

Appendix A
**Python Source Code for Taverna Outputs Manipulation in Data
Preparation Process of ADR-Protein Associations and Example**

A.1 Python Source Code

```
def filecount(dir_name):
    # return the number of files in directory dir_name
    try:
        return len([f for f in os.listdir(dir_name) if os.path.isfile(os.path.join(dir_name,
f))])
    except Exception, e:
        return None

import os, time, shutil, sys

# Edit followed 3 lines (question sign)
total_protein = ???
folder_name = "???\\"
path = "???\\"
#output_file = open('C:\Users\parwin\Downloads\testout.txt','w')
#print path

for i in range(total_protein):
    pmid_path = path+folder_name+"pmid\\"+str(i)+".text"
    #print pmid_path
    pmid_file = open(pmid_path)
    pmid = pmid_file.readlines()
    #print output
    protein_path = path+folder_name+"ValidatedProtein\\"+str(i)+"\\"
    uniprot_path = path+folder_name+"UniProtID\\"+str(i)+"\\"
    if os.listdir(protein_path)!=[]:
        n = filecount(protein_path)
        m = filecount (uniprot_path)
        #print n
        #print m
        if m != 0:
            for j in range(m):
                protein_file = protein_path+str(j)+".text"
                uniprot_file = uniprot_path+str(j)+".text"
                #print protein_file
                #print uniprot_file
                protein_open = open(protein_file)
                uniprot_open = open(uniprot_file)
                #print protein_open
                #print uniprot_open
                protein = protein_open.readlines()
                uniprot = uniprot_open.readlines()
                #print protein
                #print uniprot
                if protein != []:
                    for k in range(len(uniprot)):
                        output = pmid[0]+"\\t"+protein[0]+"\\t"+uniprot[k]
                        print output
```

```
#output_file.write(output)
#else:
#print "empty folder"+str(i)
```

A.2 Example

As shown in Figure A.1, mining the relevant protein with Biliary cirrhosis found three papers but only the third paper (number 17412303) had validated protein named CYP2D6 and six UniProt IDs.

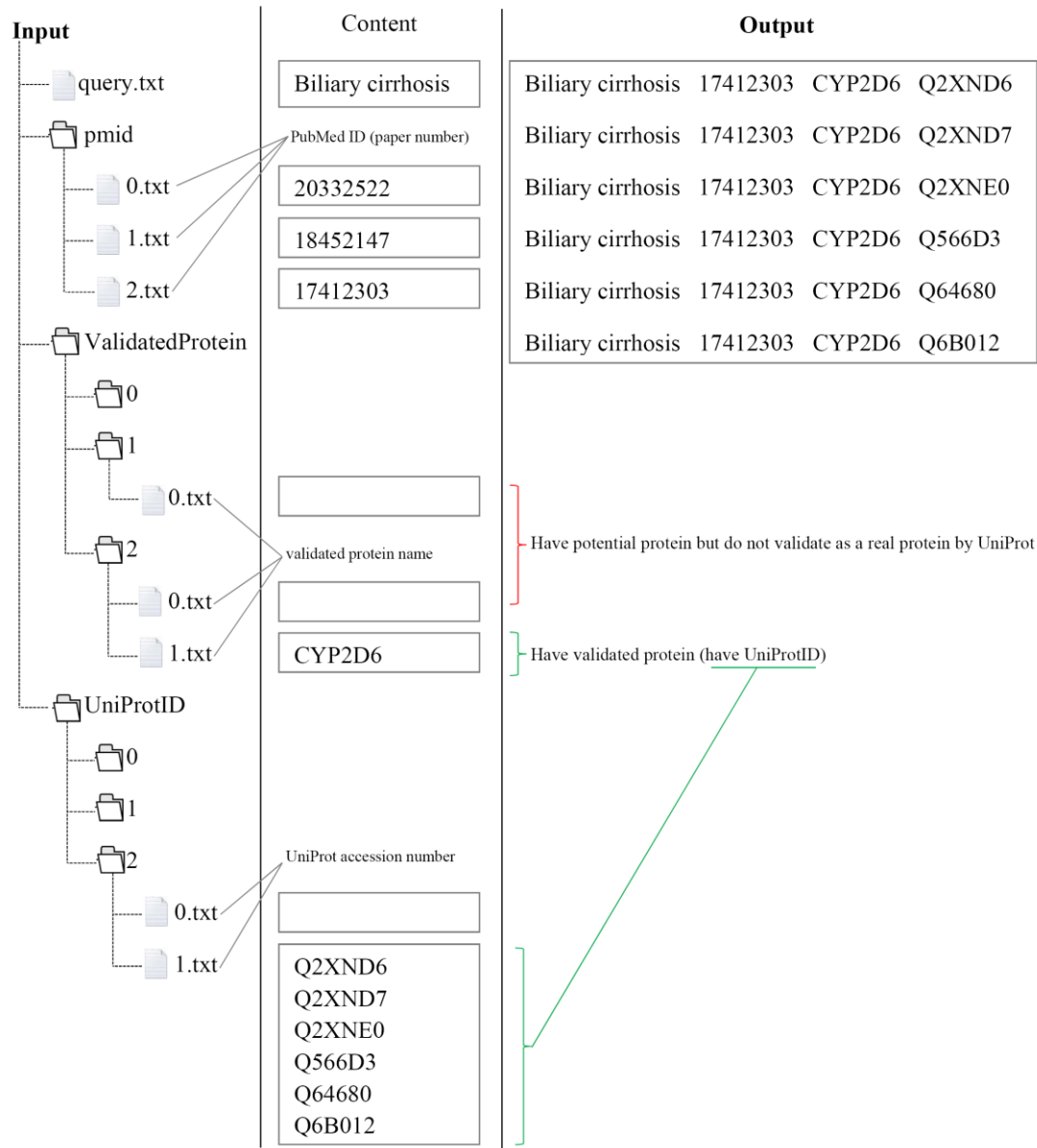


Figure A.1 Example of input and output from Python manipulating extracted ADR-Protein association using Modified BioAID_ProteinDiscovery workflow in Taverna