

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

Sarah Song and Mik Black

April 25, 2007

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` 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

```
> 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
KEGG03050	16	41.184124	1.597917e-06	2.994971e-05	0.1	0.560357
KEGG00980	13	69.188212	7.093654e-09	4.819661e-07	0.1	0.560357
KEGG04010	96	36.703744	4.550407e-06	6.183392e-05	0.1	0.560357
KEGG04210	44	27.299352	5.138148e-05	5.585643e-04	0.1	0.560357
KEGG04620	46	39.267674	2.481105e-06	4.167543e-05	0.1	0.560357
KEGG04660	43	33.338131	1.042993e-05	1.288443e-04	0.1	0.560357
KEGG04662	39	46.757474	4.717016e-07	1.137680e-05	0.1	0.560357
KEGG04920	30	57.385632	5.693675e-08	2.578980e-06	0.1	0.560357
KEGG05120	38	71.943029	4.512199e-09	3.503699e-07	0.1	0.560357
KEGG05212	47	30.445183	2.198784e-05	2.598132e-04	0.1	0.560357
KEGG05220	51	40.088079	2.052357e-06	3.718503e-05	0.1	0.560357
KEGG04510	87	46.068490	5.460168e-07	1.196885e-05	0.1	0.560357
KEGG04512	32	50.317318	2.257251e-07	6.457476e-06	0.1	0.560357
KEGG04640	70	115.400542	1.210776e-11	6.581128e-09	0.1	0.560357
KEGG04810	90	52.782602	1.378974e-07	4.996907e-06	0.1	0.560357
KEGG01032	11	16.083606	1.559644e-03	1.211055e-02	0.1	0.560357
KEGG04060	86	50.940814	1.990119e-07	6.009564e-06	0.1	0.560357
KEGG04612	53	75.705838	2.477228e-09	3.503699e-07	0.1	0.560357
KEGG04630	56	38.326364	3.092383e-06	4.638829e-05	0.1	0.560357
KEGG04650	70	45.580783	6.060466e-07	1.220053e-05	0.1	0.560357
KEGG04670	63	102.785352	5.549838e-11	1.508297e-08	0.1	0.560357
KEGG04720	38	14.226109	2.944604e-03	2.088827e-02	0.1	0.560357
KEGG01510	27	14.133816	3.040922e-03	2.119080e-02	0.1	0.560357
KEGG05040	21	13.809328	3.406880e-03	2.344047e-02	0.1	0.560357
KEGG04020	62	42.299762	1.243043e-06	2.413040e-05	0.1	0.560357
KEGG03022	13	23.820339	1.381779e-04	1.390853e-03	0.1	0.560357
KEGG00230	54	18.234794	7.681197e-04	6.325888e-03	0.1	0.560357
KEGG00240	32	58.978644	4.234863e-08	2.301844e-06	0.1	0.560357
KEGG05010	16	5.728149	7.547169e-02	4.143673e-01	0.1	0.560357
KEGG04742	10	9.165107	1.889037e-02	1.128329e-01	0.1	0.560357
KEGG00500	18	18.283394	7.561735e-04	6.323312e-03	0.1	0.560357
KEGG04080	73	38.688571	2.840182e-06	4.410772e-05	0.1	0.560357
KEGG04530	41	33.404722	1.025619e-05	1.288443e-04	0.1	0.560357
KEGG04730	36	37.999497	3.340337e-06	4.777968e-05	0.1	0.560357
