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

Sarah Song and Mik Black

April 30, 2008

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` annotation package. Both packages can be downloaded from www.bioconductor.org.

```
> library(pcot2)
> library(multtest)
> library(hu6800)
> 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 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` 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  109
```

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
KEGG04010	101	37.5591256	3.708122e-06	5.198359e-05	0.1	0.5690818
KEGG04060	87	50.9617543	1.981749e-07	7.119095e-06	0.1	0.5690818
KEGG04350	28	20.1492856	4.185715e-04	3.590798e-03	0.1	0.5690818
KEGG04520	36	23.8052982	1.387873e-04	1.329519e-03	0.1	0.5690818
KEGG05210	44	27.5538698	4.789705e-05	5.098132e-04	0.1	0.5690818
KEGG05212	48	30.3992536	2.225552e-05	2.558373e-04	0.1	0.5690818
KEGG05220	52	40.1093857	2.042326e-06	3.353923e-05	0.1	0.5690818
KEGG00190	42	14.2095556	2.961639e-03	1.979383e-02	0.1	0.5690818
KEGG04810	92	49.5919537	2.616479e-07	8.846358e-06	0.1	0.5690818
KEGG04514	73	30.8382642	1.983140e-05	2.326234e-04	0.1	0.5690818
KEGG04670	56	37.4219523	3.831246e-06	5.243084e-05	0.1	0.5690818
KEGG00564	10	45.8715971	5.694580e-07	1.258880e-05	0.1	0.5690818
KEGG00590	20	44.3957904	7.829082e-07	1.451594e-05	0.1	0.5690818
KEGG04370	37	32.2815602	1.364532e-05	1.668714e-04	0.1	0.5690818
KEGG04664	38	61.3798116	2.735820e-08	1.572474e-06	0.1	0.5690818
KEGG04730	36	37.9994969	3.340337e-06	4.799837e-05	0.1	0.5690818
KEGG04912	38	15.9191093	1.648417e-03	1.184332e-02	0.1	0.5690818
KEGG04510	85	65.8760447	1.241745e-08	7.930235e-07	0.1	0.5690818
KEGG05218	32	19.2120700	5.619507e-04	4.614199e-03	0.1	0.5690818
KEGG00280	21	40.9446790	1.687207e-06	2.852237e-05	0.1	0.5690818
KEGG00240	32	58.9786441	4.234863e-08	2.212803e-06	0.1	0.5690818
KEGG04360	33	39.5598135	2.318480e-06	3.701664e-05	0.1	0.5690818
KEGG03022	12	23.6751657	1.441801e-04	1.358537e-03	0.1	0.5690818
KEGG00260	12	9.0142092	2.002974e-02	1.211847e-01	0.1	0.5690818
KEGG00330	13	17.4711336	9.844744e-04	7.657064e-03	0.1	0.5690818
KEGG03320	20	53.2014932	1.269935e-07	5.213743e-06	0.1	0.5690818
KEGG04310	42	37.0402727	4.197126e-06	5.482712e-05	0.1	0.5690818
KEGG05221	41	42.9729541	1.070094e-06	1.922065e-05	0.1	0.5690818
KEGG04330	15	14.4138200	2.758517e-03	1.910265e-02	0.1	0.5690818
KEGG05120	37	66.0776488	1.199519e-08	7.930235e-07	0.1	0.5690818
KEGG00230	52	19.2543394	5.544749e-04	4.614199e-03	0.1	0.5690818
KEGG04612	56	82.6896309	8.571762e-10	1.642272e-07	0.1	0.5690818
KEGG00561	16	69.2425821	7.029794e-09	5.824626e-07	0.1	0.5690818
KEGG04512	31	48.4096545	3.337579e-07	9.591747e-06	0.1	0.5690818
KEGG05216	22	29.2717954	3.003285e-05	3.384717e-04	0.1	0.5690818
KEGG00020	12	12.2075129	6.036512e-03	3.812771e-02	0.1	0.5690818
KEGG04340	11	6.0731284	6.534459e-02	3.755829e-01	0.1	0.5690818
KEGG04916	33	16.5292411	1.343613e-03	1.002950e-02	0.1	0.5690818
KEGG04640	69	114.3585231	1.366329e-11	7.853287e-09	0.1	0.5690818
KEGG04650	69	45.4328122	6.256100e-07	1.331791e-05	0.1	0.5690818
KEGG04662	39	46.7574737	4.717016e-07	1.190747e-05	0.1	0.5690818
KEGG04610	15	73.3638672	3.589230e-09	4.560615e-07	0.1	0.5690818
KEGG04070	31	25.7760641	7.869364e-05	7.666264e-04	0.1	0.5690818

