PCOT2: Principal Coordinates and Hotelling's T^2 for the analysis of microarray data

Sarah Song and Mik Black

May 2, 2019

1 Overview

`pcot2` is an R-package for the analysis of groups of genes in microarray experiments. It utilizes inter-gene correlation information to detect significant alterations in the activities of gene sets. Incorporating additional (usually functional) information into the data analysis process allows gene interactions to be investigated in a statistical framework. One of the reasons that gene set analysis is becoming important is that it is suitable for detecting small coordinated changes in expression of groups of genes which are functionally related, which may not be considered significant in a single gene analysis. This vignette gives a tutorial-style introduction to the functions in the `pcot2` package. These functions are used for testing and visualizing changes in expression activity for groups of genes.

2 Example: ALL/AML data

In this example the ALL/AML leukemia data set of Golub *et al.*(1999) is used to illustrate the functionality of the `pcot2` package. This data set contains 38 bone marrow samples obtained from adult leukemia patients, 11 relating to acute myeloid leukemia (AML, class 1) and 27 relating to acute lymphoblastic leukemia (ALL, class 0). Gene expression levels were measured using Affymetrix high density oligonucleotide arrays containing 6817 human genes, of which 3051 genes were considered suitable for analysis by Golub et al.(1999) after preprocessing. This data set is available as part of the `multtest` package and gene sets are defined as KEGG pathways using the `hu6800.db` annotation package. Both packages can be downloaded from www.bioconductor.org.

```
> library(pcot2)
> library(multtest)
> library(hu6800.db)
> set.seed(1234567)
```

3 The `pcot2` function

The `pcot2` function implements the PCOT2 testing method, which is a two-stage permutation-based approach for testing changes in activity in pre-specified

gene sets. The function requires at least three inputs: gene expression data, sample class labels, and a gene category indicator matrix. The gene expression data should be in the form of a matrix with no missing values. Data pre-processing (e.g. normalization) must therefore take place before running the PCOT2 analysis.

```
> data(golub)
> rownames(golub) <- golub.gnames[,3]
> colnames(golub) <- golub.cl
```

The class labels represent two distinct experimental conditions (e.g., AML and ALL).

```
> golub.cl
```

```
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
```

The gene category indicator matrix is designed to indicate presence or absence of genes in the pre-defined gene categories (e.g., gene pathways). The indicator matrix contains rows representing gene identifiers for genes present in the expression data, and columns representing pre-defined group names. The values 1 or 0 indicate the presence or absence of a gene in a particular group.

In this example, the `hu6800.db` annotation package is used to define the KEGG (http://www.genome.jp/kegg/pathway.html) pathways for all of 3051 genes in the data. The `getImat` function is used to generate an indicator matrix which includes 65 KEGG pathways containing at least 10 of the total 3051 genes.

```
> KEGG.list <- as.list(hu6800PATH)
> imat <- getImat(golub, KEGG.list, ms=10)
> colnames(imat) <- paste("KEGG", colnames(imat), sep="")
> dim(imat)
```

```
[1] 3051  156
```

Permutations are used to produce p-values based on the null distribution of the T^2 statistic. By default `pcot2` will automatically run 1000 permutations. In order to minimize the time taken to build this vignette, only 10 permutations have been performed.

```
> results <- pcot2(golub, golub.cl, imat, iter=10)
```

```
Comparison: 0-1
```

The output from the `pcot2` function can contain information on either all pathways or just significantly differentially expressed pathways, based on the value of α used in the function, where α determines the significance threshold for the permutation p-values. For each KEGG pathway, the number of genes in the pathway is listed, along with Hotelling's T^2 statistic. These are followed by parametric p-values for the test statistic, both raw and adjusted. The last two columns provide raw and adjusted permutation-based p-values. The default adjustment method is the false discovery rate controlling method of Benjamini and Yekutieli (2001).

```
> results$res.sig

[1] Num         T2          P.nor       P.adj       P.permu     P.permu.adj
<0 rows> (or 0-length row.names)

> results$res.all
```

