

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

Sarah Song and Mik Black

April 15, 2011

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

```
> 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 | 56 | 52.995959 | 1.322251e-07 | 3.368299e-06 | 0.1 | 0.5865721 |
| KEGG04360 | 30 | 35.193509 | 6.570324e-06 | 8.614753e-05 | 0.1 | 0.5865721 |
| KEGG04010 | 97 | 40.757244 | 1.760778e-06 | 2.660833e-05 | 0.1 | 0.5865721 |
| KEGG04910 | 55 | 23.685638 | 1.437380e-04 | 1.307707e-03 | 0.1 | 0.5865721 |
| KEGG03410 | 14 | 40.040059 | 2.075157e-06 | 3.083647e-05 | 0.1 | 0.5865721 |
| KEGG04650 | 58 | 53.920463 | 1.103493e-07 | 3.036523e-06 | 0.1 | 0.5865721 |
| KEGG05322 | 45 | 72.465218 | 4.146772e-09 | 3.081016e-07 | 0.1 | 0.5865721 |
| KEGG04962 | 14 | 25.229090 | 9.194586e-05 | 9.108663e-04 | 0.1 | 0.5865721 |
| KEGG04510 | 79 | 52.030331 | 1.600398e-07 | 3.963605e-06 | 0.1 | 0.5865721 |
| KEGG04270 | 44 | 24.494250 | 1.135788e-04 | 1.077294e-03 | 0.1 | 0.5865721 |
| KEGG04810 | 84 | 42.380783 | 1.220748e-06 | 2.015566e-05 | 0.1 | 0.5865721 |
| KEGG04060 | 83 | 57.649662 | 5.419198e-08 | 1.932680e-06 | 0.1 | 0.5865721 |
| KEGG04062 | 87 | 70.607587 | 5.610503e-09 | 3.573047e-07 | 0.1 | 0.5865721 |
| KEGG03050 | 23 | 26.894107 | 5.749360e-05 | 6.030670e-04 | 0.1 | 0.5865721 |
| KEGG04110 | 57 | 46.327670 | 5.167040e-07 | 1.023751e-05 | 0.1 | 0.5865721 |
| KEGG03320 | 18 | 55.009039 | 8.939526e-08 | 2.748410e-06 | 0.1 | 0.5865721 |
| KEGG04971 | 35 | 20.675694 | 3.555077e-04 | 2.962308e-03 | 0.1 | 0.5865721 |
| KEGG04972 | 35 | 45.037150 | 6.812727e-07 | 1.265449e-05 | 0.1 | 0.5865721 |
| KEGG04976 | 20 | 19.403696 | 5.288905e-04 | 4.248228e-03 | 0.1 | 0.5865721 |
| KEGG05110 | 30 | 24.810971 | 1.036597e-04 | 1.004586e-03 | 0.1 | 0.5865721 |
| KEGG04146 | 20 | 32.548726 | 1.274396e-05 | 1.556491e-04 | 0.1 | 0.5865721 |
| KEGG00190 | 43 | 14.212036 | 2.959080e-03 | 2.077390e-02 | 0.1 | 0.5865721 |
| KEGG01100 | 316 | 68.363930 | 8.140692e-09 | 4.428958e-07 | 0.1 | 0.5865721 |
| KEGG05010 | 68 | 17.102359 | 1.111225e-03 | 8.325687e-03 | 0.1 | 0.5865721 |
| KEGG05012 | 43 | 10.731022 | 1.040403e-02 | 6.871201e-02 | 0.1 | 0.5865721 |
| KEGG05016 | 70 | 31.545201 | 1.649643e-05 | 1.935269e-04 | 0.1 | 0.5865721 |
