

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

Sarah Song and Mik Black

October 28, 2009

## 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 pre-processing. 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](http://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

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

```
> golub.cl
```

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.

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

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.

Comparison: 0-1

2

```
> 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	51	53.578119	1.179668e-07	3.311923e-06	0.1	0.566029
KEGG04360	30	35.193509	6.570324e-06	9.223105e-05	0.1	0.566029
KEGG04010	98	40.421733	1.901071e-06	3.103060e-05	0.1	0.566029
KEGG04910	55	23.685638	1.437380e-04	1.363328e-03	0.1	0.566029
KEGG03410	14	40.040059	2.075157e-06	3.310234e-05	0.1	0.566029
KEGG04650	60	54.634383	9.608742e-08	3.065521e-06	0.1	0.566029
KEGG05322	46	71.425708	4.907946e-09	3.917203e-07	0.1	0.566029
KEGG04510	79	52.030331	1.600398e-07	4.320311e-06	0.1	0.566029
KEGG04270	43	24.402333	1.166417e-04	1.153070e-03	0.1	0.566029
KEGG04810	84	42.380783	1.220748e-06	2.196957e-05	0.1	0.566029
KEGG04520	34	21.222398	3.005300e-04	2.636685e-03	0.1	0.566029
KEGG04670	53	34.386677	8.020671e-06	1.103827e-04	0.1	0.566029
KEGG04060	84	57.482933	5.590852e-08	2.308285e-06	0.1	0.566029
KEGG04062	87	70.607587	5.610503e-09	3.937877e-07	0.1	0.566029
KEGG03050	22	20.987918	3.229199e-04	2.798145e-03	0.1	0.566029
KEGG04110	57	46.327670	5.167040e-07	1.098976e-05	0.1	0.566029
KEGG03320	18	55.009039	8.939526e-08	3.013283e-06	0.1	0.566029
KEGG05110	30	24.810971	1.036597e-04	1.054438e-03	0.1	0.566029
KEGG00190	43	14.212036	2.959080e-03	2.119293e-02	0.1	0.566029
KEGG01100	310	69.329251	6.929252e-09	4.421341e-07	0.1	0.566029
KEGG05010	70	16.994598	1.151439e-03	8.965259e-03	0.1	0.566029
KEGG05012	43	10.731022	1.040403e-02	6.967526e-02	0.1	0.566029
KEGG05016	70	31.545201	1.649643e-05	2.105173e-04	0.1	0.566029
KEGG04142	52	54.964931	9.015688e-08	3.013283e-06	0.1	0.566029
KEGG03420	15	15.484007	1.909975e-03	1.441469e-02	0.1	0.566029
KEGG04144	48	32.894544	1.166969e-05	1.545410e-04	0.1	0.566029
KEGG04020	57	34.108620	8.595950e-06	1.160248e-04	0.1	0.566029
KEGG04666	44	46.457167	5.026857e-07	1.098976e-05	0.1	0.566029