	Num	T2	P.nor	P.adj	P.permu	P.permu.adj
KEGG04080	55	50.687230	2.094494e-07	5.110114e-06	0.1	0.6057398
KEGG04360	30	35.193509	6.570324e-06	8.244092e-05	0.1	0.6057398
KEGG04010	95	41.205795	1.590086e-06	2.397578e-05	0.1	0.6057398
KEGG04910	52	22.332273	2.147277e-04	1.886002e-03	0.1	0.6057398
KEGG03410	14	40.040059	2.075157e-06	2.985624e-05	0.1	0.6057398
KEGG04650	58	53.908656	1.106031e-07	2.991339e-06	0.1	0.6057398
KEGG05322	41	72.008342	4.464687e-09	3.267863e-07	0.1	0.6057398
KEGG04962	14	25.229090	9.194586e-05	9.073947e-04	0.1	0.6057398
KEGG04510	78	53.483090	1.201785e-07	3.015871e-06	0.1	0.6057398
KEGG04270	40	39.972837	2.107525e-06	2.985624e-05	0.1	0.6057398
KEGG04810	82	43.451630	9.627080e-07	1.565867e-05	0.1	0.6057398
KEGG04060	84	57.657377	5.411391e-08	1.901179e-06	0.1	0.6057398
KEGG04062	86	70.943758	5.309744e-09	3.331192e-07	0.1	0.6057398
KEGG03050	21	22.055650	2.333540e-04	1.989904e-03	0.1	0.6057398
KEGG04110	57	46.327670	5.167040e-07	9.656018e-06	0.1	0.6057398
KEGG04971	33	19.064533	5.888905e-04	4.618177e-03	0.1	0.6057398
KEGG04972	33	43.960548	8.609302e-07	1.454182e-05	0.1	0.6057398
KEGG04976	20	19.403696	5.288905e-04	4.223059e-03	0.1	0.6057398
KEGG05110	30	24.810971	1.036597e-04	1.011629e-03	0.1	0.6057398
KEGG04146	18	37.574387	3.694684e-06	4.772242e-05	0.1	0.6057398
KEGG00190	42	14.032358	3.150661e-03	2.213838e-02	0.1	0.6057398
KEGG01100	311	69.111510	7.184788e-09	3.944101e-07	0.1	0.6057398
KEGG05010	65	17.300928	1.040963e-03	7.748318e-03	0.1	0.6057398
KEGG05012	41	11.792235	7.023285e-03	4.638128e-02	0.1	0.6057398
KEGG05016	67	17.363594	1.019776e-03	7.655494e-03	0.1	0.6057398
KEGG04142	50	62.211277	2.357599e-08	9.003187e-07	0.1	0.6057398
KEGG03420	15	15.484007	1.909975e-03	1.363882e-02	0.1	0.6057398
KEGG04141	60	45.799088	5.783566e-07	1.058299e-05	0.1	0.6057398
KEGG03018	13	5.431912	8.549321e-02	4.972889e-01	0.1	0.6057398
KEGG04144	52	37.787256	3.512529e-06	4.604678e-05	0.1	0.6057398
KEGG04020	55	32.084611	1.435281e-05	1.680853e-04	0.1	0.6057398
KEGG04666	43	46.505825	4.975229e-07	9.499687e-06	0.1	0.6057398
KEGG05100	31	33.781340	9.329447e-06	1.138092e-04	0.1	0.6057398
KEGG00350	10	5.163053	9.581005e-02	5.536325e-01	0.1	0.6057398
KEGG04514	62	30.108931	2.402932e-05	2.618860e-04	0.1	0.6057398
KEGG04530	36	31.095936	1.854001e-05	2.087707e-04	0.1	0.6057398
KEGG04670	52	33.856570	9.155178e-06	1.132563e-04	0.1	0.6057398
KEGG05160	54	53.513946	1.194556e-07	3.015871e-06	0.1	0.6057398
KEGG03430	13	22.840756	1.844695e-04	1.653303e-03	0.1	0.6057398
KEGG05200	148	68.078577	8.540243e-09	4.412405e-07	0.1	0.6057398
KEGG05210	36	30.089782	2.415145e-05	2.618860e-04	0.1	0.6057398
KEGG05213	27	26.945739	5.667387e-05	5.856228e-04	0.1	0.6057398
KEGG05416	41	21.371049	2.871904e-04	2.402341e-03	0.1	0.6057398

