Pathway Analysis

Michael Nash

October 1, 2018

Data cleaning is the process of turning 'dirty' data into 'clean' data. 'Dirty data' is not ready to analyze because it is ambiguously labeled, contains erroneous or unclear entries or is formatted in a way that makes it difficult to work with. 'Clean' data is clearly labeled and formatted and ready to analyze. Data harmonization is the process of taking data from multiple sources and combining them into a single data source in such a way that the data from the various sources are comparable and can be analyzed, visualized or otherwise understood together. In this project, I demonstrate how to clean and harmonize data from different sources in a way that is reproducable and relatively easy to follow (I hope) using R Markdown.

First, I will provide some background information about this project: Proteins are often part of biological pathways, which are chains of chemical reactions or interactions that occur in a cell to perform biological function such as producing other molecules. The investigators I worked with provided me with lists of proteins found in various sources (muscle samples, serum samples, a subset of proteins found in serum samples that were associated with phenotype measures of interest). Some proteins may be found in more than one source. The investigators were interested in the biological pathways that proteins from various sources are part of, how many proteins from various sources are represented in different pathways, and which pathways are represented by relatively large numbers of proteins from a given source. I used information about biological pathways from the KEGG database, which I accessed through an online interface called DAVID provided by the Laboratory of Human Retrovirology and Immunoinformatics (https://david.ncifcrf.gov/).

I started with the following:

NIHMS860664-supplement-Supp_info.xlsx - A spreadsheet containing a list of proteins found in serum samples in a previous study, included as a supplement in the following journal article:

Nielson, C. M., Wiedrick, J., Shen, J., Jacobs, J., Baker, E. S., Baraff, A., . . . Orwoll, E. S. (2017). Identification of Hip BMD Loss and Fracture Risk Markers Through Population-Based Serum Proteomics. Journal of Bone and Mineral Research, 32(7), 1559-1567. doi:10.1002/jbmr.3125

with full text and supplemental materials available here: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5489383/#SD1

serum_prt_table.jpg - A table containing a list of 28 proteins associated with fat mass, lean BMI or both in serum samples in the same study. I transcribed the protein names from this table into a spreadsheet.

imbsr.xlsx - A spreadsheet containing a list of proteins found in each of six muscle biopsy samples from a pilot study and their levels in the samples.

The second data source is part of an unpublished manuscript, as are my pathway analysis results. The third data source contains biological data from living human research subjects. None of these be shared publicly. In lieu of showing the results I showed my collaborators on this project and including the data I used to generate them, I created a simulated dataset consisting of three files formatted identically to those I used in my actual analysis so that I can demonstrate the methods I used.

The R source file 'make_new_sheets.R,' invoked above, generates the simulated data. Although the selection of protein names from the master list is pseudo-random, it should come out the same way every time because I set random seeds throughout the source file. In order to work, the working directory specified in the first line needs to point to a directory containing the supplement from the Nielson et al 2018 article. It will create the files in that working directory. If this is the same as the working directory for this .RMD file, the files will be there to be read and analyzed, and everything should work seamlessly. For copyright reasons, the supplement is not in this Github repository. Anyone who wants to reproduce this part of the analysis will have to download it themselves.

The files created by the script are as follows:

serum.xlsx - Intended to replace jbmr.xlsx, this spreadsheet contains a subset of the proteins in jbmr.xlsx.

serum_assoc2.xlsx - A spreadsheet containing a list of 28 proteins from among those found in serum.xlsx

muscle.xlsx - Intended to replace imbsr.xls, this spreadsheet contains six sheets, each with a list of proteins randomly picked from a different subset of the proteins in jbmr.xlsx.

I will describe the process of exploring these data sources, as I did with the original data sources they were made to resemble. Obviously, I know what they contain because I created them. I will also be incorporating output from the DAVID tool, which constitutes another data source. From here on, I will refer to the data sources I use as if they were real data from actual experiments.

