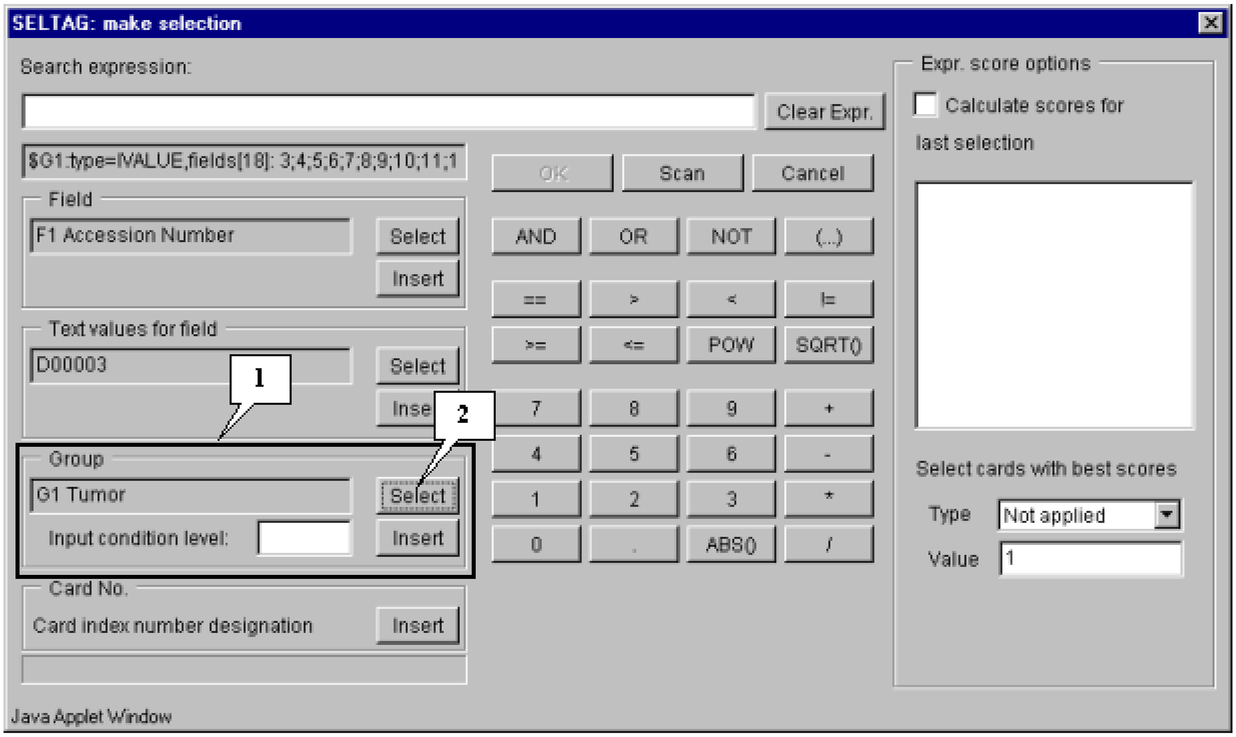


Figure 4.1.

4.2. The “Make selection” dialog will appear (fig. 4.2). For the first, select a target group to satisfy selection criteria. In the “Group” section press the “Select” button (fig. 4.2).



$G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;1

Field: F1 Accession Number

Text values for field: D00003

Group: G1 Tumor

□ Calculate scores for last selection

Type: Not applied

Value: 1

Figure 4.2.

1. The “Group” section. **2.** The “Select” button calls out the “Calculator options” dialog.

4.3. The "Calculator options" dialog with the complete list of groups in the table (fig. 4.3) will appear. In the list select a target group, for which the first part of constrains will be set. In our case, this is group «G1 Tumor». After selection press the "OK" button.

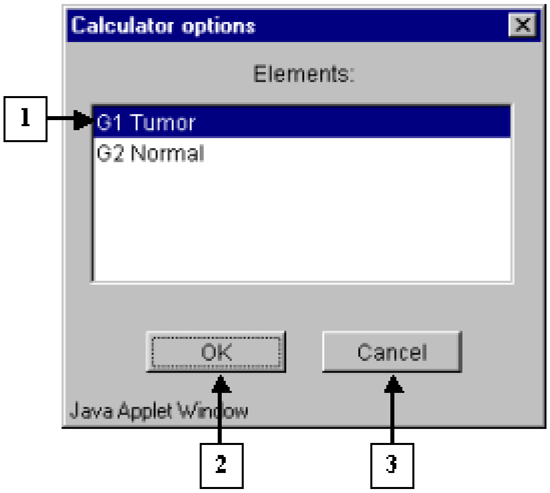


Figure 4.3.

1. Selected group. **2.** Confirmation button. **3.** Cancel button.

4.4. In the "Make selection" dialog the following changes will occur (fig. 4.4):

- In the text area, the number and the name of selected group will appear:
  G1 Tumor
- In the status bar, the information on this group will be shown:
  $G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;

4.5. In the field of additional constraints "Input condition level:" specify the threshold as a percentage share in a group:

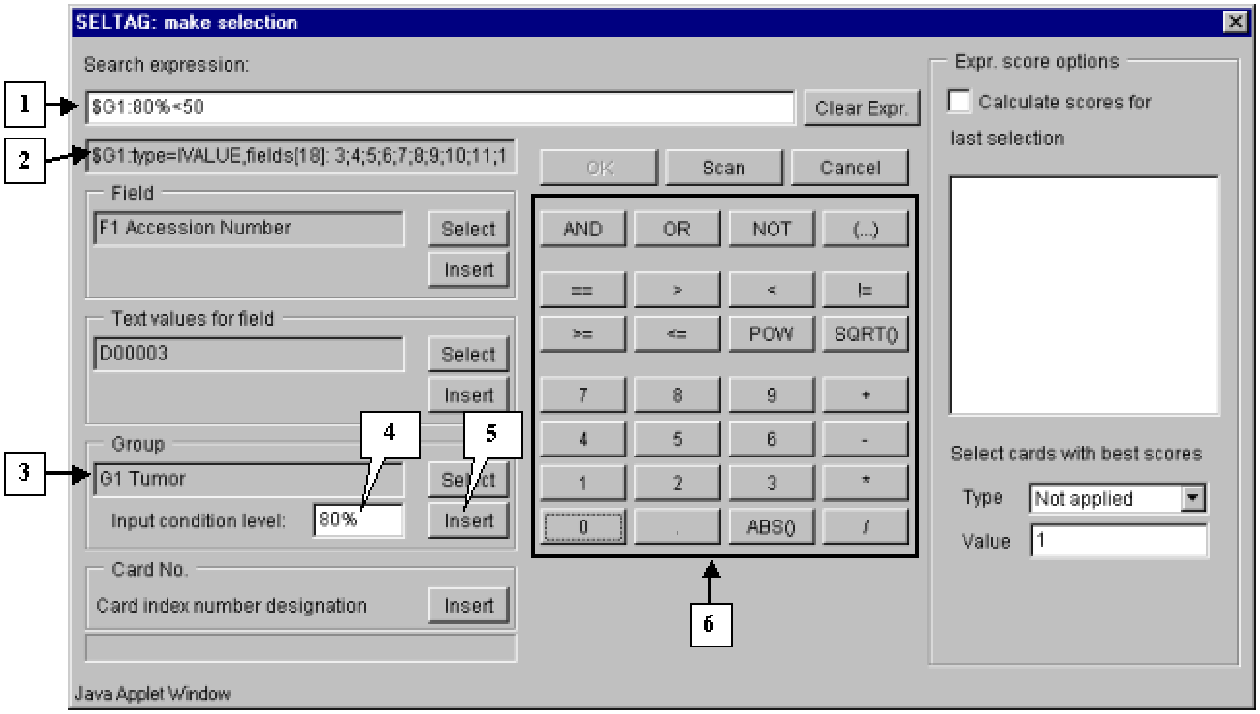
80%

4.6. The selected group ID and selection criteria should be inserted into expression line. To do this, press the "Insert" button. In the expression line the following will appear (fig. 4.4):

$G1:80%

4.7. Using buttons of query entering, specify the first part of condition in the expression line (fig. 4.4):

$G1:80%>50



SELTAG: make selection

Search expression: $G1:80%<50

$G1:type=IVALUE,fields[18]: 3;4;5;6;7;8;9;10;11;1

Field: F1 Accession Number

Text values for field: D00003

Group: G1 Tumor

Input condition level: 80%

Card No.: Card index number designation

Expr. score options

☐ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

Java Applet Window

Figure 4.4.

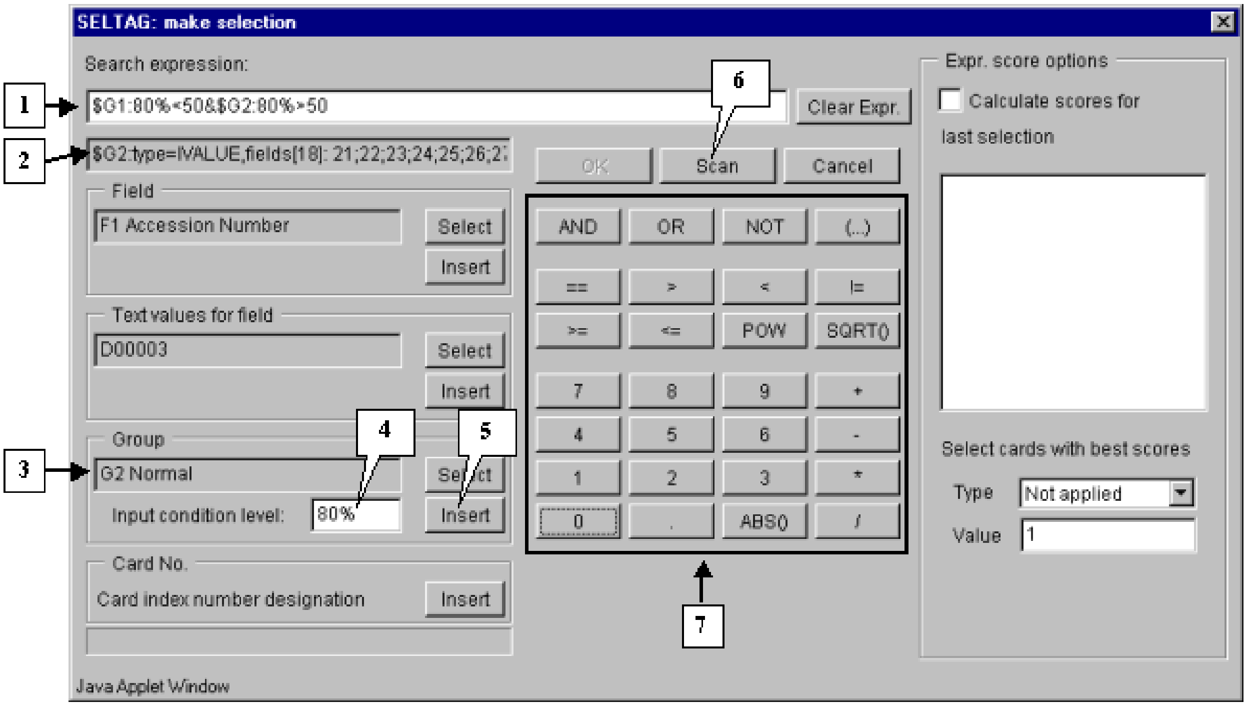
1. Expression line with selection criteria. **2.** Status bar with information on selected field. **3.** Number and ID of selected group. **4.** Field for additional constraints. **5.** The “Insert” button for insertion of the group ID with additional constraint into expression line. **6.** Query entering buttons.

4.8. Add the “&” symbol into expression line by pressing the “And” button in the query entering section.

4.9. Finally, specify the second part of selection criteria. It can be done by the same way, as described in 3.2-3.7, with the only exception: select the group G2 Normal and add a constraint for it:

$G2:80%>50

The final expression is shown on figure 4.5.



Search expression: $G1:80%<50&$G2:80%>50

$G2:type=IVALUE,fields[18]: 21;22;23;24;25;26;2

Field: F1 Accession Number

Text values for field: D00003

Group: G2 Normal

Input condition level: 80%

Card No.: Card index number designation

Expr. score options

☐ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

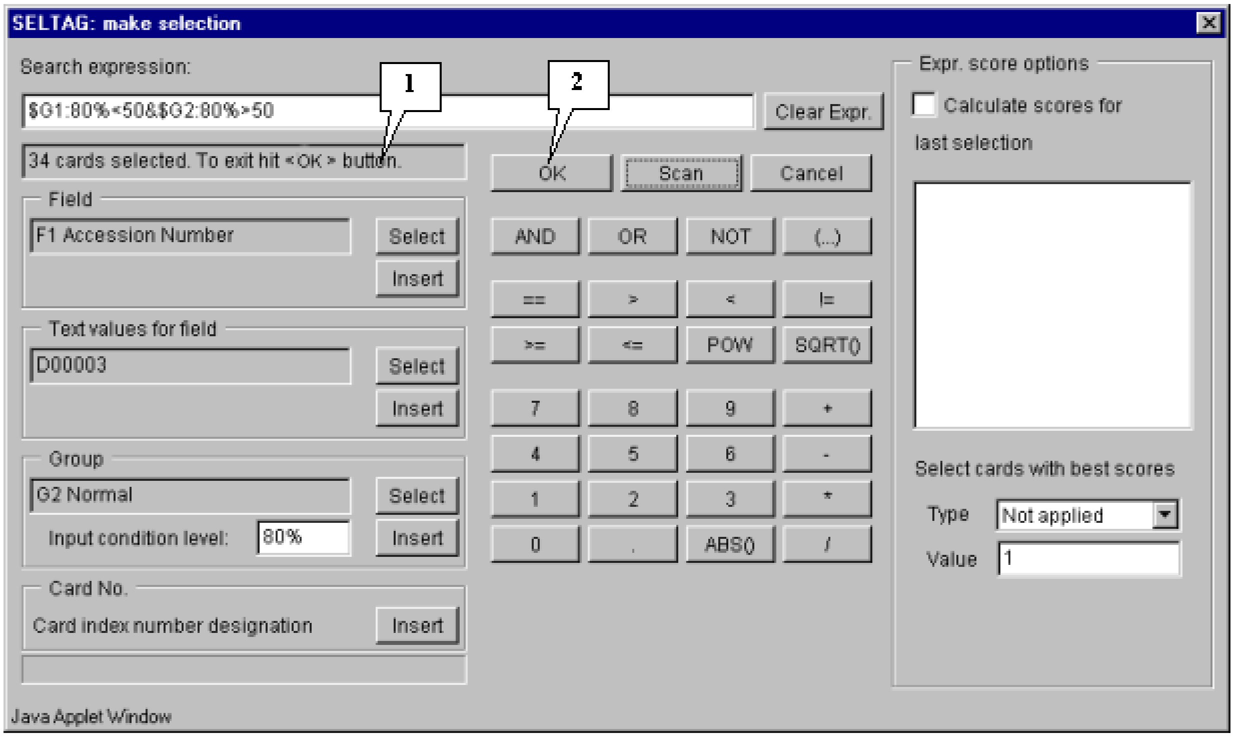
Figure 4.5.

1. Expression line with selection criteria. **2.** Status bar with information on selected field. **3.** Number and ID of selected group. **4.** Field for additional constraints. **5.** The “Insert” button for insertion of the group ID with additional constraint into expression line. **6.** Button for search start. **7.** Query entering buttons.

4.10. To start the search process press the “Scan” button (fig. 4.5).

4.11. Once the selection is finished, information on the number of found genes will be represented in the status bar (fig. 4.6), and the “OK” button will become active.

4.12. Press the “OK” button.


SELTAG: make selection

Search expression: $G1:80%<50&$G2:80%>50

34 cards selected. To exit hit <OK> button.

Field: F1 Accession Number

Text values for field: D00003

Group: G2 Normal

Input condition level: 80%

Card No.: Card index number designation

Expr. score options

☐ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

Java Applet Window

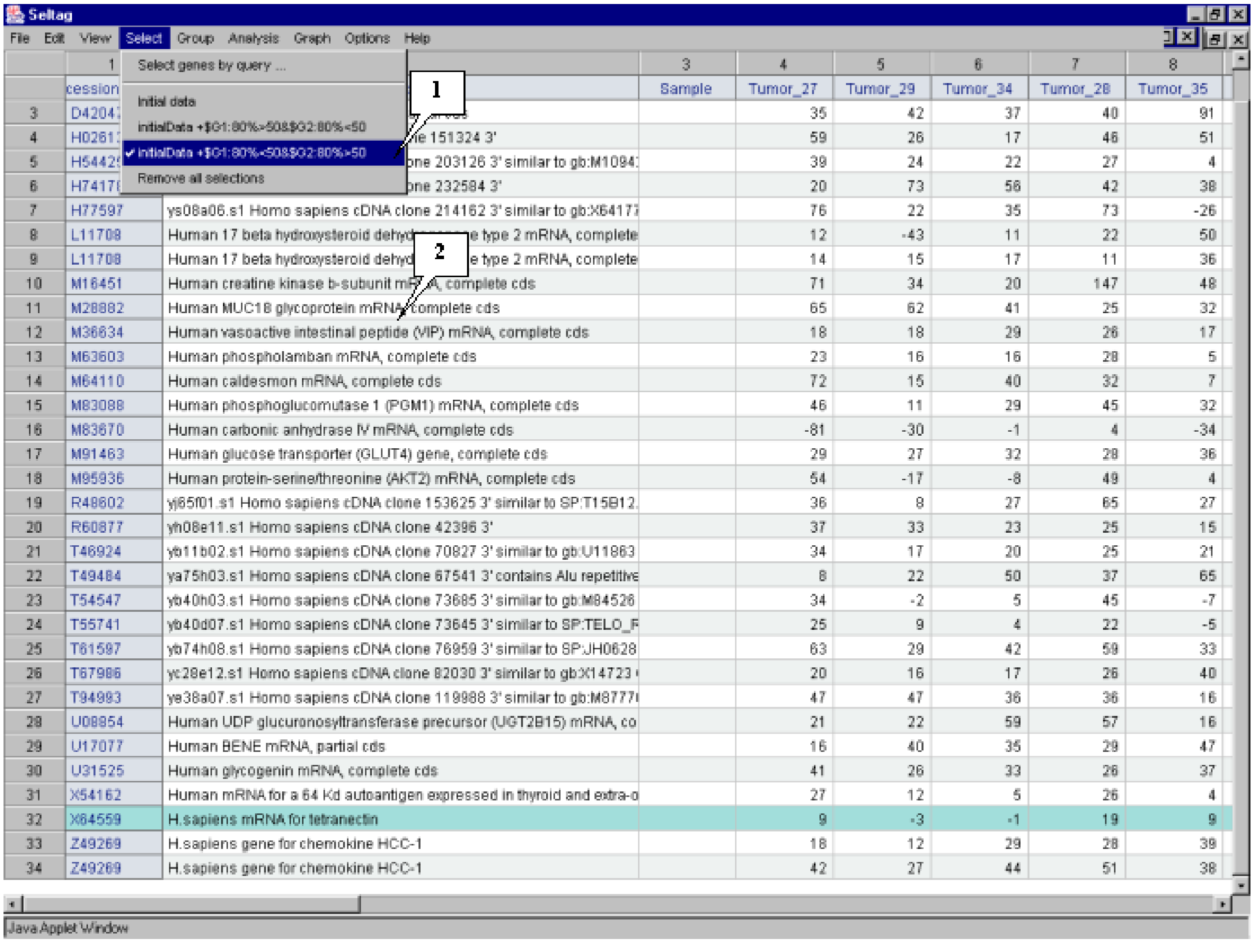
Figure 4.6.

1. Information on the number of found genes. **2.** Button for accepting the search results.

4.13. In the application main window the table with selected genes (fig. 4.7) will be represented. In the «Select» section of the main window menu an additional item with the name corresponding to selected set of genes will appear. During the project run, the obtained sets of genes can be saved and remained available by simple switching between them. To remove the list of tables use the «Remove all selections» command.

Table obtained in this example includes 34 genes. Some of selected genes are also identified as transcripts that overexpressed in normal tissues (when compared to their tumor counterparts) in the article by Notterman et al [1] (see table 2 in this article). These genes are:

M83670 Human carbonic anhydrase IV mRNA, complete cds
X64559 H. sapiens mRNA for tetranectin
T54547 H. sapiens cDNA similar to M84526 complement factor D precursor
M95936 Human protein-serine/threonine (AKT2) mRNA, complete cds
T46924 H. sapiens cDNA similar to gb:U11863 amiloride-sens amine oxidase
L11708 Human 17 b-hydroxysteroid dehydrogenase type 2 mRNA, complete cds
H54425 H. sapiens cDNA similar to gb:M10942_cds1 human metallothionein-le gene
H77597 H. sapiens cDNA similar to gb:X64177 H. sapiens mRNA for metallothionein
T67986 H. sapiens cDNA clone 82030 39 similar to gb:X14723 clusterin precursor
U17077 Human BENE mRNA, partial cds
U08854 Human UDP glucuronosyltransferase precursor (UGT2B15) mRNA, complete cds


Seltag
File Edit View Select Group Analysis Graph Options Help
Select genes by query ...
Initial data
initialData +$G1:80%>50&$G2:80%<50
✓ initialData +$G1:80%<50&$G2:80%>50
Remove all selections
1
2
Java Applet Window

	1		3	4	5	6	7	8
	cession		Sample	Tumor_27	Tumor_29	Tumor_34	Tumor_28	Tumor_35
3	D4204			35	42	37	40	91
4	H0261	ie 151324 3'		59	26	17	46	51
5	H5442	one 203126 3' similar to gb:M1094		39	24	22	27	4
6	H7417	one 232584 3'		20	73	56	42	38
7	H77597	ys08a06.s1 Homo sapiens cDNA clone 214162 3' similar to gb:X6417		76	22	35	73	-26
8	L11708	Human 17 beta hydroxysteroid dehyd e type 2 mRNA, complete		12	-43	11	22	50
9	L11708	Human 17 beta hydroxysteroid dehyd e type 2 mRNA, complete		14	15	17	11	36
10	M16451	Human creatine kinase b-subunit mR A, complete cds		71	34	20	147	48
11	M28882	Human MUC18 glycoprotein mRNA, complete cds		65	62	41	25	32
12	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds		18	18	29	26	17
13	M63603	Human phospholamban mRNA, complete cds		23	16	16	28	5
14	M64110	Human caldesmon mRNA, complete cds		72	15	40	32	7
15	M83088	Human phosphoglucomutase 1 (PGM1) mRNA, complete cds		46	11	29	45	32
16	M83670	Human carbonic anhydrase IV mRNA, complete cds		-81	-30	-1	4	-34
17	M91463	Human glucose transporter (GLUT4) gene, complete cds		29	27	32	28	36
18	M95936	Human protein-serine/threonine (AKT2) mRNA, complete cds		54	-17	-8	49	4
19	R48602	yj65f01.s1 Homo sapiens cDNA clone 153625 3' similar to SP:T15B12.		36	8	27	65	27
20	R60877	yh08e11.s1 Homo sapiens cDNA clone 42396 3'		37	33	23	25	15
21	T46924	yb11b02.s1 Homo sapiens cDNA clone 70827 3' similar to gb:U11863		34	17	20	25	21
22	T49484	ya75h03.s1 Homo sapiens cDNA clone 67541 3' contains Alu repetitive		8	22	50	37	65
23	T54547	yb40h03.s1 Homo sapiens cDNA clone 73685 3' similar to gb:M84526		34	-2	5	45	-7
24	T55741	yb40d07.s1 Homo sapiens cDNA clone 73645 3' similar to SP:TELO_F		25	9	4	22	-5
25	T61597	yb74h08.s1 Homo sapiens cDNA clone 76959 3' similar to SP:JH0628		63	29	42	59	33
26	T67986	yc28e12.s1 Homo sapiens cDNA clone 82030 3' similar to gb:X14723		20	16	17	26	40
27	T94993	ye38a07.s1 Homo sapiens cDNA clone 119988 3' similar to gb:M8777		47	47	36	36	16
28	U08854	Human UDP glucuronosyltransferase precursor (UGT2B15) mRNA, co		21	22	59	57	16
29	U17077	Human BENE mRNA, partial cds		16	40	35	29	47
30	U31525	Human glycogenin mRNA, complete cds		41	26	33	26	37
31	X54162	Human mRNA for a 64 Kd autoantigen expressed in thyroid and extra-o		27	12	5	26	4
32	X64559	H.sapiens mRNA for tetranectin		9	-3	-1	19	9
33	Z49269	H.sapiens gene for chemokine HCC-1		18	12	29	28	39
34	Z49269	H.sapiens gene for chemokine HCC-1		42	27	44	51	38

Figure 4.7.

1. Additional menu item that corresponds to obtained genes set. **2.** Table with selected genes.

4.14. To visualize expression profiles for selected genes, use the "Graph>Specified with current selection" command from the main menu (fig. 4.8).

