Bayesian Change-point (BCP) Manual

Yifan Mo, Willey Liao, Haipeng Xing

October 30, 2011

Bayesian Change-point Model (BCP) is a method for analyzing different types of ChIP-seq data. BCP is particularly useful in finding enriched segments in diffuse Histone modification (HM) data such as H3K36me3, H3K27me3, H3k9me3, etc. At the same time, it also has good performance as a peak calling method for “punctate” peaks common in ChIP-seq of transcription factor binding sites (TFBS). Due to the Bounded complexity mixture approximation (BCMIX) used in the model, especially for HM data types, BCP has a substantially reduced run time while still achieving improved performance.

1 Installation

1.1 Requirement of the system

BCP runs under Linux system and GNU Compilation Collection. Generally speaking, 1GB memory is sufficient to run BCP, but in some cases, for example high coverage data, it may need 2GB. The results in our manuscript were conducted under a high performance compute cluster (HPCC) at CSHL (dual core 64-bit processors) with different memory size (See “FAQ”).

1.2 Download and Install

Use the download link to obtain the source code package. Then you can decompress it:

Decompress:

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

You will get a folder called BCP_v1.1, enter this folder and compile the source code as follows:

```bash
$ ./make
```
After successfully compiling, you will find these files in folder BCP_v1.1:

1. Three header files: MyFun, TNT and JAMA_C.
2. Four .cpp files: cppoisson_HM.cpp, cppoisson_TF.cpp, BCP_HM.cpp and BCP_TF.cpp.
3. Three executable files: BCP_TF, BCP_HM and make.

2 Using BCP

Once finished installing, BCP it is very easy to use it. Firstly, please place your data sets (ChIP-seq data and control data) in the folder BCP_v1.1. Users then need to clarify the data type: whether it resembles punctate, transcription-factor-like binding sites or broad, diffuse HM-like enrichment. These two kinds of data are very different so we use different pre-processing procedures.

2.1 Studying Histone Modification case

Here we have 6 options for running the BCP_HM executable file. Three of them identify the basic input and output: ChIP data, control data and output results. We set them using flags 1, 2, 3, respectively, on the command line. Another three options are fragment size, window size and p-value which we set using f, w, p, respectively. Here fragment size is decided by the ChIP experiment 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 that is decided by user. We recommend 200bp as the default value because it is approximately the size of a single nucleosome. We suggest the user adjust it within the range of 100bp – 500bp. Smaller values will increase running time without substantially improving resolution in the HM case while larger values may overlook some important detail. P-values are used when calling significant segments compared to control data (default is 1e-3). Fragment size and window size must be integer values while p-values must be real. The data options (1, 2, 3) here cannot be omitted and other three options could be omitted if you just want to use their default values.

Here is an example:

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

Except the above 6 running options, we also set a option for displaying the help manual. If you want to display the help manual just use the h options without using other 6 options. There are two value for h: 0 and 1. 1 means displaying while default value 0 means NOT displaying. Here is the example:

```
$./BCP_HM -h 1
```

And the list is as below:
-1 The ChIP-seq data set you want to input.
-2 The control data set you want to input.
-f The fragment size to which we extend the reads in pre-processing data step. Default: 200bp.
-w The window size we apply the adjacent window in pre-processing data step. Default: 200bp.
-p The p-value you want to use for remove false positive based on control data. Range: 1e-2—1e-6, default is 1e-3.
-3 The results data set with 5 columns you want to output.
-h The flag indicates whether you want to output the help manual. 1 means displaying the help instructions while 0 means NOT. Default: 0. If you set the flag with 1, the program would not works except displaying the manual.

2.2 Searching for Transcription Factor Binding Sites (TFBS)

BCP_TF has 6 running options with some differences. Three options of input and output data are the same: ChIP data, control data and output data, and are again indicated by the flags 1, 2, 3, respectively. The two other options are fold enrichment and p-value indicated by, e, p, respectively. Fold enrichment is used when estimating shift size. The default value is 10 and we recommend users restrict within the range of 5 – 15. When estimating shift size we first choose some “candidate” areas which have very strong enrichment and fold enrichment decides how enrich we want to choose. p-value is also used to choose really significant peaks compared to the control data. The default p-value in the TFBS case is 1e-8. If you want adjust the number of peaks, just increase or decrease the p-value accordingly. The recommended range is between 1e-5 to 1e-15. Here we also sets a flag called m which indicates whether user wants to output multiple results corresponding to different p-values. The default is 0 means NOT outputting while 1 means outputting. More about m could be found in the end of section 3.2.2.

As in the HM case, please input integers for fold enrichment and real numbers for p-value. The three data options CANNOT omitted while other options could be omitted if their default values are preferred. Here is the example:

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

Similar to BCP_HM, we also set a option for displaying the help manual. Again using the h options without using other 6 options will give you the help manual. Here is the example:

$ ./BCP_HM -h 1

And the list is as below:
-1 The ChIP-seq data set you want to input.
-2 The control/input data set you want to input.
-e The fold enrichment parameter helps to estimate the shift size. Range: 5—12, default is 8.
-p The p-value you want to use for remove false positive based on control data. Range: 1e-5—1e-15, default is 1e-8.
-3 The results data set with 7 columns you want to output.
-m The flag indicates whether you want to output multiple results corresponding to different p-value. 1 means output the multiple results while 0 means don’t output. Default: 0.
-h The flag indicates whether you want to output the help manual. 1 means outputting the help instructions while 0 means NOT outputting. Default: 0. If you set as 1, the program would not works except displaying the manual.

