Domain calling steps

Making a DI files:

Input: Matrix file.

Bin the genome in to given bin sizes and make a matrix file for a give chromosome. The format is

```
Chrname      StartBin      EndBin      Values…..
```

Bin size – bin size used to create matrix. We use 40000.

Window size – How far we need to look at the interaction patterns of a given bin. We use 2000000.

Genome size file – its .fai file generated by samtools. It has the size of each chromosome.

Usage:

```
./DI_from_matrix.pl <matrix> <bin size> <window size> <genome size file>
```

Output: DI file for a given chromosome.

Performing HMM:

Input: One huge DI file.

Concatenate DI's for each chromosome to make "whole genome DI". Make sure column 1 is an integer. ChrX can be called 23 and henceforth.

i.e.

```
1       20000 40000 456.32
1       ..     ..     …
2       20000 40000 ….
…       ..     …      ….
…
…
…
```

Usage:

```
nice matlab < HMM_calls.m > dumpfile
```

Note:

[1] In line 9 of the script HMM_calls.m, please hardcode your input DI filename.

[2] In line 77 of the script HMM_calls.m, please hardcode your output filename.

Post-processing:

[1]
```
perl file_ends_cleaner.pl hmm_outputfile hmm_inputfile | perl converter_7col.pl > hmm_7colfile
```

[2] Split the hmm_7col based on chromosome.

[3] for each chromosome 7col file:

```
perl hmm_probablity_correcter.pl 7colfile min prob binsize | perl hmm-state_caller.pl faifile chr | perl hmm-state_domains.pl > finaldomaincalls
```

min – corrects probability if size of a cluster is <= min. we use 2

prob – checks probability in a cluster. We use 0.99

binsize – we use 40000

chr – chromosome number, say chr1