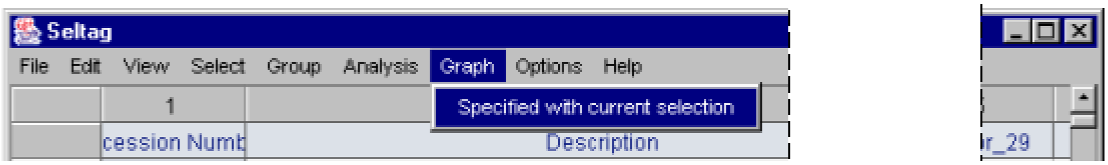
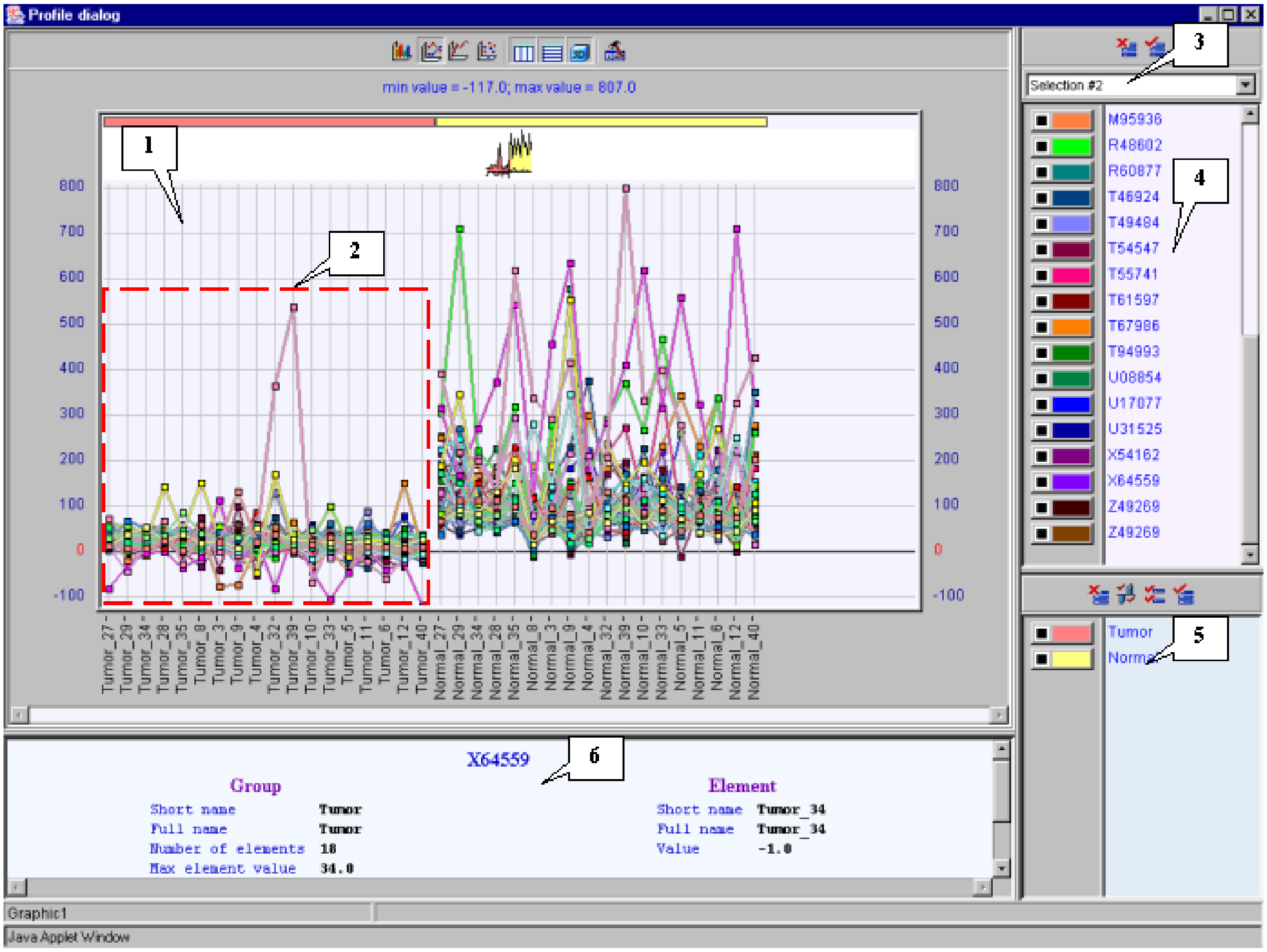

Seltag
File Edit View Select Group Analysis Graph Options Help
Specified with current selection
1
cession Numb
Description
r_29

Figure 4.8.

4.15. The "Profile dialog" window with expression profiles for selected genes (fig. 4.9) will appear.



min value = -117.0; max value = 807.0

X64559

Group		Element	
Short name	Tumor	Short name	Tumor_34
Full name	Tumor	Full name	Tumor_34
Number of elements	18	Value	-1.0
Max element value	34.0		

Figure 4.9.

1. Area of expression profiles visualization. **2.** Expression values for tumorous tissues. **3.** List of genes sets. **4.** List of genes. **5.** List of tissues groups. **6.** Information on gene, profile of which is pointed by mouse.

It is notable that expression profiles of all represented genes are lower for tumorous tissues (profiles inside the red rectangle) than for normal ones (profiles outside the red rectangle).

5. Selecting genes with highest correlation between expression profiles

From genes overexpressed in tumorous tissues we have selected the "Human gene for melanoma growth stimulatory activity (MGSA)", GenBank accession number X54489.

In current chapter it is described how to find genes having expression profiles similar to that of MGSA. As the similarity measure, the Pearson's correlation coefficient for expression profiles will be used.

To perform this task, the following steps are required:

5.1. Use the "Analysis>Correlations>Select most correlated genes" command from the main menu to call the «Select most correlated genes for specified gene set» dialog (fig. 5.1).

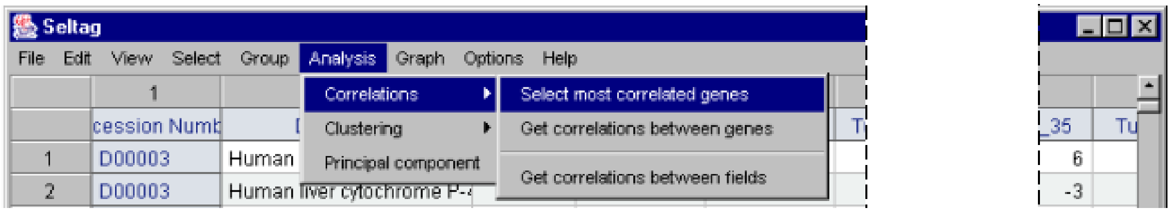
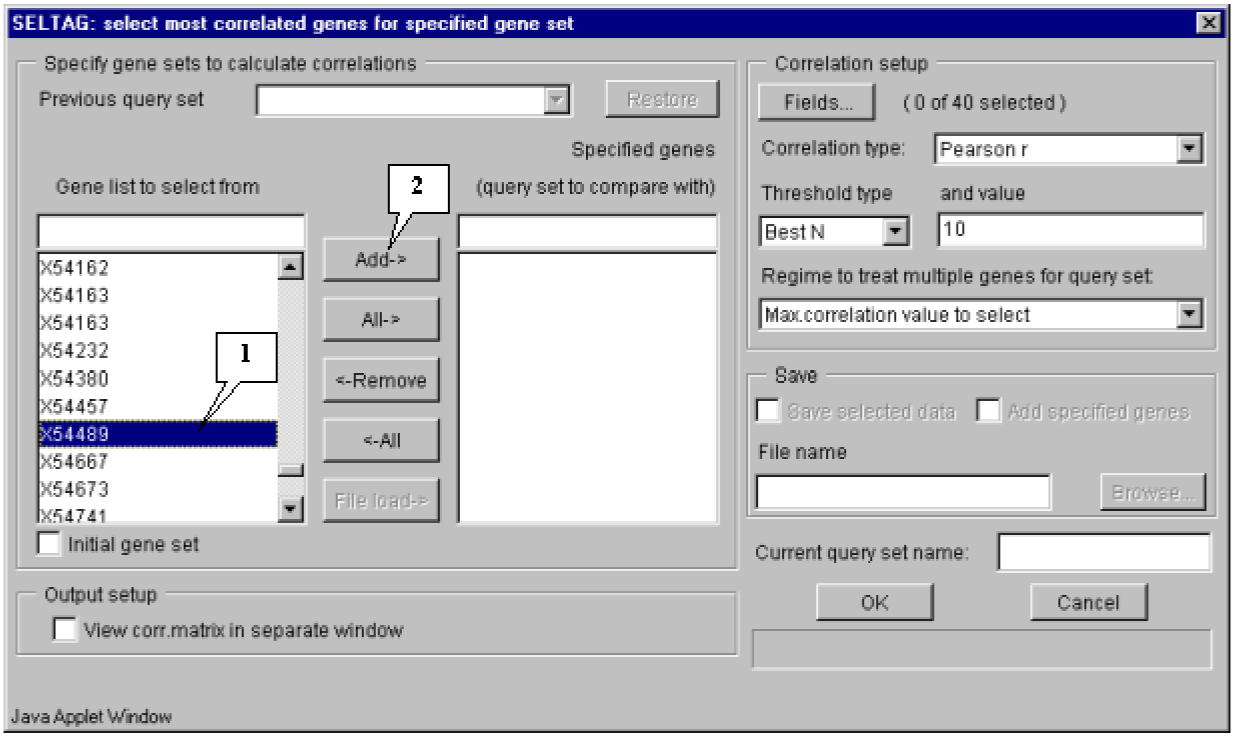

Seltag
File Edit View Select Group Analysis Graph Options Help
Correlations
Clustering
Principal component
Select most correlated genes
Get correlations between genes
Get correlations between fields
1
cession Numb
1 D00003 Human
2 D00003 Human liver cytochrome P-4
T
35 Tu
6
-3

Figure 5.1.

5.2. Specify the sets of genes to calculate correlations between them by selecting the appropriate gene, which will be used as the base for correlation analysis (in our case, this is gene "X54489" – MGSA), from the "Gene list to select from" list (fig. 5.2). Once gene is selected, press the "Add" button, and it will be relocated into "Specified genes" list (fig. 5.3).


SELTAG: select most correlated genes for specified gene set

Specify gene sets to calculate correlations

Previous query set

Gene list to select from

X54162
X54163
X54163
X54232
X54380
X54457
X54489
X54667
X54673
X54741

Specified genes (query set to compare with)

□ Initial gene set

Output setup

□ View corr.matrix in separate window

Correlation setup

Fields... (0 of 40 selected)

Correlation type: Pearson r

Threshold type: Best N and value: 10

Regime to treat multiple genes for query set: Max.correlation value to select

Save

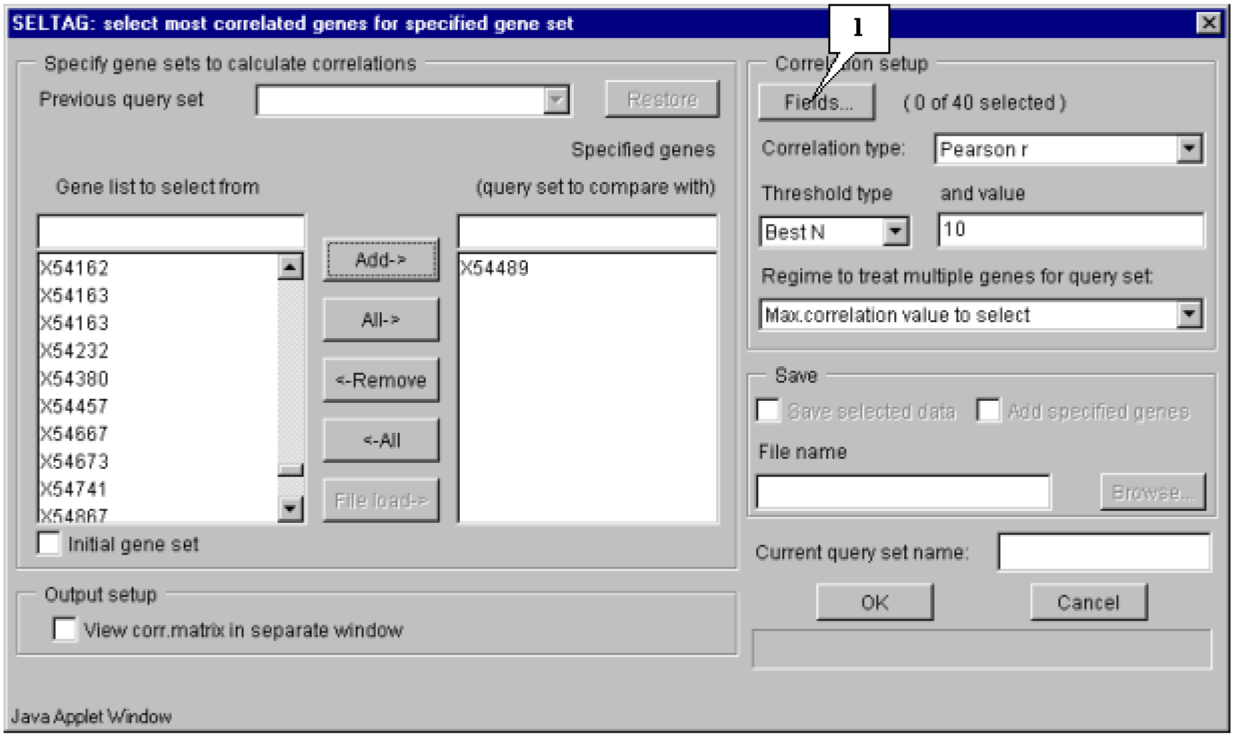
□ Save selected data □ Add specified genes

File name

Current query set name:

Figure 5.2.

1. Gene selected for adding to the "Specified genes" list. **2.** Button for adding of a gene to the "Specified genes" list.



Previous query set

Specified genes (query set to compare with)

X54489

Gene list to select from

X54162
X54163
X54163
X54232
X54380
X54457
X54667
X54673
X54741
X54867

□ Initial gene set

□ View corr.matrix in separate window

(0 of 40 selected)

Correlation type: Pearson r

Threshold type Best N and value 10

Regime to treat multiple genes for query set: Max.correlation value to select

□ Save selected data □ Add specified genes

File name

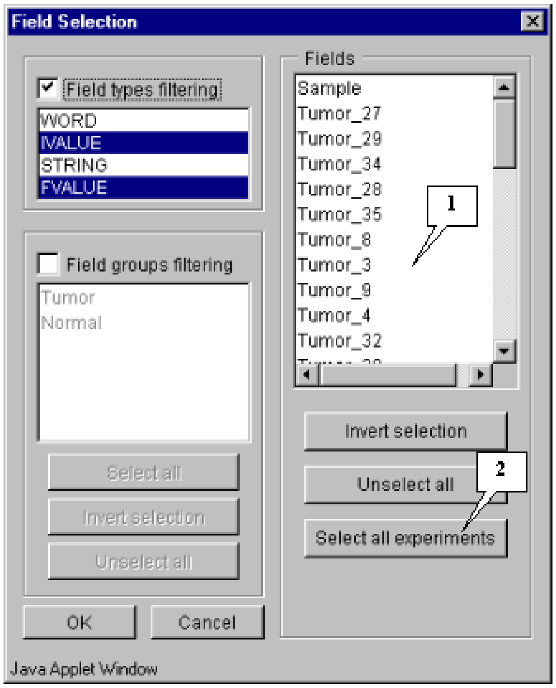
Current query set name:

Figure 5.3.

1. Button for fields selection dialog.

5.4. Specify the calculation parameters:

5.4.1. Specify fields that will be used for calculation by pressing the “Fields” button (fig. 5.3). The “Field selection” dialog (fig. 5.4.1.1) will appear.



☑ Field types filtering

WORD
IVALUE
STRING
FVALUE

□ Field groups filtering

Tumor
Normal

Sample
Tumor_27
Tumor_29
Tumor_34
Tumor_28
Tumor_35
Tumor_8
Tumor_3
Tumor_9
Tumor_4
Tumor_32

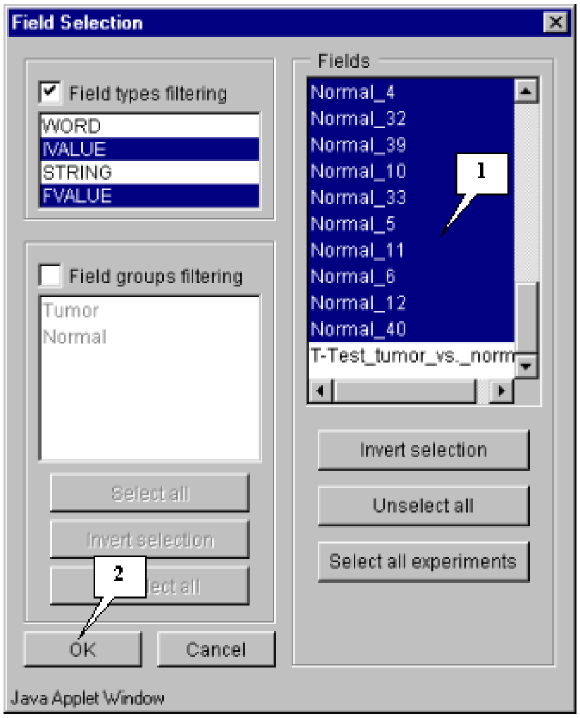
Figure 5.4.1.1.

1. List of fields. **2.** Button for selecting all fields with data on experiments.

In this example, calculation is based on expression measurements, i.e. all numeric fields, except “Sample” and “T-Test_tumor_vs._Normal”, are used.

Press the “Select all experiments” button to select all fields, and then remove selection from the mentioned fields by clicking mouse on their names (fig. 5.4.1.2).

Press the “OK” button.



Field Selection

☑ Field types filtering

WORD
IVALUE
STRING
FVALUE

□ Field groups filtering

Tumor
Normal

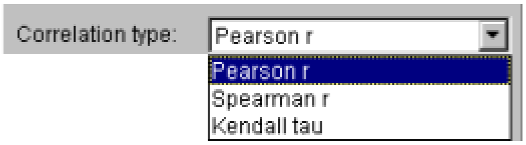
Fields

Normal_4
Normal_32
Normal_39
Normal_10
Normal_33
Normal_5
Normal_11
Normal_6
Normal_12
Normal_40
T-Test_tumor_vs._norm

Figure 5.4.1.2.

1. Selected fields with experiments data. **2.** Confirmation button.

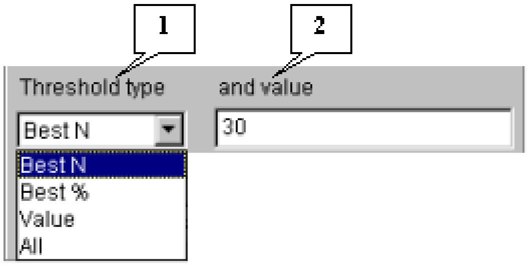
5.4.2. Select the appropriate correlation type from the “Correlation type” list (fig. 5.4.2). In this example, the Pearson’s correlation coefficient is used.



Correlation type: Pearson r

Figure 5.4.2.

5.4.3. Choose the correlation threshold type (fig. 5.4.3). If the “Best N”, “Best %” or “Value” threshold types are selected, in the “and value” field specify the threshold value. In the current example, the “Best N” type with value 30 is used. It means that after calculations 30 genes with maximal absolute values of correlation coefficients with the target gene MGSA expression profile will be selected.

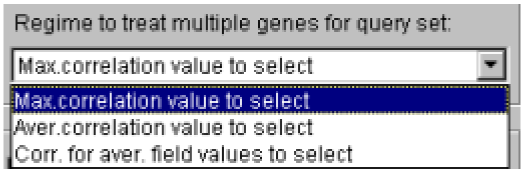


Threshold type: Best N
and value: 30

Figure 5.4.3.

1. The threshold type. **2.** The threshold value.

5.4.4. Choose the appropriate mode from the “Regime to treat multiple genes for query set” list (fig. 5.4.4). In this example, the “Max. correlation value to select” mode is used.

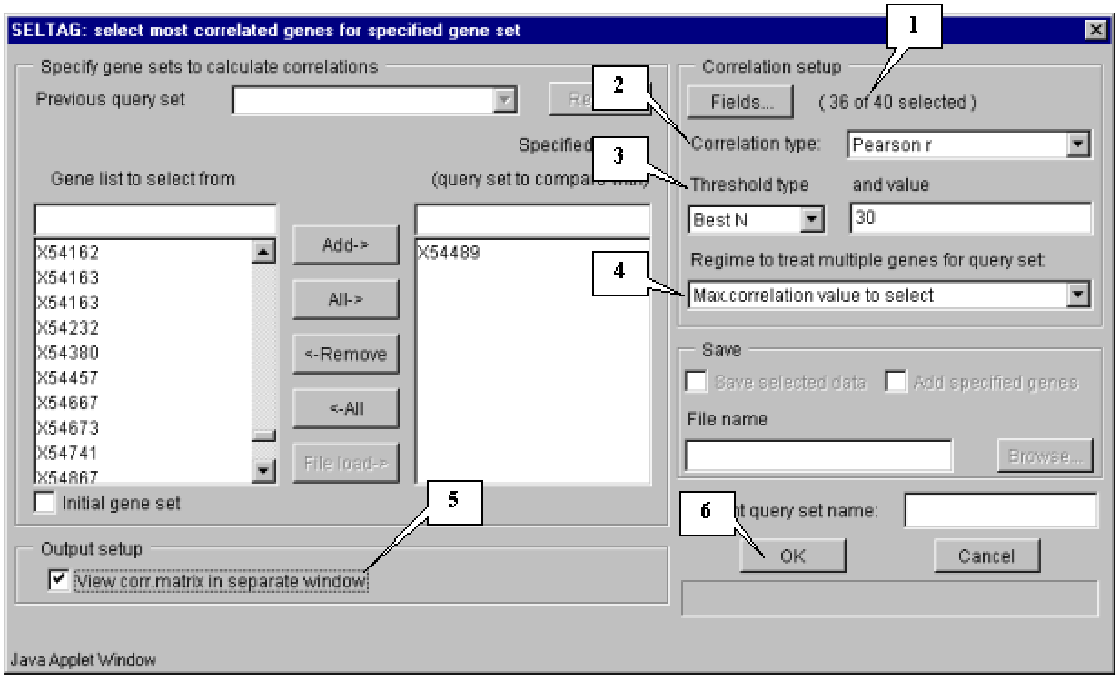


Max.correlation value to select
Aver.correlation value to select
Corr. for aver. field values to select

Regime to treat multiple genes for query set: Max.correlation value to select

Figure 5.4.4.

5.5. Specify the data output parameters. In this example, it is required to get a matrix with correlation coefficients in a separate window. To do this, check in the “View corr. matrix in separate window” checkbox (fig. 5.5).



SELTAG: select most correlated genes for specified gene set

Specify gene sets to calculate correlations

Previous query set:

Gene list to select from: X54162, X54163, X54163, X54232, X54380, X54457, X54667, X54673, X54741, X54867

Specified (query set to compare with): X54489

□ Initial gene set

Output setup

☑ View corr.matrix in separate window

Correlation setup

Fields... (36 of 40 selected)

Correlation type: Pearson r

Threshold type: Best N

and value: 30

Regime to treat multiple genes for query set: Max.correlation value to select

Save

□ Save selected data □ Add specified genes

File name:

...t query set name:

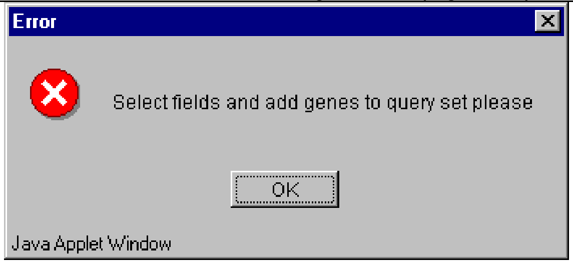
Java Applet Window

Figure 5.5.

1. Information on the number of selected fields. **2.** List of correlation types. **3.** List of correlation thresholds. **4.** List of correlation regimes. **5.** Checkbox for output of correlation matrix in a separate window. **6.** Confirmation button.

Comment. The figure 5.5 represents the selected calculation parameters.

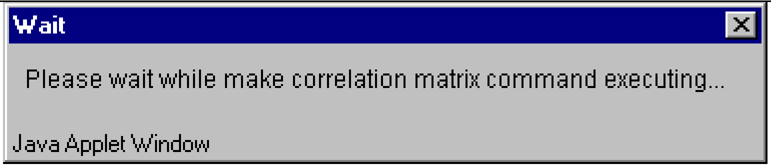
5.6. Press the "OK" button (fig. 5.5). If sets of genes are not formed or fields for calculations are not selected, the "Error" message box (fig. 5.6) will appear.



Select fields and add genes to query set please

Figure 5.6.

