CisModule (Linux or OS executable) user manual and examples

Put CisModuleU.tar/CisModuleOS.tar file in a Linux/OS directory and use "tar xvf CisModuleU.tar" or “tar xvf CisModuleOS.tar” to unpack the file. Then you will find an executable file (CisModuleU or CisModuleOS) in the same folder. In the following, we will use CisModuleOS as an illustration (it is the same for CisModuleU).

Command line

Type ./CisModuleOS, you will see a simple user manual. To run CisModule, an input sequence file must be specified. An example command is

```
./CisModuleOS -i cm_testdata.txt -o cm_testresult.txt -n 2000 -K 4 -L 200 -w 9 -W 14
```

Under this command, we are running CisModule on a data set cm_testdata.txt (which can be downloaded in the CisModule webpage). The output results will be written in the file cm_testresult.txt (-o). The program will run 2000 iterations (-n). We specify the number of motifs (TFs) to be K = 4 and the module length L = 200. The possible range for motif width is [9,14] (-w and -W). By this setting, CisModule will search both strands of sequences (default setting). If you specify “-r 0”, then only forward strand will be searched. The program will output all the sites with posterior probability > 0.5 (by default). If you want to change this threshold, you can add the option “-c x”, then only sites with posterior probability > x will be output. Under this setting, parameters (such as PWMs) are integrated out in the iterative sampling in CisModule, which usually improves the convergence and efficiency from our experiences and tests. Thus we have implemented this as the default setting. If you want to sample the parameters, i.e. under the data augmentation framework, you need to add option “-P” to your command line.

Output

1) Format for module information:
One module will be represented by one line. For example:
“>ZK721.2 [769,970] -2(775)+2(822)-1(946)”
This means that in the sequence ZK721.2, there exists one module located between positions 769 and 970. This module contains one backward site for motif 2 (-2) at position 775, one forward site for motif 2 (+2) at position 822, and one backward site for motif 1 at position 946.
2) Format for motif information:
Each motif is represented by a weight matrix followed by its sites in the sequences.
For example:

```
“>R10H      f     367       0.606
GTGCTCCCCAGAG”
```

This item says that in the forward strand of the sequence R10H at position 367 there is one motif site GTGCTCCCCAGAG whose posterior probability of being sampled as a motif site is 0.606.