User's Guide for Filtering Code

Stan Pounds

March 29, 2005

Pounds and Cheng (2005) develop two methods to use the results of the Affymetrix present-absent tests (Affymetrix 2002) in the filtering of gene expression data. The two methods are the pooled p-value and error-minimizing pooled p-value filters. For each probe, these two filters compute a pooled p-value that summarizes the present-absent p-values across a set of chips measuring the same experimental condition. The pooled p-value filter compares the pooled p-value to a significance threshold α chosen by the user; the error-minimizing pooled p-value filter compares the pooled p-value to a significance threshold α_{MTE} that minimizes an estimate of the total-error criterion (Cheng et al. 2004). Pounds and Cheng (2005) show that both filters have much greater statistical power than the popular practice of basing filtering on the number of present calls.

These two innovative filters are implemented using the commercially available S-plus (www.splus.com) or freely available R (www.r-project.org) software packages. To get started, the user should first download either the S-plus or R code from www.stjuderesearch.org/depts/biostats/filter.html. The *source* command is then used for either version to make the routines available for use. The code defines three functions: *probe.filter*, *pooled.p*, and *splosh*. The function *pooled.p* computes the pooled p-value. The function *splosh* implements the spacings loess histogram algorithm (Pounds and Cheng 2004) to estimate the total error criterion for the error-minimizing pooled p-value filter. The function *probe.filter* implements the pooled p-value or the error-minimizing pooled p-value filter.

The function *probe.filter* takes up to five arguments: *P*, *grp*, *alpha*, *delta*, and *opts*. The argument *P* is a matrix of Affymetrix present-absent p-values with rows representing samples and columns representing probe sets. The argument *grp* is a vector that gives the experimental group assignments of

each of the samples. The argument *alpha* is the threshold for determining whether to include a probe set in subsequent analysis. For each probe, a pooled p-value is computed for each experimental group. All probe sets with at least one pooled p-value less than or equal the *alpha* for the corresponding experimental group will be included in subsequent analysis. The default for *alpha* is to determine the threshold that minimizes an estimate of the total-error criterion for each experimental group. The argument *delta* specifies that the spacings-loess histogram should be applied to pooled p-values over the interval (*delta*,1-*delta*). Therefore, *delta* must be in the range (0,0.5). The default for *delta* is 0.01. The argument *opts* is a list of options giving the values of the tuning parameters for the spacings-loess histogram. Defaults for these options are provided.

The function *probe.filter* returns a list with the following four components: *include*, *uniq.grps*, *alpha* and *pooled.p*. The returned component *include* is a Boolean vector that indicates whether or not the corresponding probe set should be included in subsequent analysis. Probe sets corresponding to entries of *include* with a value of TRUE should be included in subsequent analyses. Probe sets corresponding to entries of *include* with a value of FALSE should be excluded from subsequent analyses. The returned component *uniq.grps* gives the names of the experimental groups provided in the argument *grp*. The entries of returned component *uniq.grp* indicates which experimental group the entries of the returned vector *alpha* and rows of the returned matrix *pooled.p* correspond to. The returned component *alpha* is either a scalar or a vector. The component *alpha* contains either the user-provided value of the argument *alpha* or the threshold determined to minimize the total error criterion for each of the experimental groups listed in *uniq.grp*. The returned component *pooled.p* is a matrix of pooled p-values for each experimental group (rows) and each probe set (columns).

To implement the pooled p-value filter at the $\alpha = 0.05$ level, one would issue a command such as

```
filt.res<-probe.filter(P,grp,alpha=0.05)
```

and then conduct subsequent analyses on the the subset of probe sets such that *filt.res$include* == *TRUE*. To implement the error-minimizing pooled p-value filter, one would issue a command such as

```
filt.res<-probe.filter(P,grp)
```

and then conduct subsequent analyses on the subset of probe sets such that *filt.res$include == TRUE*.

The function *pooled.p* takes a matrix P of present-absent p-values (rows represent samples and columns represent probe sets) and returns a vector of pooled p-values, one for each probe set.

The function *splosh* implements the spacings loess histogram. For more information, see Pounds and Cheng (2004).

Questions regarding the use of these functions should be directed to Stan Pounds at stanley.pounds@stjude.org.

Works Cited

Affymetrix (2002) *Statistical Algorithms Description Document.* (www.affymetrix.com)

Pounds, S. and C. Cheng (2004). Improving False Discovery Rate Estimation. *Bioinformatics* **20**, 1737-1745.

Pounds, S. and C. Cheng (2005). Statistical Development and Evaluation of Gene Expression Data Filters. *Journal of Computational Biology*, in press.