KEGG00350	12	5.486352	8.355107e-02	4.654166e-01	0.1	0.566029
KEGG04514	62	30.108931	2.402932e-05	2.810934e-04	0.1	0.566029
KEGG04530	36	31.095936	1.854001e-05	2.282945e-04	0.1	0.566029
KEGG03430	13	22.840756	1.844695e-04	1.659932e-03	0.1	0.566029
KEGG05200	152	68.412505	8.074670e-09	4.722847e-07	0.1	0.566029
KEGG05210	41	25.797211	7.822372e-05	8.097214e-04	0.1	0.566029
KEGG05213	28	26.480816	6.452479e-05	7.076313e-04	0.1	0.566029
KEGG04120	29	12.630167	5.181351e-03	3.600659e-02	0.1	0.566029
KEGG04210	41	25.794077	7.829317e-05	8.097214e-04	0.1	0.566029
KEGG05014	25	31.407616	1.709570e-05	2.142689e-04	0.1	0.566029
KEGG04115	24	37.099129	4.138379e-06	6.051310e-05	0.1	0.566029
KEGG00510	14	11.111469	9.027707e-03	6.212089e-02	0.1	0.566029
KEGG04916	32	14.931124	2.307207e-03	1.704604e-02	0.1	0.566029
KEGG05215	47	53.971118	1.092671e-07	3.286810e-06	0.1	0.566029
KEGG04310	44	41.315269	1.551142e-06	2.655388e-05	0.1	0.566029

KEGG04350	24	24.218857	1.230198e-04	1.182803e-03	0.1	0.566029
KEGG05130	21	9.677190	1.550796e-02	9.895150e-02	0.1	0.566029
KEGG05410	31	18.282987	7.562727e-04	6.101260e-03	0.1	0.566029
KEGG00010	37	9.063638	1.964873e-02	1.199215e-01	0.1	0.566029
KEGG01061	36	10.673205	1.063189e-02	7.039874e-02	0.1	0.566029
KEGG01062	38	9.775479	1.493544e-02	9.617273e-02	0.1	0.566029
KEGG01063	34	6.912828	4.624011e-02	2.617325e-01	0.1	0.566029
KEGG01064	36	8.775639	2.198176e-02	1.318672e-01	0.1	0.566029
KEGG01065	45	7.361699	3.854195e-02	2.217350e-01	0.1	0.566029
KEGG01066	40	10.726066	1.042335e-02	6.967526e-02	0.1	0.566029
KEGG01070	50	7.595956	3.507399e-02	2.034512e-01	0.1	0.566029
KEGG04620	48	49.019006	2.942818e-07	7.611291e-06	0.1	0.566029
KEGG04630	55	40.667327	1.797269e-06	3.003476e-05	0.1	0.566029
KEGG05212	43	25.787083	7.844841e-05	8.097214e-04	0.1	0.566029
KEGG04640	62	127.436052	3.174794e-12	1.114156e-09	0.1	0.566029
KEGG00980	10	66.696592	1.079104e-08	5.826131e-07	0.1	0.566029
KEGG00983	12	44.930783	6.971132e-07	1.397963e-05	0.1	0.566029
KEGG00240	30	74.320240	3.081965e-09	3.605263e-07	0.1	0.566029
KEGG00480	14	89.964548	3.026550e-10	5.310657e-08	0.1	0.566029
KEGG00590	17	41.335666	1.543998e-06	2.655388e-05	0.1	0.566029
KEGG00860	14	45.065971	6.770464e-07	1.397655e-05	0.1	0.566029
KEGG00030	15	13.506746	3.790243e-03	2.687152e-02	0.1	0.566029