| KEGG04142 | 53 | 61.157011 | 2.847713e-08 | 1.103909e-06 | 0.1 | 0.5865721 |
| KEGG03420 | 15 | 15.484007 | 1.909975e-03 | 1.395831e-02 | 0.1 | 0.5865721 |
| KEGG04141 | 63 | 41.998038 | 1.329932e-06 | 2.121540e-05 | 0.1 | 0.5865721 |
| KEGG03018 | 13 | 5.431912 | 8.549321e-02 | 4.982017e-01 | 0.1 | 0.5865721 |
| KEGG04144 | 52 | 36.689760 | 4.565751e-06 | 6.167842e-05 | 0.1 | 0.5865721 |
| KEGG04020 | 57 | 34.108620 | 8.595950e-06 | 1.110733e-04 | 0.1 | 0.5865721 |
| KEGG04666 | 43 | 46.505825 | 4.975229e-07 | 1.008150e-05 | 0.1 | 0.5865721 |
| KEGG05100 | 31 | 33.781340 | 9.329447e-06 | 1.171555e-04 | 0.1 | 0.5865721 |
| KEGG00350 | 10 | 5.163053 | 9.581005e-02 | 5.546964e-01 | 0.1 | 0.5865721 |
| KEGG04514 | 62 | 30.108931 | 2.402932e-05 | 2.678036e-04 | 0.1 | 0.5865721 |
| KEGG04530 | 36 | 31.095936 | 1.854001e-05 | 2.146764e-04 | 0.1 | 0.5865721 |
| KEGG04670 | 52 | 33.856570 | 9.155178e-06 | 1.166094e-04 | 0.1 | 0.5865721 |
| KEGG05160 | 55 | 53.505956 | 1.196423e-07 | 3.137407e-06 | 0.1 | 0.5865721 |
| KEGG03430 | 13 | 22.840756 | 1.844695e-04 | 1.596806e-03 | 0.1 | 0.5865721 |
| KEGG05200 | 149 | 68.145490 | 8.444726e-09 | 4.428958e-07 | 0.1 | 0.5865721 |
| KEGG05210 | 36 | 29.839909 | 2.580669e-05 | 2.840615e-04 | 0.1 | 0.5865721 |
| KEGG05213 | 28 | 26.480816 | 6.452479e-05 | 6.537458e-04 | 0.1 | 0.5865721 |

| | | | | | | |
|-----------|----|------------|--------------|--------------|-----|-----------|
| KEGG05416 | 41 | 21.371049 | 2.871904e-04 | 2.415622e-03 | 0.1 | 0.5865721 |
| KEGG04120 | 29 | 12.630167 | 5.181351e-03 | 3.499726e-02 | 0.1 | 0.5865721 |
| KEGG04974 | 29 | 8.159227 | 2.801678e-02 | 1.676474e-01 | 0.1 | 0.5865721 |
| KEGG04210 | 41 | 25.794077 | 7.829317e-05 | 7.843300e-04 | 0.1 | 0.5865721 |
| KEGG05142 | 55 | 49.447678 | 2.694878e-07 | 6.322961e-06 | 0.1 | 0.5865721 |
| KEGG05014 | 23 | 31.606850 | 1.623516e-05 | 1.930013e-04 | 0.1 | 0.5865721 |
| KEGG05130 | 24 | 8.239661 | 2.713860e-02 | 1.634898e-01 | 0.1 | 0.5865721 |
| KEGG05131 | 31 | 55.243538 | 8.545878e-08 | 2.721220e-06 | 0.1 | 0.5865721 |
| KEGG04115 | 24 | 37.099129 | 4.138379e-06 | 5.676516e-05 | 0.1 | 0.5865721 |
| KEGG04916 | 32 | 14.931124 | 2.307207e-03 | 1.658937e-02 | 0.1 | 0.5865721 |