I began by looking at the three data sources I started out with in more detail.

serum.xlsx

This file contains the following sheets:

```
## [1] "T1. DAVID bone loss list"        "T2. DAVID background list"
## [3] "T3.DAVID GO Cellular Enrichment"
```

The second sheet, entitled “T2. DAVID background list,” contains the list of proteins found in serum samples. The other sheets contain results from other types of proteomic analyses and are not of interest. T2 contains a single column. Here are a few of the 2934 entries:

```
## [1] "# of rows in T2: 1847"
```

```
## # A tibble: 6 x 1
##   `UniProtKB Name`
##   <chr>
## 1 KIF5A_HUMAN
## 2 DLGP5_HUMAN
## 3 MYG_HUMAN
## 4 RECQ4_HUMAN
## 5 PSA6_HUMAN
## 6 H10_HUMAN
```

These are protein names of some sort and appear to have two parts. The first seems to consist of five digits and a letter. The second appears to be an organism name. They are separated by an underscore.

```
## [1] "Organism suffix counts:"
```

```
## HUMAN
##  1847
```

```
## [1] "length of KB name (minus organism suffix):"
```

```
##    2    3    4    5
##    2   75  571 1199
```

All KB names are followed by ’_HUMAN’, suggesting they were all identified as human proteins. The length of the part before this varies from two to five characters.

muscle.xlsx

This file contains the following sheets:

```
##              X1       X2       X3       X4       X5       X6
## name   sample_1 sample_2 sample_3 sample_4 sample_5 sample_6
## length     1335     1344     1460     1501     1599     1612
```

There are six sheets, named with the IDs of the six muscle samples, starting with ‘sample_1’. Each sheet has 1335 to 1612 rows, with each row corresponding to a protein. They look like this:

```
## # A tibble: 6 x 4
##   Probability Protein              Indistinguishable_Proteins `Log(dSIn)`
##         <dbl> <chr>                <chr>                            <dbl>
## 1          1. sp|Q43333|YS027_HUMAN -                               -23.8
## 2          1. sp|Q49293|K2C6B_HUMAN -                               -15.0
## 3          1. sp|Q77477|CNDP2_HUMAN -                               -17.6
## 4          1. sp|P20289|TSP2_HUMAN  -                               -14.2
## 5          1. sp|Q68541|MTPN_HUMAN  sp|P41714|PEDF_HUMAN             -20.7
## 6          1. sp|P84691|CUL5_HUMAN  -                               -27.0
```

‘Probability’ refers to the probability that a given protein is actually present in the sample and not a false positive. ‘Log(dSIn)’ is a measure of the concentration of that protein in a log scale. I’m not going to use the ‘Probability’ or ‘Log(dSIn)’ columns for anything in this analysis. The investigators I am working with asked for pathway analysis with proteins found in the samples, not proteins found in high concentrations or with high probability in the samples. The protein name in ‘Protein’ seems like it could actually be two names in two different formats stuck together. ‘Indistinguishable_Proteins’ is blank for most rows, but sometimes contains one or more alter egos for a given protein, like so:

```
## [1] "sp|P41714|PEDF_HUMAN"
## [2] "sp|Q82369|ROBO2_HUMAN; sp|Q59359|CAC1H_HUMAN; sp|P52521|FRYL_HUMAN"
## [3] "sp|P77375|CATL1_HUMAN; sp|P56712|TF3C1_HUMAN; sp|P17940|PPE1_HUMAN"
## [4] "sp|P11835|CK081_HUMAN"
## [5] "sp|Q79153|1A26_HUMAN"
## [6] "sp|P35764|OPSB_HUMAN"
```

The data cleaning/analysis will proceed as follows:

1) Get all protein names into a common format. Determine ‘overlap’ among sources of proteins.
2) Determine pathways for each protein using the KEGG database.
3) Create a table containing all proteins from all sources and listing sources and pathways for each protein.
4) Generate reports about pathways represented in various sources and proteins in specific pathways found in various sources as requested by the investigators.

I will begin with muscle protein names. As shown above, the names in the spreadsheet each consist of three strings separated with ‘|’ characters. In all the examples we have seen thus far, the first consists of the letters ‘sp’. Maybe they all say this. I will call this part a prefix. The second part seems to consist of a letter followed by some digits. I will call this a numeric ID. The third seems to be some letters and digits followed by ’_HUMAN’. I will call this a protein name.

