Script Name:

`evaluate_orthoDBv11.ipynb`

Purpose

Evaluation of Zen OrthoDB Mapping: Benchmarking UniProtKB The UniProt-Plants dataset, downloaded on April 29, 2024, was analyzed to determine the percentage of entries in the TrEMBL and Swiss-Prot accessions that contain OrthoDB (v.11) annotations. This dataset was used as the ground truth for comparison with Zen mapping. The script 'extract_orthodb.py' was used to extract UniProtKB accession IDs and their corresponding OrthoDB annotations from the UniProt data files (uniprot_sprot_plants.dat and uniprot_trembl_plants.dat). It generates two files: sp_acc2orthodb.txt (for Swiss-Prot) and tr_acc2orthodb.txt (for TrEMBL). The comparison between UniProt-OrthoDB and Zen mapping was conducted using the Jupyter notebook script 'evaluate_orthoDBv11.ipynb'. Scripts are publicly accessible on GitHub at 'https://github.com/aghelfi/HayaiAnnotation/zen_odbv11_evaluation'.

Author

Andrea Ghelfi

Date

November 20, 2024

License

GNU GPL-3.0 License

Software

R version 4.4

In [ ]:
```r
library(data.table)
```

In [ ]:
```r
# OrthoDB to parental from OrthoDB v.11.
og_pairs <- fread("odb11v0_OG_pairs.tab", header = FALSE, sep = "\t")
colnames(og_pairs) <- c("OrthoDB", "parental_OrthoDB")
```

In [ ]:
```r
# Select Eukaryotes to match UniProtKB - OrthoDB taxon level
og_pairs_euk <- og_pairs[grep("at2759", og_pairs$parental_OrthoDB), ]
```

In [ ]:
```r
# Inferred OrthoDB using Zen (OrthoDB v.11 and UniProtKB dataset download
inferred_ogs <- fread("uniprot2orthodb.tsv", header = TRUE, sep = "\t")
# Remove accessions without assigned OrthoDB
inferred_ogs <- inferred_ogs[!is.na(inferred_ogs$OrthoDB), ]
```

In [ ]:
```
# Extracted OrthoDB from UniProt-Plants: Swiss-Prot
all_sp <- fread("sp_acc2orthodb.txt", header = FALSE, sep = "\t")
colnames(all_sp) <- c("AC", "UK_OrthoDB")
print("Total accession in SwissProt")
nrow(all_sp)
```

In [ ]:
```
[1] "Total accession in SwissProt"
[1] 44592
```

In [ ]:
```
# Extracted OrthoDB from UniProt-Plants: TrEMBL
all_tr <- fread("tr_acc2orthodb.txt", header = FALSE, sep = "\t")
colnames(all_tr) <- c("AC", "UK_OrthoDB")
```

In [ ]:
```
print("Total accession in TrEMBL")
nrow(all_tr)
```

In [ ]:
```
[1] "Total accession in TrEMBL"
[1] 19023625
```

In [ ]:
```
# Add parental levels for the inferred ogs
par_inf_ogs_euk <- merge(inferred_ogs, og_pairs_euk, by = "OrthoDB")
```

In [ ]:
```
# Add database name for UniProtKB-Plants
all_sp$db <- "sp"
all_tr$db <- "tr"
```

In [ ]:
```
# Merge both SwissProt and TrEMBL databases
all_uni <- rbind(all_sp, all_tr)
colnames(all_uni) <- c("AC", "UK_OrthoDB", "db")
```

In [ ]:
```
print("Total accession in UniProtKB-Plants")
print(nrow(all_uni))
```

In [ ]:
```
[1] "Total accession in UniProtKB-Plants"
[1] 19068217
```

In [ ]:
```
# Remove accessions without assigned OrthoDB in UniProtKB
uni <- all_uni[!is.na(all_uni$UK_OrthoDB), ]
print("Total of UniProt entries with OrthoDB ID")
print(nrow(uni))
```

In [ ]:
```
[1] "Total of UniProt entries with OrthoDB ID"
[1] 2729135
```

In [ ]:
```
# Percentage of UniProt accessions with OrthoDB in UniProtKB-Plants
print("Percentage of UniProt accessions with OrthoDB in UniProtKB-Plants"
print(nrow(uni)/nrow(all_uni)*100)
```

In [ ]:
```
[1] "Percentage of UniProt accessions with OrthoDB in UniProtKB-Plants"
[1] 14.31248
```