KEGG00230	49	25.818426	7.775525e-05	8.097214e-04	0.1	0.566029
KEGG00071	18	39.257416	2.487030e-06	3.879081e-05	0.1	0.566029
KEGG04920	27	62.446658	2.260875e-08	1.027908e-06	0.1	0.566029
KEGG00620	14	24.286911	1.206120e-04	1.175760e-03	0.1	0.566029
KEGG04930	21	19.258351	5.537710e-04	4.572689e-03	0.1	0.566029
KEGG04664	36	62.245608	2.343224e-08	1.027908e-06	0.1	0.566029
KEGG04722	56	56.071300	7.296574e-08	2.695416e-06	0.1	0.566029
KEGG04912	35	15.060709	2.206856e-03	1.647808e-02	0.1	0.566029
KEGG00280	19	38.660972	2.858611e-06	4.268916e-05	0.1	0.566029
KEGG00310	12	28.018168	4.216839e-05	4.773706e-04	0.1	0.566029
KEGG00380	15	103.491944	5.077894e-11	1.188017e-08	0.1	0.566029
KEGG00640	14	47.605074	3.946596e-07	9.233403e-06	0.1	0.566029
KEGG00650	12	18.081508	8.071233e-04	6.365174e-03	0.1	0.566029
KEGG00020	14	13.152966	4.297080e-03	3.016017e-02	0.1	0.566029
KEGG04012	38	23.225345	1.645928e-04	1.520049e-03	0.1	0.566029
KEGG05220	48	38.786725	2.775650e-06	4.235135e-05	0.1	0.566029
KEGG00564	10	42.516575	1.184323e-06	2.187495e-05	0.1	0.566029
KEGG05340	25	148.792814	3.701484e-13	2.597982e-10	0.1	0.566029
KEGG00500	12	28.113816	4.108093e-05	4.726839e-04	0.1	0.566029
KEGG05120	34	65.157949	1.405379e-08	7.045727e-07	0.1	0.566029
KEGG04660	50	10.494546	1.136995e-02	7.458219e-02	0.1	0.566029
KEGG00410	12	46.645514	4.830102e-07	1.093591e-05	0.1	0.566029
KEGG05221	39	35.710984	5.788211e-06	8.291033e-05	0.1	0.566029
KEGG04340	11	6.073128	6.534459e-02	3.669104e-01	0.1	0.566029
KEGG05218	31	20.513822	3.737548e-04	3.199141e-03	0.1	0.566029
KEGG04512	26	24.645916	1.087092e-04	1.090006e-03	0.1	0.566029
KEGG05222	48	43.992713	8.548920e-07	1.621698e-05	0.1	0.566029
KEGG04610	13	71.766981	4.642947e-09	3.917203e-07	0.1	0.566029

KEGG03030	19	22.769488	1.884236e-04	1.674051e-03	0.1	0.566029
KEGG04622	20	53.826381	1.123894e-07	3.286810e-06	0.1	0.566029
KEGG00970	16	23.403392	1.561698e-04	1.461491e-03	0.1	0.566029
KEGG04370	35	31.024253	1.889009e-05	2.285948e-04	0.1	0.566029
KEGG04662	45	44.427951	7.774477e-07	1.515755e-05	0.1	0.566029
KEGG00051	16	26.636897	6.176816e-05	6.881522e-04	0.1	0.566029
KEGG00052	15	19.849740	4.596460e-04	3.886922e-03	0.1	0.566029
KEGG04540	35	9.106446	1.932494e-02	1.189799e-01	0.1	0.566029
KEGG04070	30	22.848678	1.840354e-04	1.659932e-03	0.1	0.566029
KEGG04720	36	8.649082	2.309800e-02	1.373893e-01	0.1	0.566029
