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

Sarah Song and Mik Black

April 21, 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 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 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  118
```

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  57  52.9762487 1.327386e-07 5.239453e-06     0.1   0.5794051
KEGG04360  33  39.5598135 2.318480e-06 3.754461e-05     0.1   0.5794051
KEGG04010 106  41.9142710 1.355177e-06 2.445326e-05     0.1   0.5794051
KEGG04910  58  26.3048560 6.778921e-05 7.281689e-04     0.1   0.5794051
KEGG03410  16  31.9695622 1.478395e-05 1.905475e-04     0.1   0.5794051
KEGG04650  68  45.9773494 5.567382e-07 1.255746e-05     0.1   0.5794051
KEGG05322  55  51.7830882 1.681140e-07 5.903051e-06     0.1   0.5794051
KEGG04510  86  68.3488680 8.161280e-09 6.442836e-07     0.1   0.5794051
KEGG04810  90  48.7638352 3.101761e-07 8.904190e-06     0.1   0.5794051
KEGG04520  36  23.8052982 1.387873e-04 1.369552e-03     0.1   0.5794051
KEGG04670  57  37.1667546 4.071950e-06 5.807999e-05     0.1   0.5794051
KEGG04060  88  51.5084222 1.775912e-07 5.903051e-06     0.1   0.5794051
KEGG03050  23  26.8941073 5.749360e-05 6.483959e-04     0.1   0.5794051
KEGG04110  51  46.1841153 5.327283e-07 1.246094e-05     0.1   0.5794051
KEGG03320  20  53.2014932 1.269935e-07 5.239453e-06     0.1   0.5794051
KEGG05110  33  25.2478777 9.145353e-05 9.315745e-04     0.1   0.5794051
KEGG00190  42  14.2095556 2.961639e-03 1.989817e-02     0.1   0.5794051
KEGG05010  72  16.5669890 1.326828e-03 9.743721e-03     0.1   0.5794051
KEGG05012  45  12.3080613 5.820370e-03 3.750881e-02     0.1   0.5794051
KEGG04020  62  42.2997618 1.243043e-06 2.308957e-05     0.1   0.5794051
KEGG00350  11   5.1429840 9.663123e-02 5.497982e-01     0.1   0.5794051
KEGG04514  72  25.5917134 8.291894e-05 8.584850e-04     0.1   0.5794051
KEGG04530  38  31.5999193 1.626431e-05 2.054350e-04     0.1   0.5794051
KEGG03430  13  22.8407563 1.844695e-04 1.713264e-03     0.1   0.5794051
KEGG05210  43  27.6220143 4.700716e-05 5.397717e-04     0.1   0.5794051
KEGG05213  29  26.2924394 6.802606e-05 7.281689e-04     0.1   0.5794051
KEGG04120  32  14.6056854 2.581106e-03 1.811224e-02     0.1   0.5794051
KEGG04210  43  26.3121240 6.765097e-05 7.281689e-04     0.1   0.5794051
KEGG05014  27  34.9064115 7.051758e-06 9.475635e-05     0.1   0.5794051
KEGG04115  24  37.0991286 4.138379e-06 5.807999e-05     0.1   0.5794051
KEGG04916  33  16.5292411 1.343613e-03 9.753570e-03     0.1   0.5794051
KEGG05215  49  54.5238031 9.816234e-08 4.428184e-06     0.1   0.5794051
KEGG04310  47  41.7662229 1.401042e-06 2.457862e-05     0.1   0.5794051
KEGG04350  27  23.1228690 1.696573e-04 1.599214e-03     0.1   0.5794051
KEGG00010  37   8.4805948 2.467792e-02 1.484322e-01     0.1   0.5794051
KEGG05040  21  13.8093276 3.406880e-03 2.264864e-02     0.1   0.5794051
KEGG05050  11   7.6809163 3.389911e-02 1.964132e-01     0.1   0.5794051
KEGG04620  50  47.5933174 3.956320e-07 1.041092e-05     0.1   0.5794051
KEGG04630  59  45.6261682 6.001766e-07 1.307043e-05     0.1   0.5794051
KEGG05212  47  28.7540432 3.452416e-05 4.037739e-04     0.1   0.5794051
KEGG04640  68 123.0170433 5.129008e-12 3.239233e-09     0.1   0.5794051
KEGG01032  10  16.6323921 1.298268e-03 9.646155e-03     0.1   0.5794051
KEGG00980  13  69.1882122 7.093654e-09 6.400011e-07     0.1   0.5794051
```

