#Voor het geval dat elo een hekel heeft aan shebanglines

#!/bin/bash

# Additional information:

# =======================

# Dit script is geschreven door Sebastiaan de Vriend en

# kan gebruikt worden bij de opdracht van bapgc door

# Jesse Kerkvliet.

# Het script is geschreven om de bapgc pipeline te starten.

# In het begin wordt er een temp en result directory gemaakt.

# Daarna wordt gevraagd welke pathway ge-analyseerd

# kan worden. Het resultaat wordt geschreven in outputfiles en

# de tijdelijke bestanden worden opgeslagen in de directroy temp.

# Ook wordt er aan de user gevraagd of deze packages wilt

# Downloaden die gebruikt worden in de pipeline.

# Als de gebruiker geen packages heeft, wordt er aangeraden

# om hier ja op te beantwoorden.

# Opties voor pathways.

# Deze line wordt maximaal 3, voor de verschillende pathways.

# Declaratie folder bestanden

#create new directories

# Starten pipeline. Verander directory om lokale scripts aan te

# roepen.

# Additional information:

# =======================

# Dit script is geschreven door Sebastiaan de Vriend en

# kan gebruikt worden bij de opdracht van bapgc door

# Jesse Kerkvliet.

# Dit script roept losse script aan voor de BAPGC pipeline

# en zorgt ervoor dat alle tusenbestanden op correcte manier beheerd

# worden. Het script kan aangeroepen worden met:

# ./BAPGC-Pipeline.sh ResultaatDirectory TijdelijkeDirectory j/n Pathwaynaam.

#

# Het script loopt alle tussenstappen af voor het maken van een verslag

# en de daarbij behorende data te genereren met diverse

# scripts. Op het laatst wordt er een phylogenetische boom

# gemaakt met ClustalW door fasta bestanden te mergen in een enkele

# fasta file.

# Declaratie bestanden en namen.

# Optie om alle packages te downloaden. Hiervoor zijn root rechten nodig.

# Maken van multi fasta met de volgende twee pipes:

# Functie download alle onderstaande packages. Dit is handig voor bij

# een schone installatie. De functie heeft root rechten nodig

# om alle files te installeren.

# Additional information:

# =======================

# Dit script is geschreven door Sebastiaan de Vriend en

# kan gebruikt worden bij de opdracht van bapgc door

# Jesse Kerkvliet

#

# Het script download de benodigde packages die gebruikt

# worden bij de bapgc opdract.

# Alle gevonden medegereguleerde genen worden ingeladen.

# De honderd bovenste hits van de lijst worden genomen en de

# ID's worden omgezet naar Entrez-gene ID's

# De annotaties worden opgehaald. Hierin zitten de Entrez-gene ID,

# functie en EC-nummer. Alle annotaties met een EC-nummer worden

# opgehaald. De hypothetical, unknowns en uncharaterizeds worden

# gezocht en verwijdert en de functie om een venndiagram te genereren

# wordt aangeroepen.

#De functie splitst de ID's op een punt en de unieke lijst

#Wordt teruggegeven

#De functieloze enzymen worden gezocht,

#Een logical object wordt gemaakt met alle enzymen die een EC-nummer hebben

#(dat zijn ze allemaal, het typeert een enzym)

#De twee logische klassen worden vergeleken en een venndiagram wordt gemaakt

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Het script genereert een venndiagram met enzymen zonder bekende functie.

# en slaat deze op als afbeelding

# De functie leest tree.ph in en visualiseerd deze met ggtree

# De visualisatie wordt opgeslagen als tree.png

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Het script maakt een visualisatie van tree.ph en slaat

