HiC-3DViewer: a novel tool to visualize Hi-C data in 3D space (User Manual)

Djekidel Mohamed Nadhir, Mengjie Wang, Juntao Gao, Michael Q. Zhang

September 11, 2015

Contents

1 Introduction

The expending arsenal of chromatin conformation study techniques developed in the last decade helped shed some light on the different aspects in which chromatin cross-talk can influence gene regulation. Hi-

C is one of the widely adopted chromatin conformation study techniques as it enables the quantitative measurement of the genome-wide chromatin interactions.

Many efforts have been made to infer the chromatin 3D structure from the Hi-C contact map but not many tools have been developed to visualize and analyze these data interactively. Here we present *HiC-3DViewer*, a simple user friendly tool for Hi-C data display and analysis. *HiC-3DViewer* was built on the top of the `Flask` micro-framework which is a very light-weight framework that enables the establishment of a small footprint web-server without any additional configuration.

2 Architecture

HiC-3DViewer is based on a client-server architecture with different roles as described below:

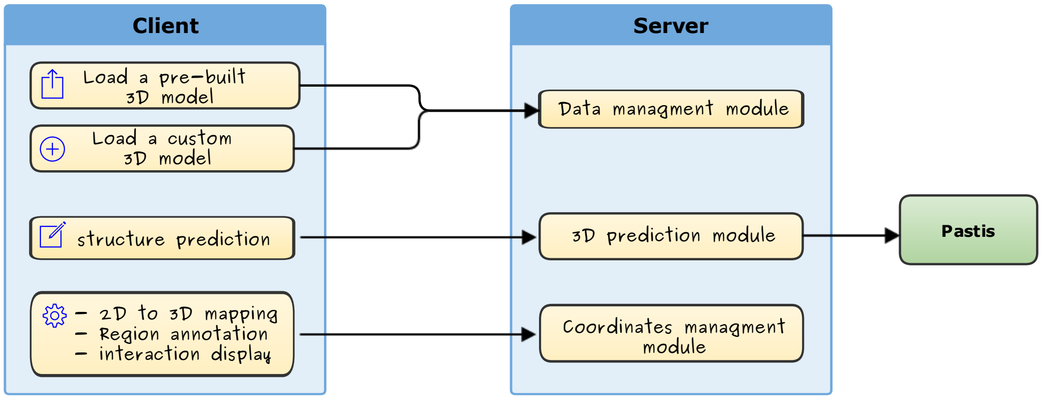


Figure 1: General architecture of the HiC-3DViewer. It is based on a client-server architecture, where the client is responsible for the rendering, while the server handles data storage and processing

*) **Server-side** : is written in `Python` using the *Flask* micro-framework. The server-side handles different tasks and is responsible for fetching pre-existing and user loaded data, 3D model prediction and the conversion between genomic and bin coordinates. Two types of data can be found on the server side.

- **pre-built data** that comes with the tool and contains the Hi-C frequency matrix and the predicted 3D models for different species (such as yeast, human and Drosophila) in addition to the cytoband of some species (such as the cow, dog, ...etc). Pre-built data are stored in the `Data` folder , one for each species.
- **user data** : uploaded into the `tmp` folder and are named after the user's session ID.

```
/home/nadhir/Documents/HiC Viewer/Hic3DViewer/hicViewer/data
 Cow
 Dog
 Drosophila
 Horse
 Human
 Mouse
 Pig
 Rat
 Yeast
/home/nadhir/Documents/HiC Viewer/Hic3DViewer/hicViewer/tmp
 2e331eed-8668-4e67-b61b-76ccab56b021
  model
  yeast_hic_matrix
 3c0a4892-9a4b-46e7-9dd8-694f986690d7
     model
```

*) **Client-side**: is responsible for the 3D model and the 2D Hi-C map visualization. It is based on the *Threejs* and *D3js* javascript libraries and runs on all major browsers supporting HTML5 and WebGL (FireFox is recommended).