KEGG04730	33	56.104042	7.251313e-08	2.695416e-06	0.1	0.566029
KEGG00561	12	88.191090	3.878922e-10	5.445044e-08	0.1	0.566029
KEGG00330	21	71.283630	5.022943e-09	3.917203e-07	0.1	0.566029
KEGG00520	15	8.466957	2.481070e-02	1.463364e-01	0.1	0.566029
KEGG05310	21	32.129242	1.418916e-05	1.844264e-04	0.1	0.566029
KEGG05320	25	15.995128	1.606747e-03	1.225801e-02	0.1	0.566029
KEGG05330	24	19.655395	4.885585e-04	4.082232e-03	0.1	0.566029
KEGG04612	41	47.655231	3.905391e-07	9.233403e-06	0.1	0.566029
KEGG04940	24	9.486543	1.668563e-02	1.045646e-01	0.1	0.566029
KEGG05332	24	10.138221	1.300856e-02	8.454067e-02	0.1	0.566029
KEGG05214	39	18.202500	7.761679e-04	6.190609e-03	0.1	0.566029
KEGG05219	22	48.867088	3.036379e-07	7.611291e-06	0.1	0.566029
KEGG05223	31	16.965995	1.162369e-03	8.965259e-03	0.1	0.566029
KEGG04330	16	14.667409	2.526630e-03	1.828228e-02	0.1	0.566029
KEGG00710	10	7.339073	3.889560e-02	2.219503e-01	0.1	0.566029
KEGG04150	18	11.009560	9.376387e-03	6.389380e-02	0.1	0.566029
KEGG05216	19	30.751272	2.028858e-05	2.413571e-04	0.1	0.566029
KEGG05020	21	14.773131	2.436126e-03	1.781102e-02	0.1	0.566029
KEGG04740	10	9.033627	1.987914e-02	1.202818e-01	0.1	0.566029
KEGG04742	10	9.165107	1.889037e-02	1.173336e-01	0.1	0.566029
KEGG00562	15	18.867003	6.271148e-04	5.118102e-03	0.1	0.566029
KEGG00270	12	8.220476	2.734539e-02	1.599422e-01	0.2	1.000000
KEGG00250	11	9.616124	1.587530e-02	1.003828e-01	0.2	1.000000
KEGG04260	29	1.827050	4.204849e-01	1.000000e+00	0.5	1.000000
KEGG05412	26	3.301194	2.153740e-01	1.000000e+00	0.5	1.000000
KEGG05211	31	2.628229	2.913801e-01	1.000000e+00	0.5	1.000000

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 un-pooled 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 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.

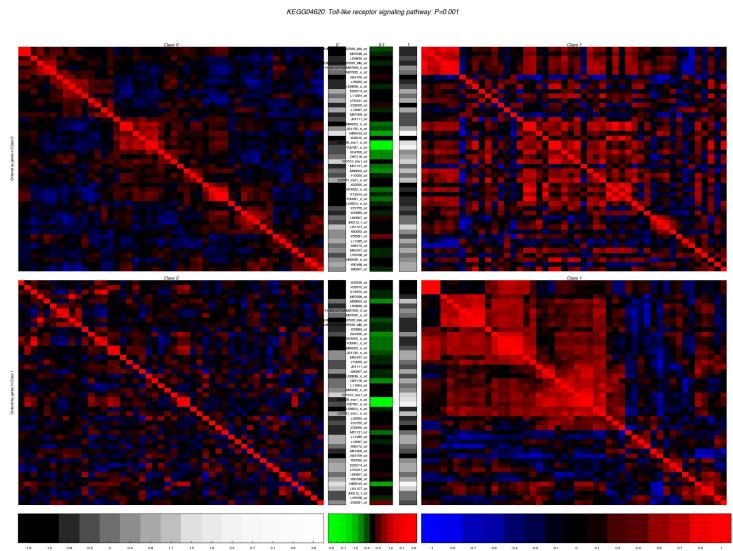


Figure 1: KEGG04620

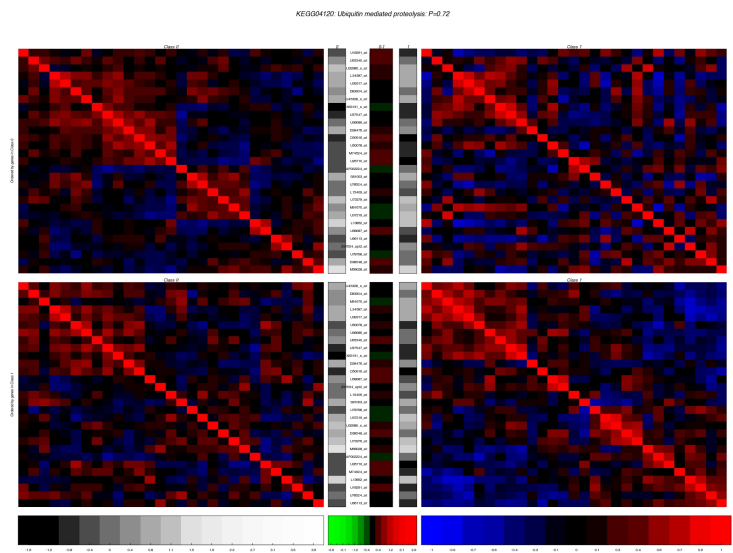


Figure 2: KEGG04120

## 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))
> newdata <- aveProbe(x = golub, ids = ids)$newx
> output <- aveProbe(x = golub, imat = imat, ids = ids)
> newdata <- output$newx
> newimat <- output$newimat
> newimat <- newimat[, apply(newimat, 2, sum) >= 10]
> dim(newdata)

[1] 2561 38

> dim(newimat)

[1] 2561 125
```

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.