In [ ]:
```
# Count occurrences of each value in the 'db' column
db_counts <- uni[, .N, by = db]
```

In [ ]:
```
# Print the result
print("Number of occurrences per database")
```

```
print(db_counts)
```

In [ ]:
```
[1] "Number of occurrences per database"
       db       N
   <char>   <int>
1:     sp   22658
2:     tr 2706477
```

In [ ]:
```
# Percentage of OrthoDB in UniProt per database
print("Percentage of OrthoDB in SwissProt")
print(db_counts$N[1]/nrow(all_sp)*100)
[1] "Percentage of OrthoDB in SwissProt"
[1] 50.8118
```

In [ ]:
```
print("Percentage of OrthoDB in TrEMBL")
print(db_counts$N[2]/nrow(all_tr)*100)
```

In [ ]:
```
[1] "Percentage of OrthoDB in TrEMBL"
[1] 14.22693
```

In [ ]:
```
# Join the inferred orthodb by Zen mapping with original data from UniPro
all_compare <- merge(par_inf_ogs_euk, uni, by = "AC", all.y = TRUE)
```

In [ ]:
```
dim(all_compare[!is.na(all_compare$parental_OrthoDB), ])
[1] 1950290       5
```

In [ ]:
```
# Removed accessions without assigned parental level for orthodb
compare <- all_compare[!is.na(all_compare$parental_OrthoDB), ]
nrow(compare)
[1] 1950290
```

In [ ]:
```
same <- compare[compare$parental_OrthoDB == compare$UK_OrthoDB,]
```

In [ ]:
```
# Correspondence Zen and Uniprot (no NAs)
print("Percentage of accessions with parental OrthoDB in UniProtKB")
print(nrow(compare)/nrow(uni) * 100)
print("Total number of accessions with OrthoDB accessions and Zen mapping
print(nrow(compare))
print("Accessions were Zen mapping matched OrthoDB in UniProtKB, consider
print(nrow(same))
print("Percentage of accessions were Zen mapping matched OrthoDB in UniPr
print(nrow(same)/nrow(compare) * 100)
```

In [ ]:
```
[1] "Percentage of accessions with parental OrthoDB in UniProtKB"
[1] 71.46184
[1] "Total number of accessions with OrthoDB accessions and Zen mapping"
[1] 1950290
[1] "Accessions were Zen mapping matched OrthoDB in UniProtKB, considerin
[1] 1943023
[1] "Percentage of accessions were Zen mapping matched OrthoDB in UniProt
[1] 99.62739
```

In [ ]:
```
# After Zen Mapping
zen_sp <- merge(inferred_ogs, all_sp, by = "AC")
print("Zen mapping in SwissProt")
print(nrow(zen_sp))
print("Zen mapping in SwissProt% ()")
```

```
print(nrow(zen_sp)/nrow(all_sp)*100)
```

In [ ]:
```
[1] "Zen mapping in SwissProt"
[1] 38765
[1] "Zen mapping in SwissProt% ()"
[1] 86.93263
```

In [ ]:
```
zen_tr <- merge(inferred_ogs, all_tr, by = "AC")
print("Zen mapping in TrEMBL")
print(nrow(zen_tr))
print("Zen mapping in TrEMBL% ()")
print(nrow(zen_tr)/nrow(all_tr)*100)
```

In [ ]:
```
[1] "Zen mapping in TrEMBL"
[1] 9170680
[1] "Zen mapping in TrEMBL% ()"
[1] 48.2068
```

In [ ]:
```
print("Zen mapping in UniProtKB")
print(nrow(zen_sp)+nrow(zen_tr))
```

In [ ]:
```
[1] "Zen mapping in UniProtKB"
[1] 9209445
```

In [ ]:
```
print("Total accession in UniProtKB-Plants")
print(nrow(all_sp)+nrow(all_tr))
```

In [ ]:
```
[1] "Total accession in UniProtKB-Plants"
[1] 19068217
```

In [ ]:
```
print("Total accession in UniProtKB-Plants (%)")
print(((nrow(zen_sp)+nrow(zen_tr))/(nrow(all_sp)+nrow(all_tr)))*100)
```

In [ ]:
```
[1] "Total accession in UniProtKB-Plants (%)"
[1] 48.29736
```

In [ ]:
```
fwrite(compare, "Zen_OrthoDB_vs_UniProt_OrthoDB_20240429.tsv", row.names
```