KEGG00982	12	57.3950181	5.683670e-08	2.761177e-06	0.1	0.5794051
KEGG00983	17	35.3186327	6.371492e-06	8.747664e-05	0.1	0.5794051
KEGG00240	32	58.9786441	4.234863e-08	2.228779e-06	0.1	0.5794051
KEGG00480	16	77.0823630	1.999325e-09	3.156692e-07	0.1	0.5794051
KEGG00590	20	44.3957904	7.829082e-07	1.545146e-05	0.1	0.5794051
KEGG00860	15	51.6866136	1.713804e-07	5.903051e-06	0.1	0.5794051
KEGG00030	15	13.5067464	3.790243e-03	2.493472e-02	0.1	0.5794051
KEGG00230	52	19.2543394	5.544749e-04	4.547786e-03	0.1	0.5794051
KEGG00071	19	39.1834195	2.530215e-06	3.994903e-05	0.1	0.5794051
KEGG04920	27	62.4466583	2.260875e-08	1.427859e-06	0.1	0.5794051
KEGG00620	16	21.6691227	2.622921e-04	2.333112e-03	0.1	0.5794051
KEGG04930	17	17.1390667	1.097868e-03	8.340893e-03	0.1	0.5794051
KEGG04664	38	61.3798116	2.735820e-08	1.570737e-06	0.1	0.5794051
KEGG04912	38	15.9191093	1.648417e-03	1.169731e-02	0.1	0.5794051
KEGG00280	21	40.9446790	1.687207e-06	2.879887e-05	0.1	0.5794051
KEGG00310	14	28.9221170	3.299301e-05	3.931469e-04	0.1	0.5794051
KEGG00380	17	94.7362550	1.578930e-10	4.985879e-08	0.1	0.5794051
KEGG00640	16	49.1083172	2.889241e-07	8.904190e-06	0.1	0.5794051
KEGG00650	14	19.2545623	5.544358e-04	4.547786e-03	0.1	0.5794051
KEGG00020	14	13.1529665	4.297080e-03	2.797760e-02	0.1	0.5794051
KEGG04012	39	21.8088717	2.514144e-04	2.268302e-03	0.1	0.5794051
KEGG05220	51	38.9558382	2.668076e-06	4.109822e-05	0.1	0.5794051
KEGG00260	12	9.0142092	2.002974e-02	1.216328e-01	0.1	0.5794051
KEGG00564	10	42.5165749	1.184323e-06	2.266549e-05	0.1	0.5794051
KEGG05340	27	90.3005167	2.888710e-10	6.081232e-08	0.1	0.5794051
KEGG00500	13	30.3529470	2.252890e-05	2.789835e-04	0.1	0.5794051
KEGG05120	36	64.7251734	1.514880e-08	1.063028e-06	0.1	0.5794051
KEGG04660	45	17.1317622	1.100512e-03	8.340893e-03	0.1	0.5794051
KEGG01030	19	16.3235154	1.439122e-03	1.032818e-02	0.1	0.5794051
KEGG00410	13	46.6612263	4.814060e-07	1.169357e-05	0.1	0.5794051
KEGG03420	17	17.4935061	9.772938e-04	7.675880e-03	0.1	0.5794051
KEGG05221	41	37.6997502	3.586233e-06	5.267188e-05	0.1	0.5794051
KEGG04340	11	6.0731284	6.534459e-02	3.751680e-01	0.1	0.5794051
KEGG05218	32	19.2120700	5.619507e-04	4.550011e-03	0.1	0.5794051
KEGG04512	31	48.4096545	3.337579e-07	9.164580e-06	0.1	0.5794051
KEGG05222	53	44.6894003	7.345227e-07	1.496416e-05	0.1	0.5794051
KEGG04610	15	73.3638672	3.589230e-09	4.533567e-07	0.1	0.5794051
KEGG03030	21	22.3959686	2.106656e-04	1.928205e-03	0.1	0.5794051
KEGG00970	16	23.4033917	1.561698e-04	1.494383e-03	0.1	0.5794051
KEGG04370	37	32.2815602	1.364532e-05	1.795359e-04	0.1	0.5794051
KEGG04662	41	47.0274568	4.455674e-07	1.125595e-05	0.1	0.5794051
KEGG00051	18	24.8562802	1.023173e-04	1.025693e-03	0.1	0.5794051
KEGG00052	15	19.8497404	4.596460e-04	3.870535e-03	0.1	0.5794051
KEGG04540	41	10.8912732	9.799036e-03	6.067251e-02	0.1	0.5794051
KEGG04070	31	25.7760641	7.869364e-05	8.283182e-04	0.1	0.5794051
KEGG04720	39	14.3691931	2.801602e-03	1.923213e-02	0.1	0.5794051
KEGG04730	35	45.4503098	6.232622e-07	1.312074e-05	0.1	0.5794051
KEGG00561	16	69.2425821	7.029794e-09	6.400011e-07	0.1	0.5794051
KEGG00330	13	17.4711336	9.844744e-04	7.675880e-03	0.1	0.5794051
KEGG05310	27	19.9578105	4.443562e-04	3.792349e-03	0.1	0.5794051

