BCP Manual

Yifan Mo,Willey Liao

September 14, 2011

Bayesian Change-point Model (BCP) is a method for analysis different type of ChIP-seq data.It is greatly encouraged to use BCP when studying diffuse data like many Histone modification (HM) data sets like H3K36me3, H3K27me3,H3k9me3and so on. In the same time, it also has good performance on studying "punctuate" peaks like transcription factor binding sites (TFBS) and relative small segments of Histone modification mark like H3K4me3. Owe to the Bounded complexity mixture approximation (BCMIX) in the model, especially in HM study, BCP could largely decrease the running time and have better results.

1 Installation

1.1 Requirement of the system

BCP runs under Linux System with GNU compilation,(is preferred.)Generally speaking 1GB memory is enough. While in some case,like large coverage data,it might need 2GB. The results in our paper is conducted under BlueHelix system (HPCC) in CSHL (dual core 64-bit processors running at 2.0GHz and 2GB of memory).

1.2 Download and Install

In the download link,please download the source code package.Then you could decompress it:
Decompress:

```
$ tar -zxf BCP_v1.1.tar.gz
```

Then you will get a folder called BCP_v1.1, enter this folder and compile the source code like below:
Compile for HM:

```
$ g++ BCP_HM.cpp -o HM.out
```

Compile for TFBS:

```
$ g++ BCP_TF.cpp -o TF.out
```

After successfully compiling,you could find these files in folder BCP_v1.1:

1. Three header files: MyFun,TNT and JAMA_C.
2. Three *.cpp* files:cppoisson.cpp,BCP_HM.cpp and BCP_TF.cpp.
3. Two executable files:TF.out and HM.out.

2 Using BCP

Once finish installing BCP it is very easy to use it. Here user first need to clarify the data type: whether it is searching transcription factor binding sites or studying Histone Modification. As these two kinds of data appear very different so we use different pre-processing procedure.

2.1 Studying Histone Modification case

Here we have 6 options for running HM.out executable file. Three of them are indicating the data need to input or output:ChIP-seq data, control data and output data. Respectively, we set the number $1, 2, 3$ in the command line.Another three options are fragment size, window size and p-value which we set command line as f, w, p respectively. Here fragment size is decided by the ChIP experiment such as sonication size (default value is 200bp). We first extend the small reads to the length of fragment size based on the plus and minus strand. Window size is a resolution parameter thus is decided by user. We recommend 200bp as the default value as it is approximate size of a single nucleosome. p-value is used when called significant segments compare to the control data(default is $1e-3$). Here please input integer for fragment size and window size while p-value needs real number.Please notice the data option cannot be omitted.
Here is an example to run:

```
./HM.out -1 data.bed -2 control.bed -f 200 -w 200 -p 0.001 -3 results_HM.bed
```

2.2 Searching for Transcription Factor Binding Sites (TFBS)

Similar to HM.out, TF.out has 6 options too but with little different. Three options of input and output data are the same: ChIP-seq data, control data and output data and indicated by the number $1, 2, 3$ in the command line respectively. Another three options are read size, fold enrichment and p-value indicated by command line as r, e, p respectively.Here read size is the length of the small read in the data such as 36bp(default value),72bp and so on. Most of the data the read length is fix but in some case it might be not so we still list it here as one adjustable parameter.(If the read length in your data are constant please use that

length as read size.) Fold enrichment is used when estimating shift size. The default value is 10 and we recommend user adjust it in the range of $5-15$. When estimating shift size we first choose some "candidate" areas which have very strong enrichment and fold enrichment decide how enrich we want to choose. And p-value is also used to choose really significant peaks compared to the control data. The default p-value in TFBS case is $1e-8$. If you want to have more peaks just increase the p-value while you want really "high" peaks you could decrease it. The recommend range is $1e-6$ to $1e-12$. Again the same with HM case please input integer for read size and fold enrichment while p-value needs real number and data option are required.
Here is the example to run:

```
./TF.out -1 data.bed -2 control.bed -r 36 -e 10 -p 0.00000001 -3 results_TF.bed
```

Compared to other methods, BCP has less parameters to choose. In many cases we find it is not easy for user to find a optimal combinations among many parameter setting. So BCP limits the number of options but also can provide good performance.

