Set Cover Tools

Release 1.0

Jul 29, 2019

Contents

CHAPTER 1

Relaxed Integer Linear Programming

1.1 Usage

Step 0. Download the GLPK library

https://www.gnu.org/software/glpk/

Step 1. Compile the C code

```
g++ -std=c++11 RIPL_simplex2_07_31.cc -o msdc -L/PATH/glpk/lib -I/PATH/glpk/include -
↪lglpk
```

Step 2. Run the RILP method

```
./msdc foreground.list background.list motif_mapping.fimo
```

You can also use the python wrapper:

```
python MSDC_wrapper.py [input_matrix] [output_file_name] [pos_k] [neg_j]
```

Below is an example:

```
python MSDC_wrapper.py AP1.Motif_Selection_Paper.csv AP1_2_2 2 2
```

1.1.1 Parameters

`pos_k`: the maximal uncovered percent of foreground set. Possible values are integers from 1 to 10, where 1 means 10%.

`neg_j`: the maximal covered percent of background set. Possible values are integers from 1 to 10, where 1 means 10%.

CHAPTER 2

Tabu search for motif selection

2.1 Usage

Step 0. Download the GLPK library

https://projects.coin-or.org/metslib

Step 1. Compile the C code

```
g++ -I/PATH/metslib-0.5 main.cc
```

Step 2. Run the RILP method

```
tabu_MS -in input_matrix -out output_file_name -t 0.2
```

2.2 Parameters

-in inputFile *-out* outputFile *-a* alpha (between [0, 1]) *-b* beta (default is the number of motifs) *-t* tenure (between (0, 1)) *-max* maxIteration (default 1000) *-d* delta (between [0,1], default 0.05) *-p* penalty (default 10000) *-iter* iterations (default 1, the times of running the whole program.)

CHAPTER 3

Motivation

De novo motif discovery algorithms find statistically over-represented sequence motifs that may function as transcription factor binding sites. Current methods often report large numbers of motifs, making it difficult to perform further analyses and validation. The motif selection problem seeks to identify a minimal set of regulatory motifs that characterize sequences of interest (e.g. ChIP-Seq binding regions).

CHAPTER 4

What is motif selection?

The motif selection problem seeks to identify a minimal set of regulatory motifs that characterize sequences of interest (e.g. ChIP-seq binding regions). The output motifs represent putative binding sites for primary transcription factors (ChIP-ed factors) and co-factors.

CHAPTER 5

Our solution to motif selection

5.1 The greedy set cover solution

This is our first generation solution. For more details, see: https://github.com/RamiOran/SeqCov

Contrast to the first generation solution, our current methods introduce a background set, which ideally, we wish to (1) cover as much of the foreground as possible, (2) cover as little of the background as possible, and (3) select the smallest set of motifs. See below:

5.2 The tabu search method

Tabu search is a metaheuristic local search method. It starts with a randomly generated initial solution then searches its neighborhood. Traditional local search methods, such as hill climbing, update current solution if they find a better solution in the neighborhood and thus result in local optima. In contrast, tabu search alleviates this issue by employing two strategies: (1) tabu search accepts non-improving moves when better moves are unavailable in the neighborhood of current solution and (2) tabu search uses a short-term memory structure, called tabu list, to store recently visited solutions and prevent selecting solutions that are visited previously.

5.3 The relaxed integer linear programming method

This algorithm contains two steps. The first step is to model the motif selection problem as a linear programming problem and obtain an optimal solution via GLPK. The second step is to solve the integer-version problem through a randomized algorithm.