| KEGG05215 | 47 | 53.971118 | 1.092671e-07 | 3.036523e-06 | 0.1 | 0.5865721 |
| KEGG04310 | 44 | 41.315269 | 1.551142e-06 | 2.426284e-05 | 0.1 | 0.5865721 |
| KEGG04350 | 23 | 24.300792 | 1.201271e-04 | 1.120171e-03 | 0.1 | 0.5865721 |
| KEGG03013 | 42 | 12.521968 | 5.387307e-03 | 3.611479e-02 | 0.1 | 0.5865721 |
| KEGG04145 | 75 | 146.861808 | 4.447553e-13 | 1.982696e-10 | 0.1 | 0.5865721 |
| KEGG04520 | 32 | 22.207678 | 2.229148e-04 | 1.892843e-03 | 0.1 | 0.5865721 |
| KEGG05410 | 31 | 18.282987 | 7.562727e-04 | 5.863346e-03 | 0.1 | 0.5865721 |
| KEGG05414 | 32 | 10.341813 | 1.204386e-02 | 7.838090e-02 | 0.1 | 0.5865721 |
| KEGG00010 | 37 | 9.063638 | 1.964873e-02 | 1.216569e-01 | 0.1 | 0.5865721 |
| KEGG04380 | 71 | 32.680575 | 1.232254e-05 | 1.525923e-04 | 0.1 | 0.5865721 |
| KEGG04620 | 48 | 49.019006 | 2.942818e-07 | 6.727656e-06 | 0.1 | 0.5865721 |
| KEGG04630 | 54 | 41.009056 | 1.662694e-06 | 2.555932e-05 | 0.1 | 0.5865721 |
| KEGG05140 | 56 | 41.989367 | 1.332522e-06 | 2.121540e-05 | 0.1 | 0.5865721 |
| KEGG05145 | 70 | 46.045074 | 5.487504e-07 | 1.063609e-05 | 0.1 | 0.5865721 |
| KEGG05212 | 42 | 26.528157 | 6.367518e-05 | 6.525532e-04 | 0.1 | 0.5865721 |
| KEGG04640 | 62 | 117.560134 | 9.449774e-12 | 2.808440e-09 | 0.1 | 0.5865721 |
| KEGG00980 | 10 | 66.696592 | 1.079104e-08 | 5.345098e-07 | 0.1 | 0.5865721 |
| KEGG00983 | 12 | 44.930783 | 6.971132e-07 | 1.268447e-05 | 0.1 | 0.5865721 |
| KEGG00240 | 30 | 74.320240 | 3.081965e-09 | 2.747848e-07 | 0.1 | 0.5865721 |
| KEGG00480 | 14 | 89.964548 | 3.026550e-10 | 5.396881e-08 | 0.1 | 0.5865721 |
| KEGG00590 | 16 | 39.391204 | 2.410915e-06 | 3.523847e-05 | 0.1 | 0.5865721 |
| KEGG00860 | 15 | 49.760861 | 2.527749e-07 | 6.091121e-06 | 0.1 | 0.5865721 |
| KEGG00030 | 15 | 13.506746 | 3.790243e-03 | 2.619644e-02 | 0.1 | 0.5865721 |
| KEGG00230 | 51 | 25.179015 | 9.327181e-05 | 9.138480e-04 | 0.1 | 0.5865721 |
| KEGG00071 | 18 | 39.257416 | 2.487030e-06 | 3.576467e-05 | 0.1 | 0.5865721 |
| KEGG04920 | 27 | 62.446658 | 2.260875e-08 | 9.496335e-07 | 0.1 | 0.5865721 |
| KEGG05150 | 32 | 56.063892 | 7.306855e-08 | 2.412858e-06 | 0.1 | 0.5865721 |
| KEGG00620 | 14 | 24.286911 | 1.206120e-04 | 1.120171e-03 | 0.1 | 0.5865721 |
