User Manual for SuperRec

July 2017

Contents

1 Introduction

SuperRec reconstructs 3D genome structures from chromosomal contact maps generated with Hi-C technology. It accepts as input a set of contact frequencies between loci pairs, and returns a 3D structure that best explains the input.

2 Installation

SuperRec is provided as a statically-linked executable program. The standalone program was compiled using Ubuntu 15.04 libraries, and has been tested to run on many 64-bit Linux operation systems including CentOS. SuperRec can be downloaded from http://www.cs.cityu.edu.hk/~shuaicli/. The source codes is available upon request.

3 Input and Output

3.1 Input

An input file for SuperRec should contain a sparse contact matrix specified in plain text. There should be three values (first loci number, second loci number, and contact frequency) in each line, separated by spaces. The loci number should be an integer value. If you have a dataset with n loci in dense contact map format (i.e. $n \times n$ matrix), you can use the simple script `matrix2list.py` to convert it into a sparse matrix input file for SuperRec, like this:

```
> python matrix2list.py yourdata.txt > contact.txt
```

An example input file for SuperRec:

```
0 22 36.0
0 21 54.0
0 23 34.0
0 24 37.0
0 25 70.0
0 26 36.0
0 27 43.0
0 29 47.0
0 31 46.0
0 33 46.0
...
```

3.2 Output

SuperRec gives its output structure as a list of 3D coordinates in plain text.

The coordinates for each loci is given on a single line. Hence, a structure of n loci will result in an output of exactly 100 lines. The i-th line states the 3D coordinates of the i-th loci.

example:

```
1.178 1.796 0.474
0.903 1.428 0.973
1.432 1.557 1.191
1.078 1.864 0.881
0.554 1.608 1.221
1.053 1.971 1.347
0.932 2.314 1.671
1.322 2.542 1.249
...
```

You can use the provided Python script `marker.py` to plot the structure. The script requires the Python package matplotlib. For more advanced visualization, consider the use of PyMOL.

4 Options and examples

Suppose the input file name is `contact.txt` in this section.

4.1 Minimum command

The simplest way to run SuperRec is to just provide it with the file name of a contact map:

```
> SuperRec contact.txt
```

In which case, the output will be stored in a file called `contact.txt.SuperRec.txt`.

4.2 `-o` outputfilename

You can specify the name of the output file, by using the `-o` switch.

```
> SuperRec contact.txt -o result
```

In which case, the output will be stored in the file `result.SuperRec.txt`.

4.3 `-p` pivots

As described in the paper, SuperRec makes use of pivots to approximate the shortest-path distances. By default, SuperRec uses 100 pivots (i.e. 100 loci as pivots to approximate the shortest-path distances), a number which performed well in our test datasets.

Using the `-p` switch, you may change the number of pivots used by SuperRec to any positive integer value. For example, the following command makes SuperRec use 1000 loci as pivots.

```
> SuperRec contact.txt -o result -p 1000
```

4.4 `-c` cluster size

The two methods, iMDS and sMDS, implemented in SuperRec both work on random subsets or clusters of loci. By default, when the number of loci is equal or smaller than 1000, SuperRec performs iMDS and sMDS without subsetting. For larger datasets however, SuperRec will randomly group loci into small clusters that contain at most 1000 loci each. For example, when the input contains 1900 loci, the cluster sizes may be 1000 and 900. Suppose the number of common loci between overlapped clusters in iMDS is r (see SuperRec paper for definition of overlap; also see below for the default value of r), the cluster size should be at least $r+1$. In practice, 1000 is good enough for both accuracy and speed.

The following command would set the cluster size to 900.

```
> SuperRec contact.txt -o result -c 900
```

4.5 `-r` overlap

The `-r` switch can be used to change the number of overlapped loci in iMDS. The default value is 50. For a 3D structure, the value should be at least 4. As described in our paper, we recommend using 50 for this value.