To start with, I will read the compound protein names (i.e. the format shown above, with multiple strings separated with a ‘|’ character) in the ‘Protein’ from each of the six sheets in IMBSR and compile a list of all compound names found in any sample. Later on, I will want to determine which of these are found in all samples and which are found in only some.

```
## [1] "# of unique compound protein names in muscle samples: 1849"
```

The next step is to make separate lists for the three substrings. I can do this by splitting the strings in the list and putting the elements in a data frame.

```
## [1] "Rows with missing values:"
```

```
##      X1        X2   X3
## 1803 sp UBB_BOVIN <NA>
```

There is one compound name with a missing component, which has ’_BOVIN’ in its second string instead of “_HUMAN“. It seems like this part of the protein name is a species identifier. In this case, this protein appears to be a cow protein, not a human protein. Even if this protein was in one of the samples and has been identified correctly, it wasn't produced by the body of the research subject who produced the sample and isn't part of any biological pathway in a human being. I will remove it from the table and move on.

```
## [1] "Summary counts for prefix and organism suffix values:"
```

```
##  prefix       org
##  sp:1848   HUMAN:1848
```

All remaining entries have the 'sp' prefix before the first name, and all are identified as human proteins. I will denote muscle protein names with the species identifer removed as short names and muscle protein names with the species identifer intact as long names.

```
##   prefix num_ID    longname shortname   org
## 1     sp Q43333 YS027_HUMAN     YS027 HUMAN
## 2     sp Q49293 K2C6B_HUMAN     K2C6B HUMAN
## 3     sp Q77477 CNDP2_HUMAN     CNDP2 HUMAN
## 4     sp P20289  TSP2_HUMAN      TSP2 HUMAN
## 5     sp Q68541  MTPN_HUMAN      MTPN HUMAN
## 6     sp P84691  CUL5_HUMAN      CUL5 HUMAN
```

```
## [1] "length of numeric IDs:"
```

```
##    6
## 1848
```

```
## [1] "length of 'short names':"
```

```
##    2    3    4    5
##    1   70  567 1210
```

All numeric IDs have the same length. Short names range from 2 to 5 characters.

The six character numeric IDs (one letter and five digits) in the actual muscle sample proteomics results are UniProt Accession identifiers, a type of protein/gene identifier. Because I do not actually use them for anything, the numeric IDs in the simulated data are just random strings of characters that look like UniProt Accession IDs. Likewise, the 'Probability' and 'Log(dSIn)' values in the simulated muscle protein data are random numbers and don't mean anything.

Which of the muscle proteins found in at least one sample are found in all samples?

```
##    Mode   FALSE    TRUE
## logical    1357     491
```

Indistinguishable Proteins

There are also proteins listed in the muscle protein spreadsheets as 'Indistinguishable Proteins'. As explained above, these are possible alternate identities for some of the proteins detected. They appear in the same format as muscle proteins' primary identities.

```
## [1] "Rows with missing values:"
```

```
## [1] prefix    num_ID    longname  shortname org
## <0 rows> (or 0-length row.names)
```

```
## [1] "prefix and species identifier:"
```

```
##  sp
## 489
```

```
## HUMAN
##   489
```

```
## [1] "length of numeric IDs:"
##   6
## 489
```

```
## [1] "length of 'short names':"
```

```
##   3   4   5
##  23 155 311
```

None have missing components. All are from humans. All prefixes are ‘sp’. Numeric IDs are all 6 characters long. Short names range from 2 to 5 characters.

serum_assoc2.xlsx

Here are the 28 proteins in serum associated with phenotype measures.

```
##  [1] "CABL2" "FRMD7" "REG1B" "SRBS2" "FMN1"  "SEM3F" "RAB35" "K2C8"
##  [9] "A2ML1" "FGF9"  "TGFB1" "VAX2"  "RARA"  "RG18L" "KCIP1" "CBPC2"
## [17] "MOD5"  "GTF2I" "S26A5" "LAC"   "TSP4"  "KV402" "NOMO1" "DI3L1"
## [25] "AGRIN" "SPEF2" "RSPO4" "CATB"
```

I now needed to convert all names to a common format. To summarize, I had… 1) serum protein names that look like this: KIF5A_HUMAN, DLGP5_HUMAN, MYG_HUMAN, RECQ4_HUMAN 2) Muscle protein names that look like this: YS027_HUMAN, K2C6B_HUMAN, CNDP2_HUMAN, TSP2_HUMAN 3) Muscle numeric IDs that look like this: Q43333, Q49293, Q77477, P20289 4) Serum phenotype-associated protein names that look like this: CABL2, FRMD7, REG1B, SRBS2

It would be convenient if the serum protein names and muscle protein names that end in ’_HUMAN’ were equivalent. If so, we would expect some of them to match.

```
## [1] "Muscle names that match serum names:"
```

```
##    Mode   FALSE    TRUE
## logical     547    1301
```

Many of the do.

If serum phenotype-associated protein names are also in the same format as serum protein names, then the names in the first collection should all be found in the second.

```
## [1] "Serum phenotype-associated protein names that match serum protein names:"
```

```
##    Mode    TRUE
## logical      28
```

They are.