KEGG04910	59	26.200968	6.979813e-05	7.438924e-04	0.1	0.560357
KEGG05210	44	27.553870	4.789705e-05	5.313115e-04	0.1	0.560357
KEGG05214	39	19.240182	5.569671e-04	5.045623e-03	0.1	0.560357
KEGG05218	30	18.611644	6.804589e-04	5.870808e-03	0.1	0.560357
KEGG00190	43	14.212036	2.959080e-03	2.088827e-02	0.1	0.560357
KEGG00010	36	8.520856	2.429027e-02	1.419665e-01	0.1	0.560357
KEGG00030	15	13.506746	3.790243e-03	2.575215e-02	0.1	0.560357
KEGG00051	18	18.678736	6.659944e-04	5.838690e-03	0.1	0.560357
KEGG00710	12	6.022369	6.673974e-02	3.701647e-01	0.1	0.560357

KEGG04514	74	31.010802	1.895656e-05	2.289726e-04	0.1	0.560357
KEGG04350	26	23.712919	1.425931e-04	1.409199e-03	0.1	0.560357
KEGG04520	37	23.992055	1.314175e-04	1.347764e-03	0.1	0.560357
KEGG04310	42	37.040273	4.197126e-06	5.849570e-05	0.1	0.560357
KEGG05216	24	9.623252	1.583196e-02	9.561558e-02	0.1	0.560357
KEGG04610	15	73.363867	3.589229e-09	3.503699e-07	0.1	0.560357
KEGG00760	11	53.597908	1.175117e-07	4.913311e-06	0.1	0.560357
KEGG04540	40	10.819786	1.006419e-02	6.310966e-02	0.1	0.560357
KEGG04912	38	15.919109	1.648417e-03	1.244432e-02	0.1	0.560357
KEGG00590	19	47.576677	3.970128e-07	1.027595e-05	0.1	0.560357
KEGG04916	33	16.529241	1.343613e-03	1.058428e-02	0.1	0.560357
KEGG04370	33	18.512589	7.024245e-04	5.965628e-03	0.1	0.560357
KEGG04664	38	61.379812	2.735820e-08	1.652272e-06	0.1	0.560357
KEGG04340	11	6.073128	6.534459e-02	3.661630e-01	0.1	0.560357
KEGG00071	19	39.183420	2.530215e-06	4.167543e-05	0.1	0.560357
KEGG00280	22	45.846527	5.725182e-07	1.196885e-05	0.1	0.560357
KEGG00310	15	33.903265	9.048742e-06	1.171050e-04	0.1	0.560357
KEGG00380	23	74.540039	2.976464e-09	3.503699e-07	0.1	0.560357
KEGG00410	13	46.661226	4.814060e-07	1.137680e-05	0.1	0.560357
KEGG00640	17	48.892360	3.020600e-07	8.209179e-06	0.1	0.560357
KEGG00650	16	15.980992	1.614410e-03	1.235925e-02	0.1	0.560357
KEGG00562	14	19.021299	5.970409e-04	5.319989e-03	0.1	0.560357
KEGG04070	31	25.776064	7.869364e-05	8.225700e-04	0.1	0.560357
KEGG05030	15	28.150202	4.067506e-05	4.605995e-04	0.1	0.560357
KEGG00350	15	4.806800	1.115512e-01	6.063325e-01	0.1	0.560357
KEGG00561	17	58.313748	4.788850e-08	2.366329e-06	0.1	0.560357
KEGG00564	11	45.847691	5.723758e-07	1.196885e-05	0.1	0.560357
KEGG04940	36	34.444566	7.906105e-06	1.048130e-04	0.1	0.560357
KEGG03320	20	53.201493	1.269935e-07	4.930490e-06	0.1	0.560357