KEGG04120	29	13.923334	3.273264e-03	2.281732e-02	0.1	0.6057398
KEGG04974	27	6.422258	5.654610e-02	3.343652e-01	0.1	0.6057398
KEGG04210	41	25.794077	7.829317e-05	7.814393e-04	0.1	0.6057398
KEGG05142	55	49.447678	2.694878e-07	6.228875e-06	0.1	0.6057398
KEGG05014	23	31.606850	1.623516e-05	1.876277e-04	0.1	0.6057398
KEGG05131	31	55.243538	8.545878e-08	2.532835e-06	0.1	0.6057398
KEGG04115	24	37.099129	4.138379e-06	5.267872e-05	0.1	0.6057398
KEGG04916	30	13.646748	3.607485e-03	2.494910e-02	0.1	0.6057398
KEGG05215	45	55.198131	8.620647e-08	2.532835e-06	0.1	0.6057398
KEGG04310	43	41.149434	1.610537e-06	2.397578e-05	0.1	0.6057398
KEGG04350	23	24.300792	1.201271e-04	1.139100e-03	0.1	0.6057398
KEGG03013	37	13.343180	4.016233e-03	2.732355e-02	0.1	0.6057398
KEGG04145	73	148.214096	3.910205e-13	1.995124e-10	0.1	0.6057398
KEGG04520	32	22.207678	2.229148e-04	1.938526e-03	0.1	0.6057398
KEGG05410	31	18.282987	7.562727e-04	5.776099e-03	0.1	0.6057398
KEGG05414	32	10.341813	1.204386e-02	7.778234e-02	0.1	0.6057398
KEGG00010	37	9.063638	1.964873e-02	1.198467e-01	0.1	0.6057398
KEGG04380	71	32.680575	1.232254e-05	1.482625e-04	0.1	0.6057398
KEGG04620	48	49.019006	2.942818e-07	6.627548e-06	0.1	0.6057398
KEGG04630	55	41.053652	1.645933e-06	2.409434e-05	0.1	0.6057398
KEGG05140	56	41.989367	1.332522e-06	2.053306e-05	0.1	0.6057398
KEGG05145	69	47.765451	3.816431e-07	7.981091e-06	0.1	0.6057398
KEGG05212	41	26.814390	5.878394e-05	6.003636e-04	0.1	0.6057398
KEGG04640	63	123.041575	5.115131e-12	1.497578e-09	0.1	0.6057398
KEGG00980	10	66.696592	1.079104e-08	5.265563e-07	0.1	0.6057398
KEGG00983	11	44.129914	8.296347e-07	1.428798e-05	0.1	0.6057398
KEGG00240	30	74.320240	3.081965e-09	2.706960e-07	0.1	0.6057398
KEGG00480	14	89.964548	3.026550e-10	4.867064e-08	0.1	0.6057398
KEGG00590	16	39.391204	2.410915e-06	3.361208e-05	0.1	0.6057398
KEGG00860	15	49.760861	2.527749e-07	6.000484e-06	0.1	0.6057398
KEGG00030	15	13.506746	3.790243e-03	2.600825e-02	0.1	0.6057398
KEGG00230	50	24.125069	1.264215e-04	1.178643e-03	0.1	0.6057398
KEGG00071	18	39.257416	2.487030e-06	3.413148e-05	0.1	0.6057398
KEGG03320	18	55.009039	8.939526e-08	2.532835e-06	0.1	0.6057398
KEGG04920	27	62.446658	2.260875e-08	9.003187e-07	0.1	0.6057398
KEGG05150	32	56.063892	7.306855e-08	2.468375e-06	0.1	0.6057398
KEGG00620	14	24.286911	1.206120e-04	1.139100e-03	0.1	0.6057398
KEGG04930	21	19.258351	5.537710e-04	4.381888e-03	0.1	0.6057398
KEGG04664	36	62.245608	2.343224e-08	9.003187e-07	0.1	0.6057398
KEGG04722	53	55.570416	8.027457e-08	2.532835e-06	0.1	0.6057398
KEGG04912	33	13.323658	4.044143e-03	2.732355e-02	0.1	0.6057398
KEGG00280	19	38.660972	2.858611e-06	3.804217e-05	0.1	0.6057398
KEGG00310	12	28.018168	4.216839e-05	4.409220e-04	0.1	0.6057398
KEGG00380	15	103.491944	5.077894e-11	1.115007e-08	0.1	0.6057398
KEGG00640	14	47.605074	3.946596e-07	8.061360e-06	0.1	0.6057398
KEGG04012	37	23.566720	1.488433e-04	1.347757e-03	0.1	0.6057398
KEGG05220	45	38.900652	2.702687e-06	3.652048e-05	0.1	0.6057398
KEGG00564	14	58.921920	4.279369e-08	1.566111e-06	0.1	0.6057398
KEGG05340	26	146.639555	4.543033e-13	1.995124e-10	0.1	0.6057398
KEGG00500	12	28.113816	4.108093e-05	4.347267e-04	0.1	0.6057398

