

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

Sarah Song and Mik Black

October 22, 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 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.

```
> 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


```
> 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	5.292152e-06	0.1	0.5799124
KEGG04360	33	39.5598135	2.318480e-06	3.792223e-05	0.1	0.5799124
KEGG04010	102	39.3554517	2.431011e-06	3.876877e-05	0.1	0.5799124
KEGG04910	59	26.2009678	6.979813e-05	7.420747e-04	0.1	0.5799124
KEGG03410	16	31.9695622	1.478395e-05	1.886147e-04	0.1	0.5799124
KEGG04650	68	45.9773494	5.567382e-07	1.210864e-05	0.1	0.5799124
KEGG04510	85	65.8760447	1.241745e-08	8.801264e-07	0.1	0.5799124
KEGG04810	90	48.7638352	3.101761e-07	8.993748e-06	0.1	0.5799124
KEGG04520	36	23.8052982	1.387873e-04	1.383327e-03	0.1	0.5799124
KEGG04670	57	37.1667546	4.071950e-06	5.820352e-05	0.1	0.5799124
KEGG04060	87	50.9617543	1.981749e-07	6.653500e-06	0.1	0.5799124
KEGG03050	12	46.5088461	4.972042e-07	1.174698e-05	0.1	0.5799124
KEGG04110	51	46.1841153	5.327283e-07	1.210864e-05	0.1	0.5799124
KEGG03320	20	53.2014932	1.269935e-07	5.292152e-06	0.1	0.5799124
KEGG05110	33	25.2478777	9.145353e-05	9.260086e-04	0.1	0.5799124
KEGG00190	42	14.2095556	2.961639e-03	2.031441e-02	0.1	0.5799124
KEGG04020	62	42.2997618	1.243043e-06	2.265547e-05	0.1	0.5799124
KEGG00350	14	4.6547487	1.190821e-01	6.782400e-01	0.1	0.5799124
KEGG04514	72	25.5917134	8.291894e-05	8.531338e-04	0.1	0.5799124
KEGG04530	39	31.3633894	1.729318e-05	2.163017e-04	0.1	0.5799124
KEGG03430	13	22.8407563	1.844695e-04	1.756324e-03	0.1	0.5799124
KEGG05210	43	27.6220143	4.700716e-05	5.354650e-04	0.1	0.5799124
KEGG05213	29	26.2924394	6.802606e-05	7.354928e-04	0.1	0.5799124
KEGG04120	32	14.6056854	2.581106e-03	1.849996e-02	0.1	0.5799124
KEGG04210	44	27.2993520	5.138148e-05	5.750252e-04	0.1	0.5799124
KEGG04115	24	37.0991286	4.138379e-06	5.820352e-05	0.1	0.5799124
KEGG04916	33	16.5292411	1.343613e-03	1.008347e-02	0.1	0.5799124
KEGG05215	49	53.5662225	1.182413e-07	5.292152e-06	0.1	0.5799124
KEGG04310	42	37.0402727	4.197126e-06	5.820352e-05	0.1	0.5799124
KEGG04350	28	20.1492856	4.185715e-04	3.657648e-03	0.1	0.5799124
KEGG00380	19	88.6534920	3.634677e-10	7.728579e-08	0.1	0.5799124
KEGG00010	36	8.5208560	2.429027e-02	1.489890e-01	0.1	0.5799124
KEGG01510	27	14.1338165	3.040922e-03	2.063634e-02	0.1	0.5799124
KEGG05010	16	5.7281493	7.547169e-02	4.337268e-01	0.1	0.5799124
KEGG05040	21	13.8093276	3.406880e-03	2.287644e-02	0.1	0.5799124
KEGG05050	11	7.6809163	3.389911e-02	2.020969e-01	0.1	0.5799124
KEGG04620	51	48.0579425	3.590588e-07	9.543538e-06	0.1	0.5799124
KEGG04630	59	45.6261682	6.001766e-07	1.235016e-05	0.1	0.5799124
KEGG05212	48	30.3992536	2.225552e-05	2.730169e-04	0.1	0.5799124
KEGG04640	68	123.0170433	5.129008e-12	3.271813e-09	0.1	0.5799124
KEGG01032	10	16.6323921	1.298268e-03	9.859166e-03	0.1	0.5799124
KEGG00980	13	69.1882122	7.093654e-09	7.541779e-07	0.1	0.5799124
KEGG00982	12	57.3950181	5.683670e-08	2.793859e-06	0.1	0.5799124