KEGG00980	13	69.1882122	7.093654e-09	5.824626e-07	0.1	0.5690818
KEGG00350	14	4.6547487	1.190821e-01	6.645157e-01	0.1	0.5690818
KEGG04660	43	33.3381307	1.042993e-05	1.303226e-04	0.1	0.5690818
KEGG00380	19	88.6534920	3.634677e-10	1.044556e-07	0.1	0.5690818
KEGG04110	51	46.1841153	5.327283e-07	1.224791e-05	0.1	0.5690818
KEGG00220	12	38.2376153	3.157719e-06	4.653771e-05	0.1	0.5690818
KEGG01510	27	14.1338165	3.040922e-03	2.009010e-02	0.1	0.5690818
KEGG05010	16	5.7281493	7.547169e-02	4.252850e-01	0.1	0.5690818
KEGG04940	35	33.7846816	9.321633e-06	1.190627e-04	0.1	0.5690818
KEGG04120	32	14.6056854	2.581106e-03	1.809206e-02	0.1	0.5690818
KEGG04020	62	42.2997618	1.243043e-06	2.165051e-05	0.1	0.5690818
KEGG04540	41	10.8912732	9.799036e-03	6.056148e-02	0.1	0.5690818
KEGG05214	41	20.5232602	3.726642e-04	3.245412e-03	0.1	0.5690818
KEGG05215	49	53.5662225	1.182413e-07	5.213743e-06	0.1	0.5690818
KEGG04530	39	31.3633894	1.729318e-05	2.070760e-04	0.1	0.5690818
KEGG04115	24	37.0991286	4.138379e-06	5.482712e-05	0.1	0.5690818
KEGG03050	12	46.5088461	4.972042e-07	1.190747e-05	0.1	0.5690818
KEGG04080	57	44.7224048	7.292848e-07	1.430100e-05	0.1	0.5690818
KEGG05040	21	13.8093276	3.406880e-03	2.225206e-02	0.1	0.5690818
KEGG04210	44	27.2993520	5.138148e-05	5.369577e-04	0.1	0.5690818
KEGG04620	51	48.0579425	3.590588e-07	9.827484e-06	0.1	0.5690818
KEGG04920	30	57.3856318	5.693675e-08	2.727141e-06	0.1	0.5690818
KEGG05222	54	44.6152606	7.464344e-07	1.430100e-05	0.1	0.5690818
KEGG05110	15	13.1124654	4.359517e-03	2.784146e-02	0.1	0.5690818
KEGG00500	16	18.0915387	8.045080e-04	6.422350e-03	0.1	0.5690818
KEGG00010	36	8.5208560	2.429027e-02	1.454311e-01	0.1	0.5690818
KEGG00030	15	13.5067464	3.790243e-03	2.447784e-02	0.1	0.5690818
KEGG00051	17	26.6553960	6.144979e-05	6.307082e-04	0.1	0.5690818
KEGG00710	12	6.0223686	6.673974e-02	3.798038e-01	0.1	0.5690818
KEGG04910	59	26.2009678	6.979813e-05	6.916906e-04	0.1	0.5690818
KEGG04630	58	45.1667624	6.624832e-07	1.359919e-05	0.1	0.5690818
KEGG00860	15	51.6866136	1.713804e-07	6.566983e-06	0.1	0.5690818
KEGG00071	19	39.1834195	2.530215e-06	3.930537e-05	0.1	0.5690818
KEGG00310	13	28.9303577	3.291981e-05	3.638732e-04	0.1	0.5690818
KEGG00410	13	46.6612263	4.814060e-07	1.190747e-05	0.1	0.5690818
KEGG00640	16	49.1083172	2.889241e-07	9.225869e-06	0.1	0.5690818
KEGG00650	16	15.9809922	1.614410e-03	1.174581e-02	0.1	0.5690818
KEGG04720	38	14.2261093	2.944604e-03	1.979383e-02	0.1	0.5690818
