scanRBP

What is scanRBP?

scanRBP loads RNA-protein binding motif PWM and computes the log-odds scores for all the loaded RBPs across a given genomic sequence + draws a heatmap of the scores.

The scores can be described as follows (biopython docs):

Unset
```
Here we can see positive values for symbols more frequent in the motif than in
the background and negative for symbols more frequent in the background. 0.0
means that it's equally likely to see a symbol in the background and in the
motif.

Using the background distribution and PWM with pseudo-counts added, it's easy
to compute the log-odds ratios, telling us what are the log odds of a
particular symbol to be coming from a motif against the background.
```

For more information, see the biopython docs.

Installation

The easiest way to install **scanRBP** is to simply run:

Unset
```
pip install scanRBP
```

Note that on some systems, **pip** is installing the executable scripts under `~/.local/bin`. However this folder is not in the PATH which will result in “`command not found”` if you try to run `”scanRBP”` on the command line. To fix this, please execute `“export PATH="$PATH:~/.local/bin”` (and add this to your `.profile`). Another suggestion is to install inside a virtual environment (using `virtualenv`).

If you would like instead to **install the latest developmental version** from this repository:

Unset

```
# clone scanRBP GitHub repository
git clone https://github.com/grexor/scanRBP.git

# build and install
./build.sh
```

Quick Start

scanRBP quick start:

Unset

```
Usage for single sequence: scanRBP sequence output [options]
     * one sequence provided on the command line, generates output.png/pdf +
output.tab

Usage for processing FASTA file: scanRBP filename.fasta [options]
     * one heatmap/matrix will be generated per sequence
     * output name of the files will be sequence ids provided in the fasta file

Options:
     -annotate               Annotate each heatmap cell with the number
     -xlabels                Display sequence (x-labels), default False
     -only_protein TARDBP    Only analyze binding for the specific protein /
search by name
     -all_protein TARDBP     Additionally to one motif per protein (for all
proteins), also include all motifs (PWMs) for this specific protein (search by
name)
                             (note that one protein can have several PWMs)
     -figsize "(10,20)"      Change matplotlib/seaborn figure size for the
heatmap, example width=10, height=20
     -heatmap title          Make heatmap (png+pdf) with title
     -output_folder folder   Store all results to the output folder (default:
current folder)
     -nonzero                All negative vector values are set to 0, not
enabled by default
```

Examples:

Unset

```
# random sequence: produce binding scores and a heatmap
# output: example1_PWM.tab
# (log-odds vectors for all proteins for the given command line sequence)
# output: example1.png/pdf
# (heatmap image with clustering of protein binding vectors)

./scanRBP
AAAGCGGCGACTTATTATATCCCCATATATTATATCTTCTTCTCTTATATATAAACCAGAGATAGATGTGTGTGGTGG
example1 -heatmap example1

# input can be a multi-sequence fasta file
./scanRBP data.fasta
```

Database

Currently, **scanRBP** is using the mCross PWM database of 112 RBPs from the paper:

Feng H, Bao S et al.
Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites
Molecular Cell, 2019

Unset

```
# to download PWMs
wget http://zhanglab.c2b2.columbia.edu/data/mCross/eCLIP_mCross_PWM.tgz
--no-check-certificate
tar xfz eCLIP_mCross_PWM.tg
```

Additional PWM datasets

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-023-02913-0
https://static-content.springer.com/esm/art%3A10.1186%2Fs13059-023-02913-0/MediaObjects/13059_2023_2913_MOESM6_ESM.txt

CLIP datasets

# bedGraph files
https://www.encodeproject.org/metadata/?status=released&internal_tags=ENCORE&assay_title=eCLIP&biosample_ontology.term_name=K562&biosample_ontology.term_name=HepG2&type=Experiment&files.analyses.status=released&files.preferred_default=true

Gene Annotation

Gene metadata (names, aliases) donwloaded from:
https://www.ncbi.nlm.nih.gov/gene/?term=human[organism]