3 Installation & Requirement

3.1 Required modules

On the server side, the following python modules are required:

- python (version ≥ 2.7)
- flask (version $\geq 0.10.1$)
- bxpython (version $\geq 0.5.0$)
- numpy (version ≥ 1.3)
- matplotlib (version ≥ 1.4)
- scipy (version ≥ 0.7)
- scikit-learn (version ≥ 0.13)
- pastis (custom version downloaded from our repository)

All packages are available on the *Anaconda* scientific python distribution. The *bxpython* package can be downloaded from Binstart website.

3.2 Some notes on Pastis installation

Pastis is based on the interior point optimization C++ library *IPOPT*. This library can be installed if users are interested in predicting 3D models using *Pastis* implemented algorithms. The python package *pyipopt* interfacing with *IPOPT* should also be installed in this case. However, if the user is just interested in the 3D visualization of a model he already predicted, it can directly upload to *HiC-3DViewer*in one of the specified formats in section 5.4.1.

If *pastis* was installed without installing *IPOPT*, the tool can be used without the 3D prediction capability. Otherwise, please go ahead to compile the `src/MDS` and `src/PM` in the *pastis* package source code after *IPOPT* is installed.

4 Launching Hi-C3DViewer

After downloading *HiC-3DViewer* from the bitbucket repository, you just need to extract it. On a Linux machine for example you can extract it using:

```
$ tar -zxvf hicViewer.tar.gz
```

Then go to the root directory and run the init script in the hicViewer directory as shown bellow:

```
$ cd hicViewer
$ python __init__.py
 * Running on http://0.0.0.0:5000/
 * Restarting with reloader
```

You can then open *HiC-3DViewer* by typing *http://localhost:5000* if you are on you local machine, or you can access it from another machine by typing *http://Machine_IP:5000*. A user-friendly interface should be displayed as the one in Fig.2.

If want to use *HiC-3DViewer* with APACHE, you need to install the `wsgi` APACHE module.

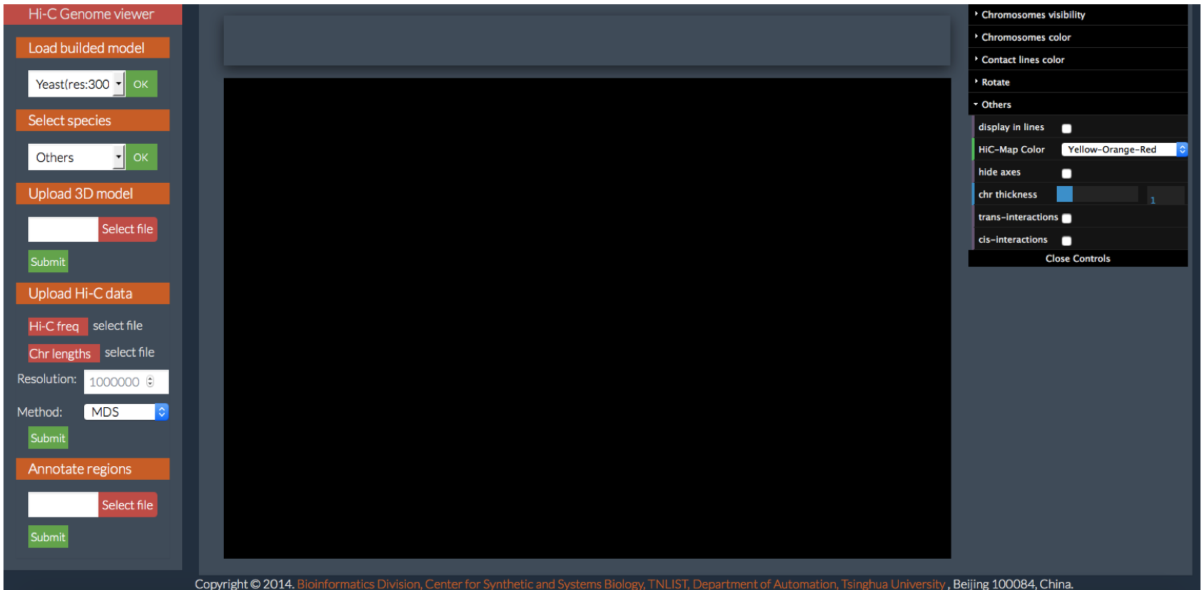
```
<VirtualHost ${YOUR_IP}:80>
  ServerName www.mywebsite.com
  ServerAdmin admin@mywebsite.com
  DocumentRoot /home/web/html
  <Directory /home/web/html>
    WSGIApplicationGroup %{GLOBAL}
    Order allow,deny
    Allow from all
  </Directory>
  WSGIDaemonProcess {Hic3DViewer_PATH} user={MYACCOUT}
        group=webuser processes=2 threads=15 python-path={PYTHON_PATH}

  Alias /static {Hic3DViewer_PATH}/hicViewer/static

  WSGIProcessGroup bioinfo.au.tsinghua.edu.cn/member/nadhir/HiC3DViewer
  WSGIScriptAlias /member/nadhir/HiC3DViewer {Hic3DViewer_PATH}/hic3dviewer.wsgi

  <Directory {Hic3DViewer__PATH}>
    WSGIApplicationGroup %{GLOBAL}
    Order allow,deny
    Allow from all
```

```
 </Directory>
 <Directory {Hic3DViewer_PATH}/hicViewer/static>
  WSGIApplicationGroup %{GLOBAL}
  Order allow,deny
  Allow from all
 </Directory>
 <Directory {Hic3DViewer_ROOT_PATH}/hicViewer/tmp>
   WSGIApplicationGroup %{GLOBAL}
   Order allow,deny
   Allow from all
 </Directory>
</VirtualHost>
```