5.7. During the data processing the "Wait" message (fig. 5.7) appears, and once the process is over it disappears.

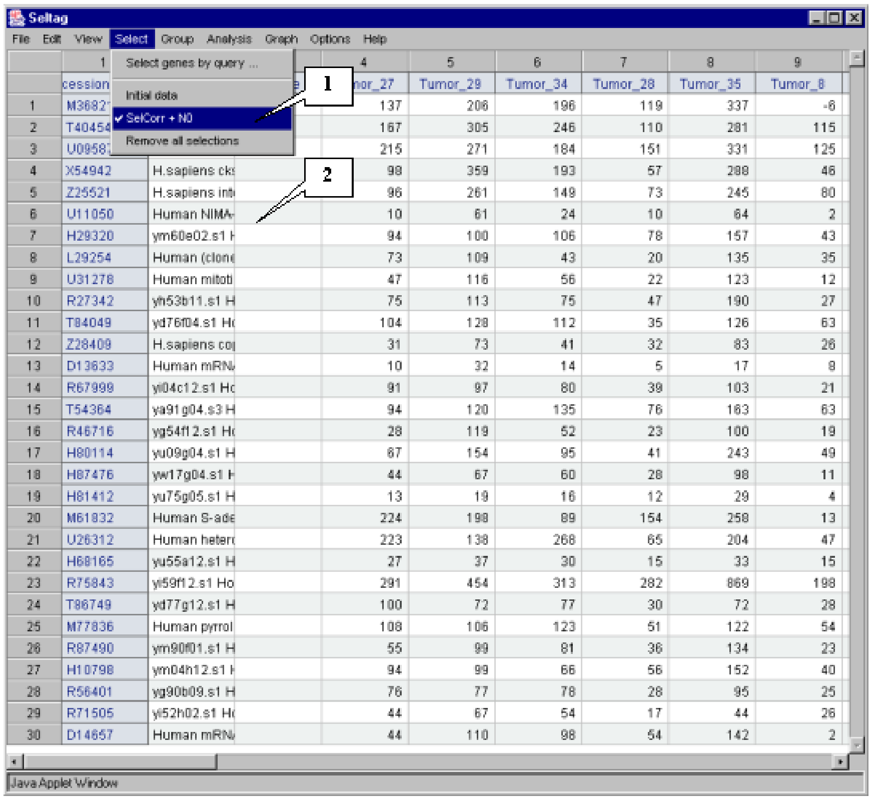


Please wait while make correlation matrix command executing...

Figure 5.7.

5.8. Results of data processing:

5.8.1. In the main window the table with selected genes will be represented (fig. 5.8.1). Into the "Select" menu section the new "SelCorr + N0" item corresponding to current set of genes will be added.



	1			4	5	6	7	8	9
	cession		e	nor_27	Tumor_29	Tumor_34	Tumor_28	Tumor_35	Tumor_8
1	M3682			137	206	196	119	337	-6
2	T40454			167	305	246	110	281	115
3	U0958			215	271	184	151	331	125
4	X54942	H.sapiens ck		98	359	193	57	288	46
5	Z25521	H.sapiens int		96	261	149	73	245	80
6	U11050	Human NIMA-		10	61	24	10	64	2
7	H29320	ym60e02.s1 H		94	100	106	78	157	43
8	L29254	Human (clone		73	109	43	20	135	35
9	U31278	Human mitoti		47	116	56	22	123	12
10	R27342	yh53b11.s1 H		75	113	75	47	190	27
11	T84049	yd76f04.s1 H		104	128	112	35	126	63
12	Z28409	H.sapiens co		31	73	41	32	83	26
13	D13633	Human mRN		10	32	14	5	17	8
14	R67999	yi04c12.s1 H		91	97	80	39	103	21
15	T54364	ya91g04.s3 H		94	120	135	76	163	63
16	R46716	yg54f12.s1 H		28	119	52	23	100	19
17	H80114	yu09g04.s1 H		67	154	95	41	243	49
18	H87476	yw17g04.s1 H		44	67	60	28	98	11
19	H81412	yu75g05.s1 H		13	19	16	12	29	4
20	M61832	Human S-ade		224	198	89	154	258	13
21	U26312	Human heter		223	138	268	65	204	47
22	H68165	yu55a12.s1 H		27	37	30	15	33	15
23	R75843	yi59f12.s1 Ho		291	454	313	282	869	198
24	T86749	yd77g12.s1 H		100	72	77	30	72	28
25	M77836	Human pyrrol		108	106	123	51	122	54
26	R87490	ym90f01.s1 H		55	99	81	36	134	23
27	H10798	ym04h12.s1 H		94	99	66	56	152	40
28	R56401	yg90b09.s1 H		76	77	78	28	95	25
29	R71505	yi52h02.s1 H		44	67	54	17	44	26
30	D14657	Human mRN		44	110	98	54	142	2

Figure 5.8.1.

1. The "Select" menu item corresponding to obtained set of genes. **2.** Table with obtained set of genes.

5.8.2. If to open the “View group data” dialog (using the “Group>View” command from the main menu - fig. 5.8.2.1), then one can see that in the list of experiments the new «group from SelCorr +N0» item with fields used for calculations (fig. 5.8.2.2) has appeared.

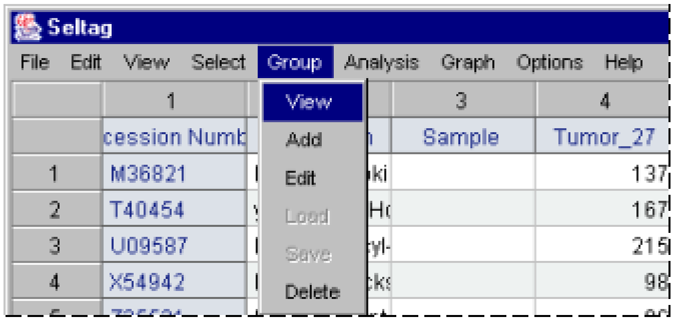
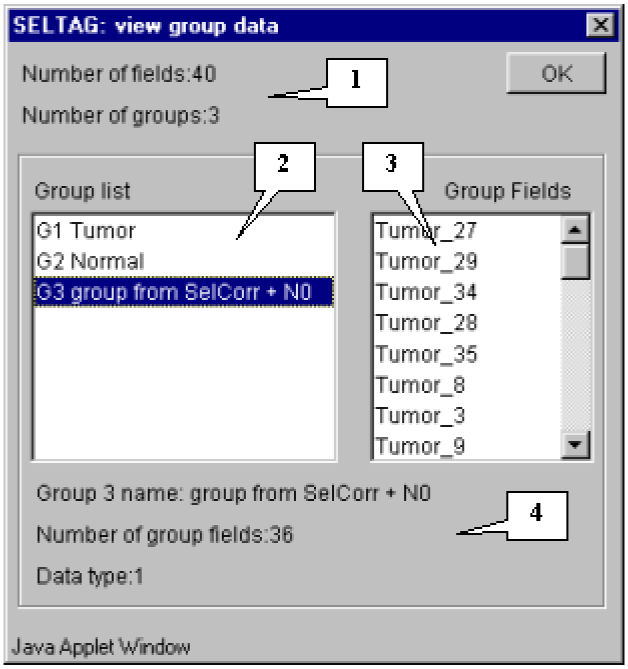


Figure 5.8.2.1.



Number of fields:40

Number of groups:3

G1 Tumor
G2 Normal
G3 group from SelCorr + N0

Tumor_27
Tumor_29
Tumor_34
Tumor_28
Tumor_35
Tumor_8
Tumor_3
Tumor_9

Group 3 name: group from SelCorr + N0

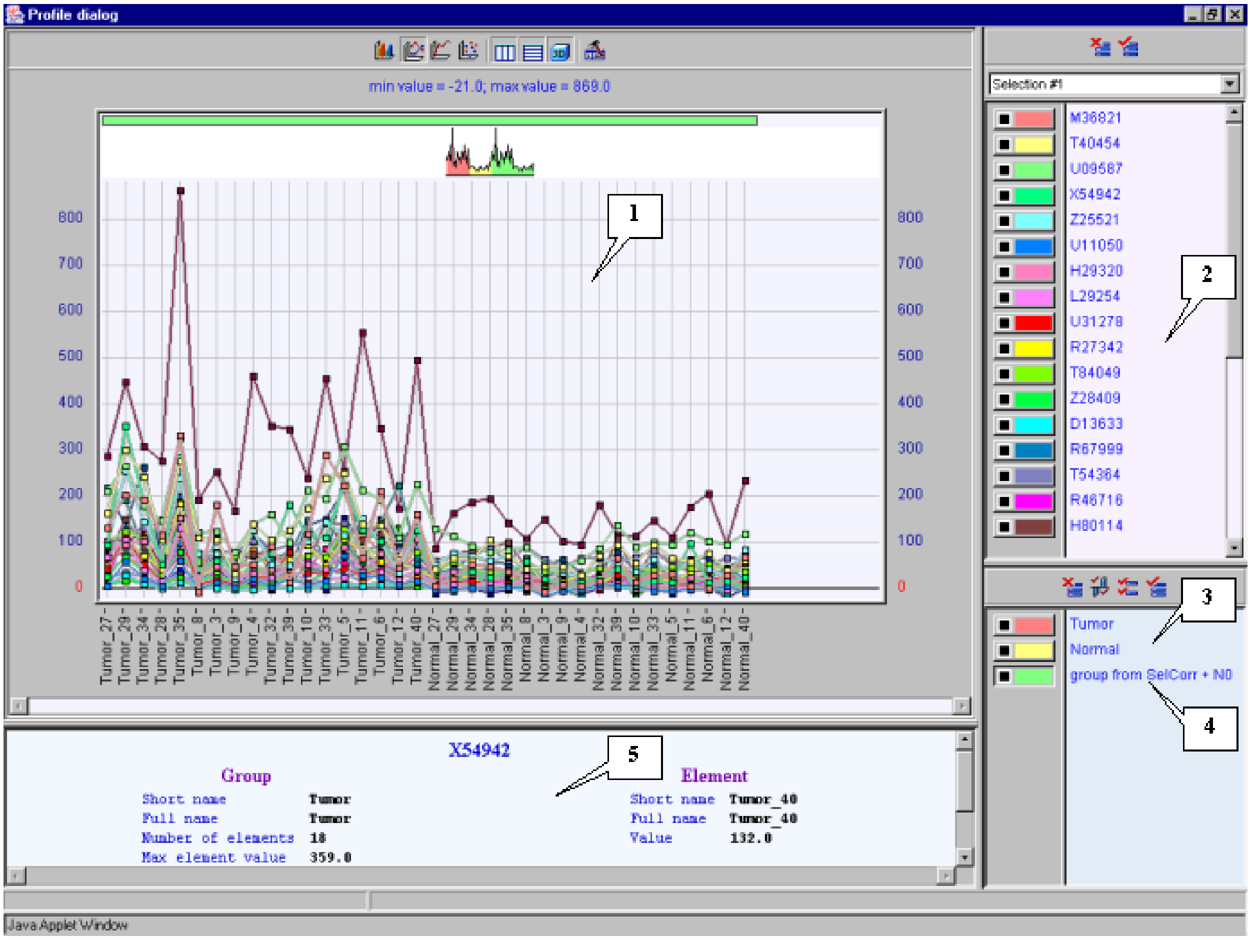
Number of group fields:36

Data type:1

Figure 5.8.2.2.

1. Information on the number of fields and groups in a project. **2.** List of groups. **3.** List of fields in the selected group. **4.** Information on the selected group.

5.8.3. The «Profile dialog» window (fig. 5.8.3) with expression profiles for selected genes will appear.



🟥	M36821
🟨	T40454
🟩	U09587
🟩	X54942
🟦	Z25521
🟦	U11050
🟪	H29320
🟪	L29254
🟥	U31278
🟨	R27342
🟩	T84049
🟩	Z28409
🟦	D13633
🟦	R67999
🟪	T54364
🟪	R46716
🟫	H80114

🟥	Tumor
🟨	Normal
🟩	group from SelCorr + N0

X54942

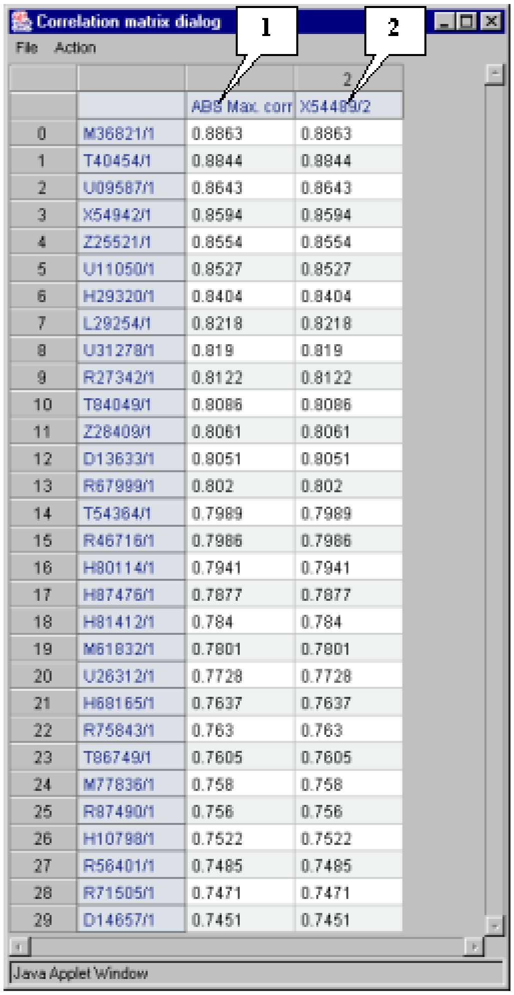
Group		Element	
Short name	Tumor	Short name	Tumor_40
Full name	Tumor	Full name	Tumor_40
Number of elements	18	Value	132.0
Max element value	359.0		

Figure 5.8.3.

1. Area of expression profiles visualization. **2.** List of genes. **3.** List of groups. **4.** Group with fields, used for calculations. **5.** Information on gene, profile of which is pointed by mouse.

5.8.4. Since the «View corr. matrix in separate window» checkbox in the «Select most correlated genes for specified gene set» dialog was checked in, the correlation matrix's window with correlation coefficients obtained in data processing (fig. 5.8.4) will appear.

The figure 5.8.4 shows that genes with GenBank accession numbers M36821, T40454, U09587, X54942 и Z25521 have expression profiles, which correlate with that of target X54489 gene in larger degree than others.



		1	2
		ABS Max. corr	X54489/2
0	M36821/1	0.8863	0.8863
1	T40454/1	0.8844	0.8844
2	U09587/1	0.8643	0.8643
3	X54942/1	0.8594	0.8594
4	Z25521/1	0.8554	0.8554
5	U11050/1	0.8527	0.8527
6	H29320/1	0.8404	0.8404
7	L29254/1	0.8218	0.8218
8	U31278/1	0.819	0.819
9	R27342/1	0.8122	0.8122
10	T84049/1	0.8086	0.8086
11	Z28409/1	0.8061	0.8061
12	D13633/1	0.8051	0.8051
13	R67999/1	0.802	0.802
14	T54364/1	0.7989	0.7989
15	R46716/1	0.7986	0.7986
16	H80114/1	0.7941	0.7941
17	H87476/1	0.7877	0.7877
18	H81412/1	0.784	0.784
19	M61832/1	0.7801	0.7801
20	U26312/1	0.7728	0.7728
21	H68165/1	0.7637	0.7637
22	R75843/1	0.763	0.763
23	T86749/1	0.7605	0.7605
24	M77836/1	0.758	0.758
25	R87490/1	0.756	0.756
26	H10798/1	0.7522	0.7522
27	R56401/1	0.7485	0.7485
28	R71505/1	0.7471	0.7471
29	D14657/1	0.7451	0.7451

Figure 5.8.4.

1. Column with maximal coefficients of correlation. **2.** Column with obtained coefficients of correlation.

6. Clustering genes by their expression profiles.

In this chapter, the task of clustering genes by similarity of their expression profiles will be considered. For the first, genes with expression in tumorous tissues that significantly (by Student's criterion) differs from that in normal tissues (genes for which the value "T-Test_tumor_vs._normal" is lesser than 0.0001) should be selected.

To perform this task, the following steps are required:

6.1. Use the "Select>Select genes by query…" command from the main menu (fig. 6.1).

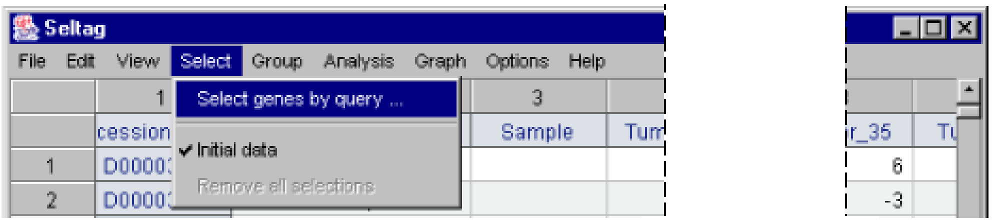
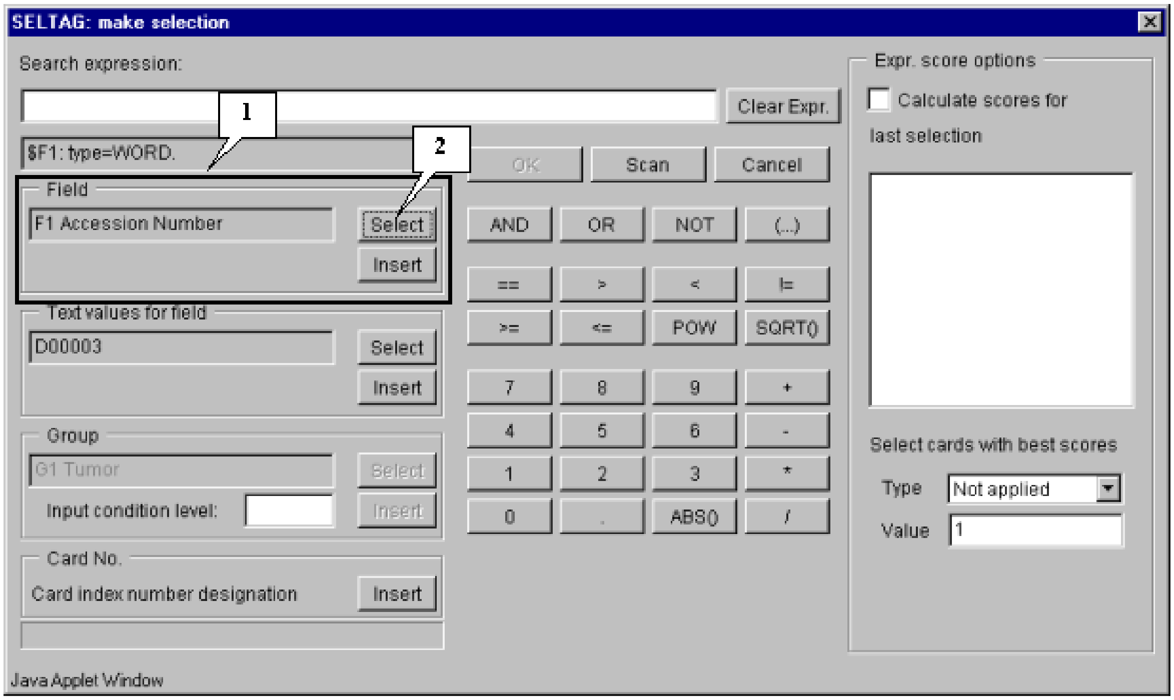


Figure 6.1.

6.2. The "Make selection" dialog (fig. 6.2) will appear. For the first, choose a field, which will be used for selection. In the "Field" section press the "Select" button (fig. 6.2).



$F1: type=WORD.

Field: F1 Accession Number

Text values for field: D00003

Group: G1 Tumor

Input condition level:

Type: Not applied

Value: 1

Figure 6.2.

1. The "Field" section. **2.** The "Select" button for calling the "Calculator options" dialog.

6.3. The "Calculator options" dialog with complete list of fields in the table (fig. 6.3) will appear. In the list select a field, which will be used for search. In our case it is the field: «F40 T-Test_tumor_vs._normal».
Press the "OK" button.

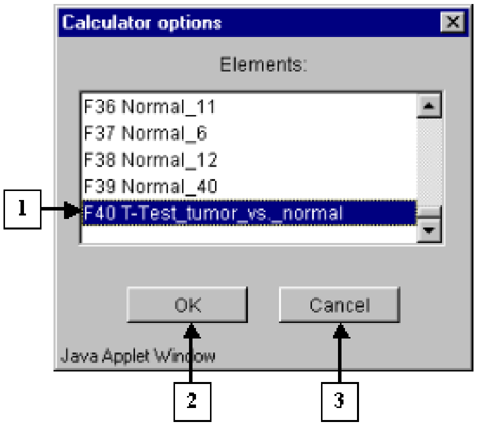


Figure 6.3.

1. Selected field. **2.** Confirmation button. **3.** Cancel button.

6.4. In the "Make selection" dialog the following changes will occur (fig. 6.4):

- In the text area, the number and the name of selected field will appear: F40 T-Test_tumor_vs._normal
- In the status bar, the information on this field will be shown: $F40: type=FVALUE.

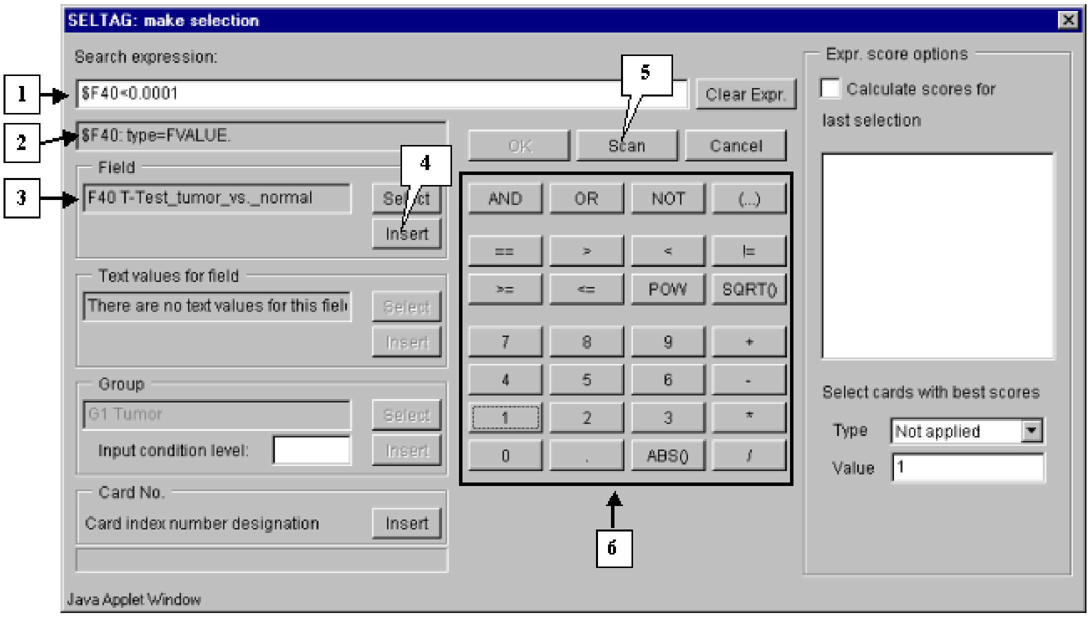
6.5. Selected field ID should be inserted into expression line. To do this, press the "Insert" button. In the expression line the following will appear:

$F40

6.6. Using buttons of query entering specify the condition in the expression line (fig. 6.4):

$F40<0.0001

6.7. To start the search process press the "Scan" button (fig. 6.4).



Search expression: $F40<0.0001

$F40: type=FVALUE.

Field: F40 T-Test_tumor_vs._normal

Text values for field: There are no text values for this fiel

Group: G1 Tumor

Input condition level:

Card No.: Card index number designation

Expr. score options

□ Calculate scores for last selection

Select cards with best scores

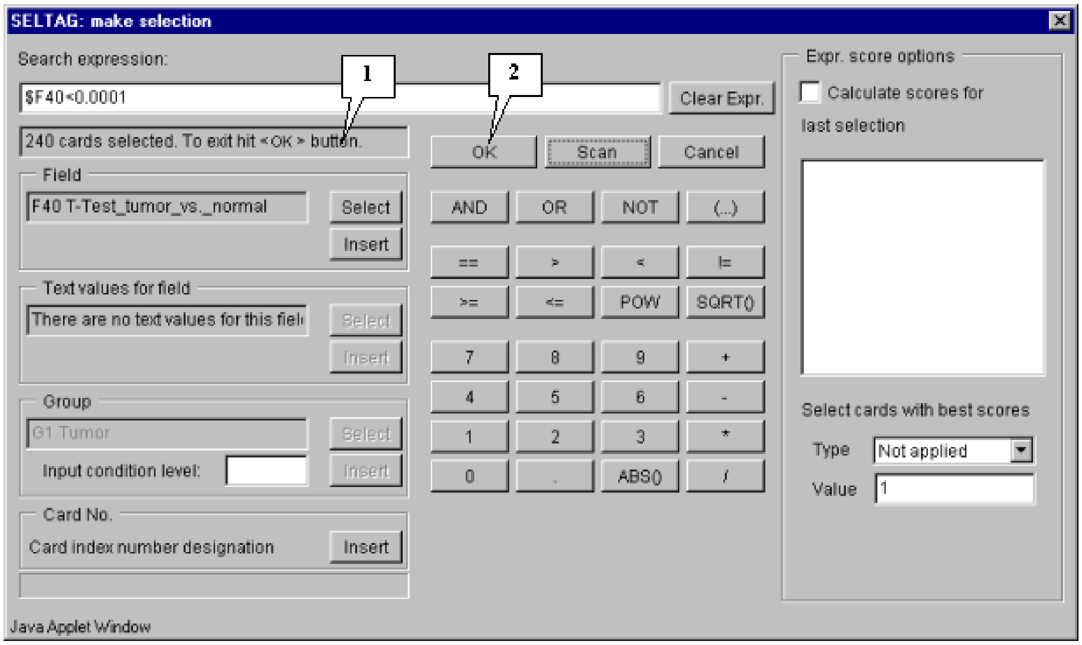
Type: Not applied

Value: 1

Figure 6.4.

