Make new IdentDB

Tobias Opialla

5 July 2019

# read current DB

data = scan('Z:/Exchange/\_kempa\_mtx/\_\_\_SILVIA/\_Maui-DBs/KempaHMDBKEGG\_Ident.msp', what=character(), sep = "\n", blank.lines.skip = FALSE)  
#data = scan("Kempa\_Ident.MSP", what=character(), sep = "\n", blank.lines.skip = FALSE)  
  
#fieldids=gsub("(^.\*\\:)(.\*)","\\1",grep(":",data,value=T))  
fieldids=gsub("(^.\*\\: )(.\*)","\\1",grep(":",data,value=T))  
  
  
  
data\_single = paste(data, collapse="\n", sep="")  
data = unlist(strsplit(data\_single, split="(?<!Original)(?i)Name: ",perl=T))  
data = paste("Name: ", data, sep="")  
data=data[-1] #1st entry is always empty  
  
  
mydataframe=data.frame(Name=gsub("(Name: )(.\*?\n)(.\*)","\\2",data),  
 Author=gsub("(.\*Author: )(.\*?\n)(.\*)","\\2",data),  
 #Formula=gsub("(.\*Formula: )(.\*?\n)(.\*)","\\2",data),  
 #MW=gsub("(.\*MW: )(.\*?\n)(.\*)","\\2",data),  
 ID=gsub("(.\*ID: )(.\*?\n)(.\*)","\\2",data),  
 #Comment=sub("(.\*Comment: )(.\*?\n)(.\*)","\\2",data),  
 #CASNO=gsub("(.\*CASNO: )(.\*?\n)(.\*)","\\2",data),  
 #HMDBID=gsub("(.\*HMDBID: )(.\*?\n)(.\*)","\\2",data),  
 #KEGGID=gsub("(.\*KEGGID: )(.\*?\n)(.\*)","\\2",data),  
 #OriginalName=gsub("(.\*OriginalName: )(.\*?\n)(.\*)","\\2",data),  
 #querystring=gsub("(.\*querystring: )(.\*?\n)(.\*)","\\2",data),  
 # Synon=gsub("(.\*Synon: )(.\*?\n)(.\*)","\\2",data),  
   
 Num.Peaks=gsub("(.\*Num Peaks: )(.\*)","\\2",data), #Num Peaks is always last entry in MSP format  
 stringsAsFactors=F)  
  
#clean the import, in case field is not in db entry, whole entry is returned  
for(mycolname in colnames(mydataframe)){  
   
 mydataframe[[mycolname]][nchar(mydataframe[[mycolname]])==nchar(data)]="\n"  
}

# read new RI-entries

df\_new\_RIs=read.table('./New\_RIs\_2019-07-02.tsv',header = T,sep="\t",stringsAsFactors = F)  
df\_new\_RIs$Name2match=paste0(df\_new\_RIs$Name,"\n")

# merge tables (without touching original “mydataframe”)

df=full\_join(df\_new\_RIs,mydataframe,by=c("Name2match" = "Name"),suffix=c("\_new","\_old"))

# create new names

df$Name\_new=paste0(gsub("(.\*\_RI:)(\\d+)(\_.\*)","\\1",df$Name2match),  
 df$RI\_new\_mean,  
 gsub("(.\*\_RI:)(\\d+)(\_.\*)","\\3",df$Name2match)  
 )  
df$Name\_new[is.na(df$RI\_new\_mean)]=df$Name2match[is.na(df$RI\_new\_mean)]

# replace names in old db

mydataframe$Name=df$Name\_new[match(mydataframe$Name,df$Name2match)]  
mydataframe$Author=paste0("MauiSILVIAPHJLK","\_RI\_restek\n")

# write new db file

##write msp file  
data2wrt="mydataframe"  
filename="Kempa\_Restek\_column\_Ident.msp"  
  
mypaste=function(x){paste(paste(names(x),x,sep=": "),sep="\n")}  
  
myfilename=function(myfile='file.msp'){  
 fn=myfile  
 if(grepl('.msp$',fn)){  
 fn.tail=".msp"  
 fn.short=gsub('.msp$','',fn)  
 }else{  
 lf.tail=''  
 fn.short=fn  
 }  
 if(file.exists(fn)){  
 myfile=paste0(fn.short,'\_',gsub(':','-',gsub(' ','\_',Sys.time())),fn.tail)  
 }  
 return(myfile)  
}  
  
  
if(colnames(get(data2wrt))[length(colnames(get(data2wrt)))]=="Num.Peaks"){ #check that columns in df have correct order  
 mystring=paste(apply(get(data2wrt),1,FUN=mypaste),collapse="")  
 mystring=gsub("Num.Peaks","Num Peaks",mystring)  
 cat(mystring,file=myfilename(filename),sep="\n")  
}else{print("#Num Peaks has to be always last entry in MSP format, adjust column order!")}