KEGG04930	18	17.4669584	9.858206e-04	7.657064e-03	0.1	0.5690818
KEGG05060	12	14.2363324	2.934135e-03	1.979383e-02	0.1	0.5690818
KEGG04150	18	11.0095598	9.376387e-03	5.857925e-02	0.1	0.5690818
KEGG04742	10	9.1651073	1.889037e-02	1.155071e-01	0.1	0.5690818
KEGG05050	11	7.6809163	3.389911e-02	1.988192e-01	0.1	0.5690818
KEGG00252	15	20.6683819	3.563113e-04	3.150738e-03	0.1	0.5690818
KEGG00360	11	38.9524791	2.670169e-06	4.038790e-05	0.1	0.5690818
KEGG01030	19	16.3235154	1.439122e-03	1.060472e-02	0.1	0.5690818
KEGG01032	10	16.6323921	1.298268e-03	9.818539e-03	0.1	0.5690818
KEGG00562	14	19.0212991	5.970409e-04	4.833278e-03	0.1	0.5690818
KEGG00480	11	72.7397595	3.967321e-09	4.560615e-07	0.1	0.5690818
KEGG04012	39	21.8088717	2.514144e-04	2.293748e-03	0.1	0.5690818

```
KEGG04740  10  14.8878889 2.341753e-03 1.661699e-02     0.1   0.5690818
KEGG00052  15  19.8497404 4.596460e-04 3.885176e-03     0.1   0.5690818
KEGG05213  29  26.2924394 6.802606e-05 6.859565e-04     0.1   0.5690818
KEGG05219  24  48.8277747 3.061100e-07 9.260192e-06     0.1   0.5690818
KEGG05223  32  17.1073827 1.109387e-03 8.501937e-03     0.1   0.5690818
KEGG00620  16  21.6691227 2.622921e-04 2.355599e-03     0.1   0.5690818
KEGG00970  16  23.4033917 1.561698e-04 1.447776e-03     0.1   0.5690818
KEGG05030  15  28.1502025 4.067506e-05 4.411116e-04     0.1   0.5690818
KEGG05211  34   3.4775298 1.991449e-01 1.000000e+00     0.2   1.0000000
KEGG05130  26   4.3772397 1.342465e-01 7.348685e-01     0.2   1.0000000
KEGG05131  26   4.3772397 1.342465e-01 7.348685e-01     0.2   1.0000000
KEGG00251  13   6.4285217 5.640018e-02 3.274473e-01     0.2   1.0000000
KEGG00510  13   8.0054388 2.978054e-02 1.764643e-01     0.2   1.0000000
KEGG01430  35   2.7477594 2.760440e-01 1.000000e+00     0.4   1.0000000
KEGG04320  10   1.2012903 5.630263e-01 1.000000e+00     0.7   1.0000000
KEGG00530  10   0.3799859 8.321461e-01 1.000000e+00     0.8   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

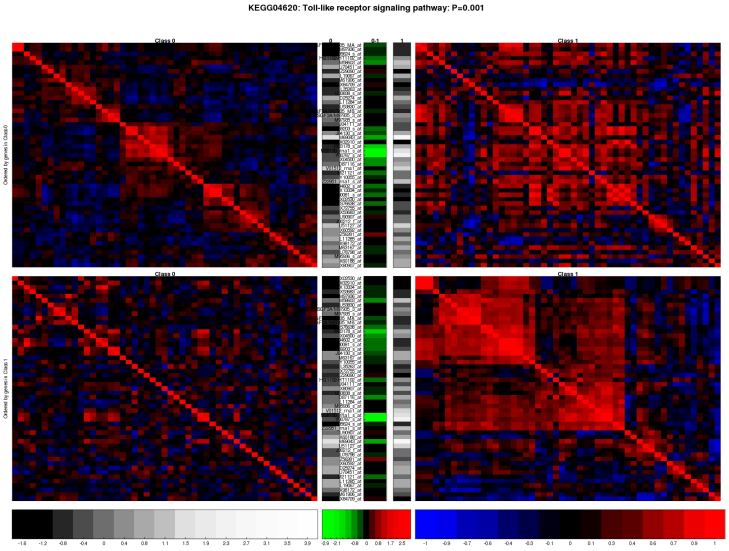

Figure 1: KEGG04620

`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)
> 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.

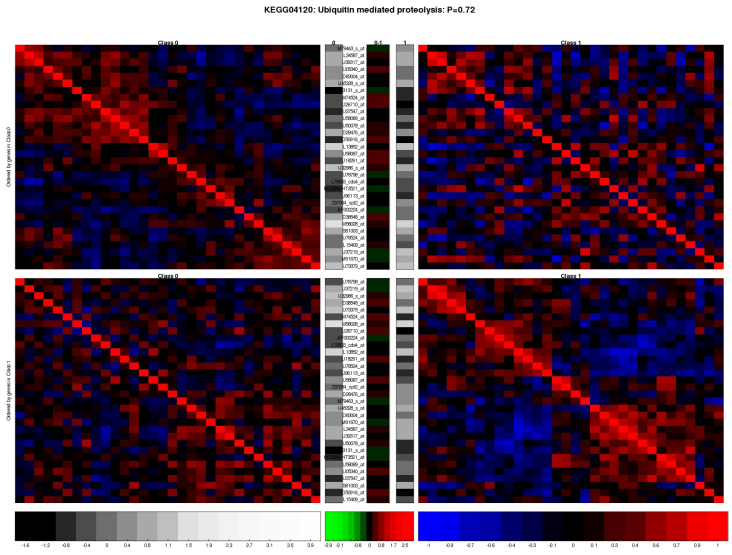

Figure 2: KEGG04120

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))
> 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] 2755   38

> dim(newimat)

[1] 2755  105
```

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.