Load builded model: Yeast(res:300
Select species: Others
Resolution: 1000000
Method: MDS
HiC-Map Color: Yellow-Orange-Red
chr thickness: 1

Copyright © 2014. Bioinformatics Division, Center for Synthetic and Systems Biology, TNLIST, Department of Automation, Tsinghua University , Beijing 100084, China.

Figure 2: Main user interface of HiC-3DViewer. It has three main parts : the top-left panel for launching 3D prediction and for data upload. The middle part for 3D visualization and the top-right panel to customize the display

5 Available features

The main features of *HiC-3DViewer* can be summarized as follows:

5.1 Loading existing models

HiC-3DViewer comes with some pre-calculated 3D models (Fig.3) for some species such as budding yeast, human and drosophila predicted using the *Pastis* package.

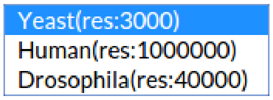


Figure 3: pre-calculated models in Hi-C3DViewer. Each entry indicates the species name and the resolution of the Hi-C heatmap. Here the Yeast Hi-C data with a 3kb resolution is selected for display

These models can be accessed from the top scroll-bar on the left-side menu of the user interface (Fig.4).

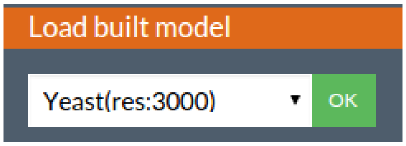
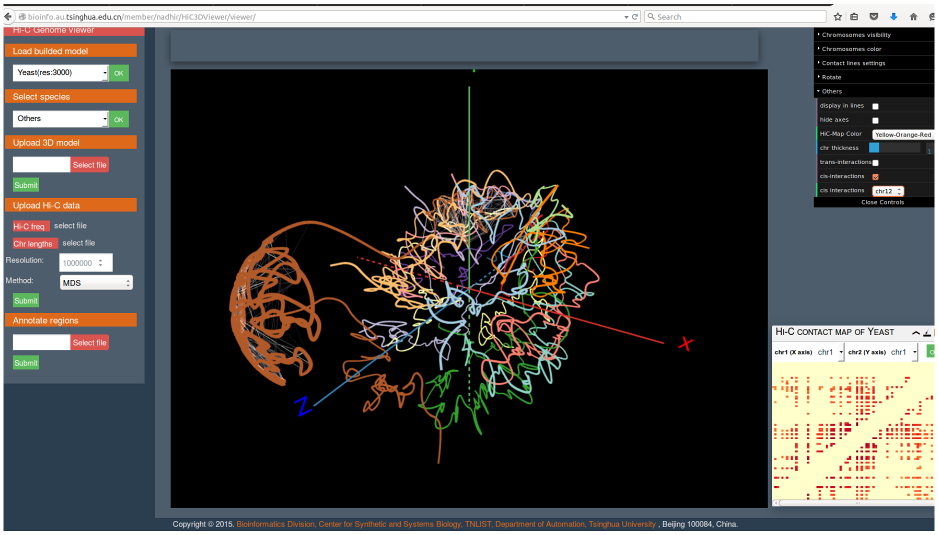


