Enrichment analysis

This project contains code for performing domain enrichment analysis on a list of proteins and quantifying the disordered region composition of the proteins.

Setup

1. Check if perl is installed. It should be available on MacOS and Linux by default.

```
which perl
```

If not, install it from https://www.perl.org.

2. Install the cpanminus package.

```
cpan App::cpanminus
```

To make the `cpanm` command available, do one of the following:

- restart your terminal
- if you use zsh, run `source ~/.zshrc`
- if you use bash, run `source ~/.bashrc`

3. Install perl dependencies

```
cpanm String::Util List::MoreUtils Text::CSV_XS
Text::NSP::Measures::2D::Fisher::right JSON::Parse Statistics::Basic
Statistics::Robust::Scale
```

Domain enrichment

More information can be found in the file `domain-enrichment/README.md`.

1. Create directories for storing input data and output

```
mkdir data && mkdir output
```

2. Download Pfam data and unzip

```
wget
ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/proteomes/9606.tsv.g
```

```
z -O data/9606.tsv.gz && gunzip data/9606.tsv.gz
```

If `wget` is not installed, use curl:

```
curl
ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/proteomes/9606.tsv.g
z -o data/9606.tsv.gz && gunzip data/9606.tsv.gz
```

3. Download reviewed UniProt entries and unzip

```
wget
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/c
omplete/uniprot_sprot.dat.gz -O data/uniprot_sprot.dat.gz && gunzip
data/uniprot_sprot.dat.gz
```

If `wget` is not installed, use curl:

```
curl
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/c
omplete/uniprot_sprot.dat.gz -o data/uniprot_sprot.dat.gz && gunzip
data/uniprot_sprot.dat.gz
```

4. Give perl scripts execution permissions

```
chmod +x domain-enrichment/domain-extraction.pl domain-
enrichment/enrichment.pl
```

5. Extract known protein domains from reviewed UniProt entries

```
./domain-enrichment/domain-extraction.pl -p data/9606.tsv -u
data/uniprot_sprot.dat
```

This will create a file called `pfam-domains.txt` in the `output` directory.

6. Calculate the domain enrichment for a list of genes. The input list must have two columns, containing the gene name and the UniProt ID. You'll need to create this for the genes you are interested in. Example lists are in the `lists` directory. A background list can also be passed in with the `-b` flag. See the `domain-enrichment/README.md` file for more information.

```

```

```
./domain-enrichment/enrichment.pl -g gene-list.txt -t output/pfam-
domains.txt
```

This will create a file called `enrichment.txt` in the `output` directory.

Region composition

More information can be found in the file `region-composition/README.md`.

1. Give perl scripts execution permissions

```
chmod +x region-composition/fractional-summary.pl region-composition/pfam-
fetch.pl
```

2. Fetch region information for UniProt entries. This takes a while as Pfam's API is slow.

```
./region-composition/pfam-fetch.pl -u data/uniprot_sprot.dat
```

This will create two files in the `output` directory:

- `pfam-motifs.txt`: list motifs for each gene name/UniProt ID with no duplicate motifs per gene
- `pfam-motifs-complete.txt`: all motifs (including duplicates), along with start and end positions of the motif and length

3. Compute fractional summary. This will produce stats on a motif in question for both the total proteome and a list of query genes. It will determine the mean and median fractional protein length for the region of interest (ROI) (a protein may have multiple instances of the region, so this is the fraction of the protein that has it, summing up the lengths of the separate regions), the mean and median longest region (for each protein the longest region is retained) and the mean and median total amino acid count with that motif (counts all amino acids comprising the motif of interest per protein). A background list can be passed in with the `-b` flag. See the `region-composition/README.md` file for more information.

```
./region-composition/fractional-summary.pl -l genes.txt -m disorder -p
output/pfam-motifs-complete.txt
```

- `histograms.txt`:
  - bins for fractional protein length that is the region of interest
  - bins go from 0-5% to 95-100%
  - first column is for background and second for genes of interest
- `<region>-summary.txt`:
  - <region> will be replaced by the region/motif name being summarized
  - summary includes mean and median for background and genes of interest

- `wilcox_*_*.txt`:
    - list of values for all background proteins or genes of interest for performing Wilcoxian test

Visualization

1. Install R and necessary packages `cowplot`, `ggplot2` and `reshape2`.
2. The values in the `wilcox_*_*.txt` files can be used to make box or violin plots with `region-composition/boxplot.R`.

For example, to create plots for the longest region, create a text file as below with the columns containing the values from the `wilcox` files

```
Background  List1  List2
0.02        0.5    0.17
0.57        0.34   0.18
...         ...     ...
```

Specify the name of the file in the script and the column headings to use.

Significance tests

To test for significance between the fractional region length of a query list versus the background (or the longest region instead of fraction), the `wilcox_*.*.txt` files can be used for performing a Wilcox test using the script `region-composition/wilcox_test.R`

To run, open R and specify file names to import and compare.