KEGG05120	34	65.157949	1.405379e-08	6.496717e-07	0.1	0.6057398
KEGG05323	46	85.827246	5.429581e-10	5.961155e-08	0.1	0.6057398
KEGG03040	33	19.525470	5.089551e-04	4.101163e-03	0.1	0.6057398
KEGG04660	50	10.494546	1.136995e-02	7.397396e-02	0.1	0.6057398
KEGG00410	12	46.645514	4.830102e-07	9.427529e-06	0.1	0.6057398
KEGG05221	38	42.585990	1.166149e-06	1.829027e-05	0.1	0.6057398
KEGG04340	10	5.894040	7.040922e-02	4.122801e-01	0.1	0.6057398
KEGG05218	30	21.265471	2.965986e-04	2.457634e-03	0.1	0.6057398
KEGG04512	26	24.645916	1.087092e-04	1.049250e-03	0.1	0.6057398
KEGG05146	49	71.444582	4.892881e-09	3.305791e-07	0.1	0.6057398
KEGG05222	46	43.526104	9.470522e-07	1.565867e-05	0.1	0.6057398
KEGG04610	14	72.954692	3.832568e-09	3.060210e-07	0.1	0.6057398
KEGG03030	19	22.769488	1.884236e-04	1.671684e-03	0.1	0.6057398
KEGG04622	20	53.826381	1.123894e-07	2.991339e-06	0.1	0.6057398
KEGG00970	15	24.086482	1.278498e-04	1.178643e-03	0.1	0.6057398
KEGG03015	18	46.828887	4.646349e-07	9.274986e-06	0.1	0.6057398
KEGG04970	36	63.607321	1.841606e-08	8.087621e-07	0.1	0.6057398
KEGG04370	35	31.024253	1.889009e-05	2.100202e-04	0.1	0.6057398
KEGG04662	45	44.427951	7.774477e-07	1.374197e-05	0.1	0.6057398
KEGG00051	16	26.636897	6.176816e-05	6.235905e-04	0.1	0.6057398
KEGG00052	14	22.065051	2.326937e-04	1.989904e-03	0.1	0.6057398
KEGG04114	40	21.595025	2.682605e-04	2.265570e-03	0.1	0.6057398
KEGG04914	32	18.719680	6.573272e-04	5.064433e-03	0.1	0.6057398
KEGG04070	28	20.870886	3.347487e-04	2.747826e-03	0.1	0.6057398
KEGG04720	33	9.202007	1.862232e-02	1.160029e-01	0.1	0.6057398
KEGG04730	30	88.207412	3.870011e-10	4.867064e-08	0.1	0.6057398
KEGG00561	12	88.191090	3.878922e-10	4.867064e-08	0.1	0.6057398
KEGG00330	20	70.508466	5.702581e-09	3.339137e-07	0.1	0.6057398
KEGG00520	15	8.466957	2.481070e-02	1.502883e-01	0.1	0.6057398
KEGG04672	24	44.399449	7.822850e-07	1.374197e-05	0.1	0.6057398
KEGG05144	32	81.917469	9.608954e-10	9.377513e-08	0.1	0.6057398
KEGG05310	21	32.129242	1.418916e-05	1.680853e-04	0.1	0.6057398
