ChromHMM tutorial

This tutorial explain step by step how to setup and run ChromHMM on the provided example, or with pre-aligned ChIP-seq datasets in .bam format.

Prerequisites for this tutorial

- **Required:**

    1. Java virtual machine (http://java.com/)
    2. ChromHMM software (http://compbio.mit.edu/ChromHMM/ChromHMM.zip)

- **Optionally, if you want to use it on your data:**

    1. Raw or aligned reads for different histone modifications for example from the ENCODE portal (https://www.encodeproject.org/)
    2. Bedtools (https://github.com/arq5x/bedtools2)

Installation

1) Open a terminal and check if you have java installed:

```
myhost:~ luca$ java -showversion
java version "1.6.0_65"
Java(TM) SE Runtime Environment (build 1.6.0_65-
b14-466.1-11M4716)
Java HotSpot(TM) 64-Bit Server VM (build 20.65-
b04-466.1, mixed mode)
.
.
.
```

2) Download the chromHMM .zip archive from here: http://compbio.mit.edu/ChromHMM/ChromHMM.zip and decompress it.

Testing ChromHMM on the provided example

1) Open a terminal and go the decompressed folder, for example:

```
cd ~/Downloads/ChromHMM
```

2) Run the provided example on Human data:

```
java -mx4000M -jar ChromHMM.jar LearnModel
SAMPLEDATA_HG18 OUTPUTSAMPLE 10 hg18
```

3) Open the output html page `webpage_10.html`, inside the newly created folder `OUTPUTSAMPLE`, with a modern browser (for example safari, chrome, firefox or internet explorer) and explore the output.

Visualize the segmentation

1) Launch IGV from the Broad website clicking the Launch button from this webpage:

   https://www.broadinstitute.org/software/igv/download

2) Select the Human Genome hg18 form the top left corner

3) Drag and drop the following .bed file created by the ChromHMM in the IGV main window:

   a. K562_10_dense.bed

   b. GM12878_10_dense.bed

Run ChromHMM starting from aligned files (.bam)

To run ChromHMM starting from aligned .bam files, for different histone modifications, (for example downloaded from the ENCODE portal) some additional steps are required before running the LearnModel command showed in the previous example.

1) Let's assume you have already downloaded/created some .bam files in the folder ~/data for different histone marks/cell types.

2) Using the bedtools for each .bam file you need to convert it to a .bed file with the following command:

```
bedtools bamtobed -i cell1_mark1.bam >
~/data/cell1_mark1.bed
```

3) Create an empty file called `cellmarkfiletable.txt` in `~/data`. Each row of this file should contain the following information: cell type (for example K562), the name of the mark (for example H3k4me1), the filename of the .bed file for that mark and the optional input or control .bed.

   It should look similar to this:

```
cell1 mark1 cell1_mark1.bed cell1_control.bed
cell1 mark2 cell1_mark2.bed cell1_control.bed
cell2 mark1 cell2_mark1.bed cell2_control.bed
cell2 mark2 cell2_mark2.bed cell2_control.bed
```

   Note: you can put marks for different cell types in this file. In this way you will learn a model for all the cell types simultaneously.

4) Now we can binarize the tracks converted to .bed with this command:

```
java –mx4000M –jar ChromHMM.jar BinarizeBed –b
200  CHROMSIZES/hg18 ~/data/
~/data/cellmarkfiletable.txt ~/binarizedData
```

5) Finally we can learn the model and generate the output, as done before with the provided example, with this command:

```
java –mx4000M -jar ChromHMM.jar LearnModel
~/binarizedData MYOUTPUT 10 hg18
```

Here we are assuming that the genome of reference is hg18, but other genomes are available. See the documentation for more details:

http://compbio.mit.edu/ChromHMM/ChromHMM_manual.pdf