I now have everything in one format. I will now compile a master list of all protein names to submit to the DAVID tool.

Based on the heading of the column of protein names in the serum protein spreadsheet, I selected ‘UNIPROT_ID’ as the type of identifier. 4491 of the 4767 names were identified as human proteins. Some were identified as being from unknown species. I was reasonably sure that all these were human proteins, both because they were labeled as such, as because they came from human serum or muscle samples.

I obtained four types of output from DAVID: a functional annotation chart, a list of protein/gene names not found in the functional annotation chart, a gene list report, and a list of unmapped protein/gene names not found in the DAVID database. The code in this .RMD file does not replicate this part of the analysis. Anyone who wishes to do so will need to visit the website mentioned above, manually submit the file “protein_names_sim.txt” created by this script, and save the results as “all_genes_chart_sim.txt”, “not_in_output_sim.txt”, “gene_list_report_sim.txt” and “unmapped_sim.txt”. These files are also included in the repository for convenience.

The gene list report indicates the species of each gene. As any students of molecular biology are probably aware, each protein is coded by a single gene, and each gene codes for a single protein. Thus, protein identifiers are also gene identifiers.

```
##              Homo sapiens
##          223         2216
```

These are a few of the genes from unknown species.

```
## [1] "STUB1_HUMAN" "HV101_HUMAN" "HV104_HUMAN" "HV301_HUMAN" "FTHFD_HUMAN"
## [6] "CA063_HUMAN"
```

I also have list of protein names which could not be matched to entries in the DAVID gene database. Here are the first few:

```
## [1] "RICH2_HUMAN" "F27E1_HUMAN" "LV302_HUMAN" "KV124_HUMAN" "LV204_HUMAN"
## [6] "GP116_HUMAN"
```

Perhaps not surprisingly, the proteins that failed to map to are the same ones for whom the species could not be determined.

```
##
##             Human Unknown
##   Matched    2216       0
##   Unmatched     0     223
```

Hopefully, the investigators can determine why these gene names are not being recognized by DAVID and rectify this issue.

Each row in the a functional annotation chart corresponds to one of 297 pathways containing proteins in the list I submitted. Here are the first few:

```
## [1] hsa04610:Complement and coagulation cascades
## [2] hsa04512:ECM-receptor interaction
## [3] hsa04145:Phagosome
## [4] hsa04510:Focal adhesion
## [5] hsa00010:Glycolysis / Gluconeogenesis
## [6] hsa05150:Staphylococcus aureus infection
## 286 Levels: hsa00010:Glycolysis / Gluconeogenesis ...
```

Each row also contains a list of proteins among those I submitted found in that pathway. Here is the beginning of the protein list for the first row in the table, corresponding to 'hsa01200:Carbon metabolism'.

```
## [1] "TFPI1_HUMAN, PROC_HUMAN, UROK_HUMAN, CO2_HUMAN, CO7_HUMAN, CR2_HUMAN, CO8G_HUMAN, C4BPB_HUMAN"
```

The last type of output is a list of all the protein/gene names that were not found in a biological pathway. Here is a sample:

```
## [1] "OASL_HUMAN"  "AKA12_HUMAN" "AKP13_HUMAN" "AKAP9_HUMAN" "TARSH_HUMAN"
## [6] "ADEC1_HUMAN"
```

One would expect to find all of these in the list of protein names obtained from the various sources and submitted to DAVID. In fact, there is one entry from the list of proteins not in pathways returned by DAVID which is not in the list of protein names I compiled:

```
## [1] ""
```

```
## [1] "NACAM_HUMAN, NACA_HUMAN"
```

This name is actually two names stuck together with a comma and space in between. That's why it doesn't match any of the protein names in my list. Fortunately, there are proteins in my list called "NACAM_HUMAN" and "NACA_HUMAN". I will have to manually add these protein to the list of proteins not in pathways.