KEGG01030	18	12.723491	5.010394e-03	3.321196e-02	0.1	0.560357
KEGG00620	16	21.669123	2.622921e-04	2.501191e-03	0.1	0.560357
KEGG00860	15	51.686614	1.713804e-07	5.479597e-06	0.1	0.560357
KEGG04360	32	39.327865	2.446638e-06	4.167543e-05	0.1	0.560357
KEGG04740	10	14.887889	2.341753e-03	1.743632e-02	0.1	0.560357
KEGG04930	18	17.466958	9.858206e-04	7.879987e-03	0.1	0.560357
KEGG04110	51	51.947868	1.626862e-07	5.479597e-06	0.1	0.560357
KEGG05130	27	9.739467	1.514257e-02	9.247963e-02	0.1	0.560357
KEGG05131	27	9.739467	1.514257e-02	9.247963e-02	0.1	0.560357
KEGG00330	13	17.471134	9.844744e-04	7.879987e-03	0.1	0.560357
KEGG00220	12	38.237615	3.157719e-06	4.638829e-05	0.1	0.560357
KEGG00260	12	9.014209	2.002974e-02	1.183379e-01	0.1	0.560357
KEGG00340	10	29.388643	2.910738e-05	3.366215e-04	0.1	0.560357
KEGG00360	11	38.952479	2.670169e-06	4.268707e-05	0.1	0.560357
KEGG00252	15	20.668382	3.563113e-04	3.339167e-03	0.1	0.560357
KEGG00970	16	23.403392	1.561698e-04	1.515813e-03	0.1	0.560357
KEGG00150	10	11.673533	7.335679e-03	4.746763e-02	0.1	0.560357
KEGG04330	15	14.413820	2.758517e-03	2.026191e-02	0.1	0.560357
KEGG00052	15	19.849740	4.596460e-04	4.234557e-03	0.1	0.560357
KEGG00480	11	72.739759	3.967321e-09	3.503699e-07	0.1	0.560357
KEGG05050	11	7.680916	3.389911e-02	1.960185e-01	0.1	0.560357

KEGG04150	18	11.009560	9.376387e-03	5.995883e-02	0.1	0.560357
KEGG05110	15	13.112465	4.359517e-03	2.925431e-02	0.1	0.560357
KEGG05060	12	14.236332	2.934135e-03	2.088827e-02	0.1	0.560357
KEGG00510	13	10.809932	1.010133e-02	6.310966e-02	0.1	0.560357
KEGG00020	12	12.207513	6.036512e-03	3.953161e-02	0.2	1.000000
KEGG00251	13	6.428522	5.640018e-02	3.226959e-01	0.2	1.000000
KEGG04120	12	6.182388	6.244511e-02	3.535605e-01	0.2	1.000000
KEGG01031	10	1.390378	5.152243e-01	1.000000e+00	0.4	1.000000
KEGG00530	11	1.535339	4.814905e-01	1.000000e+00	0.4	1.000000
KEGG01430	34	2.615432	2.930746e-01	1.000000e+00	0.5	1.000000
KEGG04320	10	1.201290	5.630263e-01	1.000000e+00	0.7	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 principle 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.

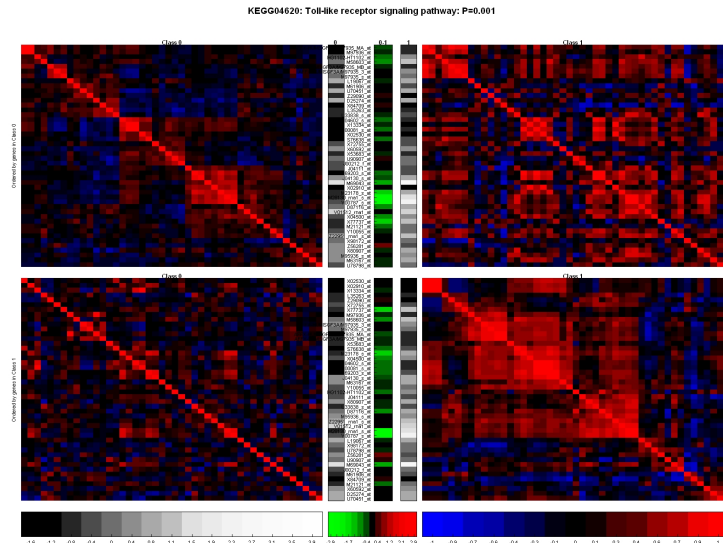


Figure 1: KEGG04620

```
> 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.

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

- [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.