1. Expression line with selection criteria. **2.** Status bar with information on selected field. **3.** Number and ID of selected field. **4.** The "Insert" button for insertion of the field's number into expression line. **5.** Button for search start. **6.** Query entering buttons.

6.8. Once the selection is finished, information on the number of found genes will be represented in the status bar (fig. 6.5), and the “OK” button will become active.



SELTAG: make selection

Search expression: $F40<0.0001

240 cards selected. To exit hit <OK> button.

Field: F40 T-Test_tumor_vs._normal

Text values for field: There are no text values for this fiel

Group: G1 Tumor

Input condition level:

Card No.: Card index number designation

Expr. score options

□ Calculate scores for last selection

Select cards with best scores

Type: Not applied

Value: 1

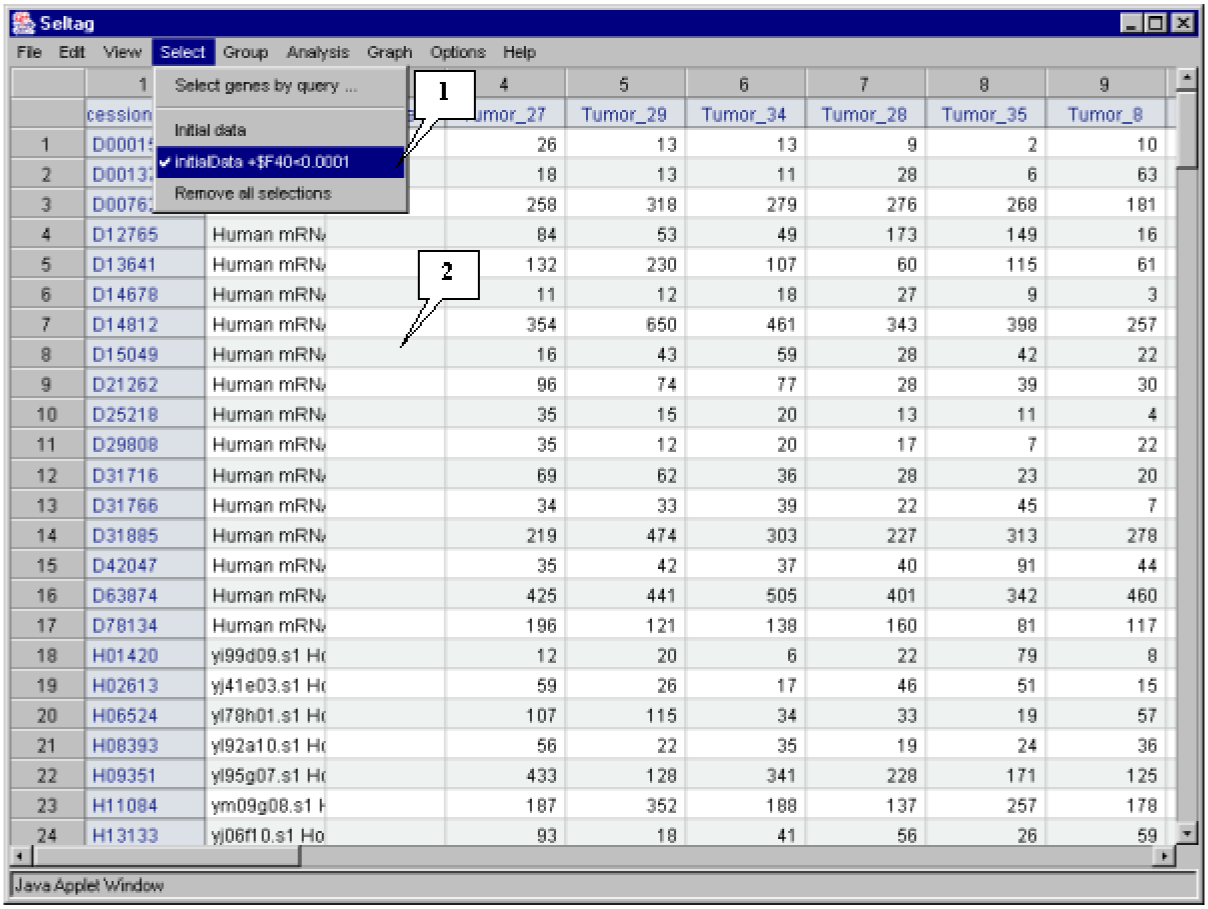
Java Applet Window

Figure 6.5.

1. Information on the number of found genes. **2.** Button for accepting the search results.

6.9. Press the “OK” button.

6.10. In the application main window the table with selected genes (fig. 6.6) will be represented. In the «Select» section of the main window menu an additional item with the name corresponding to selected set of genes will appear. During the project run, the obtained sets of genes can be saved and remained available by simple switching between them. To remove the list of tables use the «Remove all selections» command.



	1			4	5	6	7	8	9
	cession			umor_27	Tumor_29	Tumor_34	Tumor_28	Tumor_35	Tumor_8
1	D0001			26	13	13	9	2	10
2	D0013			18	13	11	28	6	63
3	D0076			258	318	279	276	268	181
4	D12765	Human mRN		84	53	49	173	149	16
5	D13641	Human mRN		132	230	107	60	115	61
6	D14678	Human mRN		11	12	18	27	9	3
7	D14812	Human mRN		354	650	461	343	398	257
8	D15049	Human mRN		16	43	59	28	42	22
9	D21262	Human mRN		96	74	77	28	39	30
10	D25218	Human mRN		35	15	20	13	11	4
11	D29808	Human mRN		35	12	20	17	7	22
12	D31716	Human mRN		69	62	36	28	23	20
13	D31766	Human mRN		34	33	39	22	45	7
14	D31885	Human mRN		219	474	303	227	313	278
15	D42047	Human mRN		35	42	37	40	91	44
16	D63874	Human mRN		425	441	505	401	342	460
17	D78134	Human mRN		196	121	138	160	81	117
18	H01420	yi99d09.s1 H		12	20	6	22	79	8
19	H02613	yj41e03.s1 H		59	26	17	46	51	15
20	H06524	yl78h01.s1 H		107	115	34	33	19	57
21	H08393	yl92a10.s1 H		56	22	35	19	24	36
22	H09351	yl95g07.s1 H		433	128	341	228	171	125
23	H11084	ym09g08.s1 H		187	352	188	137	257	178
24	H13133	yj06f10.s1 Ho		93	18	41	56	26	59

Figure 6.6.

1. Additional menu item that corresponds to obtained genes set. **2.** Table with selected genes.

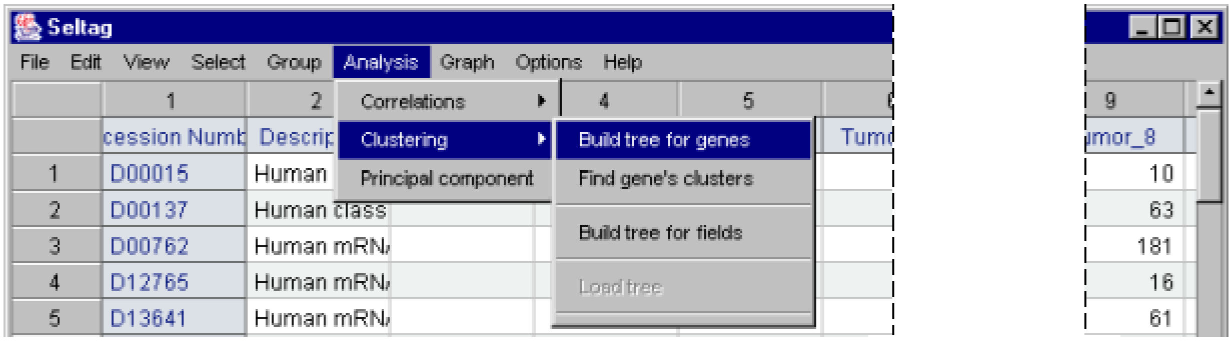
As a result, the set of 240 genes with significant differences in expression between normal and tumorous tissues have been obtained. The further analysis will be performed for this set of genes.

7. Building the tree of genes.

One of the approaches for revealing the clusters of genes with similar expression profiles is the hierarchical clustering [2]. This approach is based on building a binary tree for genes using a specified metrics of distance between their expression profiles. Each tree node binds two descending nodes and branches lengths correspond to distances between expression profiles.

To perform this task, the following steps are required:

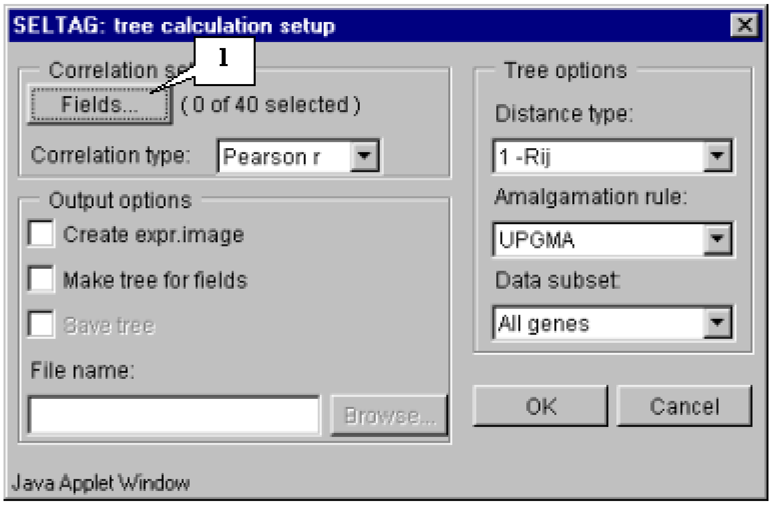
7.1. Use the “Analysis>Clustering>Build tree for genes” command from the main menu (fig. 7.1).



	1	2	4	5	9
	cession Numb	Descrip		Tumo	ımor_8
1	D00015	Human			10
2	D00137	Human class			63
3	D00762	Human mRN/			181
4	D12765	Human mRN/			16
5	D13641	Human mRN/			61

Figure 7.1.

7.2. The “Tree calculation setup” dialog (fig. 7.2) will appear. For the first, choose the fields that will be used for calculations by pressing the “Fields” button.



(0 of 40 selected)

Correlation type: Pearson r

☐ Create expr.image

☐ Make tree for fields

☐ Save tree

File name:

Distance type: 1 -Rij

Amalgamation rule: UPGMA

Data subset: All genes

Figure 7.2.

1. The "Fields" button.

7.3. The “Field selection” dialog (fig. 7.3) with the list of fields will appear.

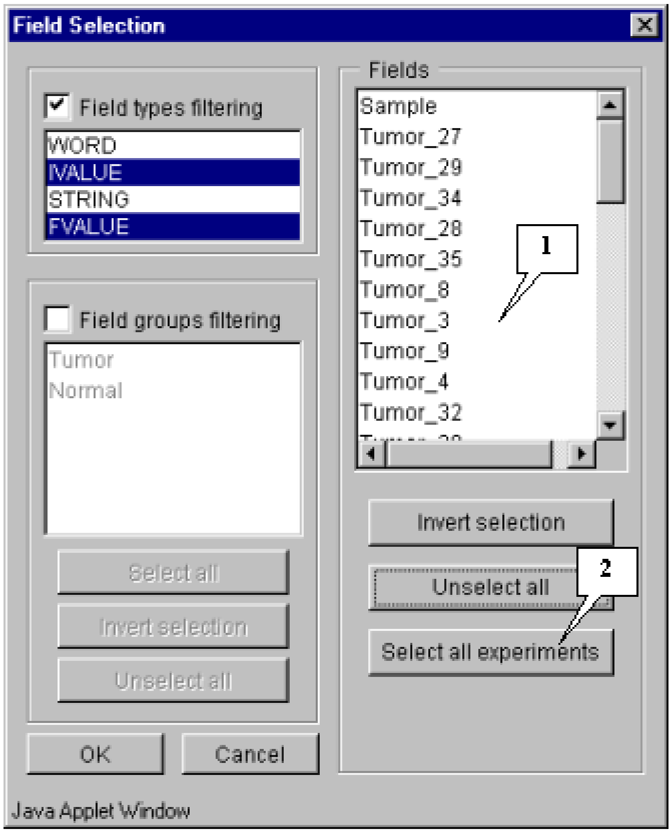


Figure 7.3.

1. List of fields. **2.** Button for selection of all fields that are expression data.

7.4. In this example, all fields, except the “Sample” and “T-Test_tumor_vs_Normal” ones, are used for calculations. Press the “Select all experiments” button (fig. 7.3) to select all fields and then remove selection from the appropriate ones (“Sample” and “T-Test_tumor_vs_Normal”) (fig. 7.4).

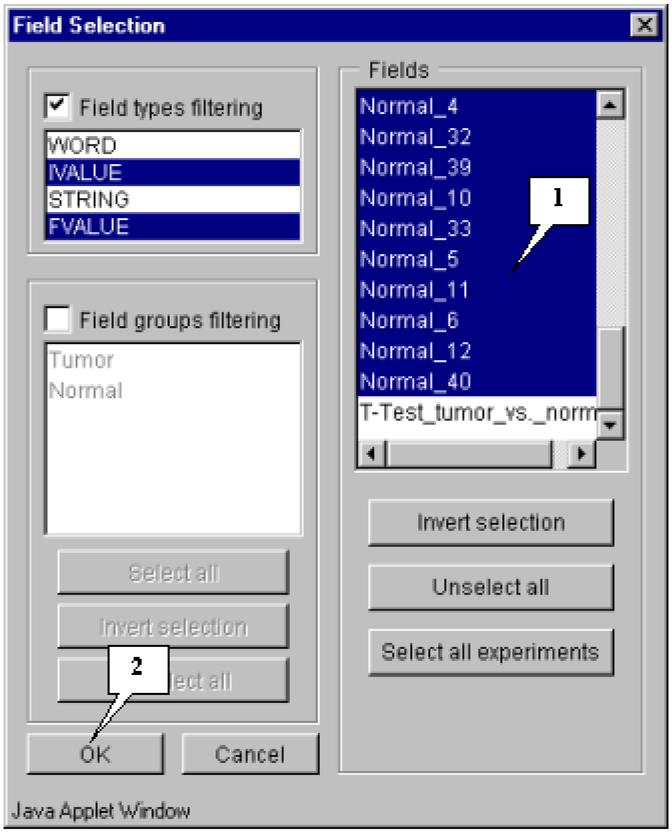


Figure 7.4.

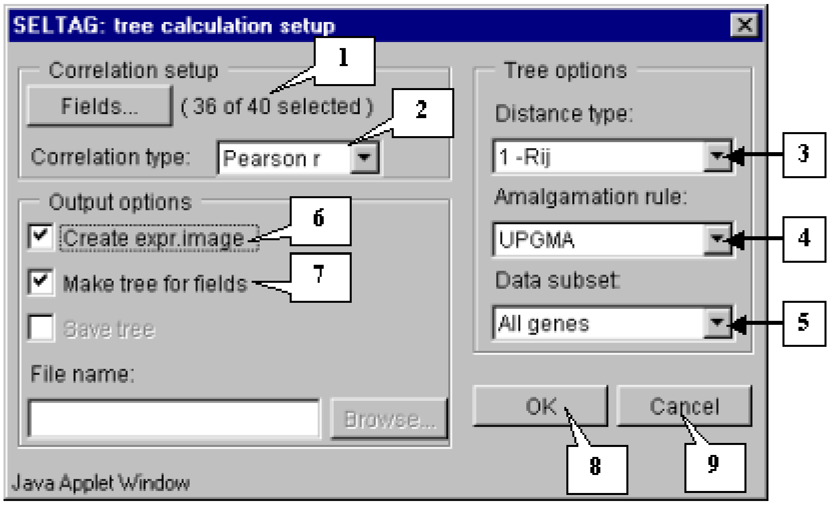
1. Selected fields. **2.** Confirmation button.

7.5. Press the “OK” button. In the “Tree calculation setup” dialog, in the area near the “Fields” button, the information on the number of selected fields will be represented (fig. 7.5). After this, do the following:

- Choose the type of correlation from the “Correlation type” list. In this example, the Pearson's correlation coefficient is used.
- Choose the type of distances that are calculated on the base of Rij correlation coefficients from the «Distance type» list. In this example, the “1-Rij” distance is used. It means that the coefficient of correlation between pair of genes *i* and *j* is higher the distance is smaller.
- Choose the type of nodes amalgamation from the «Amalgamation rule» list. In this example, the “UPGMA” type is used.
- Choose the subset of genes for building a tree from the «Data subset» list. In this example, the “all genes” subset (all genes from the current table) is used.
- Specify the parameters for data output.

- Check in the «Create expr. image» checkbox to obtain the diagram of expression matrix after calculation.
- Check in the «Make tree for fields» checkbox. In this case, the tree of similarity between experiments values in expression matrix is calculated, and visualization of the expression diagram is occurred in accordance to tree of fields.

The described settings are shown in the figure 7.5.



SELTAG: tree calculation setup

Correlation setup

(36 of 40 selected)

Correlation type: Pearson r

Output options

☑ Create expr.image

☑ Make tree for fields

□ Save tree

File name:

Tree options

Distance type: 1 -Rij

Amalgamation rule: UPGMA

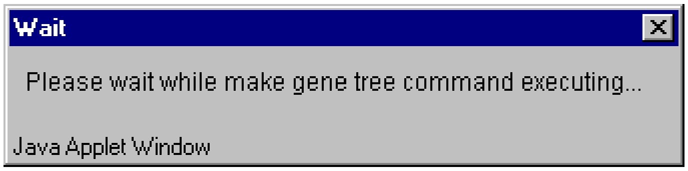
Data subset: All genes

Java Applet Window

Figure 7.5.

1. Information on the number of selected fields. **2.** List of correlation coefficients. **3.** List of distance types. **4.** List of nodes amalgamation types. **5.** List of genes subsets for building a tree. **6.** Checkbox for diagram of expression matrix output. **7.** Checkbox that regulates the order of experiments in accordance to tree of fields. **8.** Confirmation button. **9.** Cancel button.

7.6. Press the "OK" button. The "Wait" message box (fig. 7.6.) will appear for the duration of calculation process.



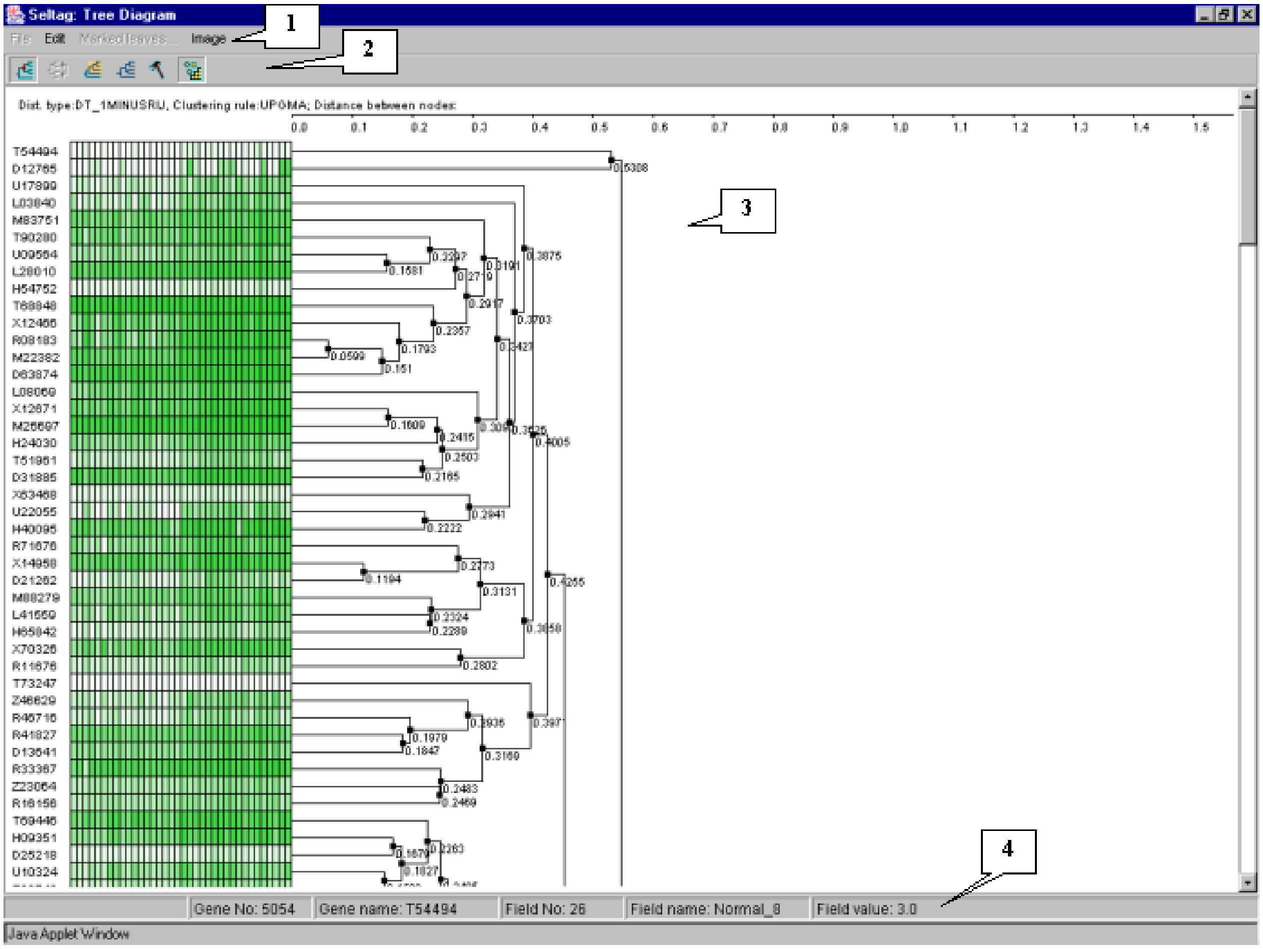
Wait

Please wait while make gene tree command executing...

Java Applet Window

Figure 7.6.

7.7. The “Tree Diagram” window with the tree of genes and expression matrix (fig. 7.7) will appear.



Gene No: 5054 | Gene name: T54494 | Field No: 26 | Field name: Normal_8 | Field value: 3.0

Figure 7.7.

1. Main menu. **2.** Toolbar. **3.** Diagram representation area. **4.** Status bar.

7.8. For more illustrative representation of the expression matrix, change the appropriate parameters. Use the "Image>Image setup" command from the main menu (fig. 7.8).

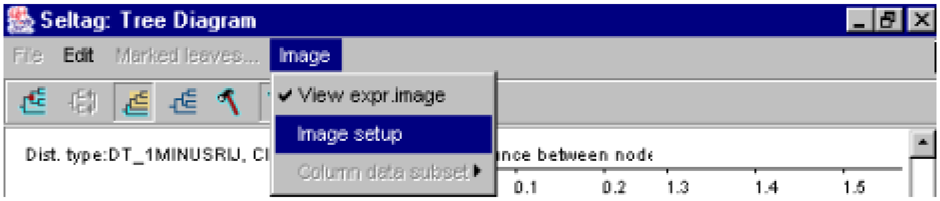


Figure 7.8.

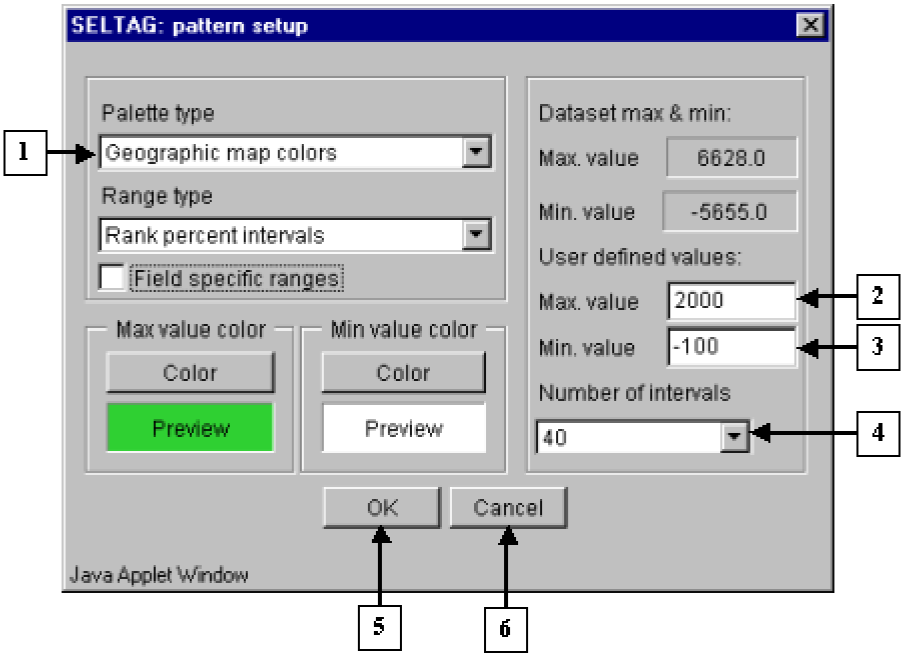
7.9. The “Pattern setup” dialog (fig. 7.9) will appear. In this window, the following settings should be changed:

- In the “Palette type” list choose the “Geographic map colors” palette type.
- In the “User defined values:” field set the value «2000» for maximum and «–100» for minimum.

