Smooth Data

Lori Shepherd and Jonathan Dare

July 10, 2007

Contents

Please download and utilize example1. This code will also work if the complete example was run from the load.in.example1 data.

We have provided files that would have been created during the process of making an aCGHroster object. This includes the samples' array image object, mapping files, design files, and the RData aCGHroster object.

In the directory smoothing.example (or in the load.in.example1 directory if that example has already been run) begin an R session and load the package library and the aCGHroster object. If defaults were kept in the loadin process, this object will be located in the RData directory with the name aCGHroster.RData.

```
> library(aCGHplus)
> load("RData/aCGHroster.RData")
```

This example set has 200 samples that need to be smoothed. This can be done in one of three ways:

1. individually load each object and smooth data for each sample (NOT RECOMMENDED)
2. take all samples at once and smooth data
3. Batch. This means breaking up the samples into group and smoothing data

Smoothing this data set may be done all together, however since there are 200 samples, it is recommended to break the data set up into Batch calls. We will break this example up into 4 batch jobs, which means 50 samples will be smoothed per batch call. The following batch calls may be run in the same R session in succession or the user may open any number of additional R session in the same directory as the first (remember to load the package and aCGHroster object in each session) to run the following calls:

```
> SmoothBatch(aCGHroster, overwrite = T, BatchDX = NA, nbatchjobs = 4,
+     ibatchjob = 1, time.order = F, map.name = NA, thetas = seq(0.5,
+         7.5, length = 11), cvLevel = 10, LossF = "L2", useResid = T,
+     BCoption = "none", lambdas = seq(0, 1, length = 2), DesignFlag = F,
+     lib.loc = NA, cyC.label = "array.images$matrix$cy3$Signal.Mean",
+     cyT.label = "array.images$matrix$cy5$Signal.Mean", cyC.BC.label = "array.images$ma
+     cyT.BC.label = "array.images$matrix$cy5$Background.Mean",
+     cy.grid.label = "array.images$matrix$Grid", exclude.rule = "array.images$matrix$cy
+     output.label = "smooth2D.noDesign", pname = "auto", plt = T,
+     vrb = T, DesignList = c("Plate", "Pin", "PlateRow", "PlateCol",
+         "Repetition"), noDesignExit = F, weightSex.loess = 1/10,
+     excludeSex.smooth = F, excludeSex.design = F, weightNonMap.loess = 1/10,
+     excludeNonMap.smooth = T, excludeNonMap.design = T, ilambda.default = 1,
+     saveAll = T)
> SmoothBatch(aCGHroster, overwrite = T, BatchDX = NA, nbatchjobs = 4,
+     ibatchjob = 2, time.order = F, map.name = NA, thetas = seq(0.5,
+         7.5, length = 11), cvLevel = 10, LossF = "L2", useResid = T,
+     BCoption = "none", lambdas = seq(0, 1, length = 2), DesignFlag = F,
+     lib.loc = NA, cyC.label = "array.images$matrix$cy3$Signal.Mean",
+     cyT.label = "array.images$matrix$cy5$Signal.Mean", cyC.BC.label = "array.images$ma
+     cyT.BC.label = "array.images$matrix$cy5$Background.Mean",
+     cy.grid.label = "array.images$matrix$Grid", exclude.rule = "array.images$matrix$cy
+     output.label = "smooth2D.noDesign", pname = "auto", plt = T,
+     vrb = T, DesignList = c("Plate", "Pin", "PlateRow", "PlateCol",
+         "Repetition"), noDesignExit = F, weightSex.loess = 1/10,
+     excludeSex.smooth = F, excludeSex.design = F, weightNonMap.loess = 1/10,
+     excludeNonMap.smooth = T, excludeNonMap.design = T, ilambda.default = 1,
+     saveAll = T)
> SmoothBatch(aCGHroster, overwrite = T, BatchDX = NA, nbatchjobs = 4,
+     ibatchjob = 3, time.order = F, map.name = NA, thetas = seq(0.5,
+         7.5, length = 11), cvLevel = 10, LossF = "L2", useResid = T,
+     BCoption = "none", lambdas = seq(0, 1, length = 2), DesignFlag = F,
+     lib.loc = NA, cyC.label = "array.images$matrix$cy3$Signal.Mean",
+     cyT.label = "array.images$matrix$cy5$Signal.Mean", cyC.BC.label = "array.images$ma
+     cyT.BC.label = "array.images$matrix$cy5$Background.Mean",
+     cy.grid.label = "array.images$matrix$Grid", exclude.rule = "array.images$matrix$cy
+     output.label = "smooth2D.noDesign", pname = "auto", plt = T,
+     vrb = T, DesignList = c("Plate", "Pin", "PlateRow", "PlateCol",
+         "Repetition"), noDesignExit = F, weightSex.loess = 1/10,
+     excludeSex.smooth = F, excludeSex.design = F, weightNonMap.loess = 1/10,
+     excludeNonMap.smooth = T, excludeNonMap.design = T, ilambda.default = 1,
+     saveAll = T)
> SmoothBatch(aCGHroster, overwrite = T, BatchDX = NA, nbatchjobs = 4,
+     ibatchjob = 4, time.order = F, map.name = NA, thetas = seq(0.5,
+         7.5, length = 11), cvLevel = 10, LossF = "L2", useResid = T,
+     BCoption = "none", lambdas = seq(0, 1, length = 2), DesignFlag = F,
```

```
+     lib.loc = NA, cyC.label = "array.images$matrix$cy3$Signal.Mean",
+     cyT.label = "array.images$matrix$cy5$Signal.Mean", cyC.BC.label = "array.images$ma
+     cyT.BC.label = "array.images$matrix$cy5$Background.Mean",
+     cy.grid.label = "array.images$matrix$Grid", exclude.rule = "array.images$matrix$cy
+     output.label = "smooth2D.noDesign", pname = "auto", plt = T,
+     vrb = T, DesignList = c("Plate", "Pin", "PlateRow", "PlateCol",
+         "Repetition"), noDesignExit = F, weightSex.loess = 1/10,
+     excludeSex.smooth = F, excludeSex.design = F, weightNonMap.loess = 1/10,
+     excludeNonMap.smooth = T, excludeNonMap.design = T, ilambda.default = 1,
+     saveAll = T)
```