# deze op als tree.png. Het script kan aangeroepen worden met:

# ./generateTree.R Werkmap

# De functie is een wat langere functie. Dit is omdat de input data

# in tien stukken wordt gehakt om het zo mogelijk te maken om

# het script op een computer te draaien zonder dat er memmory

# errors opkomen.

# De functie leest als eerste FoundMotifs.csv in en zet deze in

# de variable foundMotifs. Daarna worden 2500 genen geladen met

# de functie getMatches en worden deze opgeslagen. Daarna worden de

# genen opgehaald met de functie getGenes. Als laaste wordt het

# totale resultaat samengevoegd en weggeschreven naar

# AllCoregulatedGenes.csv

# input: 2

# left: nummer van gen om links mee te beginnen

# right: nummer van gen om mee te eindigen.

#

# De functie haalt de data van FoundMotifs op en gebruikt

# deze samen met genome om alle transcripten op te zoeken

# die voor de range van left en right gelden. Daarna worden motifs

# opgezocht voor de mens die uit FoundMotifs komen en daar worden

# matchende genen voor gezocht door searchSeq. Als laaste

# wordt er een dataframe terug gegeven met de gevonden genen en

# daarbij de scores.

# Maakt motifList m.b.v JASPAR2014. De opties voor mens en PWM worden

# Meegegeven

# Input: 2

# dfMotifs: Dataframe met Motifs en de daarbij behorende score en genen.

# amount: : integer met lengte van aantal motifs van pathway.

#

# De functie kijkt per subset motifs of de score hoger is dan 97%

# daarna worden alle unieke motifs toegevoegd aan de vector found.

# Daarna worden alle unieke motifs terug gegeven.

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

# Het script zoekt naar genen die matchen voor motifs uit het bestand

# FoundMotifs.csv en slaat deze op.

# Het script kan als volgt worden aangeroepen:

# ./GenesFromMotifs.R Werkmap

# Functie haalt informatie van pathway op en zet deze in een csv file.

#

# lstPathwayInfo: Pathway object van kegg

# strPathwayName: String met pathway naam

# strGeneIDs: Char string. 1e element wordt genomen.

# 2e element wordt overgeslagen.

# dfGenes: Omgezette geneid's.

# png: Variable met pathway afbeelding.

#

# Na het omzetten van de geneid's worden deze in een dataframe gestopt

# zodat de pathwaynaam als kolom meegegeven kan worden.

# Het bestand wordt opgeslagen als csv bestand.

# Daarnaast wordt de afbeelding van de pathway gedownload en

# opgeslagen.

# Additional information:

# =======================

# Dit script is geschreven door Sebastiaan de Vriend en

# kan gebruikt worden bij de opdracht van bapgc door

# Jesse Kerkvliet

#

# Het script download de gegevens van een opgegven pathway

# en zet deze in de opgegeven directory in args 1.

#Gene database en genoom worden gedefinieerd

#Alle gevonden genen worden geladen en de eerste 10 worden meegenomen

#De ID's worden omgezet naar Entrez Gene ID's

# Alle exons van de genen worden geladen,

# hiervan wordt de sequentie geladen

# De sequenties worden getransleerd naar AA sequenties

#De hydrofobiciteit wordt geteld en de barplot wordt gegenereerd

#De functie splitst de genen op een punt en geeft de

#Entrez ID’s terug

#Een set van unieke ID's wordt gegenereerd en een lege matrix wordt gemaakt

#Voor ieder ID in de unieke ID list wordt een set van eiwitsequenties

#gemaakt, horend bij dat ID.

#Ieder aminozuur wordt geteld. De set van hydrofobische AA's wordt geteld

#en de set van hydrofiele AA's wordt geteld.

#Een volledige som van het eiwit wordt berekend. Alle sets worden

#in de matrix opgeslagen. Deze wordt teruggegeven.

# De aantallen matrix wordt omgedraaid (rijen worden kolommen en v.v)

# Een figuur wordt gemaakt, en een plot wordt gegenereerd en weggeschreven

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Een barplot van de hydrofobiciteit van de beste 10 medegereguleerde

# genen wordt gegenereerd en opgeslagen in een afbeelding.

#Gene database en genoom worden gedefinieerd

#De medegereguleerde genen worden ingeladen

#en de 20 beste hits worden meegenomen en omgezet naar Entrez ID's

#Voor ieder gen worden de exonen en de transcripten opgehaald.

#De lengte hiervan wordt berekend en aan een matrix toegevoegd.

#De matrix met lengten wordt omgezet naar een dataframe en een plot

#Wordt weggeschreven naar een PNG bestand.

#De functie splitst de genen op een punt en geeft de

#Entrez ID’s terug

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Een scatterplot van de lengten van de intronen en exonen