- In the “Number of intervals” list select the value «40».

The described settings are shown in the figure 7.9.

Press the “OK” button. The matrix palette will change in accordance with selected settings.



SELTAG: pattern setup

Palette type: Geographic map colors

Range type: Rank percent intervals

□ Field specific ranges

Max value color: Color, Preview

Min value color: Color, Preview

Dataset max & min:

Max. value 6628.0

Min. value -5655.0

User defined values:

Max. value 2000

Min. value -100

Number of intervals 40

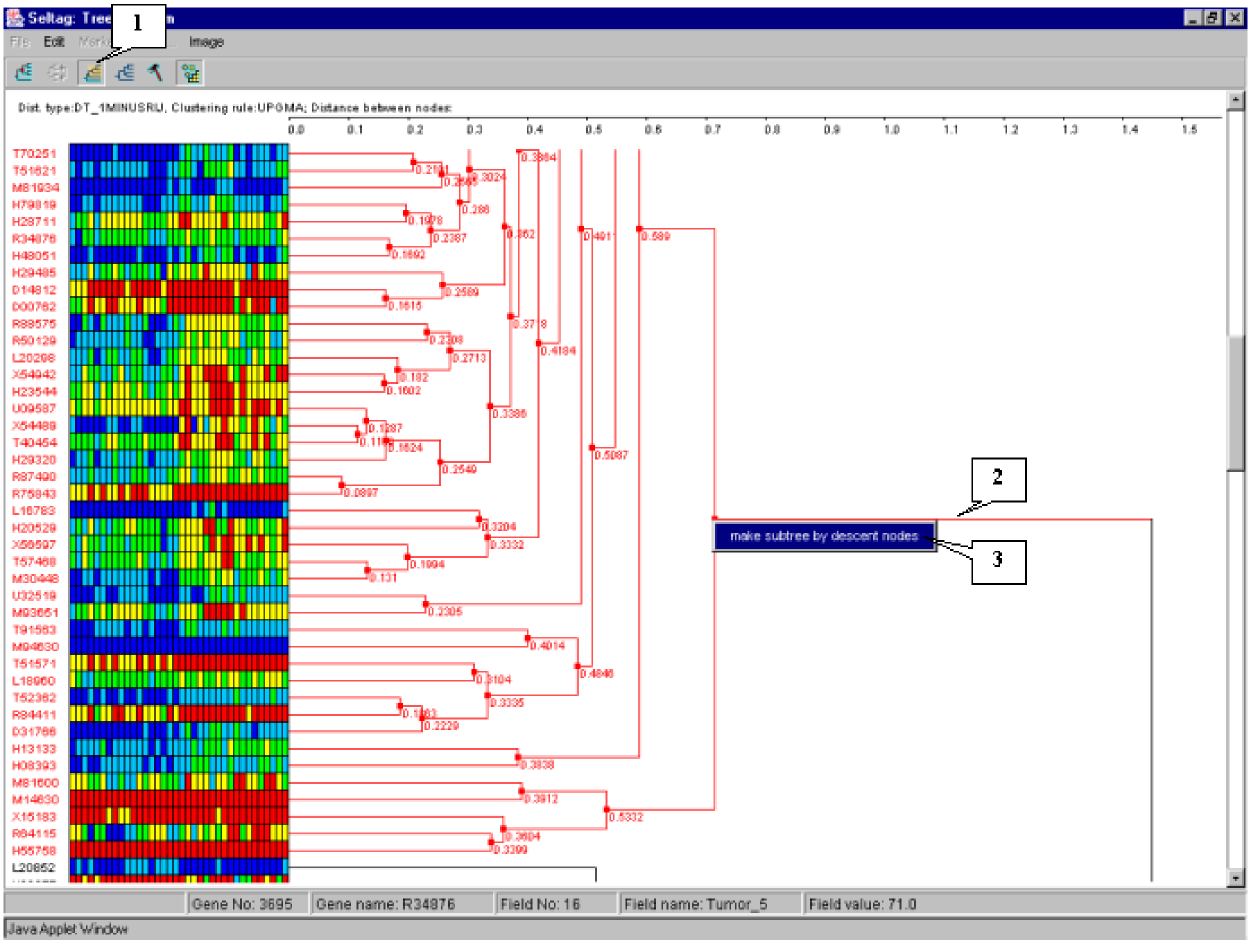
OK Cancel

Java Applet Window

Figure 7.9.

1. List of palettes. **2.** User-defined maximal value. **3.** User-defined minimal value. **4.** List of intervals. **5**. Confirmation button. **6.** Cancel button.

7.10. It is illustrative on the tree diagram, that genes are divided into two large clusters. To continue analysis for the one of them only, press the button to turn on the mode for selection of descending nodes by mouse clicking, and click on the root node of the cluster. All nodes and branches of the selected cluster (cluster 1 for the further) will be highlighted by red (fig. 7.10). Click on the selected cluster by the right mouse button and choose the appeared "Make subtree by descent nodes" command (fig. 7.10).



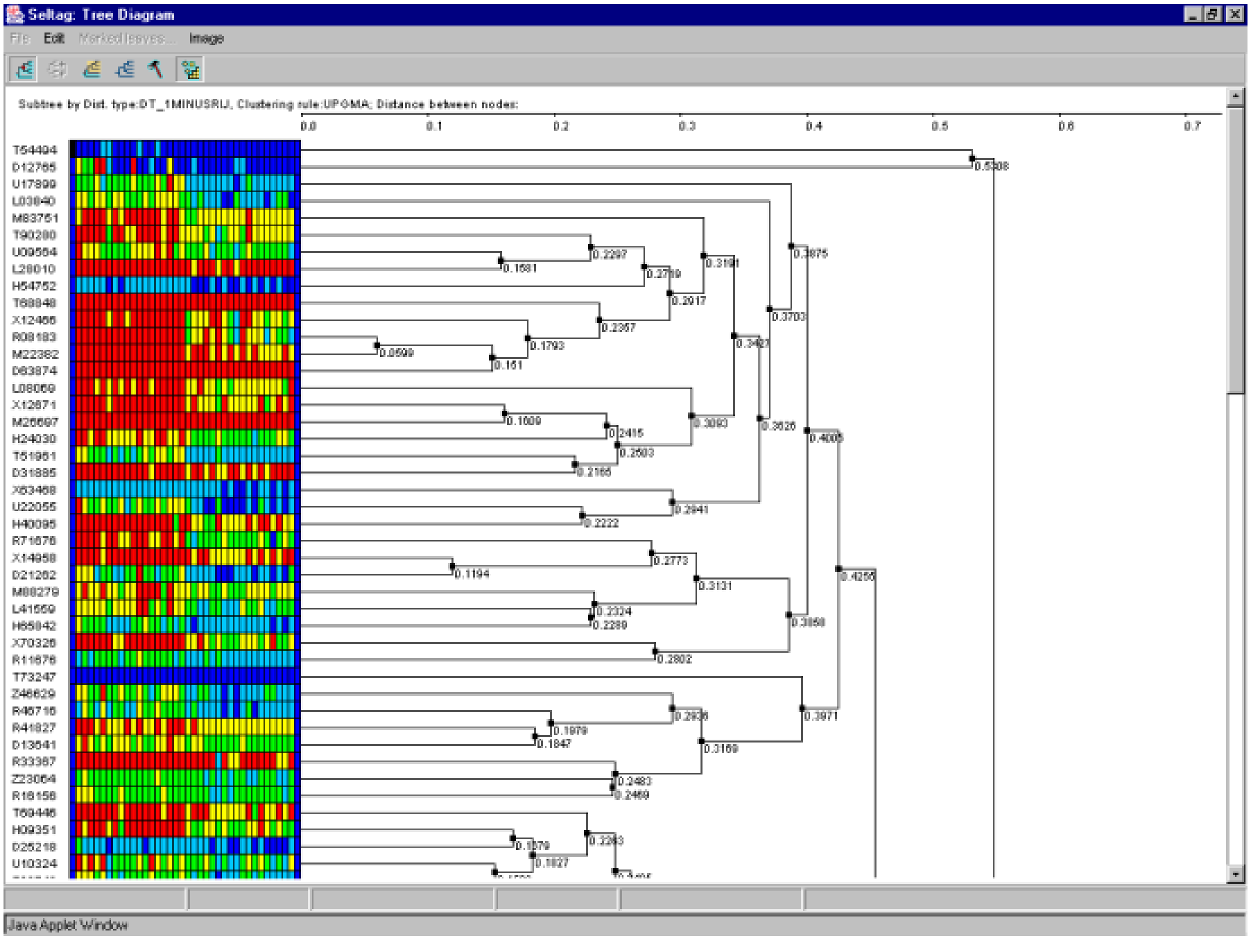
Dist. type:DT_1MINUSRU, Clustering rule:UPGMA; Distance between nodes:

Gene No: 3695 | Gene name: R34876 | Field No: 16 | Field name: Tumor_5 | Field value: 71.0

Figure 7.10.

1. Button for descent nodes selection mode. **2.** Selected node. **3.** Command for building the tree of selected cluster.

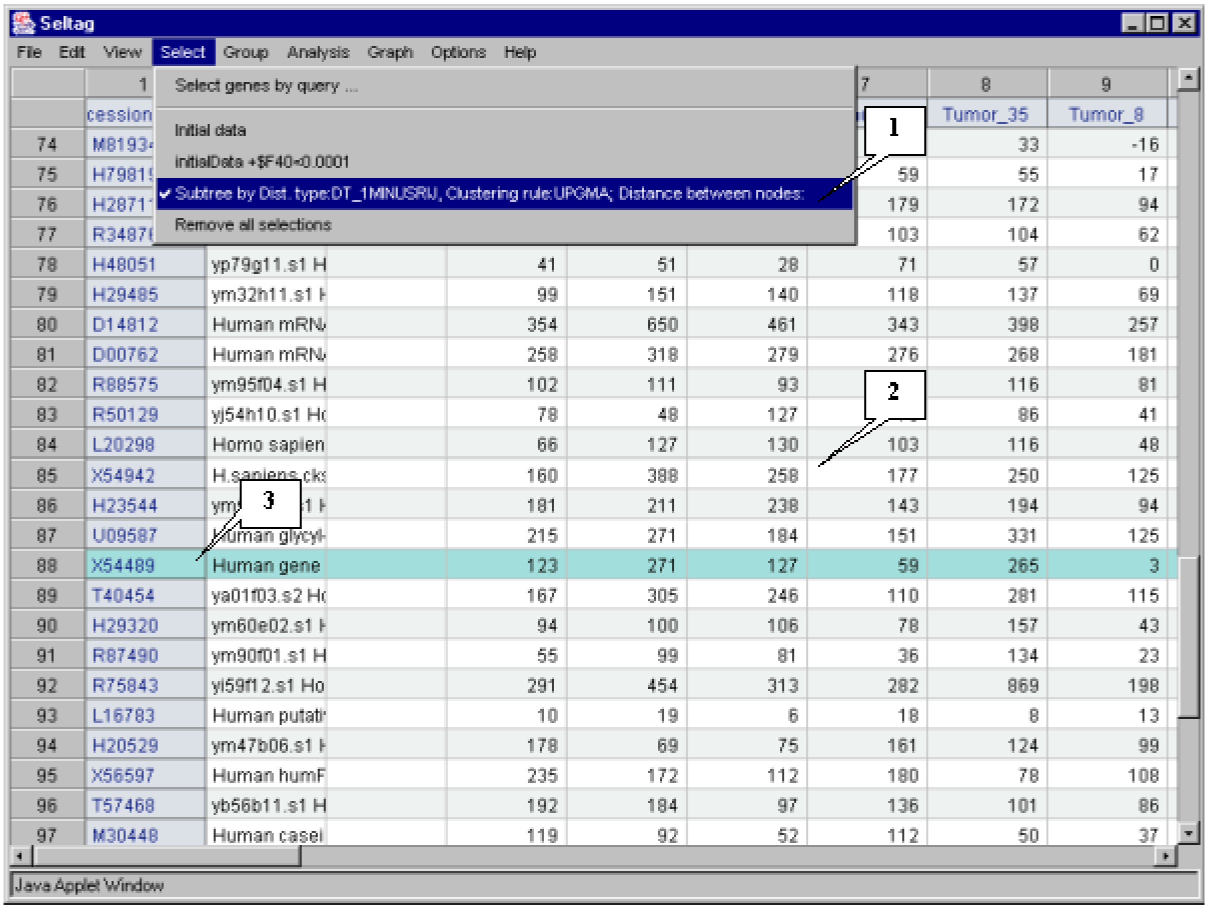
7.11. The window with tree and expression matrix for cluster 1 will appear (fig. 7.11). It is notable that color expression diagram is characterized by division of experiments (all columns except the first and the last ones) into two groups. The left part of diagram for the cluster of genes, shown on figure 7.11, has the higher expression level than the right one. In this case, the left part represents the expression values for the tumorous tissues. Such a division is a result of additional clustering of tissues simultaneously with that of genes (option 7 of panel on figure 7.5)



Subtree by Dist. type:DT_1MINUSRIJ, Clustering rule:UPGMA; Distance between nodes:

Figure 7.11.

7.12. The table with genes from the cluster 1 (fig. 7.12) will be represented in the main window. It is illustrative that genes X54489 and X54942 with high expression in tumorous tissues have been included in this table.

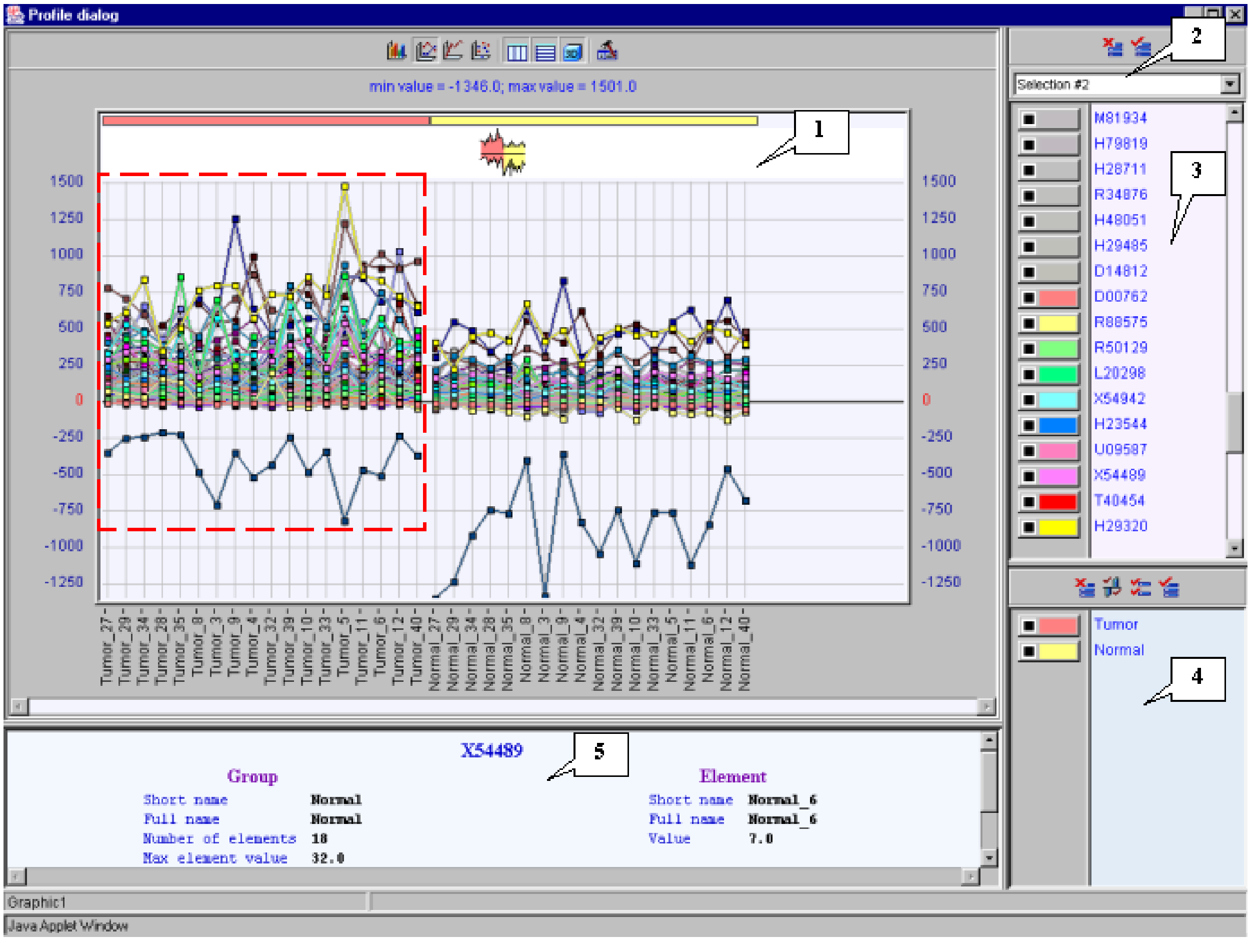


	1						7	8	9
	cession							Tumor_35	Tumor_8
74	M8193							33	-16
75	H7981						59	55	17
76	H2871						179	172	94
77	R3487						103	104	62
78	H48051	yp79g11.s1 H		41	51	28	71	57	0
79	H29485	ym32h11.s1 H		99	151	140	118	137	69
80	D14812	Human mRN		354	650	461	343	398	257
81	D00762	Human mRN		258	318	279	276	268	181
82	R88575	ym95f04.s1 H		102	111	93		116	81
83	R50129	yj54h10.s1 H		78	48	127		86	41
84	L20298	Homo sapien		66	127	130	103	116	48
85	X54942	H.sapiens cks		160	388	258	177	250	125
86	H23544	ym		181	211	238	143	194	94
87	U09587	uman glycyl		215	271	184	151	331	125
88	X54489	Human gene		123	271	127	59	265	3
89	T40454	ya01f03.s2 H		167	305	246	110	281	115
90	H29320	ym60e02.s1 H		94	100	106	78	157	43
91	R87490	ym90f01.s1 H		55	99	81	36	134	23
92	R75843	yi59f12.s1 Ho		291	454	313	282	869	198
93	L16783	Human putati		10	19	6	18	8	13
94	H20529	ym47b06.s1 H		178	69	75	161	124	99
95	X56597	Human humF		235	172	112	180	78	108
96	T57468	yb56b11.s1 H		192	184	97	136	101	86
97	M30448	Human casei		119	92	52	112	50	37

Figure 7.12.

1. Additional menu item corresponding to obtained set of genes. **2.** Table with selected genes. **3.** Selected gene.

7.13. In the main window choose the "Graph>Specified with current selection" command of the main menu. The “Profile Dialog” window with expression profiles of genes from the cluster 1 (fig. 7.13) will appear. It is illustrative that, for this cluster, the expression values in tumorous tissues (profiles inside the red rectangle) are higher than in normal ones (profiles outside the red rectangle).



M81934
H79819
H28711
R34876
H48051
H29485
D14812
D00762
R88575
R50129
L20298
X54942
H23544
U09587
X54489
T40454
H29320

Tumor
Normal

X54489

Group

Short name	Normal
Full name	Normal
Number of elements	18
Max element value	32.0

Element

Short name	Normal_6
Full name	Normal_6
Value	7.0

Figure 7.13.

1. Genes profiles visualization area. **2.** List of genes sets. **3.** List of genes. **4.** List of groups. **5.** Information on gene, profile of which is pointed by mouse.

7.14. Perform the steps described in 7.10-7.13 for the second cluster. The obtained tree, table of genes and profiles of genes for the cluster 2 are shown in the figures 7.14-7.16. It is illustrative that, for this cluster, the expression values in tumorous tissues (profiles inside the red rectangle) are lower than in normal ones (profiles outside the red rectangle) (fig. 7.16).

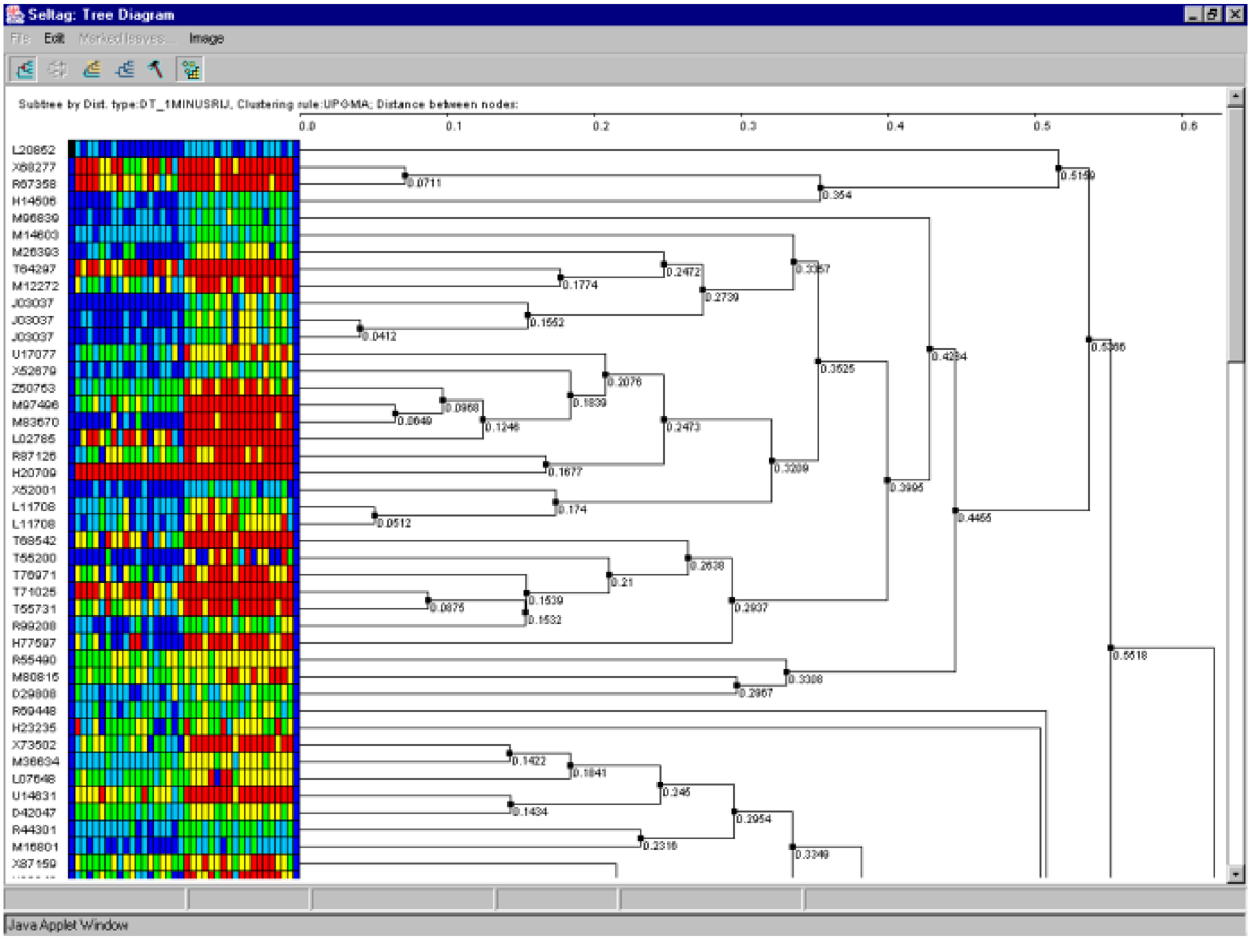
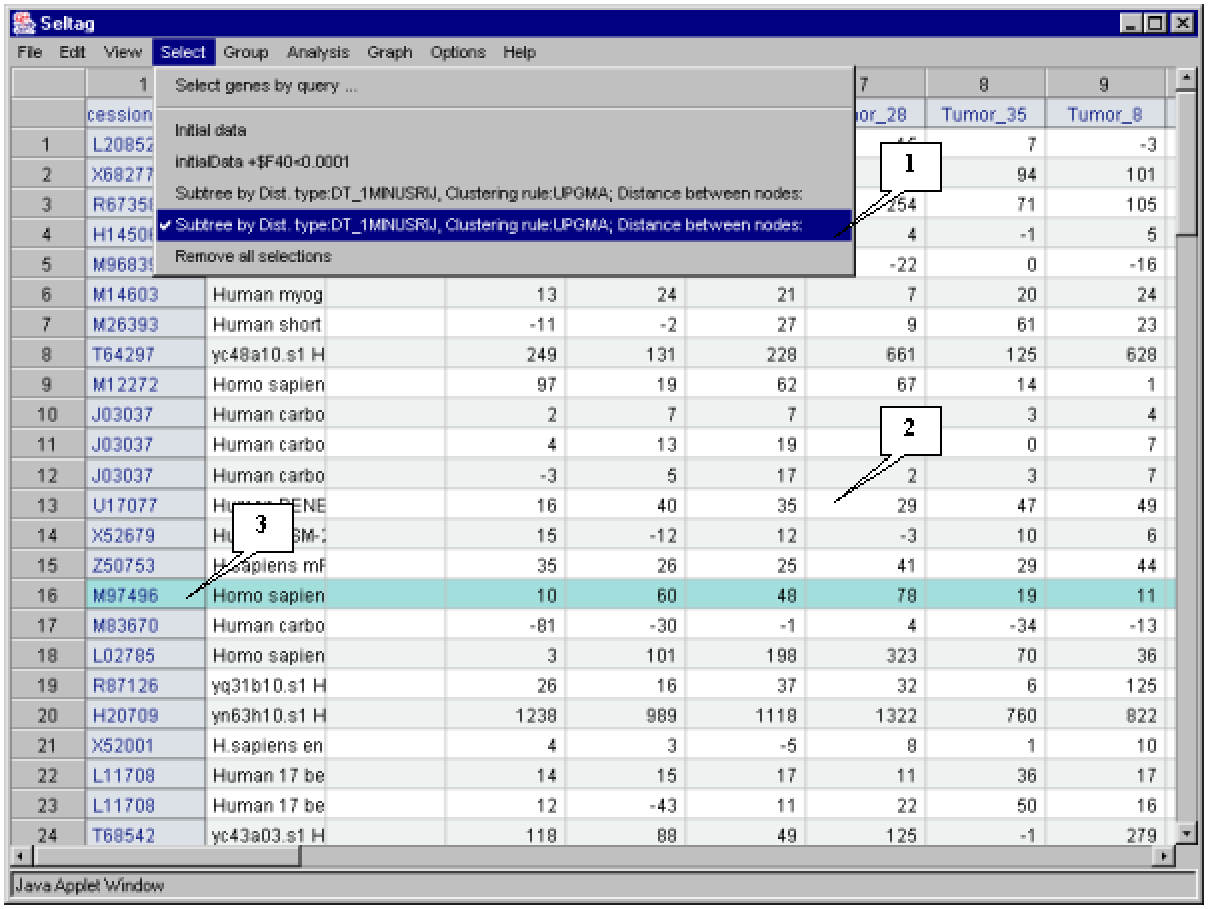