KEGG00983	17	35.3186327	6.371492e-06	8.647655e-05	0.1	0.5799124
KEGG00240	32	58.9786441	4.234863e-08	2.455850e-06	0.1	0.5799124
KEGG00480	10	52.0086356	1.607316e-07	6.031251e-06	0.1	0.5799124
KEGG00590	20	44.3957904	7.829082e-07	1.513394e-05	0.1	0.5799124
KEGG00860	15	51.6866136	1.713804e-07	6.073564e-06	0.1	0.5799124
KEGG00030	15	13.5067464	3.790243e-03	2.518552e-02	0.1	0.5799124
KEGG00230	52	19.2543394	5.544749e-04	4.653968e-03	0.1	0.5799124
KEGG00071	19	39.1834195	2.530215e-06	3.936667e-05	0.1	0.5799124
KEGG04920	30	57.3856318	5.693675e-08	2.793859e-06	0.1	0.5799124
KEGG00620	16	21.6691227	2.622921e-04	2.390244e-03	0.1	0.5799124
KEGG00710	12	6.0223686	6.673974e-02	3.870321e-01	0.1	0.5799124
KEGG04930	18	17.4669584	9.858206e-04	7.669007e-03	0.1	0.5799124
KEGG04664	38	61.3798116	2.735820e-08	1.745189e-06	0.1	0.5799124
KEGG04912	38	15.9191093	1.648417e-03	1.194922e-02	0.1	0.5799124
KEGG00280	21	40.9446790	1.687207e-06	2.989654e-05	0.1	0.5799124
KEGG00310	13	28.9303577	3.291981e-05	3.888827e-04	0.1	0.5799124
KEGG00640	16	49.1083172	2.889241e-07	8.993748e-06	0.1	0.5799124
KEGG00650	15	16.2327617	1.483513e-03	1.087746e-02	0.1	0.5799124
KEGG00020	12	12.2075129	6.036512e-03	3.969807e-02	0.1	0.5799124
KEGG04012	39	21.8088717	2.514144e-04	2.324321e-03	0.1	0.5799124
KEGG05220	52	40.1093857	2.042326e-06	3.428439e-05	0.1	0.5799124
KEGG00260	12	9.0142092	2.002974e-02	1.240490e-01	0.1	0.5799124
KEGG00564	10	45.8715971	5.694580e-07	1.210864e-05	0.1	0.5799124
KEGG05340	27	90.3005167	2.888710e-10	7.728579e-08	0.1	0.5799124
KEGG00500	16	18.0915387	8.045080e-04	6.496185e-03	0.1	0.5799124
KEGG05120	37	66.0776488	1.199519e-08	8.801264e-07	0.1	0.5799124
KEGG04660	43	33.3381307	1.042993e-05	1.386103e-04	0.1	0.5799124
KEGG01030	19	16.3235154	1.439122e-03	1.067467e-02	0.1	0.5799124
KEGG00410	13	46.6612263	4.814060e-07	1.174698e-05	0.1	0.5799124
KEGG03420	17	17.4935061	9.772938e-04	7.669007e-03	0.1	0.5799124
KEGG05221	41	42.9729541	1.070094e-06	2.007697e-05	0.1	0.5799124
KEGG04340	11	6.0731284	6.534459e-02	3.824179e-01	0.1	0.5799124
KEGG05218	32	19.2120700	5.619507e-04	4.655460e-03	0.1	0.5799124
KEGG04512	31	48.4096545	3.337579e-07	9.256757e-06	0.1	0.5799124
KEGG05222	54	44.6152606	7.464344e-07	1.487979e-05	0.1	0.5799124
KEGG04610	15	73.3638672	3.589230e-09	5.723957e-07	0.1	0.5799124
KEGG03030	21	22.3959686	2.106656e-04	1.976240e-03	0.1	0.5799124
KEGG00970	16	23.4033917	1.561698e-04	1.509413e-03	0.1	0.5799124
KEGG04370	37	32.2815602	1.364532e-05	1.776408e-04	0.1	0.5799124
KEGG04662	39	46.7574737	4.717016e-07	1.174698e-05	0.1	0.5799124
KEGG05030	15	28.1502025	4.067506e-05	4.717595e-04	0.1	0.5799124
KEGG00051	17	26.6553960	6.144979e-05	6.758456e-04	0.1	0.5799124
KEGG00052	15	19.8497404	4.596460e-04	3.909465e-03	0.1	0.5799124
KEGG04540	41	10.8912732	9.799036e-03	6.188951e-02	0.1	0.5799124
KEGG04070	31	25.7760641	7.869364e-05	8.229338e-04	0.1	0.5799124
KEGG04720	39	14.3691931	2.801602e-03	1.963903e-02	0.1	0.5799124
KEGG04730	36	37.9994969	3.340337e-06	4.955380e-05	0.1	0.5799124
KEGG00561	16	69.2425821	7.029794e-09	7.541779e-07	0.1	0.5799124
KEGG00330	13	17.4711336	9.844744e-04	7.669007e-03	0.1	0.5799124
KEGG05310	27	19.9578105	4.443562e-04	3.830493e-03	0.1	0.5799124

