Gathering a Proteomic Matrix with Batch Effects

Keith A. Baggerly

March 9, 2010

Contents

List of Figures

List of Tables

1 Executive Summary

1.1 Introduction

We're trying to pull together a protomics data matrix illustrating batch effects. For this purpose, we make use of a set of 253 SELDI spectra posted at the NCI/FDA Clinical Proteomics website, `http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`, as OvarianDataset8-7-02.zip. The spectra in this dataset were previously examined in Baggerly et al. (2004, 2005), and provide an example of complete confounding.

1.2 Methods

We downloaded the zip file and expanded it in RawData. There is one csv file per spectrum. Each file has two columns: M/Z (mass to charge ratio) and intensity. M/Z values depend on the calibration of the machine, which was not changed during the course of this experiment, so the first column is identical across all files. There are 15154 data values per spectrum. The initial analysis did no peak selection, so all M/Z values are valid measurements.

We assemble the spectral intensities into a single matrix. We also produce a common M/Z vector (mzVector) of length 15154, a batch vector of length 253, and an outcome vector of length 253.

1.3 Results

The assembled matrix and data vectors are saved as both an R Data Object (.Rda) file and as individual csv files.

2 Loading Data

```
> controlFiles <- dir(file.path("RawData", "Control"))
> cancerFiles <- dir(file.path("RawData", "Ovarian Cancer"))
> nControl <- length(controlFiles)
> nCancer <- length(cancerFiles)
> nMZValues <- 15154
> proteinMatrix <- matrix(0, nrow = nMZValues, ncol = nControl + nCancer)
> temp <- read.csv(file.path("RawData", "Control", controlFiles[1]))
> mzVector <- temp[, "M.Z"]
> for (i1 in 1:nControl) {
+     temp <- read.csv(file.path("RawData", "Control", controlFiles[i1]))
+     if (all(temp[, "M.Z"] == mzVector)) {
+         proteinMatrix[, i1] <- temp[, "Intensity"]
+     }
+ }
> for (i1 in 1:nCancer) {
+     temp <- read.csv(file.path("RawData", "Ovarian Cancer", cancerFiles[i1]))
+     if (all(temp[, "M.Z"] == mzVector)) {
+         proteinMatrix[, (i1 + nControl)] <- temp[, "Intensity"]
+     }
+ }
> colnames(proteinMatrix) <- c(substr(controlFiles, 9, 16), substr(cancerFiles,
+     16, 23))
> rm(temp, i1)
> batchVector <- c(rep("Batch 1", nControl), rep("Batch 2", nCancer))
> outcomeVector <- c(rep("Control", nControl), rep("Ovarian Cancer", nCancer))
```

3 Saving Data

3.1 As an Rda File

```
> save(list = ls(all = TRUE), file = file.path("RDataObjects", "allProtein.Rda"))
```

3.2 As csv Files

```
> write.csv(proteinMatrix, file = file.path("OutputFiles", "proteinMatrix.csv"))
> write.csv(mzVector, file = file.path("OutputFiles", "mzVector.csv"))
> write.csv(batchVector, file = file.path("OutputFiles", "batchVector.csv"))
> write.csv(outcomeVector, file = file.path("OutputFiles", "outcomeVector.csv"))
```

4 Directory Structure

```
> getwd()
```

```
[1] "/Users/kabagg/ReproRsch/BatchPaper/Analysis"
```

```
> dir()
```

```
[1] "buildProt.tex" "Figures"       "OutputFiles"   "RawData"       "RDataObjects"
[6] "RNowebSource"
```

```
> dir("RNowebSource")
```

```
[1] "buildProt.Rnw"  "buildProt.Rnw~"

> dir("RawData")

[1] "Control"                  "Ovarian Cancer"           "OvarianDataset8-7-02.zip"

> dir("RDataObjects")

[1] "allProtein.Rda"

> dir("OutputFiles")

[1] "batchVector.csv"   "csvFiles.zip"      "mzVector.csv"      "outcomeVector.csv"
[5] "proteinMatrix.csv"

> dir("Figures")

character(0)
```

5 Appendix

5.1 SessionInfo

```
> sessionInfo()

R version 2.10.1 (2009-12-14)
i386-apple-darwin8.11.1

locale:
[1] en_US.UTF-8/en_US.UTF-8/C/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
[1] tools_2.10.1
```