Compared to other methods, BCP has fewer parameters to choose. In many other methods, it can be confusing for users to find an optimal combination of parameter settings. So BCP has limited the number of options yet still provides good performance.

3 Data Format

Here we introduce the input and output data formats.

3.1 Input Data

We have two input data: ChIP data and control data. BCP requires the six-column “BED” format (http://genome.ucsc.edu/FAQ/FAQformat.html). If your data sets are in other format like “ELAND” or “BOWTIE”, they should be converted first. The 6 required columns are “chrom, chromStart, chromEnd, name, score and strand”. Please notice if your input is not 6 columns or the information is not in the right order, BCP may not run or may run in some improper way. Further, do not include “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):

<table>
<thead>
<tr>
<th>chromosome</th>
<th>start</th>
<th>end</th>
<th>trackName</th>
<th>p-value</th>
<th>shiftSize</th>
<th>strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>9796</td>
<td>9995</td>
<td>SOLEXA1_1:3:84:32:2029</td>
<td>1</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>chr1</td>
<td>9797</td>
<td>9996</td>
<td>SOLEXA1_1:3:27:2:1037</td>
<td>1</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>chr1</td>
<td>9798</td>
<td>9997</td>
<td>SOLEXA1_1:3:41:0:975</td>
<td>1</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>chr1</td>
<td>9799</td>
<td>9998</td>
<td>SOLEXA1_1:3:100:1417:2031</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>9800</td>
<td>9999</td>
<td>SOLEXA2_7:4:47:21:1111</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
3.2 Output Data

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

3.2.1 HM output data

The results of HM analysis has 5 columns. Following the example form the section “Using BCP”, we have the output file results.HM.bed,

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>len</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>14800</td>
<td>17800</td>
<td>3000</td>
<td>6.445783</td>
</tr>
<tr>
<td>chr1</td>
<td>754000</td>
<td>757600</td>
<td>3600</td>
<td>7.927614</td>
</tr>
<tr>
<td>chr1</td>
<td>768600</td>
<td>769600</td>
<td>1000</td>
<td>4.819276</td>
</tr>
<tr>
<td>chr1</td>
<td>770600</td>
<td>792200</td>
<td>21600</td>
<td>3.185302</td>
</tr>
<tr>
<td>chr1</td>
<td>856800</td>
<td>858800</td>
<td>2000</td>
<td>7.102175</td>
</tr>
<tr>
<td>chr1</td>
<td>861400</td>
<td>893200</td>
<td>31800</td>
<td>9.983955</td>
</tr>
</tbody>
</table>

These columns take on values describe below:

- **Col1**: The name of the chromosome.
- **Col2**: The starting position of the segment in the chromosome (0-based).
- **Col3**: The ending position of the segment in the chromosome (half-open).
- **Col4**: The length of the segment.
- **Col5**: The average posterior mean of the segment. Users could interpret this as the enrichment of the segment.

If users 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 use Col1, Col2, Col3, Col5 and add a “track line”.

3.2.2 TFBS output data

The results data of the TFBS case following the example in section “Using BCP”, results.TF.bed are a little more complicated than the HM case, it contains 7 columns:

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>len</th>
<th>mean</th>
<th>pval</th>
<th>scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>12805147</td>
<td>12805184</td>
<td>37</td>
<td>9.285350</td>
<td>1.110223e-16</td>
<td>12805176</td>
</tr>
<tr>
<td>chr1</td>
<td>12859715</td>
<td>12859752</td>
<td>37</td>
<td>9.251830</td>
<td>3.330669e-16</td>
<td>12859744</td>
</tr>
<tr>
<td>chr1</td>
<td>12906920</td>
<td>12907044</td>
<td>124</td>
<td>3.860053</td>
<td>9.092727e-14</td>
<td>12907010</td>
</tr>
<tr>
<td>chr1</td>
<td>12918840</td>
<td>12918939</td>
<td>99</td>
<td>18.568209</td>
<td>0.000000e+00</td>
<td>12918863</td>
</tr>
<tr>
<td>chr1</td>
<td>12973076</td>
<td>12973113</td>
<td>37</td>
<td>9.285350</td>
<td>3.330669e-16</td>
<td>12973105</td>
</tr>
</tbody>
</table>

Below is a description of these 7 columns:

- **Col1**: The name of the chromosome.
• *Col2*: The starting position of the peak in the chromosome (0-based).

• *Col3*: The ending position of the peak in the chromosome (half-open).

• *Col4*: The length of the peak.

• *Col5*: The average posterior mean of the peak. Users could interpret this as the enrichment of the peak.

• *Col6*: The p-value of the peak.

• *Col7*: The summit location of the peak.

This information can be further processed as follows:

1. When profiling in UCSC genome browser please choose *Col1, Col2, Col3, Col5* and add a “track line”.

2. The peaks can be ranked by p-values.

3. To check the motif occurrence rate or spatial resolution, the summit information in *Col8* would be most useful.

If you set the option $m$ equals to 1. You should also find many result files called *results_* such as *results_1e−9*. Because these flag tells the program outputting the multiple results corresponding to different p-values from $1e − 5$ to $1e − 10$. This will help you to determine the optimal p-value you might use based on the number of peaks in each results.