# wordt gegenereerd voor de beste 20 medegereguleerde genen.

# De functie laad een optielijst in waar PWM en de JASPAR2014

# nummer van de mens in gestopt worden. De lijst wordt gebruikt

# om een lijst met alle motifs op te halen.

# Ook wordt de variable promoseqs geladen en daarvan wordt

# bepaald hoeveel sequenties erin zitten. Daarna wordt er een

# unlist uitgevoerd die gebruikt wordt door searchSeq waar

# motifs eruit komen met een score van minimaal 80%. Dit is gedaan

# omdat er veel resultaten zijn, en er nog zat gefilterd kan worden

# op deze manier op het resultaat.

#

# De sequences en motifs worden vervolgens in een dataframe gezet en

# wordt de dataframe doorgestuurd naar getMotifs. Het resultaat

# daarvan wordt opgeslagen in FoundMotifs.csv.

# Maakt de parameters voor de volgende stap

# Maakt motifList m.b.v JASPAR2014. De opties voor mens en PWM worden

# Meegegeven.

# De motifs worden in de promotorsequenties opgezocht.

# De genen worden langs de motiflijst gehaald. Threshold 80%

# Maakt een dataframe van de searchSeq output

# Haalt de namen van de gevonden motifs uit het dataframe

# dfMotifs: Dataframe met gevonden motieven en scores per gevonden hit.

# intAmount: Integer: Lengte van aantal PromoterSequenties.

# De functie kijkt per subset motifs of de score hoger is dan 95%

# daarna worden alle unieke motifs toegevoegd aan de vector found.

# Daarna worden alle unieke motifs terug gegeven.

# De motifs die een relatieve score van 95% hebben en voor ieder gen gelden

# Worden opgeslagen in een vector. Deze wordt teruggegeven.

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Het script zoekt motifs bij bijbehorende promotorsequenties die

# geladen worden uit de R datafile promoseqs.

# Unieke motifs worden opgeslagen in: FoundMotifs.csv.

# Het script kan als volgt worden aangeroepen:

# ./MotifsFromPromoterSequences.R WerkMap

#Gene database en genoom worden gedefinieerd

#De lijst met gecoreguleerde genen wordt geladen

# De bovenste vijftig hits worden hiervan genomen en de ID's

# Worden omgezet naar Entrez Gene ID's

# De sequenties en de lengten van de genen worden in een DNAStringSetList

# en een vector opgeslagen

# De alignment wordt uitgevoerd

# De naam en sequentie van het langste gen worden opgeslagen

# Het percentage conservering van het langste gen wordt berekend

# En de resultaten worden weggeschreven

#De functie splitst de genen op een punt en geeft de

#Entrez ID’s terug

#De consensus van de alignment wordt genomen en

#De ?s worden verwijderd

# De lengte van de sequentie van het langste gen

# en van de alignment worden geteld. Als twee tekens overeen komen,

# worden deze opgeslagen. Hieruit wordt een percentage berekend.

# Het percentage wordt teruggegeven.

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Genereert een MSA en berekent het percentage conservering van

# Het langste gen

#De functie haalt alle sequenties op uit de knownGene database.

# Er wordt gekeken of ieder gen een bijbehorende sequentie heeft.

# Is dit niet het geval, wordt het gen in de blacklist gestopt.

# De genen in de blacklist worden verwijderd uit de set.

# Als er genen zonder sequentie zijn, worden deze verwijderd.

# De sequenties worden voor ieder gen opgehaald

# De sequenties worden teruggegeven

#Voor ieder gen wordt gekeken hoe veel paralogen

#aanwezig zijn. Als er meer dan vier paralogen worden

#gevonden, worden alleen de eerste vier meegegeven.

#Een matrix met hooguit vier paralogen per gen wordt teruggegeven.

#De functie splitst de ID's op een punt en de unieke lijst

#Wordt teruggegeven

#Gene database en genoom worden gedefinieerd

#De beste 10 beste medegereguleerde worden geladen

#En omgezet naar Entrez-gene ID's

# Een biomart met de ensembl genen wordt aangemaakt.

# Voor alle genen worden de

# Voor ieder gen worden hooguit vier paralogen opgehaald.

# De genen zonder paralogen worden verwijderd.

# De namen worden opgeslagen van de oorspronkelijke genen

# En de paralogen. De namen worden aan een vector met alleen

# namen toegevoegd en de sequenties worden opgehaald.

#Iedere sequentie wordt weggeschreven naar een fasta bestand.

# Additional information:

# =======================

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

#

# Het script genereert fasta bestanden voor de twintig best medegereguleerde

# genen en tot vier paralogen.

# De functie maakt gebruik van packages om promotor sequenties

# op te halen van de genen die bij een pathway horen.

# De promotor sequenties worden gezocht in de regio's 3000 upstream

# en 300 downstream voor optimale resultaten. De promotor

# sequenties die gevonden worden, worden opgeslagen in de variable

# promoseqs. Deze wordt opgeslagen als R.Datafile om in latere

# stappen gebruikt te worden.

# Gene Database (HG38)

# Het Hsa genoom

# De Entrezgene codes van de pathway inladen

# Het dataframe omzetten in een vector

# De transcripts van de lijst met genen zoeken

# De promotersequenties worden opgehaald met 3000 upstream 300

# Schrijf naar rdata file.

# Additional information:

# =======================

#

# Dit script is geschreven door Jesse Kerkvliet en

# kan gebruikt worden bij de opdracht van bapgc door

# Sebastiaan de Vriend.

# Het script zoekt promotorsequenties van genen van een ingeladen

# pathway. Het script kan als volgt aangeroepen worden:

# ./PromoterSeqsFromGenes Werkmap

# Als output wordt een Rdata file opgeslagen in de werkmap.

# Deze kan gebruikt worden voor vervolgstappen.