Subtree by Dist. type:DT_1MINUSRIJ, Clustering rule:UPGMA; Distance between nodes:

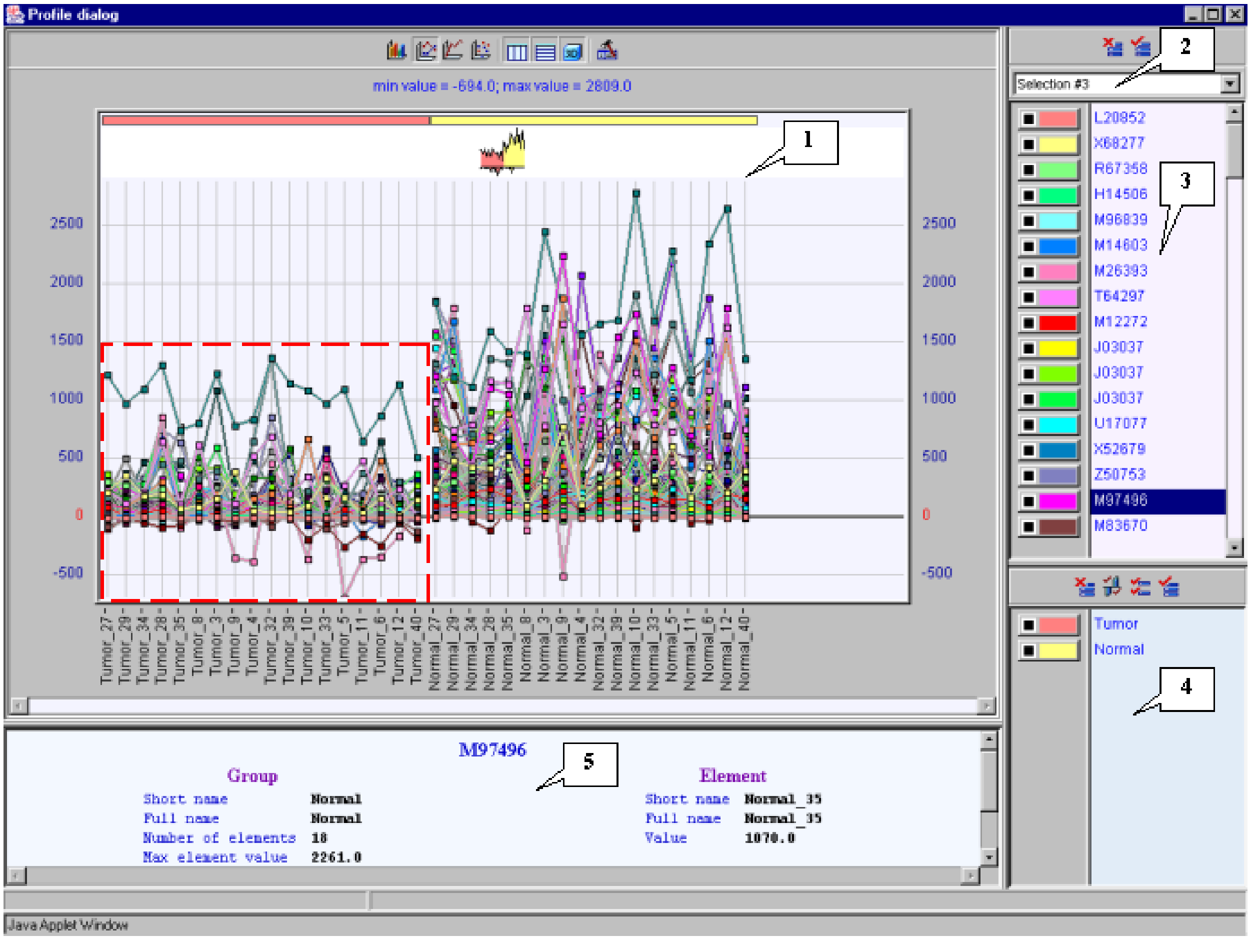
Figure 7.14.


Seltag
File
Edit
View
Select
Group
Analysis
Graph
Options
Help
Select genes by query ...
Initial data
initialData +$F40<0.0001
Subtree by Dist. type:DT_1MINUSRIJ, Clustering rule:UPGMA; Distance between nodes:
✓ Subtree by Dist. type:DT_1MINUSRIJ, Clustering rule:UPGMA; Distance between nodes:
Remove all selections
1
2
3
Java Applet Window

	1						7	8	9
	cession						or_28	Tumor_35	Tumor_8
1	L20852							7	-3
2	X68277							94	101
3	R6735						254	71	105
4	H1450						4	-1	5
5	M9683						-22	0	-16
6	M14603	Human myog		13	24	21	7	20	24
7	M26393	Human short		-11	-2	27	9	61	23
8	T64297	yc48a10.s1 H		249	131	228	661	125	628
9	M12272	Homo sapien		97	19	62	67	14	1
10	J03037	Human carbo		2	7	7		3	4
11	J03037	Human carbo		4	13	19		0	7
12	J03037	Human carbo		-3	5	17	2	3	7
13	U17077	Hu ENE		16	40	35	29	47	49
14	X52679	Hu SM-		15	-12	12	-3	10	6
15	Z50753	H.sapiens mF		35	26	25	41	29	44
16	M97496	Homo sapien		10	60	48	78	19	11
17	M83670	Human carbo		-81	-30	-1	4	-34	-13
18	L02785	Homo sapien		3	101	198	323	70	36
19	R87126	yq31b10.s1 H		26	16	37	32	6	125
20	H20709	yn63h10.s1 H		1238	989	1118	1322	760	822
21	X52001	H.sapiens en		4	3	-5	8	1	10
22	L11708	Human 17 be		14	15	17	11	36	17
23	L11708	Human 17 be		12	-43	11	22	50	16
24	T68542	yc43a03.s1 H		118	88	49	125	-1	279

Figure 7.15.

1. Additional menu item corresponding to obtained set of genes. **2.** Table with selected genes**. 3.** Selected gene.



🟥	L20852
🟨	X68277
🟩	R67358
🟩	H14506
🟦	M96839
🟦	M14603
🟪	M26393
🟪	T64297
🟥	M12272
🟨	J03037
🟩	J03037
🟩	J03037
🟦	U17077
🟦	X52679
🟪	Z50753
🟪	M97496
🟫	M83670

🟥	Tumor
🟨	Normal

M97496

Group		Element	
Short name	Normal	Short name	Normal_35
Full name	Normal	Full name	Normal_35
Number of elements	18	Value	1070.0
Max element value	2261.0		

Figure 7.16.

1. Genes profiles visualization area**. 2.** List of genes sets**. 3.** List of genes**. 4.** List of groups**. 5.** Information on gene, profile of which is pointed by mouse**.**

Thus, the hierarchical clustering analysis has allowed to select two clusters of genes that differ in their relative expression values in tumorous and normal tissues.

8. Principal components analysis.

In the current chapter, the usage of principal components [3] in analysis of 240 selected genes expression is described. During the analysis, the expression table is being represented as a cloud of dots in multidimensional space. Each coordinate of this space represents the expression in appropriate tissue (experiment), and genes are represented as dots, location of which is defined by the expression values in experiments set. In this space, the set of axes, number of which is equal to that of experiments, and which are mutually orthogonal, is being calculated. Moreover, dispersion of dots along the first axis should be absolutely maximal, the similar dispersion along the second axis should be maximal among the remaining values, etc. These axes are referred to as components. The values of dots dispersion along these axes are characterized by sets of eigenvalues for components. Directions of these axes in the space of experiments are referred to as loadings. The first *k* of components that have the maximal dots dispersion are referred to as *k* principal components. To visualize data, most commonly use *k*=2, and consider the plot of dots' projection onto plane of the first two components. Such a plot allows to illustrate the location of dots in multidimensional space as well as to reveal existing clusters of genes.

To perform this task, the following steps are required:

8.1. Use the “Analysis>Principal component” command from the main menu (fig. 8.1).

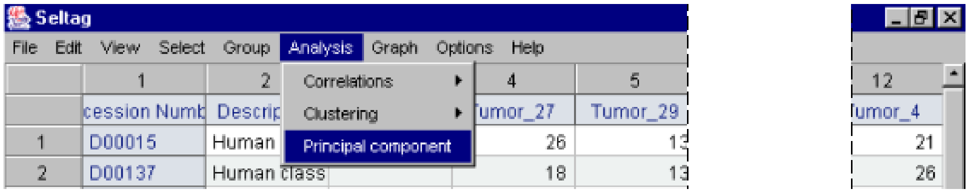
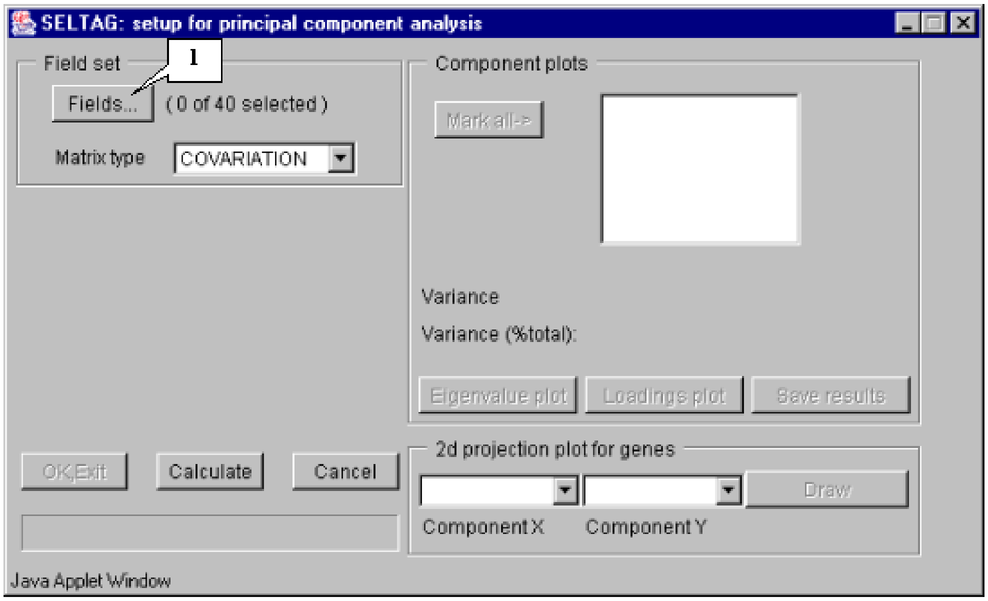


Figure 8.1.

8.2. The “Setup for principal component analysis” dialog (fig. 8.2) will appear. Press the “Fields” button (fig. 8.2).


SELTAG: setup for principal component analysis

Field set

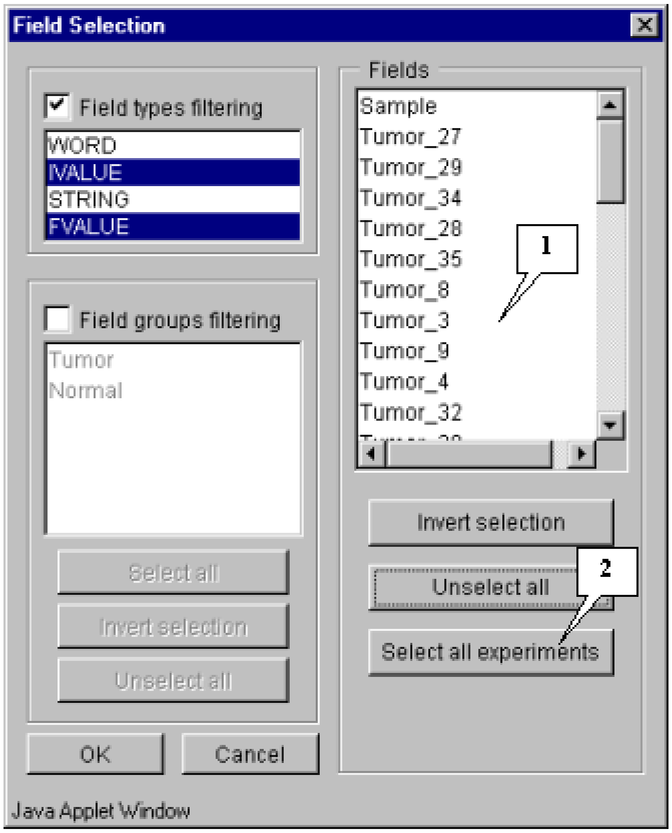
Fields... (0 of 40 selected)

Matrix type COVARIATION

Figure 8.2.

1. Button for call the “Field selection” dialog.

8.3. The “Field selection” dialog (fig. 8.3) that provides the choice of fields will appear.



Field Selection

☑ Field types filtering

WORD
IVALUE
STRING
FVALUE

□ Field groups filtering

Tumor
Normal

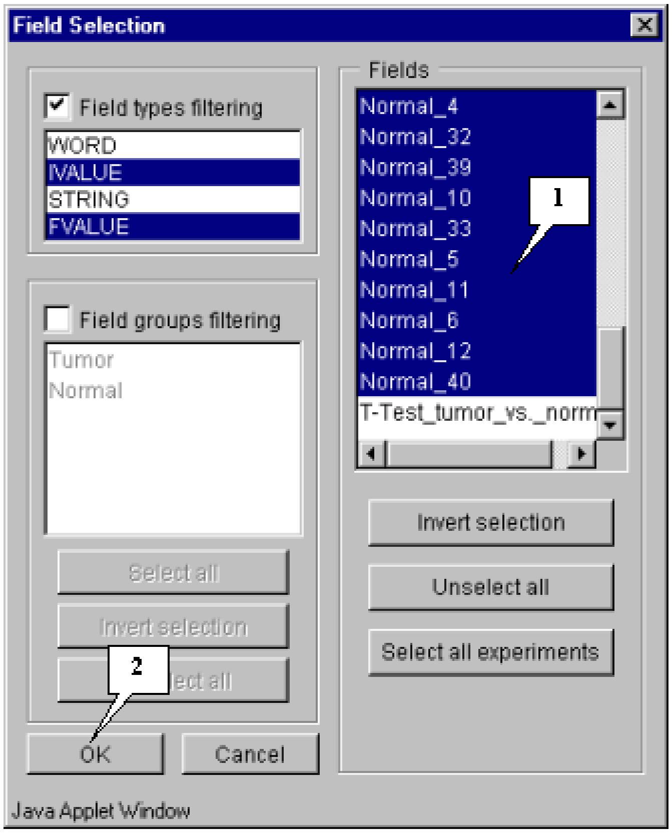
Fields

Sample
Tumor_27
Tumor_29
Tumor_34
Tumor_28
Tumor_35
Tumor_8
Tumor_3
Tumor_9
Tumor_4
Tumor_32

Figure 8.3.

1. List of fields. **2.** Button for selection of all fields that are expression data.

8.4. In this example, all fields, except the “Sample” and “T-Test_tumor_vs_Normal” ones, are used for calculations. Press the “Select all experiments” button (fig. 8.3) to select all fields and then remove selection from the appropriate ones (“Sample” and “T-Test_tumor_vs_Normal”) (fig. 8.4).



Field Selection

☑ Field types filtering

WORD
IVALUE
STRING
FVALUE

□ Field groups filtering

Tumor
Normal

Fields

Normal_4
Normal_32
Normal_39
Normal_10
Normal_33
Normal_5
Normal_11
Normal_6
Normal_12
Normal_40
T-Test_tumor_vs._norm

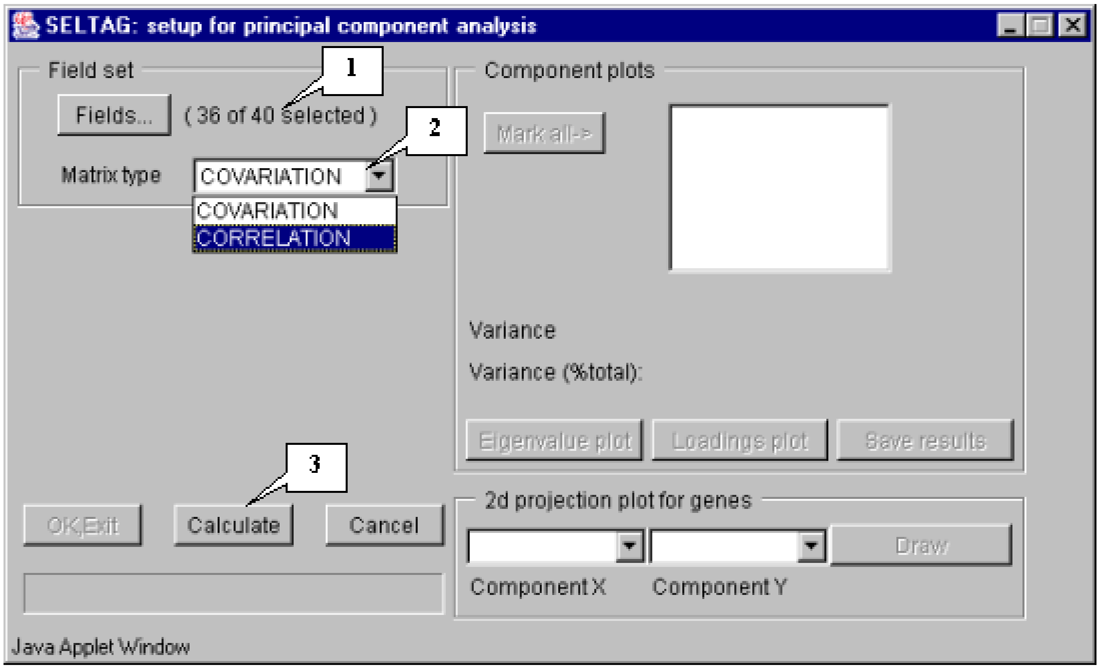
Figure 8.4.

1. Selected fields. **2.** Confirmation button.

8.5. Press the “OK” button. In the “Setup for principal component analysis” dialog, in the area near the “Fields” button, the information on the number of selected fields will be represented (fig. 8.5.1). After this, do the following:

- Choose the type of matrix from the “Matrix type” list. In this example, the correlation matrix is used.
- Press the “Calculate” button.

The “Wait” message box (fig. 8.5.2) will appear for the duration of calculation process.



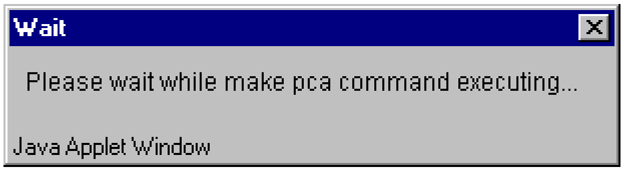
SELTAG: setup for principal component analysis

Field set: (36 of 40 selected)

Matrix type: COVARIATION

Figure 8.5.1.

1. Information on the number of selected fields. **2.** List of matrix types. **3.** Calculation start button.

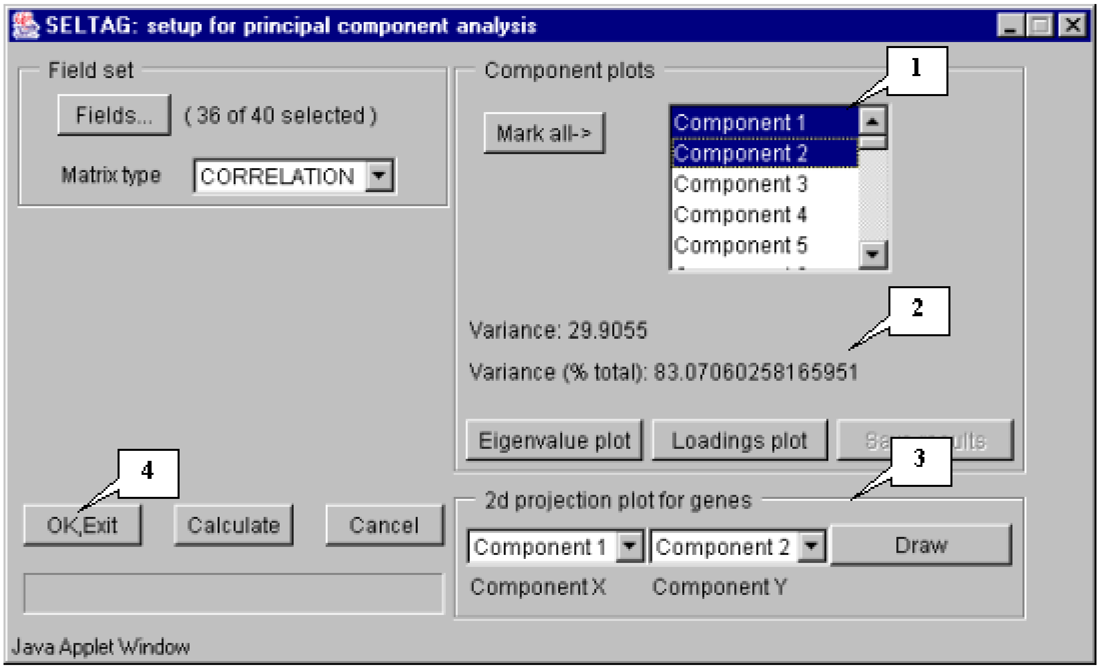


Wait

Please wait while make pca command executing...

Figure 8.5.2.

8.6. In the “Setup for principal component analysis” window, in the «Component plots» list, components (eigenvectors), numbered in descending eigenvalues order, will appear (fig. 8.6).



SELTAG: setup for principal component analysis

Field set

(36 of 40 selected)

Matrix type: CORRELATION

Component plots

Component 1
Component 2
Component 3
Component 4
Component 5

Variance: 29.9055

Variance (% total): 83.07060258165951

2d projection plot for genes

Component X: Component 1

Component Y: Component 2

Java Applet Window

Figure 8.6.

1. List of components. **2.** String of eigenvalues summation for selected in the «Component plots» list items. **3.** Projection building settings. **4.** Confirmation and exit button.

8.7. Choose the first and second components in the “2d projection plot for genes” section and press the “Draw” button (fig. 8.6).

8.8. The “Graph” dialog with plot of genes distribution in the space of 2 principal components (fig. 8.8) will appear. On the plot, the two clusters along the ordinate axis can be selected. To select one of them, click the right mouse button as shown in the figure 8.8.

