1. OVERVIEW

CISD(Chromatin Interaction Site Detector) and CISD_loop is developed for the prediction of chromain interaction sites and loops based on the characteristic nucleosome positioning pattern and Hi-C data. For CISD, it only use the MNase-seq data and could give the confident chromatin interaction sites. By calculating the FFT profiles of the MNase-seq signal as the features, the CISD predicts interaction sites in two steps: step1: find segments with periodical positioned nucleosomes , which we call high score peaks, with a logistic regression model(LRM); step2: find schromatin interaction sites from high score peaks with support vector machine(SVM). The predicted interction sites are called CISD sites. For CISD_loop, it takes CISD sites ans the Hi-C contact matrix as input, and then predict the loops with a SVM model. The predicted loops are called CISD loops.

2. ENVIRONMENTS

CISD and CISD loop request python 2.7.11 or later version, iNPS v1.1 or later version, perl v5.20.2 or later version, bedtools v2.24.0 or later version and R 3.2.2 or later version, be sure that the "e1071" package has been installed.

3. USAGE

3.1 CISD:

3.1.1 Command line:

```
$ bash CISD.sh Input1 Input2 Output
```

3.1.2 Parameters for command line:

Input1	The directory containing denosed and smoothed MNase-seq signal in .like_wig format. It is the iNPS output directory with prefix, which is the same as the -o parameter of iNPS. For more information, please refer to http://dx.doi.org/10.1038/ncomms5909.
Input2	The chromosomes that you want to choose. Different chromosomes should be seperated by comma and it is strongly recommended to do the prediction on all chromosomes. For example, if you want to do the prediction on chromosome 1 and chromosome 2, you may set this parameter as 1,2. If you want to do the prediction on all chromosomes of human, you may set this parameter as 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X.
Output	The output directory of CISD, the predicted CISD sites will be saved as CISD_site.txt

	in this directory.

3.1.3 Examples:

Suppose you have run iNPS with the command:

```
$ python3 iNPS_V1.1.2.py -i /PathA/InputFile.bed -o /PathB/Output -c chr1 -l 247249719
```

then you have the Output_chr1.like_wig file in the directory /PathB/ with the prefix 'Output'.

You may run CISD with the command:

```
$ bash CISD.sh /PathB/Output 1 /CISD_out/
```

If you have run iNPS on other chromosomes and want to run CISD on other chrosomes, you may use comma seperated chromosome number as the second parameter like:

```
$ bash CISD.sh /PathB/Output 1,2,3,4,X /CISD_out/
```

The output files of CISD are in the directory /CISD_out/, which are listed as following:

/CISD_out/CISD_site.txt	CSID sites, the predicted chromatin interaction sites given by the SVM in the step 2 of CISD algrithm.
/CISD_out/high_score_peaks/hspeaks0.50chrAll.bed	High score peaks(HSPeaks), which are segments with periodical positioned nucleosomes given by the LRM in the step 1 of CISD algrithm.
/CISD_out/wig/chr*_4.normalized.fft	Genomewide FFT profiles for MNase-seq signal. Each line represent the 0th to 49th amplitude in the frequency spectrum of the MNase-seq signal in 1kb sliding window. Note that the smooth and denoised MNase-seq signal givin in the iNPS is conbined into bins for each 10bp, thus, the 1kb sliding window is actually a 100-dimension vector rather than 1000-dimension. For more information, please refer to the manual of iNPS. The interval of the sliding window is 100bp.
/CISD_out/wig/chr*_p	Genomewide LRM score in wig format. Each line represent the LRM score in 1kb sliding window, which is the indicator of the periodical positioning pattern of nucleosomes in this window.

3.2 CISD_loop:

3.2.1 Command line:

```
bash CISD_loop.sh Input1 Input2 Input3 Output
```

3.2.2 Parameters for command line:

Input1	The output directory of CISD, which is the third parameter of CISD . Make sure you have not deleted or renamed any file in this directory.

Input2	The directory containing Hi-C contact matrix and expected reads, which are in 5kb resolution. The Hi-C contact matrix must be the same format as the format given in http://dx.doi.org/10.1016/j.cell.2014.11.021.The filename of must be the same as the name given in http://dx.doi.org/10.1016/j.cell.2014.11.021, like chr1_5kb.RAWobserved and chr1_5kb.RAWexpected. For more information, please refer to http://dx.doi.org/10.1016/j.cell.2014.11.021.
Input3	The chromosomes that you want to choose. Different chromosomes should be seperated by comma and it is strongly recommended to do the prediction on all chromosomes. For example, if you want to do the prediction on chromosome 1 and chromosome 2, you may set this parameter as 1,2. If you want to do the prediction on all chromosomes of human, you may set this parameter as 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X.
Output	The output directory of CISD_loop, the predicted loops will be saved as CISD_loop.txt in this directory.

3.2.3 Examples:

Suppose you have run CISD with the command:

```
$ bash CISD.sh    /PathB/Output    1,2,3,4,X    /CISD_out/
```

then you have the CISD_site.txt file in the directory /CISD_out/.