KEGG00252	12	21.4014845	2.845372e-04	2.495832e-03	0.1	0.5794051
KEGG04612	55	40.6888219	1.788474e-06	2.972404e-05	0.1	0.5794051
KEGG04940	34	7.7792798	3.259065e-02	1.905803e-01	0.1	0.5794051
KEGG05332	35	11.6095635	7.510092e-03	4.743010e-02	0.1	0.5794051
KEGG05214	41	20.5232602	3.726642e-04	3.224064e-03	0.1	0.5794051
KEGG05219	24	48.8277747	3.061100e-07	8.904190e-06	0.1	0.5794051
KEGG05223	32	17.1073827	1.109387e-03	8.340893e-03	0.1	0.5794051
KEGG04330	15	14.4138200	2.758517e-03	1.914446e-02	0.1	0.5794051
KEGG04150	18	11.0095598	9.376387e-03	5.863041e-02	0.1	0.5794051
KEGG00220	12	38.2376153	3.157719e-06	4.748244e-05	0.1	0.5794051
KEGG03022	12	23.6751657	1.441801e-04	1.400879e-03	0.1	0.5794051
KEGG05216	22	29.2717954	3.003285e-05	3.647556e-04	0.1	0.5794051
KEGG04740	13	11.9425590	6.647718e-03	4.240784e-02	0.1	0.5794051
KEGG00562	14	19.0212991	5.970409e-04	4.772938e-03	0.1	0.5794051
KEGG04742	10	9.1651073	1.889037e-02	1.158276e-01	0.1	0.5794051
KEGG05060	12	14.2363324	2.934135e-03	1.989817e-02	0.1	0.5794051
KEGG00510	13	8.0054388	2.978054e-02	1.757752e-01	0.2	1.0000000
KEGG05130	26	4.3772397	1.342465e-01	7.502969e-01	0.2	1.0000000
KEGG05131	26	4.3772397	1.342465e-01	7.502969e-01	0.2	1.0000000
KEGG05211	34	3.4775298	1.991449e-01	1.000000e+00	0.2	1.0000000
KEGG00251	12	8.1437353	2.818934e-02	1.679530e-01	0.2	1.0000000
KEGG01430	35	2.7477594	2.760440e-01	1.000000e+00	0.4	1.0000000
KEGG00530	11	0.3856455	8.298838e-01	1.000000e+00	0.7	1.0000000
KEGG05330	34	1.5433630	4.796928e-01	1.000000e+00	0.8	1.0000000
KEGG05320	35	0.2910241	8.685750e-01	1.000000e+00	1.0	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).

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

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

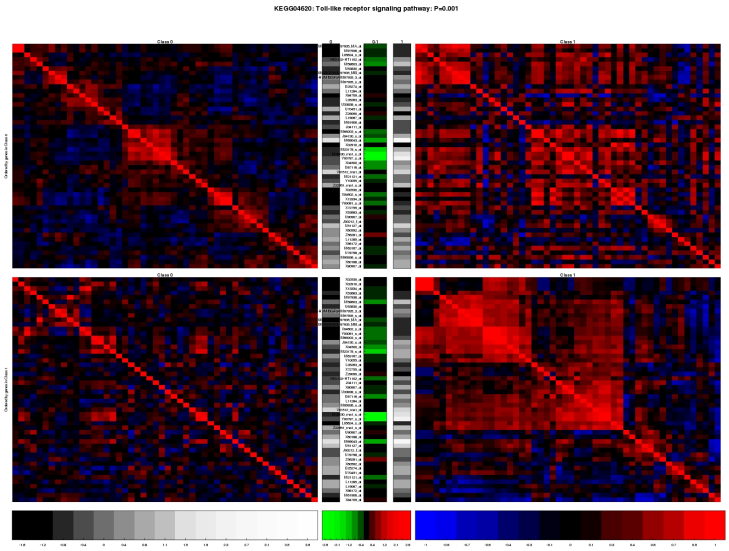

Figure 1: KEGG04620

`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.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()
+ }
```

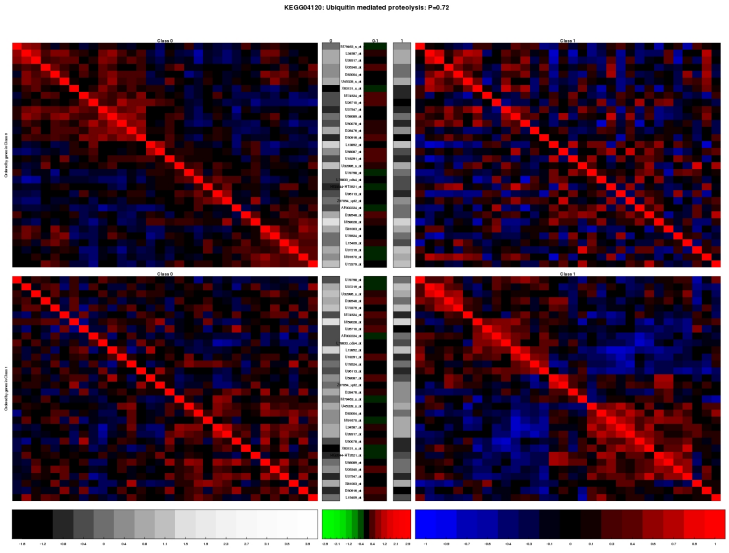

Figure 2: KEGG04120

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.

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] 2747   38

> dim(newimat)

[1] 2747  116
```

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.