The following command sets the number of overlapped loci to 100.

```
> SuperRec contact.txt -o result -r 100
```

4.6 `-it_WMDS` and `-cutoff_WMDS` (WMDS stop condition)

The sMDS algorithm implemented by SuperRec utilizes the iterative *weighted* MDS, or WMDS algorithm (SMACOF algorithm) to infer the 3D structure of a subset of loci. SuperRec terminates the WMDS computation on two different criteria: 1) when the number of iteration reaches a maximum value; and 2) when the difference between two consecutive steps is below a cutoff value.

The `-it_WMDS` switch can be used to change the maximum number of iterations (default: 50), while the `-cutoff_WMDS` can be used to change the step difference cutoff (default: 0.1).

In practice, it is hard to find a suitable step difference cutoff value that suits every dataset. For example, when the distances in the contact matrix are converted from raw counts, they may assume very small values, resulting in very small step differences even at the first few iterations. However, if the input contact matrix is rescaled to [0,1], the converted distances might become very large, and the step difference may also become large. Hence, it is advised that you discover a suitable cutoff value through trial and error. In our experiments, we set this value to 1 for binary contact maps, since the minimum distance is 1 for binary input.

The following command sets the maximum number of WMDS iterations to 20 and the step difference cutoff value to 1.

```
> SuperRec contact.txt -o result -it_WMDS 40 -cutoff_WMDS 1
```

(The `-log` option may also be helpful when setting stop conditions.)

4.7 `-s` (smooth option)

Setting the `-s` option in the argument list will cause SuperRec to smoothen the predicted structure using a linear filter. This option may be helpful when the input contact map is sparse.

The following command turns on the smooth feature.

```
> SuperRec contact.txt -o result -s
```

4.8 `-f` (fix connective option)

For very sparse matrix, the connection between neighboring loci may not be preserved. The `-f` option makes SuperRec fix the connective behavior.

The following command turns on the fix connective option:

```
> SuperRec contact.txt -o result -f
```

4.9 `-it_ref` refinement

The sMDS which SuperRec implements is an iterative algorithm. SuperRec stops the algorithm when the number of iterations exceeds 10, which works well on our datasets (100 $\sim$ 30,000 loci). You can use the `-it_ref` switch to change this default value for the number of sMDS iterations.

The following command changes the number of sMDS iterations to 15.

```
> SuperRec contact.txt -o result -it_ref 15
```

4.10 `-alpha` power-law coefficient

Recall that the distance between a loci pair i and j with a contact frequency $f_{i,j}$ is taken as $d_{i,j} = \frac{1}{f_{i,j}^{\alpha}}$.By default, SuperRec will estimated α from the input. However, you can provide α directly with the `-alpha` switch.

The following command sets α to 1.7:

```
> SuperRec contact.txt -o result -alpha 1.7
```

In practice, α values within the range [1.3, 1.75] perform well in most datasets.

4.11 `-segs` (whole genome reconstruction)

By default, SuperRec focuses on the reconstruction of a single chromosomal structure. The `-segs` switch makes SuperRec reconstruct a genome-wide contact map, by specifying the bound of each chromosome. If you do not need SuperRec to fix connective and perform smooth, you may skip **segs**.

To use this option, you will need to provide a fixed α through the `alpha` switch. A suitable α value could be the average of the α values of the individual chromosomes, or some value within the range [1.3, 1.75].

For a genome with three chromosomes, if the start positions in the contact map are respectively 1, 100, 200, you can use the following command:

```
> SuperRec contact.txt -o result -segs 1,100,200
```

4.12 `-log` log

SuperRec can produce more information if you use the `-log` option.

```
> SuperRec contact.txt -o result -log
```

This feature may be useful when deciding the stop condition of WMDS.

5 Example

We provide two datasets used in our paper: the real Hi-C contact maps, and *in silico* contact maps. You may download it from our website.

Suppose you use `contact.txt` as the input file name.

You can use the following minimal command to analyze the Hi-C contact maps.

```
> SuperRec contact.txt -o result
```

For the *in silico* maps, you will need to provide two additional arguments:

```
> SuperRec contact.txt -o result -alpha 1 -cutoff_WMDS 1
```