I assembled a data frame with a row for each protein and columns indicating the presence of various proteins in various sources. Here is a summary showing how many proteins are and are not found in each source.

```
##  Protein_Name         Serum         Serum_Assoc     Muscle_Any    
##  Length:2439        Mode :logical   Mode :logical   Mode :logical  
##  Class :character   FALSE:592       FALSE:2411      FALSE:591      
##  Mode  :character   TRUE :1847      TRUE :28        TRUE :1848     
##  Muscle_All      No_pathways      Unmapped      
##  Mode :logical   Mode :logical   Mode :logical  
##  FALSE:1948      FALSE:1186      FALSE:2216     
##  TRUE :491       TRUE :1253      TRUE :223      
```

If the investigators are interested in the associations between proteins in muscle samples and phenotypes and how these recapitulate the associations between proteins in serum samples and phenotypes (or fail to do so), they might want to focus on proteins that are found in both serum and muscle. I used the data frame assembled above to generate some tables showing the overlap between proteins found in serum and muscle samples.

```
## [1] "Overlap between proteins in serum and any muscle sample:"
```

```
## 
##         FALSE TRUE
##   FALSE    45  547
##   TRUE    546 1301
```

```
## [1] "Overlap between proteins in serum and all muscle sample:"
```

```
## 
##         FALSE TRUE
##   FALSE   451  141
##   TRUE   1497  350
```

```
## [1] "Overlap between proteins associated with phenotypes in serum and any muscle sample:"
```

```
## 
##         FALSE TRUE
##   FALSE   580 1831
##   TRUE     11   17
```

```
## [1] "Overlap between proteins associated with phenotypes in serum and all muscle sample:"
```

```
## 
##         FALSE TRUE
##   FALSE  1921  490
##   TRUE     27    1
```

I assembled the functional annotation results into a table with each row corresponding to a protein (as in the table above), each column corresponding to a pathway, and each entry indicating whether a specific protein is found in a specific pathway. This can be used to provide information about sets of proteins and pathways, such as which pathways contain the most proteins from a given source, how many proteins are in each pathway from each source, and which proteins from a given source are in a given pathway. Here are some examples in which I list numbers or names of proteins from top 5 pathways (i.e. those with the most proteins) for a given data source or the overlap between multiple sources.

```
## [1] "Total number of proteins all pathways combined:"
```

```
## [1] 3998
```

```
## [1] "Total number of unmapped proteins all pathways combined (should be zero):"
```

```
## [1] 0
## [1] "Total number of 'no pathway' proteins all pathways combined (should be zero):"

## [1] 0

## [1] "# of proteins in serum and any muscle sample in top 5 pathways:"

##             hsa01100:Metabolic pathways
##                                      97
##                      hsa04145:Phagosome
##                                      50
## hsa04514:Cell adhesion molecules (CAMs)
##                                      46
##                    hsa04144:Endocytosis
##                                      46
##              hsa05166:HTLV-I infection
##                                      44

## [1] "# of proteins in serum and all muscle samples in top 5 pathways:"

##             hsa01100:Metabolic pathways
##                                      28
##           hsa05203:Viral carcinogenesis
##                                      18
##              hsa05166:HTLV-I infection
##                                      18
##                      hsa04145:Phagosome
##                                      17
## hsa04514:Cell adhesion molecules (CAMs)
##                                      17

## [1] "# of phenotype-associated serum proteins in top 5 pathways:"

##         hsa05200:Pathways in cancer   hsa04512:ECM-receptor interaction
##                                   3                                   2
## hsa04151:PI3K-Akt signaling pathway                    hsa05144:Malaria
##                                   2                                   2
##                hsa04144:Endocytosis
##                                   2

## [1] "Names of phenotype-associated serum proteins in top 5 pathways:"

## [1] "hsa05200:Pathways in cancer"
## [1] "RARA_HUMAN"  "TGFB1_HUMAN" "FGF9_HUMAN"
## [1] "hsa04512:ECM-receptor interaction"
## [1] "AGRIN_HUMAN" "TSP4_HUMAN"
## [1] "hsa04151:PI3K-Akt signaling pathway"
## [1] "FGF9_HUMAN" "TSP4_HUMAN"
## [1] "hsa05144:Malaria"
## [1] "TGFB1_HUMAN" "TSP4_HUMAN"
## [1] "hsa04144:Endocytosis"
## [1] "TGFB1_HUMAN" "RAB35_HUMAN"
```

In addition to providing answers to specific queries about proteins and pathways, I gave them a CSV file containing information about both protein sources and pathways combining the two tables described above. Each row corresponds to a protein. The first seven columns indicate whether each protein can be found in a given source. The other 286 columns indicate whether each protein can be found in a given pathway. This way, the investigators are able answer any questions they come up with later about proteins, sources and pathways without my help. The code in this R Markdown document will generate a file called “prot_path.csv” using the other files in this repository. This is the final product: clean data that the investigators can use to conduct additional analyses as they wish.