KEGG05320	25	15.995128	1.606747e-03	1.156756e-02	0.1	0.6057398
KEGG05330	24	19.655395	4.885585e-04	3.973259e-03	0.1	0.6057398
KEGG04612	39	48.665424	3.165463e-07	6.781215e-06	0.1	0.6057398
KEGG04940	24	9.486543	1.668563e-02	1.054343e-01	0.1	0.6057398
KEGG05332	24	10.138221	1.300856e-02	8.339933e-02	0.1	0.6057398
KEGG05143	19	24.060403	1.288248e-04	1.178643e-03	0.1	0.6057398
KEGG05214	36	16.976533	1.158329e-03	8.549466e-03	0.1	0.6057398
KEGG05219	21	48.869498	3.034871e-07	6.663989e-06	0.1	0.6057398
KEGG05223	30	16.543996	1.337025e-03	9.786163e-03	0.1	0.6057398
KEGG04966	10	42.692106	1.138933e-06	1.818820e-05	0.1	0.6057398
KEGG04621	20	55.022897	8.915738e-08	2.532835e-06	0.1	0.6057398
KEGG04623	17	17.496447	9.763539e-04	7.392705e-03	0.1	0.6057398
KEGG04330	16	14.667409	2.526630e-03	1.789674e-02	0.1	0.6057398
KEGG04964	10	29.959732	2.499846e-05	2.677648e-04	0.1	0.6057398
KEGG04150	17	11.010044	9.374697e-03	6.144782e-02	0.1	0.6057398
KEGG04973	20	13.182805	4.251686e-03	2.850650e-02	0.1	0.6057398
KEGG05216	18	31.300666	1.757741e-05	2.005018e-04	0.1	0.6057398
KEGG05020	20	16.352442	1.425270e-03	1.034585e-02	0.1	0.6057398

```
KEGG04742  10   9.165107 1.889037e-02 1.168439e-01     0.1   0.6057398
KEGG00562  15  18.867003 6.271148e-04 4.874417e-03     0.1   0.6057398
KEGG05130  24   8.239661 2.713860e-02 1.632633e-01     0.2   1.0000000
KEGG00510  15   7.675775 3.396901e-02 2.029643e-01     0.2   1.0000000
KEGG00020  14  13.152966 4.297080e-03 2.859260e-02     0.2   1.0000000
KEGG04540  35   9.106446 1.932494e-02 1.186960e-01     0.2   1.0000000
KEGG00270  13   9.247003 1.830088e-02 1.148148e-01     0.2   1.0000000
KEGG00250  11   9.616124 1.587530e-02 1.010409e-01     0.2   1.0000000
KEGG04960  19   6.414720 5.672222e-02 3.343652e-01     0.2   1.0000000
KEGG00650  10   4.237641 1.426346e-01 8.188181e-01     0.3   1.0000000
KEGG05412  26   3.301194 2.153740e-01 1.000000e+00     0.4   1.0000000
KEGG05211  30   3.486340 1.983688e-01 1.000000e+00     0.4   1.0000000
KEGG04260  29   2.355025 3.299165e-01 1.000000e+00     0.5   1.0000000
```