| KEGG04930 | 21 | 19.258351 | 5.537710e-04 | 4.408361e-03 | 0.1 | 0.5865721 |
| KEGG04664 | 36 | 62.245608 | 2.343224e-08 | 9.496335e-07 | 0.1 | 0.5865721 |
| KEGG04722 | 56 | 56.071300 | 7.296574e-08 | 2.412858e-06 | 0.1 | 0.5865721 |
| KEGG04912 | 35 | 15.060709 | 2.206856e-03 | 1.599683e-02 | 0.1 | 0.5865721 |
| KEGG00280 | 19 | 38.660972 | 2.858611e-06 | 3.982356e-05 | 0.1 | 0.5865721 |
| KEGG00310 | 12 | 28.018168 | 4.216839e-05 | 4.529746e-04 | 0.1 | 0.5865721 |
| KEGG00380 | 15 | 103.491944 | 5.077894e-11 | 1.131849e-08 | 0.1 | 0.5865721 |
| KEGG00640 | 14 | 47.605074 | 3.946596e-07 | 8.377961e-06 | 0.1 | 0.5865721 |
| KEGG00650 | 10 | 4.237641 | 1.426346e-01 | 8.204612e-01 | 0.1 | 0.5865721 |
| KEGG00020 | 14 | 13.152966 | 4.297080e-03 | 2.924604e-02 | 0.1 | 0.5865721 |
| KEGG04012 | 38 | 23.225345 | 1.645928e-04 | 1.452962e-03 | 0.1 | 0.5865721 |
| KEGG05220 | 47 | 39.126096 | 2.564215e-06 | 3.628932e-05 | 0.1 | 0.5865721 |

| | | | | | | |
|-----------|----|------------|--------------|--------------|-----|-----------|
| KEGG00564 | 14 | 58.921920 | 4.279369e-08 | 1.589767e-06 | 0.1 | 0.5865721 |
| KEGG05340 | 25 | 148.792814 | 3.701484e-13 | 1.982696e-10 | 0.1 | 0.5865721 |
| KEGG00500 | 12 | 28.113816 | 4.108093e-05 | 4.466748e-04 | 0.1 | 0.5865721 |
| KEGG05120 | 34 | 65.157949 | 1.405379e-08 | 6.594849e-07 | 0.1 | 0.5865721 |
| KEGG05323 | 45 | 85.832094 | 5.425801e-10 | 6.910839e-08 | 0.1 | 0.5865721 |
| KEGG03040 | 40 | 17.132641 | 1.100194e-03 | 8.312894e-03 | 0.1 | 0.5865721 |
| KEGG04660 | 50 | 10.494546 | 1.136995e-02 | 7.453918e-02 | 0.1 | 0.5865721 |
| KEGG00410 | 12 | 46.645514 | 4.830102e-07 | 1.001504e-05 | 0.1 | 0.5865721 |
| KEGG05221 | 39 | 35.710984 | 5.788211e-06 | 7.702550e-05 | 0.1 | 0.5865721 |
| KEGG04340 | 11 | 6.073128 | 6.534459e-02 | 3.832931e-01 | 0.1 | 0.5865721 |
| KEGG05218 | 31 | 20.513822 | 3.737548e-04 | 3.085518e-03 | 0.1 | 0.5865721 |
| KEGG04512 | 26 | 24.645916 | 1.087092e-04 | 1.042194e-03 | 0.1 | 0.5865721 |
| KEGG05146 | 49 | 71.444582 | 4.892881e-09 | 3.355724e-07 | 0.1 | 0.5865721 |
| KEGG05222 | 46 | 43.526104 | 9.470522e-07 | 1.623811e-05 | 0.1 | 0.5865721 |
| KEGG04610 | 14 | 72.954692 | 3.832568e-09 | 3.081016e-07 | 0.1 | 0.5865721 |