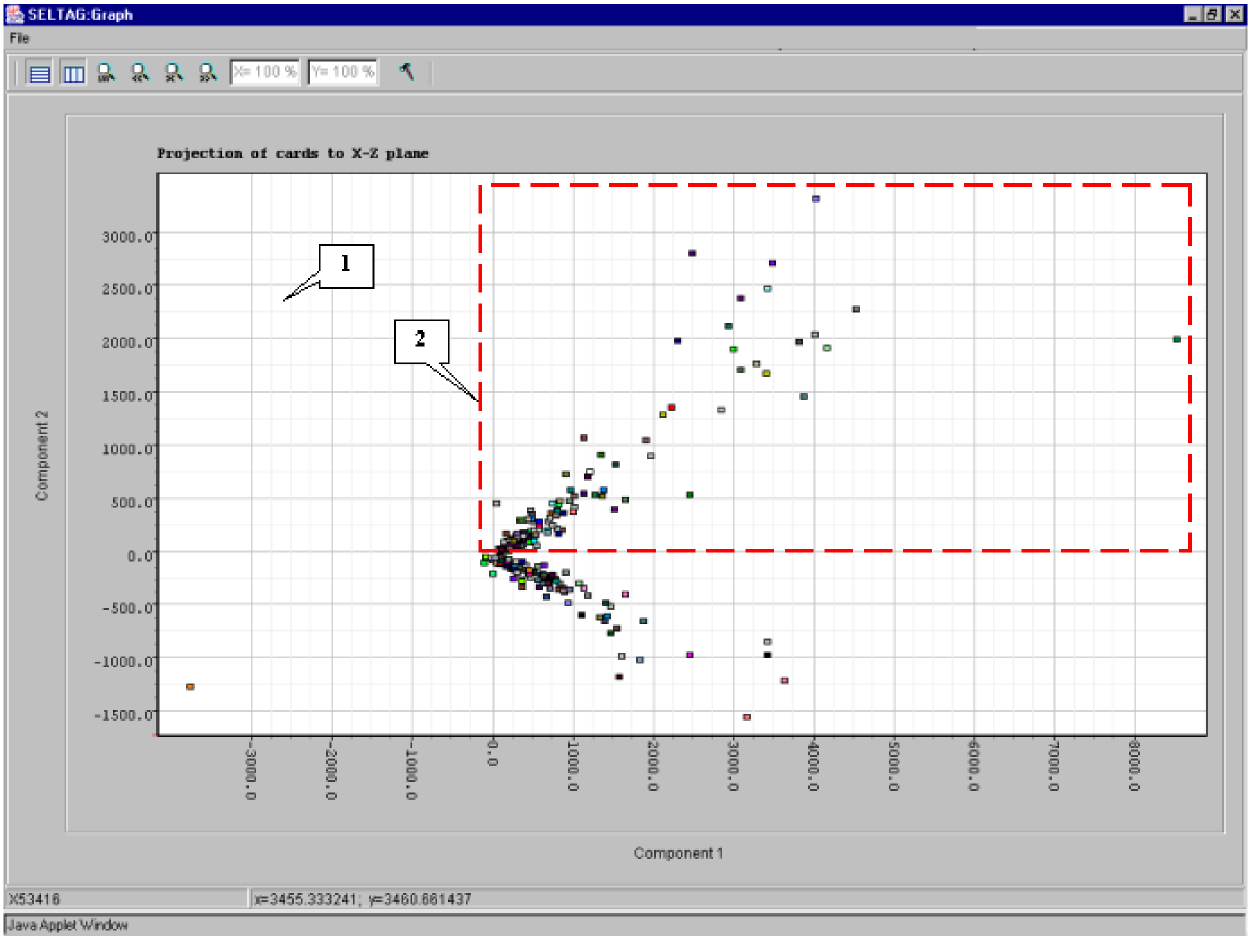


Figure 8.8.

1. Plot area. **2.** Selected cluster.

8.9. The “Operations with marked items” dialog that allows to operate the selected objects will appear. The list contains the genes of selected cluster. In order to change the color of plot objects, select all genes in the list by pressing the “<-Select all” button (fig. 8.9).

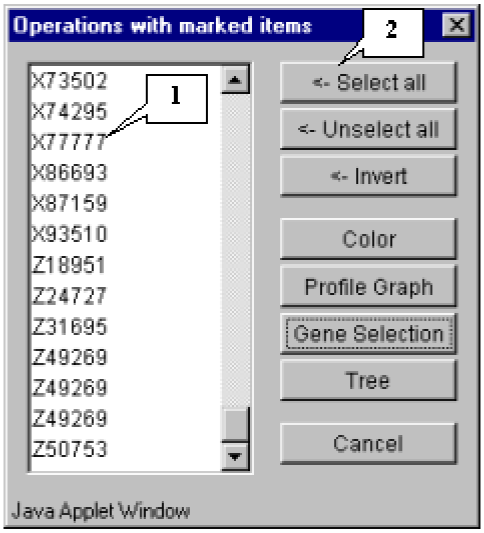


Figure 8.9.

1. List of genes that were included in the selected zone on the PCA plot. **2.** The “<-Select all” button.

8.10. All genes in the list will become selected (fig. 8.10). Press the “Color” button.

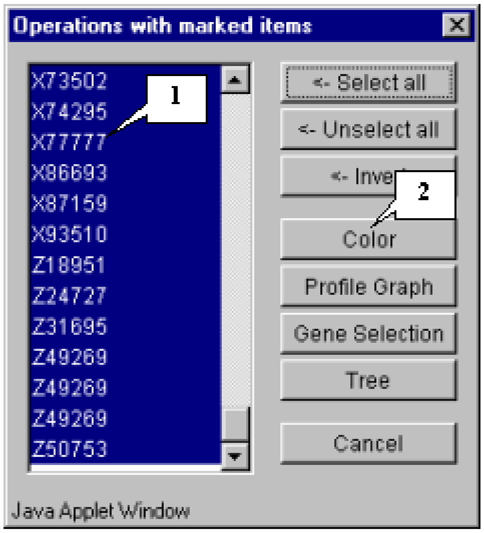
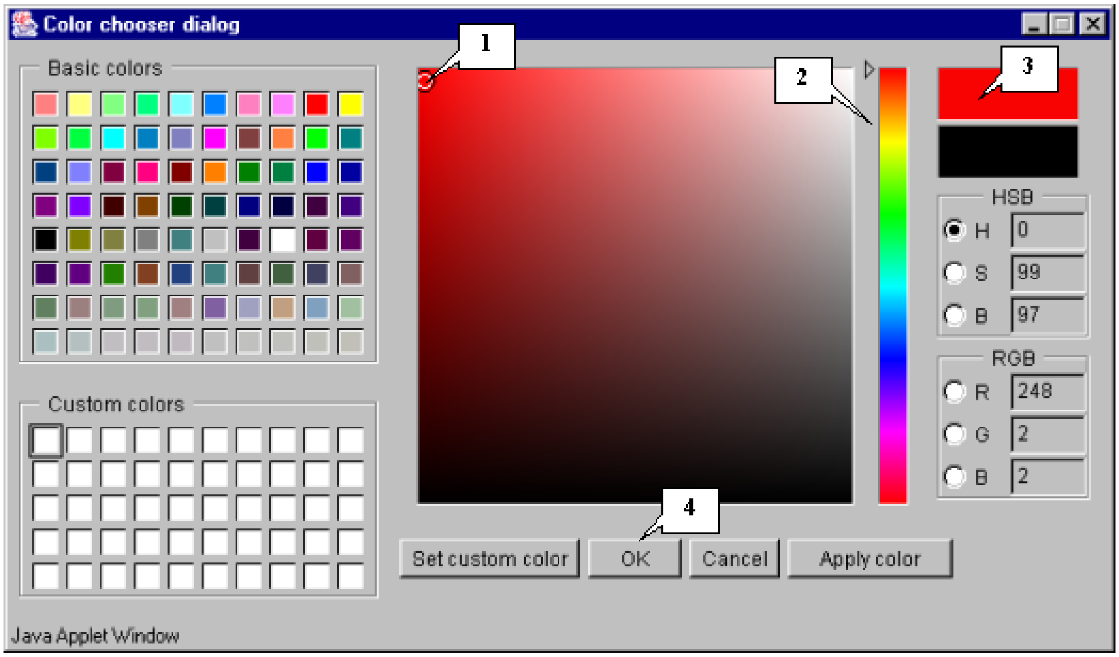


Figure 8.10.

1. Selected genes. **2.** Button for color selection dialog.

8.11. The “Color chooser dialog” window (fig. 8.11) that provides an object color selection will appear. Once selection is done, press the “OK” button.



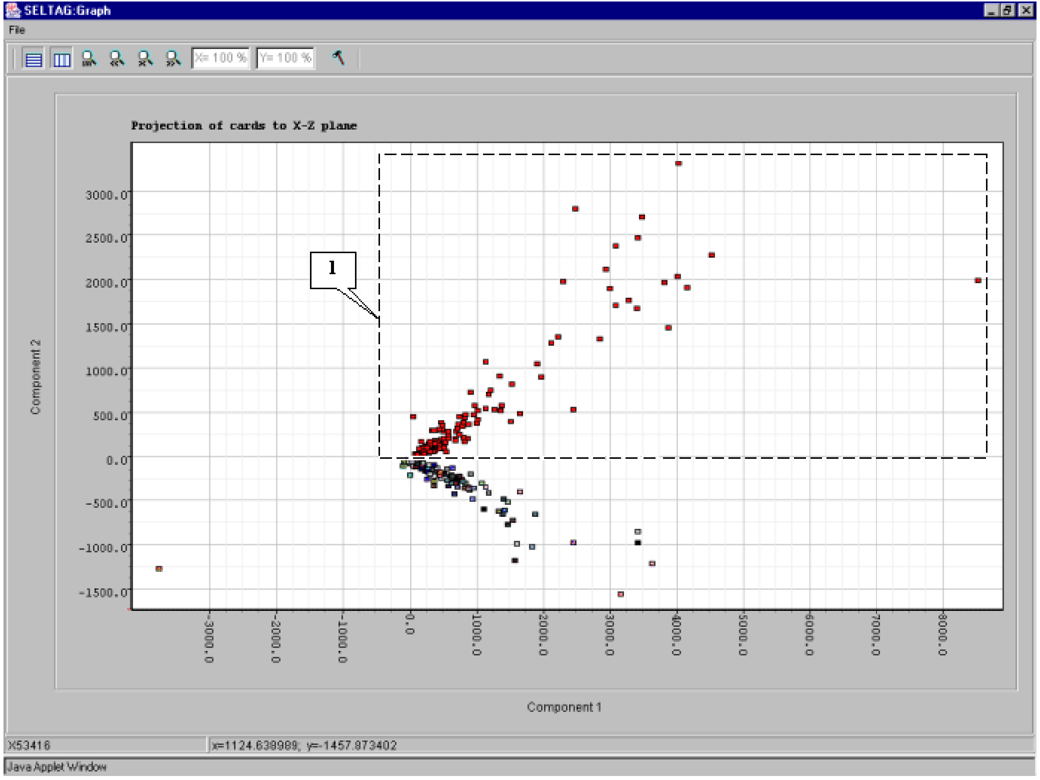
HSB
● H 0
○ S 99
○ B 97

RGB
○ R 248
○ G 2
○ B 2

Figure 8.11.

1. Color selector. **2.** Color line. **3.** New color preview area. **4.** Confirmation and exit button.

8.12. The cluster will be painted in the selected color (fig. 8.13).



X53416

x=1124.638989; y=-1457.873402

Figure 8.12.

8.13. To change the color of second cluster, repeat actions 8.8-8.11. The result is shown in fig. 8.13: the red cluster will reffered to as cluster 1, and the blue one - cluster 2.

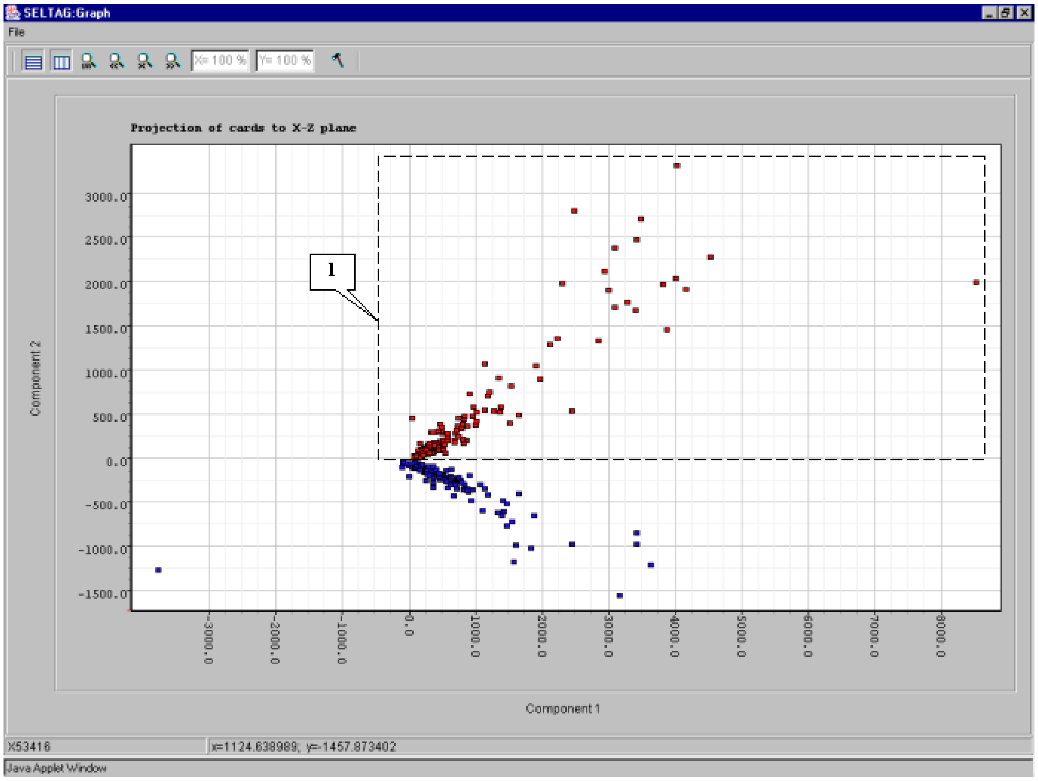
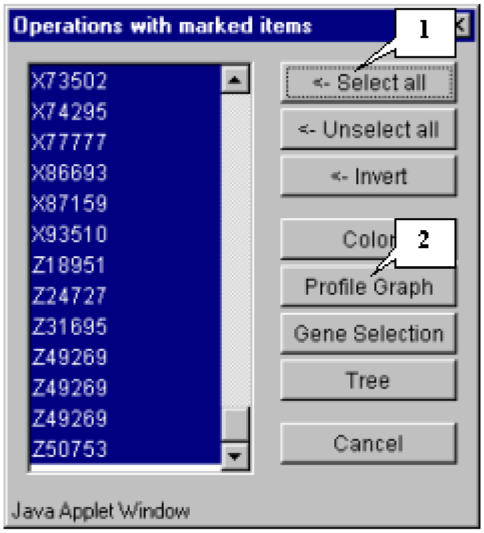


Figure 8.13.

Red color of the cluster 1. Blue color of the cluster 2.

8.14. To get the diagram of cluster 1 genes' profiles, select the cluster by drawing a rectangle at hold mouse right button as shown in figure 8.13. The “Operations with marked items” dialog will appear. Select all genes by pressing the “<-Select all” button and then press the “Profile Graph” button (fig. 8.14).

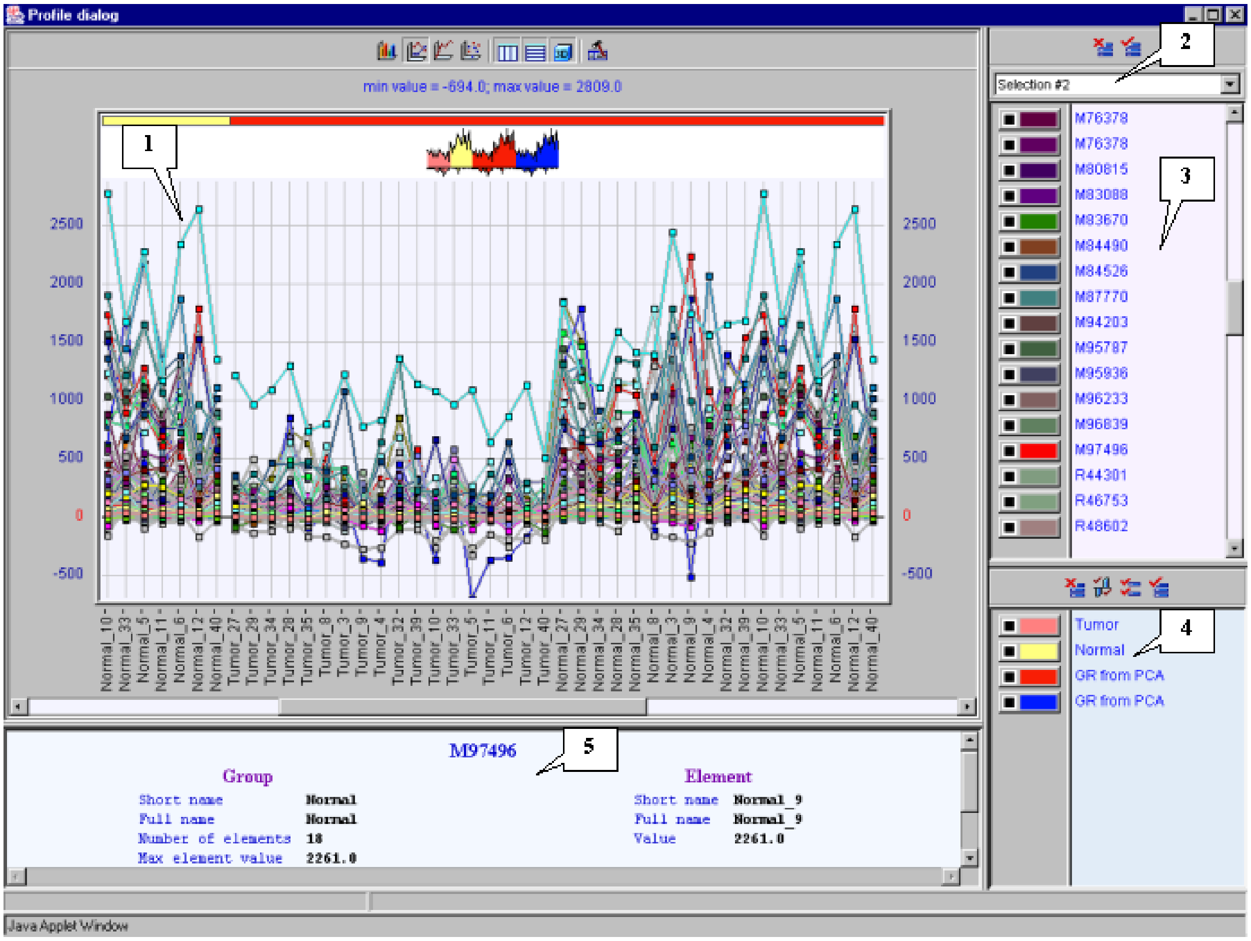


X73502
X74295
X77777
X86693
X87159
X93510
Z18951
Z24727
Z31695
Z49269
Z49269
Z49269
Z50753

Figure 8.14.

1. Select all genes button. **2.** Button for the “Graph” dialog with profiles of selected genes.

8.15. The dialog with profiles of genes in cluster 1 (fig. 8.15) will appear. It is illustrative, that genes in cluster 1 have the lower expression level in tumorous tissues than in normal ones.



M97496

Group

Short name Normal
Full name Normal
Number of elements 18
Max element value 2261.0

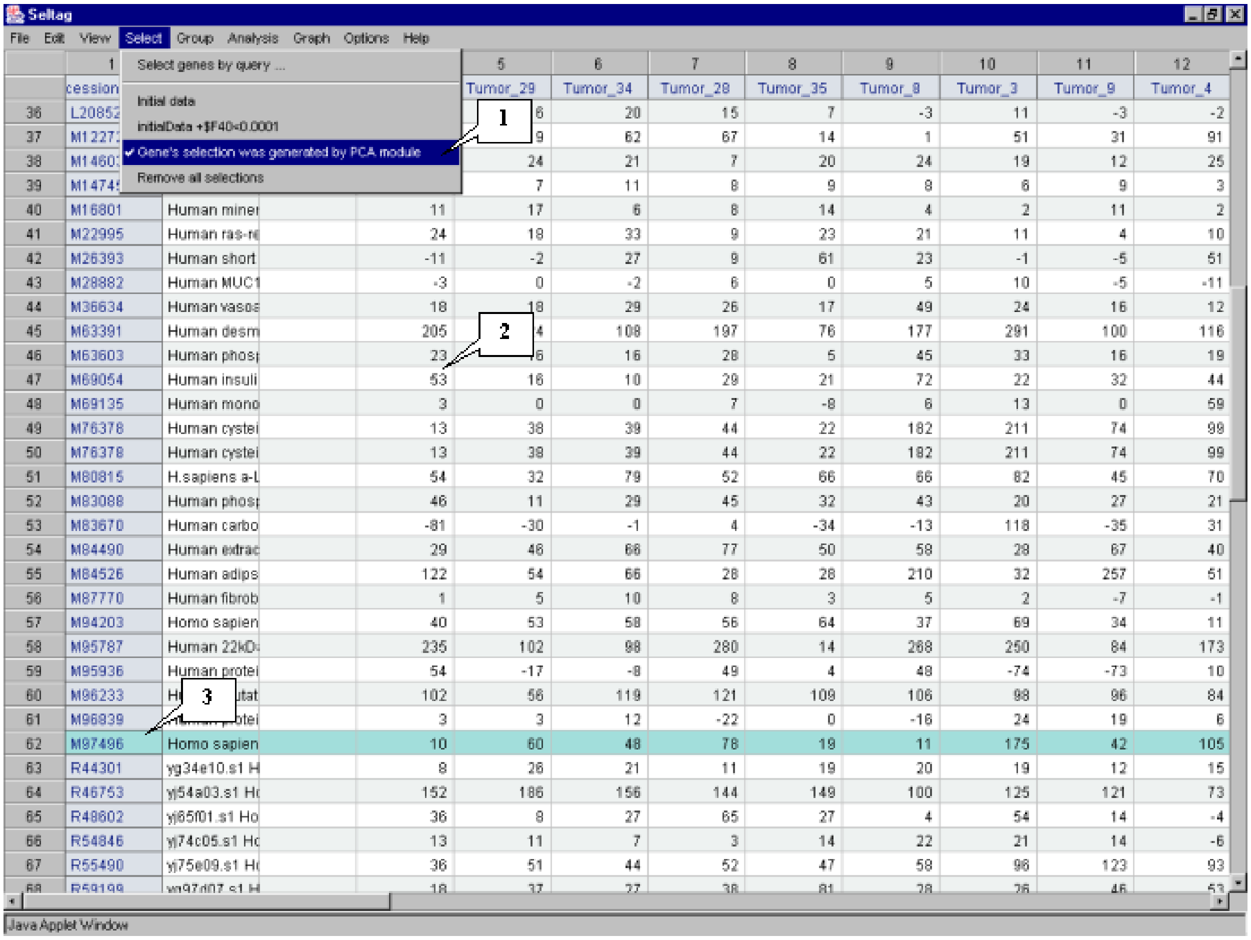
Element

Short name Normal_9
Full name Normal_9
Value 2261.0

Figure 8.15.

1. Genes profiles diagram area. **2.** List of genes sets. **3.** List of genes. **4.** List of groups. **5.** Information on gene, profile of which is pointed by mouse.