In the `pcot2` function, the T^2 statistic can be calculated in two ways, using either a pooled estimate of correlation for the two classes (default) or an unpooled estimate. And users can set *var.equal=F* if the correlation structure is assumed to differ across the two classes.

In the first step of the PCOT2 analysis, the dimensionality of the gene expression data is reduced via principal coordinates. The default dimensionality in the `pcot2` function is set as *ncomp=2*. In the second step of the PCOT2 analysis, the distances between the transformed groups are calculated via euclidean distances by default. Other distances (e.g., correlation or Spearman distances) can also be used by defining *dist.method* in the function. A permutation p-value for each category is calculated by re-arranging the sample labels. The permutations can also be performed by permuting rows (genes), using `permu='ByRow'`.

Table 1 lists computation times (in minutes) required to run 1000 permutations of the `pcot2` function on the AML/ALL data under various parameter configurations. The two machines used were a 3.2GHz Pentium 4 with 1Gb RAM running Microsoft Windows XP and R 2.1.0 (PC), and a 1.70GHz Pentium M with 256Mb of RAM running Fedora Core 3 and R 2.2.0 (Unix).

Table 1: *Computation times (minutes, 1000 permutations)*

Changes	PC machine	UNIX machine
default setting	5.6	6.8
var.equal=F	5.5	6.8
comp=8	6	7.6
dist.method="euclidean"	4.8	6
permu="ByRow"	5.6	6.8

4 The `corplot` and `corplot2` functions

The `corplot` and `corplot2` functions enable visualization of both correlation and gene expression information for a particular gene category, in particular the groups identified as being differentially expressed. The plot produced by the `corplot` function displays the pooled correlation calculated from the two classes, while the `corplot2` function produces a plot based on unpooled correlation. Gene names can be added to the plot using *add.name=T* (default). The font

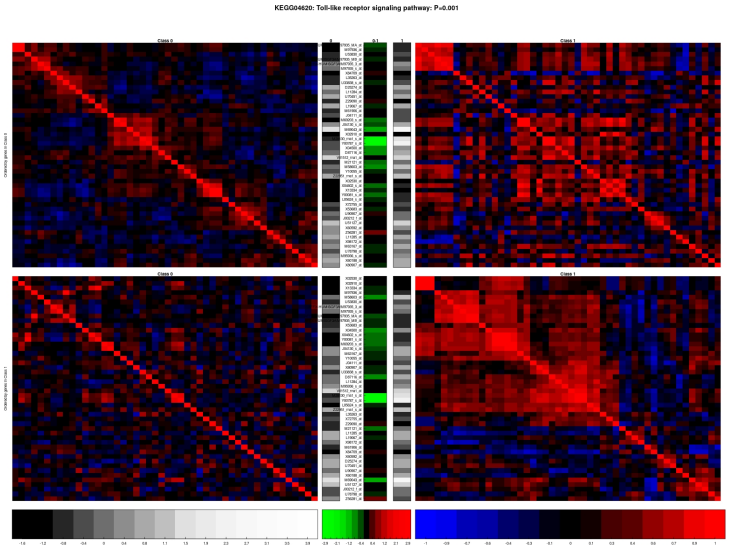

Figure 1: KEGG04620

size can be changed by setting the *font.size* argument. The *main* option specifies the title of the plot.

```
> sel <- c("04620","04120")
> pvalue <- c(0.001, 0.72)
> library(KEGG.db)
> pname <- unlist(mget(sel, env=KEGGPATHID2NAME))
> main <- paste("KEGG", sel, ": ", pname, ": ", "P=", pvalue, sep="")
> for(i in 1:length(sel)){
+ fname <- paste("corplot2-KEGG",sel[i] , ".jpg", sep="")
+ jpeg(fname, width=1600, height=1200, quality=100)
+ selgene <- rownames(imat)[imat[,match(paste("KEGG",sel,sep="")[i],colnames(imat))]==1]
+ corplot2(golub, selgene, golub.cl, main=main[i])
+ dev.off()
+ }
```