| KEGG03030 | 19 | 22.769488 | 1.884236e-04 | 1.615351e-03 | 0.1 | 0.5865721 |
| KEGG04622 | 20 | 53.826381 | 1.123894e-07 | 3.036523e-06 | 0.1 | 0.5865721 |
| KEGG00970 | 16 | 23.403392 | 1.561698e-04 | 1.392394e-03 | 0.1 | 0.5865721 |
| KEGG03015 | 21 | 48.294156 | 3.418519e-07 | 7.433942e-06 | 0.1 | 0.5865721 |
| KEGG04970 | 38 | 64.827764 | 1.488132e-08 | 6.634013e-07 | 0.1 | 0.5865721 |
| KEGG04370 | 35 | 31.024253 | 1.889009e-05 | 2.159257e-04 | 0.1 | 0.5865721 |
| KEGG04662 | 45 | 44.427951 | 7.774477e-07 | 1.367602e-05 | 0.1 | 0.5865721 |
| KEGG00051 | 16 | 26.636897 | 6.176816e-05 | 6.403703e-04 | 0.1 | 0.5865721 |
| KEGG00052 | 15 | 19.849740 | 4.596460e-04 | 3.759776e-03 | 0.1 | 0.5865721 |
| KEGG04114 | 42 | 23.562341 | 1.490349e-04 | 1.342202e-03 | 0.1 | 0.5865721 |
| KEGG04540 | 35 | 9.106446 | 1.932494e-02 | 1.204889e-01 | 0.1 | 0.5865721 |
| KEGG04914 | 33 | 18.313146 | 7.489572e-04 | 5.857565e-03 | 0.1 | 0.5865721 |
| KEGG04070 | 30 | 22.848678 | 1.840354e-04 | 1.596806e-03 | 0.1 | 0.5865721 |
| KEGG04720 | 36 | 8.649082 | 2.309800e-02 | 1.410544e-01 | 0.1 | 0.5865721 |
| KEGG04730 | 31 | 78.228678 | 1.675825e-09 | 1.660164e-07 | 0.1 | 0.5865721 |
| KEGG00561 | 12 | 88.191090 | 3.878922e-10 | 5.764010e-08 | 0.1 | 0.5865721 |
| KEGG00330 | 22 | 68.918525 | 7.419603e-09 | 4.410161e-07 | 0.1 | 0.5865721 |
| KEGG03008 | 10 | 14.128713 | 3.046345e-03 | 2.121945e-02 | 0.1 | 0.5865721 |
| KEGG00520 | 15 | 8.466957 | 2.481070e-02 | 1.504828e-01 | 0.1 | 0.5865721 |
| KEGG04672 | 24 | 44.399449 | 7.822850e-07 | 1.367602e-05 | 0.1 | 0.5865721 |
| KEGG05144 | 31 | 78.347021 | 1.645731e-09 | 1.660164e-07 | 0.1 | 0.5865721 |
| KEGG05310 | 21 | 32.129242 | 1.418916e-05 | 1.709582e-04 | 0.1 | 0.5865721 |
| KEGG05320 | 25 | 15.995128 | 1.606747e-03 | 1.183933e-02 | 0.1 | 0.5865721 |
| KEGG05330 | 24 | 19.655395 | 4.885585e-04 | 3.959942e-03 | 0.1 | 0.5865721 |
| KEGG04612 | 40 | 45.349787 | 6.368783e-07 | 1.208157e-05 | 0.1 | 0.5865721 |
| KEGG04940 | 24 | 9.486543 | 1.668563e-02 | 1.062624e-01 | 0.1 | 0.5865721 |
| KEGG05332 | 24 | 10.138221 | 1.300856e-02 | 8.404559e-02 | 0.1 | 0.5865721 |
| KEGG05143 | 19 | 24.060403 | 1.288248e-04 | 1.184112e-03 | 0.1 | 0.5865721 |