If addtional R sessions were opened, once the function is finished the sessions may be exited.Keep one R session open. We will examine the third call:

```
SmoothBatch(aCGHroster,overwrite=T,BatchDX=NA,
            nbatchjobs=4,ibatchjob=3,
            time.order=F,map.name=NA,
            thetas=seq(0.5,7.5,length=10),
            cvLevel=10,LossF="L2",useResid=T,
            BCoption="none",lambdas=seq(0,1,length=2),
            DesignFlag=F,lib.loc=NA,
            cyC.label="array.images$matrix$cy3$Signal.Mean",
            cyT.label="array.images$matrix$cy5$Signal.Mean",
            cyC.BC.label="array.images$matrix$cy3$Background.Mean",
            cyT.BC.label="array.images$matrix$cy5$Background.Mean",
            cy.grid.label="array.images$matrix$Grid",
            exclude.rule="array.images$matrix$cy3$Flag!=0",
            output.label="smooth2D.noDesign",
            pname="auto",plt=T,vrb=T,
   DesignList=c("Plate","Pin","PlateRow","PlateCol","Repetition"),
   noDesignExit=F,weightSex.loess=1/10,
            excludeSex.smooth=F, excludeSex.design=F,
            weightNonMap.loess=1/10,excludeNonMap.smooth=T,
            excludeNonMap.design=T,  ilambda.default=1,
            saveAll=T )
```

Looking at this code we can say the following:

- If files exist they will be rewritten
- BatchDX is NA, and therefore all samples in the aCGHroster will be considered
- Four Batch jobs will be run and this job will run the third section. Since there are 200 samples this means samples 100 to 150.
- time.order PLEASE IGNORE THIS OPTION. IT IS NOT FUNCTIONAL AT THIS TIME

- map.name is the name of the map file that will be used, if NA the default map file in the sample's array.image file will be used. All these sample's therefore use the default map listed in the inventory
- 10 values between .5 and 7.5 will be tested as the smoothing value as specified by thetas
- 10 percentage of the data will be removed as part of the cross validation method as specified by CVLevel
- LossF may be "L1" or "L2" represents values are used. "L2" is used in this case, meaning the sum of the difference of observed values and smoothed values squared are used. If "L1" was chosen, the sum of the absolute differenc eof the observed and smoothed values are used.
- residuals are used in analysis
- no background correction is performed as indicated by BCoption. BCoption may be none, raw, or smoothed. Raw would be the raw background is subtracted off the values. Smoothed is the smoothed background values are subtracted off the values.
- lambdas is the percentage of background percentage to used. If a range is specfied it will chose which ever yields the best result. It this case, all background or no background will be tested
- no design correction is performed. If DesignFlag was true, design information will be loaded for the sample and corrections for pin, plate, etc. performed.
- lib.loc is the path to local librarys.
- The tumor information will come from cy5 data and control information will come from cy3 data.
- The values used will be the Signal.Mean and Background.Means
- any spots that are flaged in the cy3 flag matrix (marked by no zero number) will be excluded from analysis
- all smoothed information will be stored in a smooth2D.noDesign object
- graphs will be plotted for before and after correction
- If DesignFlag is True, this lists what will be corrected
- If there isn't a design file for a sample, the program will not abort, controlled by noDesignExit
- sex chromosomes will also be smooothed as specified by excludeSex.smooth

- sex chromosomes will also be included in the design phase as specified by excludeSex.design
- non-mapped spots will be excluded from smoothing, controlled by excludeNonMap.smooth
- non-mapped spots will also be excluded from design phase, controlled by excludeNonMap.design
- the background percentage default that will be saved is to background correct
- all data of smoothing and design will be saved as controlled by saveAll

Smoothing the data must be performed before an aCGH object can be created.