8.16. In the main window, the table with genes from cluster 1 (fig. 8.16) will be represented.

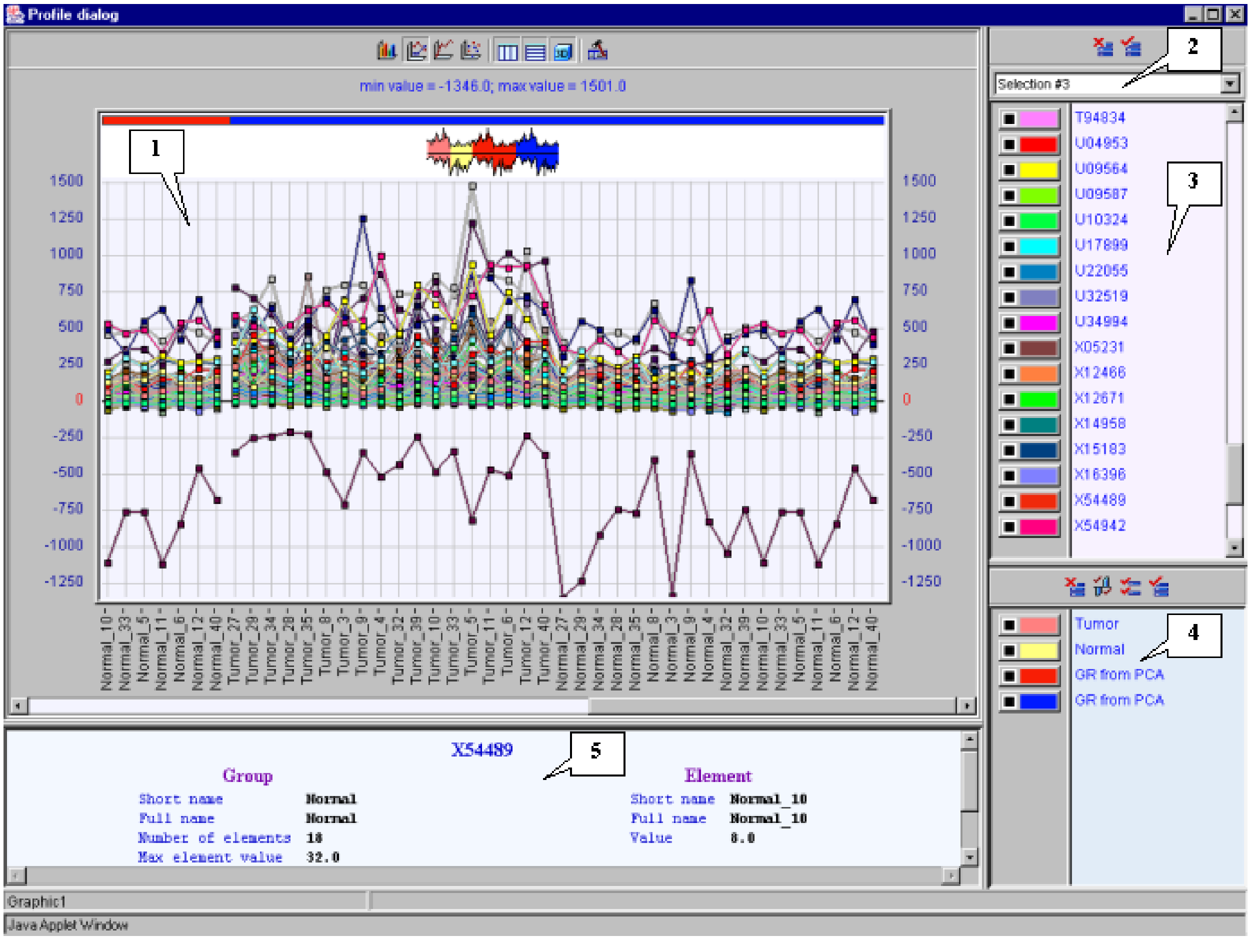


	1				5	6	7	8	9	10	11	12
	cession				Tumor_29	Tumor_34	Tumor_28	Tumor_35	Tumor_8	Tumor_3	Tumor_9	Tumor_4
36	L20852				6	20	15	7	-3	11	-3	-2
37	M1227				9	62	67	14	1	51	31	91
38	M1460				24	21	7	20	24	19	12	25
39	M1474				7	11	8	9	8	6	9	3
40	M16801	Human miner		11	17	6	8	14	4	2	11	2
41	M22995	Human ras-re		24	18	33	9	23	21	11	4	10
42	M26393	Human short		-11	-2	27	9	61	23	-1	-5	51
43	M28882	Human MUC1		-3	0	-2	6	0	5	10	-5	-11
44	M36634	Human vasoa		18	18	29	26	17	49	24	16	12
45	M63391	Human desm		205	4	108	197	76	177	291	100	116
46	M63603	Human phosp		23	6	16	28	5	45	33	16	19
47	M69054	Human insuli		53	16	10	29	21	72	22	32	44
48	M69135	Human mono		3	0	0	7	-8	6	13	0	59
49	M76378	Human cystei		13	38	39	44	22	182	211	74	99
50	M76378	Human cystei		13	38	39	44	22	182	211	74	99
51	M80815	H.sapiens a-L		54	32	79	52	66	66	82	45	70
52	M83088	Human phosp		46	11	29	45	32	43	20	27	21
53	M83670	Human carbo		-81	-30	-1	4	-34	-13	118	-35	31
54	M84490	Human extrac		29	46	66	77	50	58	28	67	40
55	M84526	Human adips		122	54	66	28	28	210	32	257	51
56	M87770	Human fibrob		1	5	10	8	3	5	2	-7	-1
57	M94203	Homo sapien		40	53	58	56	64	37	69	34	11
58	M95787	Human 22kD		235	102	98	280	14	268	250	84	173
59	M95936	Human protei		54	-17	-8	49	4	48	-74	-73	10
60	M96233	Hu utat		102	56	119	121	109	106	98	96	84
61	M96839	protei		3	3	12	-22	0	-16	24	19	6
62	M97496	Homo sapien		10	60	48	78	19	11	175	42	105
63	R44301	yg34e10.s1 H		8	26	21	11	19	20	19	12	15
64	R46753	yj54a03.s1 Ho		152	186	156	144	149	100	125	121	73
65	R48602	yj65f01.s1 Ho		36	8	27	65	27	4	54	14	-4
66	R54846	yj74c05.s1 Ho		13	11	7	3	14	22	21	14	-6
67	R55490	yj75e09.s1 Ho		36	51	44	52	47	58	98	123	93
68	R59199	yg97d07.s1 H		18	37	27	38	81	78	26	46	53

Figure 8.16.

1. Additional menu item corresponding to obtained set of genes. **2.** Table with selected genes. **3.** Selected gene.

8.17. On building the expression profiles for cluster 2 (blue cluster on fig. 8.13), it is illustrative that genes of this cluster are overexpressed in tumorous tissues if compared with normal ones (fig. 8.17). Thus, the second component represnts the relative expression of genes in tumorous/normal tissues. If the projection is positive, it means the gene is expressed presumably in normal tissues, if negative - in tumorous ones.



X54489

Group

Short name Normal
Full name Normal
Number of elements 18
Max element value 32.0

Element

Short name Normal_10
Full name Normal_10
Value 8.0

Figure 8.17.

1. Genes profiles diagram area. **2.** List of genes sets. **3.** List of genes. **4.** List of groups. **5.** Information on gene, profile of which is pointed by mouse.

8.18. To analyse the genes distribution along the abscissa axis, i.e the first component, select the genes with minimal and maximal values of the component as shown in figure 8.18. Build the expression profiles plot (described in 8.14-8.15) for the selected genes.

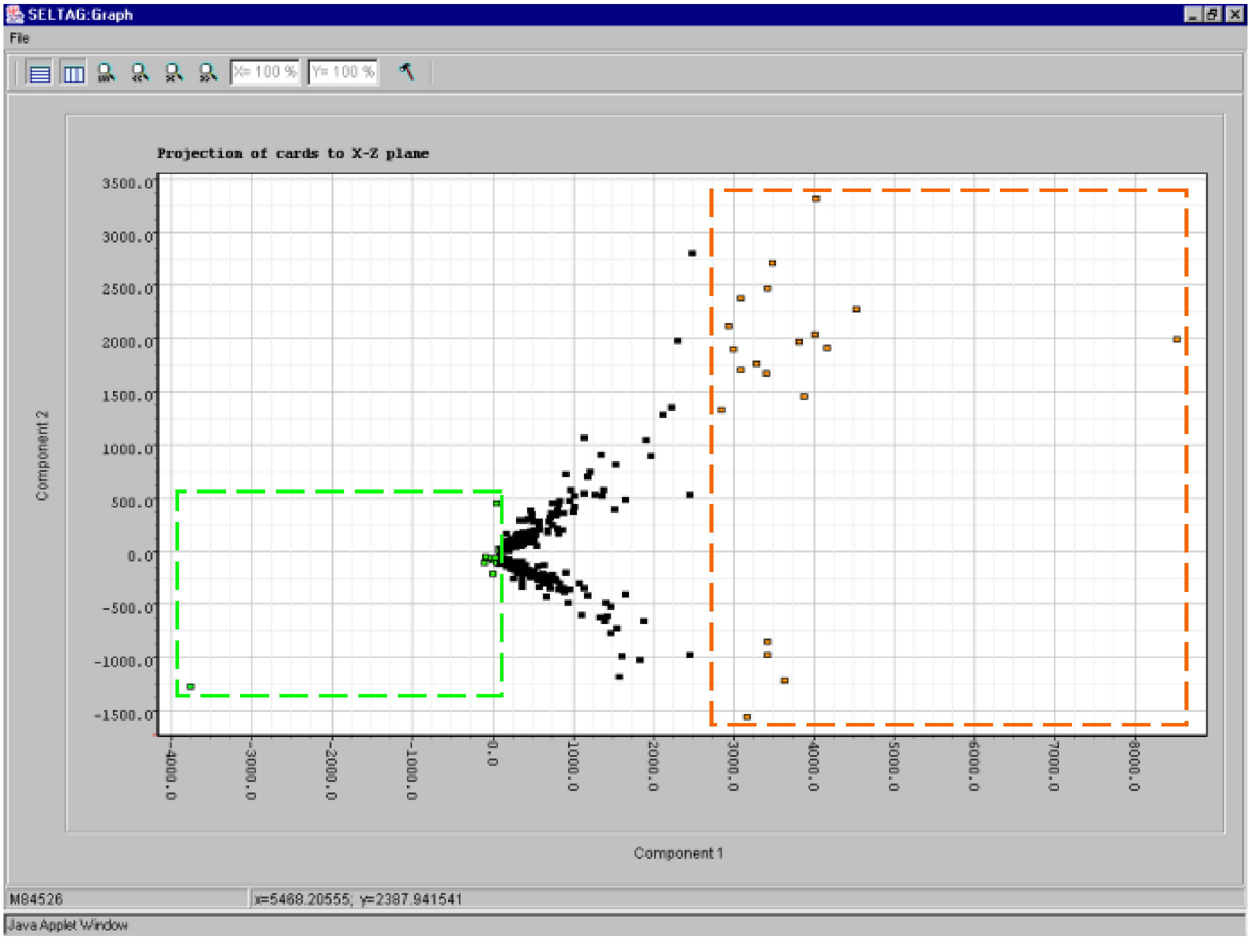
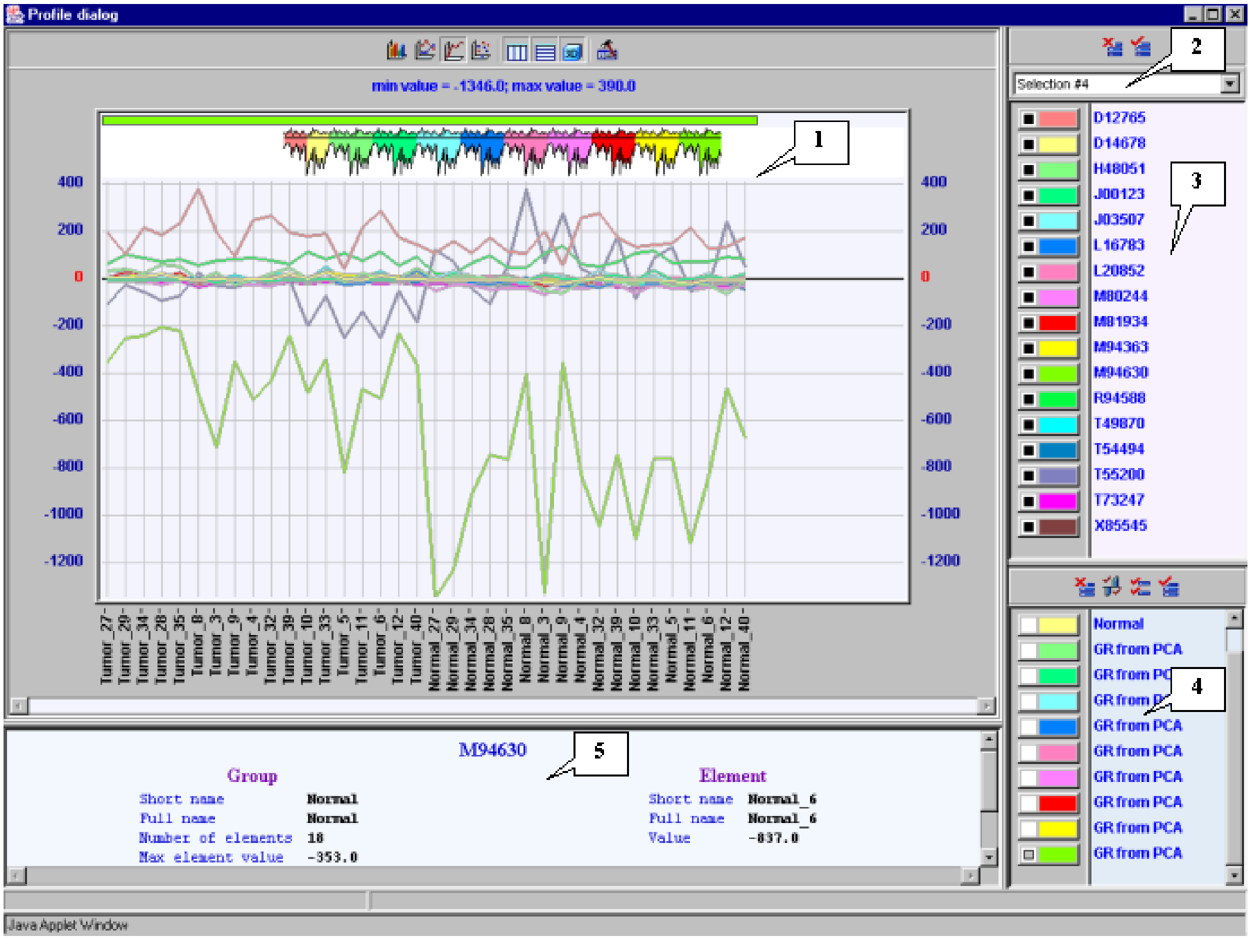


Figure 8.18.

8.19. Expression profiles plot for genes marked with green in figure 8.18 is shown on fig. 8.19.1. That for genes marked with orange in figure 8.18 is shown on fig. 8.19.2. When compared, it is illustrative, that genes of green cluster have the total expression level lower than that of orange one.



M94630

Group

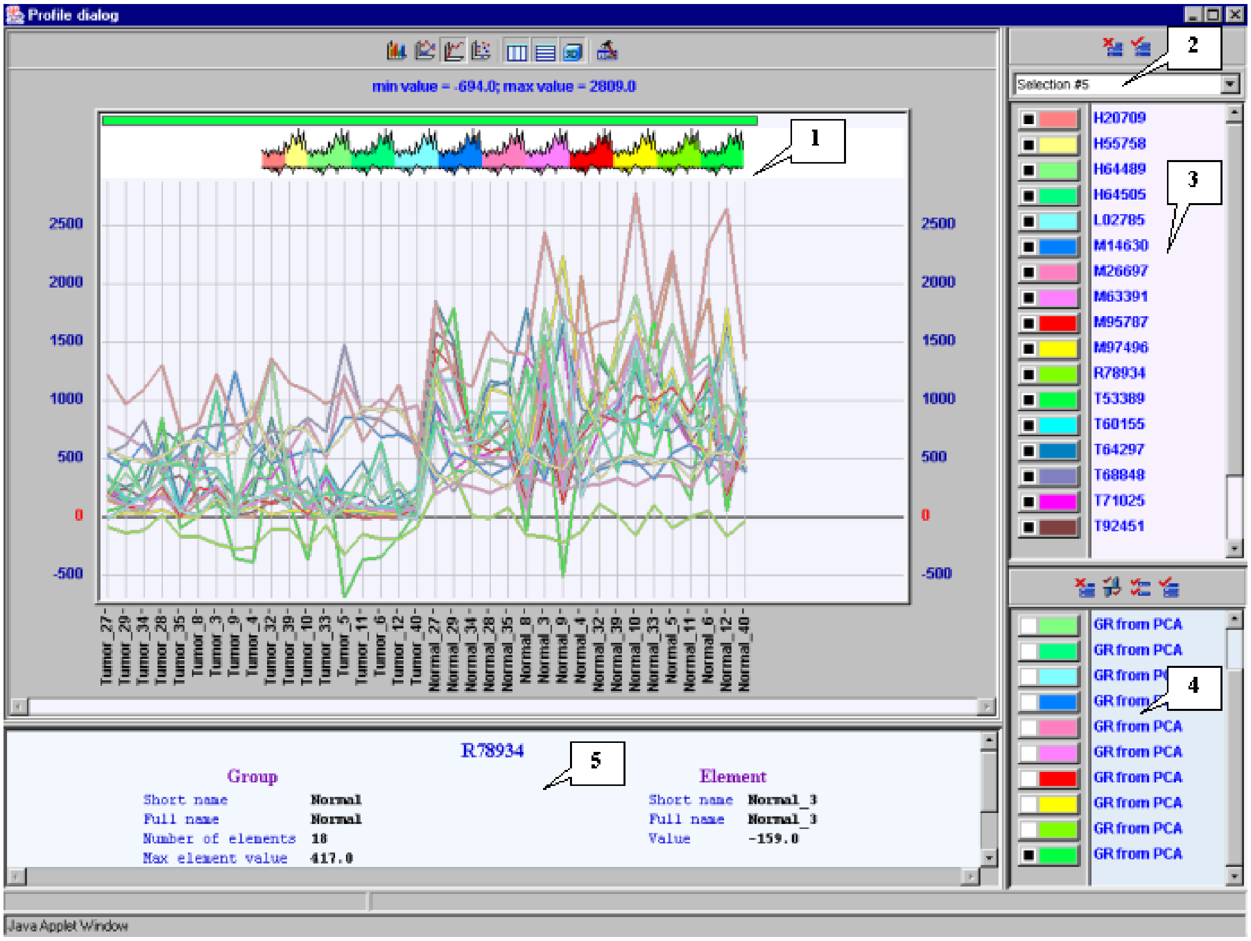
```
Short name          Normal
Full name           Normal
Number of elements  18
Max element value   -353.0
```

Element

```
Short name  Normal_6
Full name   Normal_6
Value       -837.0
```

Figure 8.19.1.

1. Genes profiles diagram area. **2.** List of genes sets. **3.** List of genes. **4.** List of groups. **5.** Information on gene, profile of which is pointed by mouse.



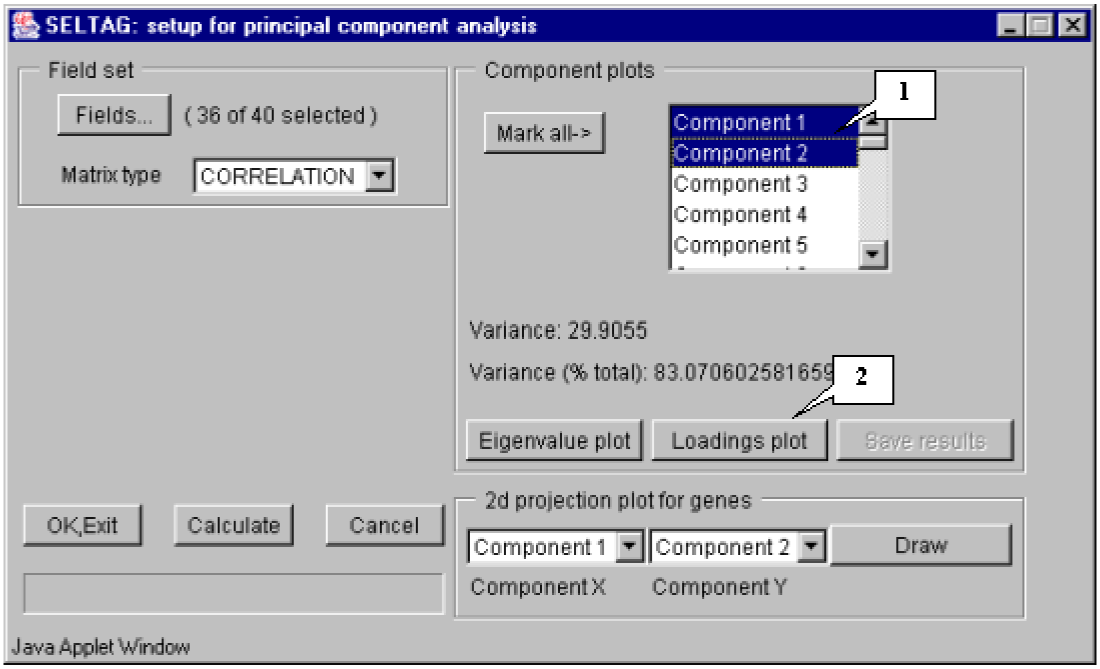
R78934

Group		Element	
Short name	Normal	Short name	Normal_3
Full name	Normal	Full name	Normal_3
Number of elements	18	Value	-159.0
Max element value	417.0		

Figure 8.19.2.

1. Genes profiles diagram area. **2.** List of genes sets. **3.** List of genes. **4.** List of groups. **5.** Information on gene, profile of which is pointed by mouse.

8.20. To analyze the plots for the first two components, select these components in the "Component plots" list in the "Setup for principal component analysis" dialog as shown on figure 8.20 and press the "Loadings plot" button.



SELTAG: setup for principal component analysis

Field set

Fields... (36 of 40 selected)

Matrix type CORRELATION

Component plots

Component 1
Component 2
Component 3
Component 4
Component 5

Variance: 29.9055

Variance (% total): 83.070602581659

2d projection plot for genes

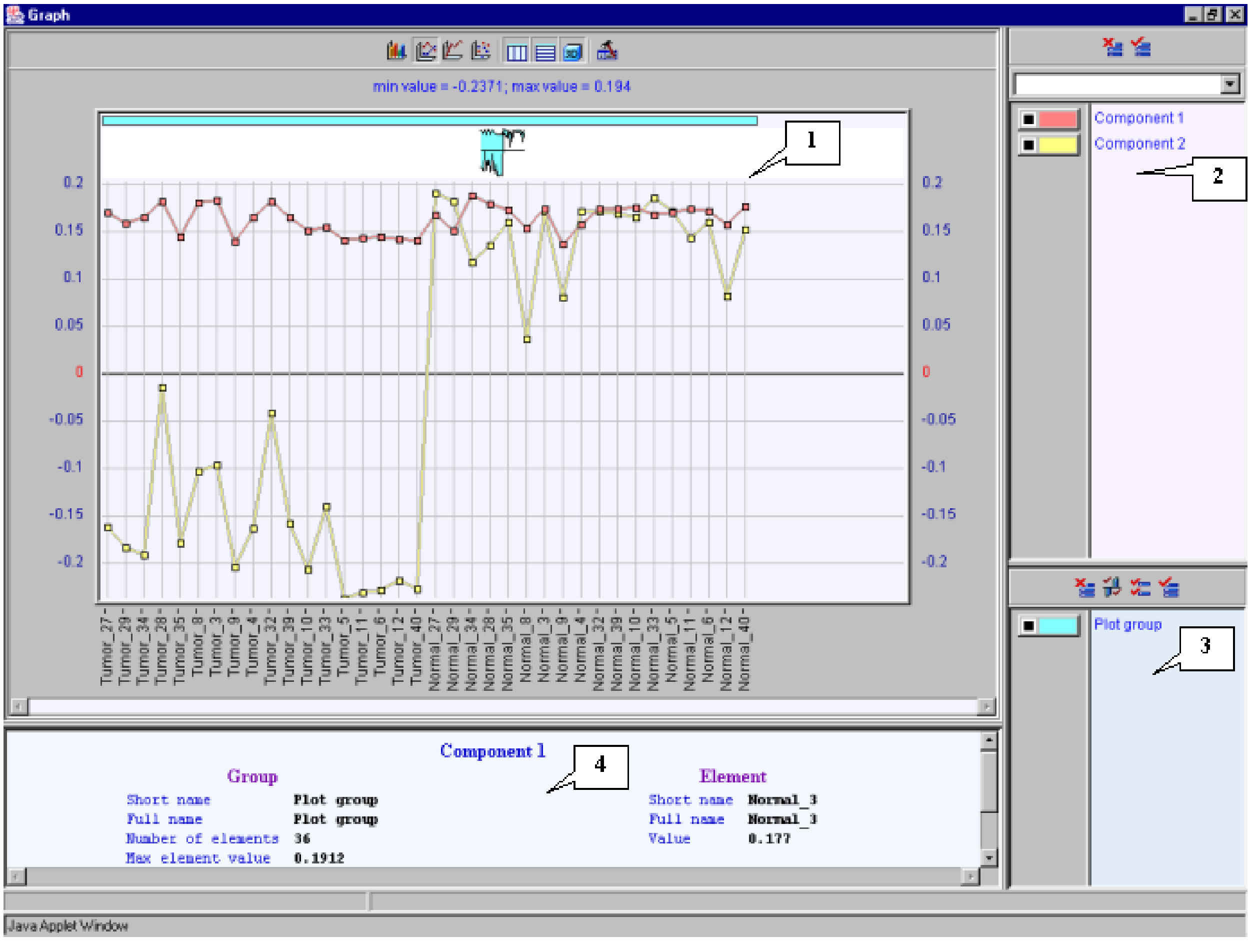
Component X: Component 1

Component Y: Component 2

Figure 8.20.

1. The “Component plots” list with two selected components. **2.** Button for the «Graph» dialog.

8.21. The «Graph» dialog with group values for each component (fig. 8.21) will appear. It is illustrative on the plot, that all coefficients for the first component are positive and approximately equal, thus it is reasonable, that the first component is responsible for total expression level. The plot for the second component shows that for tumorous tissues all coefficients are negative, and for normal one - positive.



Component 1

Group

Short name Plot group
Full name Plot group
Number of elements 36
Max element value 0.1912

Element

Short name Normal_3
Full name Normal_3
Value 0.177

Figure 8.21.

1. Components profiles diagram area. **2.** List of components. **3.** List of groups. **4.** Information on component, profile of which is pointed by mouse.

It is reasonable, that the first component represents total expression level in all tissues. Genes with higher projection value for this component are overexpressed, genes with projection value close to 0 are downexpressed. The second component is supposed to be responsible for expression diversities in tumors and normal tissues.

Thus, the analysis of the first two components for the set of 240 genes has allowed revealing two clusters of genes with different relative expression levels in tumors and normal tissues. At he same time, this analysis have showed that projection of expression value on the first component reflects the total gene expression level, while that on the second component characterizes relative gene expression in normal and tumorous tissues.

9. Resume.

Represented analysis of data from the notterman2001_set1.txt file implementing several methods has allowed to select sets of genes, which are expressed differentially in normal and tumorous tissues.

10. References.

1. Notterman DA, Alon U, Sierk AJ, and Levine AJ (2001) Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays, *Cancer Res.* **61**, 3124–3130.
2. Sneath P.H.A., Sokal R.R. (1973) Numerical Taxonomy. The principles and practice of numerical classification. San Francisco: W.H. Freeman and Co.
3. Everitt BS and Dunn G. Applied Multivariate Data Analysis. 1992 Oxford University Press, New York, NY.