| KEGG05214 | 39 | 18.202500 | 7.761679e-04 | 5.965717e-03 | 0.1 | 0.5865721 |
| KEGG05219 | 22 | 48.867088 | 3.036379e-07 | 6.768009e-06 | 0.1 | 0.5865721 |
| KEGG05223 | 31 | 16.965995 | 1.162369e-03 | 8.636297e-03 | 0.1 | 0.5865721 |
| KEGG00270 | 13 | 9.247003 | 1.830088e-02 | 1.157225e-01 | 0.1 | 0.5865721 |
| KEGG04966 | 10 | 42.692106 | 1.138933e-06 | 1.915964e-05 | 0.1 | 0.5865721 |
| KEGG04621 | 21 | 54.823971 | 9.263717e-08 | 2.753145e-06 | 0.1 | 0.5865721 |
| KEGG04623 | 17 | 17.496447 | 9.763539e-04 | 7.440231e-03 | 0.1 | 0.5865721 |

| | | | | | | |
|-----------|----|-----------|--------------|--------------|-----|-----------|
| KEGG04330 | 16 | 14.667409 | 2.526630e-03 | 1.787870e-02 | 0.1 | 0.5865721 |
| KEGG04964 | 11 | 27.049375 | 5.506555e-05 | 5.844746e-04 | 0.1 | 0.5865721 |
| KEGG04150 | 18 | 11.009560 | 9.376387e-03 | 6.238723e-02 | 0.1 | 0.5865721 |
| KEGG04973 | 20 | 13.182805 | 4.251686e-03 | 2.915968e-02 | 0.1 | 0.5865721 |
| KEGG05216 | 19 | 30.751272 | 2.028858e-05 | 2.289758e-04 | 0.1 | 0.5865721 |
| KEGG05020 | 21 | 14.773131 | 2.436126e-03 | 1.737620e-02 | 0.1 | 0.5865721 |
| KEGG04740 | 10 | 9.033627 | 1.987914e-02 | 1.222347e-01 | 0.1 | 0.5865721 |
| KEGG04742 | 10 | 9.165107 | 1.889037e-02 | 1.186088e-01 | 0.1 | 0.5865721 |
| KEGG00562 | 15 | 18.867003 | 6.271148e-04 | 4.948044e-03 | 0.1 | 0.5865721 |
| KEGG00510 | 15 | 7.675775 | 3.396901e-02 | 2.019094e-01 | 0.2 | 1.0000000 |
| KEGG00250 | 12 | 10.116799 | 1.311470e-02 | 8.412183e-02 | 0.2 | 1.0000000 |
| KEGG04960 | 19 | 6.414720 | 5.672222e-02 | 3.349201e-01 | 0.3 | 1.0000000 |
| KEGG04260 | 29 | 2.355025 | 3.299165e-01 | 1.000000e+00 | 0.4 | 1.0000000 |
| KEGG05412 | 26 | 3.301194 | 2.153740e-01 | 1.000000e+00 | 0.5 | 1.0000000 |
| KEGG05211 | 31 | 2.628229 | 2.913801e-01 | 1.000000e+00 | 0.5 | 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).

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,

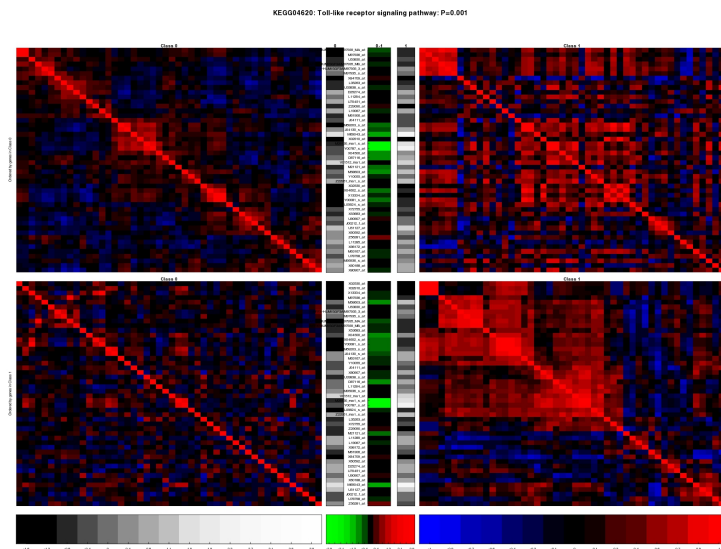


Figure 1: KEGG04620

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.

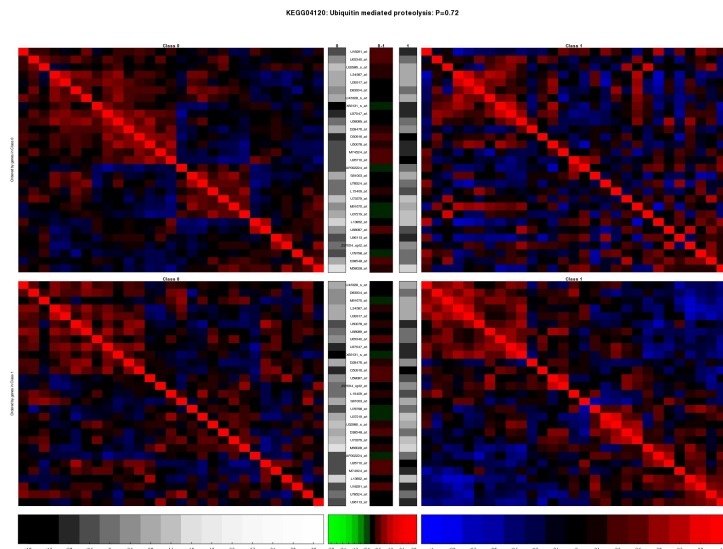


Figure 2: KEGG04120

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] 2563 38

> dim(newimat)

[1] 2563 151
```

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.