You may run CISD_loop with the command:

```
$ bash CISD.sh /CISD_out/    /Hi-C_dir/    1,2,3,4,X    /CISD_loop_out/
```

The output files of CISD_loop are in the directory /CISD_loop_out/, which are listed as following:

/CISD_loop_out/CISD_loop.txt	Predicted loops by CISD_loop, which are called CISD loops.

4. PREDICTION WITH YOUR OWN MODEL

4.1 Introduction to the model files:

CISD and CISD loop allow users to do the prediction with your own model. However, it is strongly recommeded to used default model in the ./data/ directory unless you have very confident data to generat the trainingset. There are three model files in the ./data/ directory which are named model1_LRM, model1_SVM and model2_SVM. The model1_LRM is the logistic regression model used in the step 1 of CISD algrithm; the model1_SVM is the support vector machine model used in the step 2 of CISD algrithm; the model2_SVM is the support vector machine model used in the CISD_loop algrithm.

If you have your own data to generate the trainingset to train the model which you think more confident and efficient than the default model, CISD and CISD_loop would allow you to use your own model.

4.2 How to generate your own model:

4.2.1 model1_LRM:

This model is a logistic regression model in the first step of CISD algrithm. The goal of this model is to find the segments with periodical positioned nucleosomes. Thus, the positive set should be sites flanked by strong nucleosome phasing. CTCF ChIP-seq binding sites is a good choice, but not the best choice, because we have demonstrated in our paper that part of CTCF binding sites are not engaged in chromatin interaction and have very weak nucleosome phasing flanking these bing sites. We recommand the overlapping of CTCF and Rad21 ChIA-PET anchors as the positive set, which have very strong nucleosome phasing. The negative set shold be depleted of nucleosome phasing, we recommand random sites far from(5kb away) CTCT, cohesin, ZNF143 binding sites and TSS. We recommend at least 10000 positive set and 10000 negative set. For each site in the positive or negative set, extract the 0th, 5th, 7th amplitude in the FFT profiles in the 1kb area, for more details, please see the method in the paper. The final trainingset must be in the following format:

0th amplitude	5th amplitude	6th amplitude	Flag
8.45	37.55	48.1	1
9	38.5	37.85	1
12.05	94.65	74.15	1
8.8	28.4	26.8	0
10.85	22.2	24.15	0
4.15	18.65	14	0

The first three columns in the trainingset are from the first, 6th, 7th column of the FFT profilefile; the last colume in the traingset is flag, 1 means true and 0 means false.

4.2.2 model1_SVM:

This model is a support vector machine model in the second step of CISD algrithm. The goal of this model is to find the interaction sites from the high score peaks, which are setments with periodical positioned nucleosomes reported in the step1 of CISD. Thus, the positive set should be sites that are engaged in chromatin interactions. We recommand to find ChIA-PET anchors on the high score peaks as positive set. The negative set shold be other sites not engaged in chromatin interaction, we recommand random sites that not overlapping with the ChIA-PET anchors on the high score peaks. We recommend at least 10000 positive set and 10000 negative set. For each site in the positive or negative set, extract all the amplitude in the FFT profiles in the 1kb area, for more details, please see the method in the paper. The final trainingset must be in the following format:

0th amplitude	……	49th amplitude	Flag
14.5	……	1.5	“i”
13.9	……	1.15	“i”
11.85	……	0.8	“i”

12.3	28.4	0.45	“ni”
12.05	22.2	1	“ni”
9.3	18.65	1.25	“ni”

There are 51 columns in the training set. The first 50 columns in the trainingset are from the first to the 50th column of the FFT profilefile; the last colume in the traingset is flag, “I” means interaction site and “ni” means none interaction site.

4.2.3 Model2_SVM:

This model is a support vector machine model in the CISD_loop algrithm. The goal of this model is to find the loops from random CISD site pairs, which we call random loops. Thus, the positive set should be random loops that are supported by ChIA-PET and the negative set should be random loops that are not supported by ChIA-PET. We recommend at least 5000 positive set and 5000 negative set. For each loop in the positive or negative set, extract normalized Hi-C reads in the 5kb bin and flanking bin as the feature, for more details, please see the method in the paper. Another feature is the distance between the two CISD sites in the loop. The final trainingset must be in the following format:

Hi-C reads	Distance	Flag
6.72630523097	278150	“i”
2.27205510535	62350	“i”
4.60570710425	195450	“i”
0.247308297722	555800	“ni”
2.2355409587	180550	“ni”
0.969890270541	845900	“ni”

The first colume in the trainingset is normalized Hi-C reads; the second column in the traingset is distance between two CISD sites in the loop; the third column is flag, “I” means supported by ChIA-PET and “ni” means not supported by ChIA-PET.

4.3 How to use your own model:

The way of using your model is very simple, just put your model file in the ./data/ directory to replace the original model file, then you can do the prediction with your own model. Make sure that the file name and format of your own model be consistent with the original model file.