Figure 4: Launching the display of a 3D model. The 3D model is displayed when the user clicks on OK

The time required for loading a model depends on the size of the Hi-C data and the used 3D prediction method. The fastest should be MDS and the slowest is PM2. For example, for the Drosophila 3D model, it takes longer time to upload in *HiC-3DViewer* as the Hi-C Data has a higher resolution. But generally, it doesn't take much time.


bioinfo.au.tsinghua.edu.cn/member/nadhir/HiC3DViewer/viewer/

Resolution: 1000000
Method: MDS
HiC-Map Color Yellow-Orange-Red
chr thickness 1
cis interactions chr12
chr1 (X axis) chr1
chr2 (Y axis) chr1

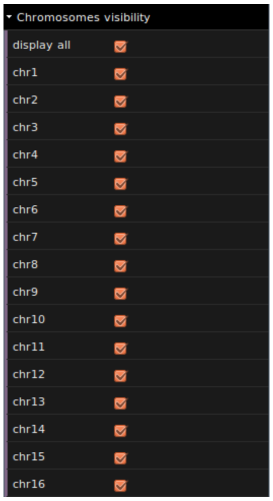
Figure 5: Example of a loaded budding yeast 3D model

When the model is loaded a couple of panels will be displayed (Fig.5). On the top-right, a control panel is displayed to enable the user perform different configuration for every chromosome. The X,Y and Z are displayed as red, green and blue lines respectively in the 3D model display region. Continuous lines represent positive coordinates while the dotted lines represent the negative ones. The (0,0,0) position refers to the center of the whole 3D space. On the bottom-right an interaction floating panel containing the Hi-C matrix is displayed. The top-left panel is always displayed. It contains different panels that enable the user to select the data to visualize, predict 3D models and upload highlight some chromatin regions.

5.2 The control panel

The control panel enables the user to customize the 3D display of every chromosome, such as the color, thikness, the display of inter- and intra-chromatin interactions and 3D rotations . For the current version, the control panel is composed of the following parts:

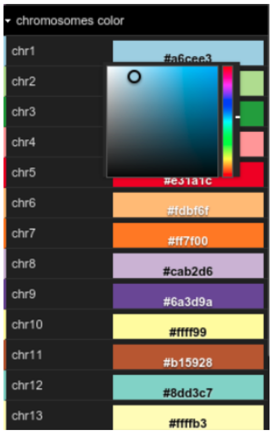
- **Chromosome visibility panel (Fig.6)**: This panel enables users to select the set of chromosomes (users can select either just one chromosome or many) to display. When the user changes the visibility of chromosomes, the camera will automatically look at the center of the visible chromosomes. The target of orbit controls will also be set to the center of the visible chromosomes.