The argument *inputP* allows users to input the p-values of individual genes calculated using other approaches, such as the limma package (Smyth *et al.*, 2004), allowing the results from both per-gene and per-pathway analysis to be printed on a single plot. To allow users to identify genes from in correlation image plots, the argument *gene.locator*=T allows the selection of interesting (e.g., highly correlated and differential expressed between two classes) genes by clicking beginning and end points on the main diagonal of the image plots. This prints the identifiers for the selected genes. Further details of this functionality are provided in the HowToUseGeneLocator.pdf document. The usage of `corplot2` is similar to that for the `corplot` function.

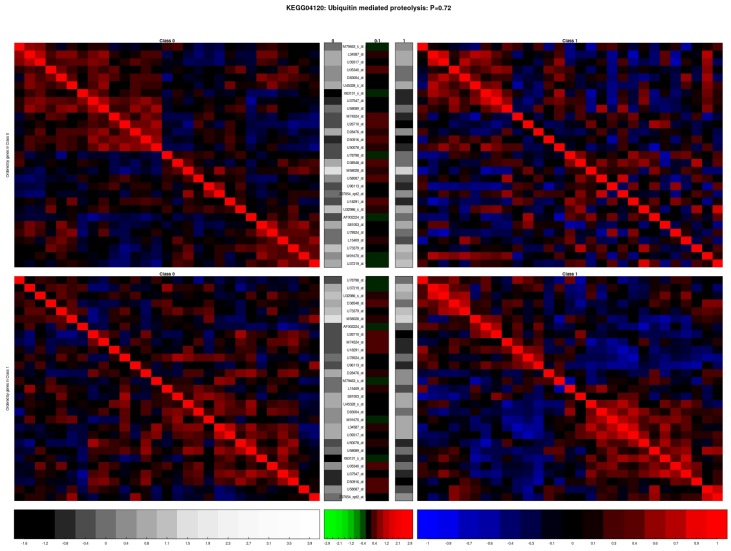

Figure 2: KEGG04120

5 The avePrObes function

In Affymetrix gene expression data, a unique gene can often link to multiple probe sets, with such genes then having a greater influence on the pathway analysis (particularly if the gene is differentially expressed). In order to solve this problem, the `aveProbe` function is provided to change the multiple probe data to the unique gene data by taking the median of the probe values. This function can be used to transform both expression data and the indicator matrix by providing a vector of unique gene identifiers.

```
> pathlist <- as.list(hu6800PATH)
> pathlist <- pathlist[match(rownames(golub), names(pathlist))]
> ids <- unlist(mget(names(pathlist), env=hu6800SYMBOL))
> #### transform data matrix only ####
> newdata <- aveProbe(x=golub, ids=ids)$newx
> #### transform both data and imat ####
> output <- aveProbe(x=golub, imat=imat, ids=ids)
> newdata <- output$newx
> newimat <- output$newimat
> newimat <- newimat[,apply(newimat, 2, sum)>=10]
> dim(newdata)

[1] 2501   38

> dim(newimat)

[1] 2501  149
```

After the multiple probe data set has been changed to the unique gene symbol data, further analysis such as testing and visualizing pathways can be done on the new data set.

References

[1] Benjamini,B.Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**, 1165-1188.

[2] Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

[3] 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. *et al.* (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, **286**, 531-537.

[4] Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, No.1, Article 3.