3 Data Format

Here we introduce the input and output data format which are important for user to run BCP successfully and understand the results.

3.1 Input Data

Here we have two input data: ChIP-seq data and control data. BCP requires the "BED" format (`http://genome.ucsc.edu/FAQ/FAQformat.html`) for both of them. If your data sets are in other format like ELAND we are sorry that you need transfer it first into BED format. (If you don't know how to transfer we could provide some help.) Further, with "BED" format BCP just need first 6 columns: "chrom,chromStart,chromEnd,name,score and strand".Please notice if your input is not 6 columns or the information is not in a right order like above BCP might not run or run in some improper way.And one more notice is please **not** include the "track line" (`http://genome.ucsc.edu/goldenPath/help/customTrack.html#TRACK`) in your data.Here is an example of the H3K36me3 input data BCP requires(more example you could see the "Sample Data and Results" in our link):

```
chr1    9796    9995    SOLEXA1_1:3:84:32:2029  1       -
chr1    9797    9996    SOLEXA1_1:3:27:2:1037   1       -
chr1    9798    9997    SOLEXA1_1:3:41:0:975    1       -
chr1    9799    9998    SOLEXA1_1:3:100:1417:2031       1       -
chr1    9800    9999    SOLEXA2_7:4:47:21:1111  1       -
```

3.2 Output Data

Here is introduction of the meaning of each column in results data. Again we introduce it in TFBS case and in HM case separately.

3.2.1 HM output data

When you open the results data of HM case (The example in section ”Using BCP” is $results_HM.bed$),you will find it has 5 columns.Here is an example:

```
chr1 14800 17799 3000 6.445783
chr1 754000 757599 3600 7.927614
chr1 768600 769599 1000 4.819276
chr1 770600 792199 21600 3.185302
chr1 856800 858799 2000 7.102175
chr1 861400 893199 31800 9.983955
```

This is part of the results from H3K36me3.Below is the explanation of these 5 columns:

- $Col1$: The name of the chromosome.
- $Col2$: The starting position of the segment in the chromosome.
- $Col3$: The ending position of the segment in the chromosome.
- $Col4$: The length of the segment.
- $Col5$: The average posterior mean of the segment.User could use it as the enrichment of the segment.

If user want to profile it in the UCSC genome browser,we recommend using ”BedGraph” format(`ftp://hgdownload.cse.ucsc.edu/apache/htdocs-rr/goldenPath/help/bedgraph.html`). Please just choose the $Col1, Col2, Col3, Col5$ and add the ”track line” to make up the new data sets.Please notice when you add the ”track line” set $visibility = 1/dense$.

3.2.2 TFBS output data

The results data of TFBS case (The example in section ”Using BCP” is $results_TF.bed$) is a little more complicated than HM case , you will find it has 8 columns. Let’s see an example:

```
chr1 1089 1125 37 47.415240 88 0.925816 1118
chr1 15889 16008 120 14.721743 55 7.097924 15917
chr1 81221 81255 35 39.004106 73 1.851632 81251
chr1 94786 94950 165 10.311333 52 7.406530 94838
chr1 130193 130247 55 7.056756 12 0.308605 130222
```

This is part of the results from CTCF. Below is the explanation of these 8 columns:

- $Col1$: The name of the chromosome.
- $Col2$: The starting position of the peak in the chromosome.
- $Col3$: The ending position of the peak in the chromosome.
- $Col4$: The length of the peak.
- $Col5$: The average posterior mean of the peak.User could use it as the enrichment of the peak.
- $Col6$: The number of reads in the peak from ChIP-seq data
- $Col7$: The estimation of number of reads in the peak from control data. Notice here this estimation is real number which represents the mean of the Poisson distribution.
- $Col8$: The summit location of the peak.

User could gain many information from this results.Below is some case:

1. When profile in UCSC genome browser please choose $Col1, Col2, Col3, Col5$ and add the "track line" to make up the new data. This is very similar with the HM case.
2. You could rank the peak by the p-value.There is a function in R called $ppois$, you could calculate the p-value using $1 - ppois(Col6, Col7)$.
3. If you want to check about the motif occurrence rate or spatial resolution you could using the summit information in $Col8$.