KEGG05322	41	68.0651582	8.559535e-09	7.800226e-07	0.1	0.5799124
KEGG00252	15	20.6683819	3.563113e-04	3.201299e-03	0.1	0.5799124
KEGG04612	55	40.6888219	1.788474e-06	3.083443e-05	0.1	0.5799124
KEGG04940	34	7.7792798	3.259065e-02	1.961292e-01	0.1	0.5799124
KEGG05332	35	11.6095635	7.510092e-03	4.839106e-02	0.1	0.5799124
KEGG05214	41	20.5232602	3.726642e-04	3.301720e-03	0.1	0.5799124
KEGG05219	24	48.8277747	3.061100e-07	8.993748e-06	0.1	0.5799124
KEGG05223	32	17.1073827	1.109387e-03	8.526289e-03	0.1	0.5799124
KEGG04330	15	14.4138200	2.758517e-03	1.955187e-02	0.1	0.5799124
KEGG04150	18	11.0095598	9.376387e-03	5.981232e-02	0.1	0.5799124
KEGG00220	12	38.2376153	3.157719e-06	4.796002e-05	0.1	0.5799124
KEGG03022	12	23.6751657	1.441801e-04	1.414969e-03	0.1	0.5799124
KEGG05216	22	29.2717954	3.003285e-05	3.614729e-04	0.1	0.5799124
KEGG04740	13	11.9425590	6.647718e-03	4.327146e-02	0.1	0.5799124
KEGG00562	14	19.0212991	5.970409e-04	4.882751e-03	0.1	0.5799124
KEGG04742	10	9.1651073	1.889037e-02	1.181396e-01	0.1	0.5799124
KEGG05060	12	14.2363324	2.934135e-03	2.031441e-02	0.1	0.5799124
KEGG00510	13	8.0054388	2.978054e-02	1.809249e-01	0.2	1.0000000
KEGG05130	26	4.3772397	1.342465e-01	7.511956e-01	0.2	1.0000000
KEGG05131	26	4.3772397	1.342465e-01	7.511956e-01	0.2	1.0000000
KEGG05211	34	3.4775298	1.991449e-01	1.000000e+00	0.2	1.0000000
KEGG00251	13	6.4285217	5.640018e-02	3.331285e-01	0.2	1.0000000
KEGG01430	35	2.7477594	2.760440e-01	1.000000e+00	0.4	1.0000000
KEGG00530	10	0.3799859	8.321461e-01	1.000000e+00	0.8	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 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).

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

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

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.

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] 2748  38

> dim(newimat)

[1] 2748 117

```

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.