Chromosomes visibility	
display all	☑
chr1	☑
chr2	☑
chr3	☑
chr4	☑
chr5	☑
chr6	☑
chr7	☑
chr8	☑
chr9	☑
chr10	☑
chr11	☑
chr12	☑
chr13	☑
chr14	☑
chr15	☑
chr16	☑

Figure 6: Chromosome visibility panel. Users can select the chromosomes to display by checking the relative checkboxes

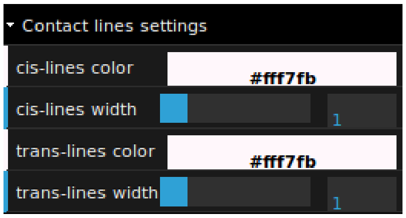
- **Chromosome color panel (Fig.7)**: This panel enables users to change the colors of the different chromosomes. Users can customize the look and feel of chromosomes by writing the hex code of the color or selecting it using the color picking panel.



chromosomes color	
chr1	#a6cee3
chr2	
chr3	
chr4	
chr5	#e31a1c
chr6	#fdbf6f
chr7	#ff7f00
chr8	#cab2d6
chr9	#6a3d9a
chr10	#ffff99
chr11	#b15928
chr12	#8dd3c7
chr13	#ffffb3

Figure 7: Chromosome color panel.Chromosome colors can be selected through the color picking panel or directly written into the text-box

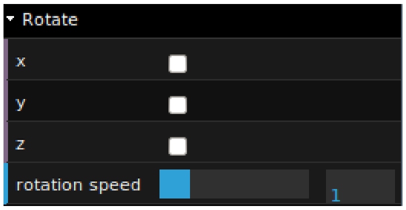
- **Contact lines settings (Fig.8)**: This panel is used to control the color and width of cis- and trans- interactions lines. The width of contact lines ranges from 0.1 to 5, default 1.



▾ Contact lines settings
cis-lines color #fff7fb
cis-lines width 1
trans-lines color #fff7fb
trans-lines width 1

Figure 8: Contact lines settings panel. Different options are available to customize the display of the contact lines

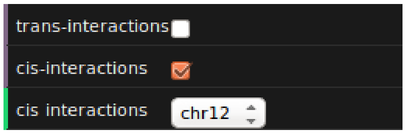
- **Rotation (Fig.9)**: This panel is used to control the rotation of the model. x,y and z means rotating around the X, Y, Z axis respectively. rotation speed can control the speed of rotation (from 0 to 5, with a default value of 1).



▾ Rotate
x □
y □
z □
rotation speed 1

Figure 9: Rotation panel. To rotate the 3D model across the X,Y and Z axis

- **Inter- and Intra- chromatin interactions (Fig.10)**: used to display inter- and intra- chromatin interactions. As the number of interactions is very large, we restrict ourselves to display a sample representing 10% or 1% of the interactions, depending on the total interactions size(Fig.10). Only trans-interaction lines between visible chromosomes can be shown. Because showing trans-interaction lines will take several seconds, the lines will not be update automatically when the user changes the visibility of chromosomes. If user want to update the trans-interaction lines, they may needs to hide then show the trans-interactions lines again by using the control panels.



trans-interactions □
cis-interactions ☑
cis interactions chr12

Figure 10: Inter- and Intra- chromosome display panel. The checkboxes can be used to display or hide interactions

- **Display in line (Fig.11)**: This panel is used to control the display mode (mesh or line). When the "display the model in lines" option is selected before loading, the model can save memory and loading time, unlike the meshe mode. The width of lines wont change when zooming in or zooming out. However, the user can adjust the width of lines or meshes using the slider chr thickness in the "Others" panel" (Fig.11).
- **Hide axes (Fig.12)**: This panel is used to control the visibility of axes, users can use it to hide or display the X,Y and Z axis.

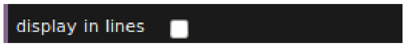

display in lines □

Figure 11: Display mode panel. When checked the chromosomes will be displayed as lines, otherwise, the whole mesh will be displayed.

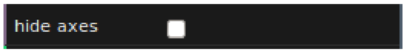

hide axes □

Figure 12: Hide axes panel. When checked the X,Y and Z axis will be hidden

- **Hi-C map color (Fig.13)**: used to control the color scale of Hi-C map. Different gradient colors are available, users can select the one that suites them. This option is useful for color blind users.

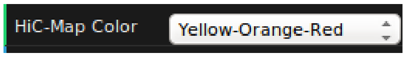

HiC-Map Color Yellow-Orange-Red

Figure 13: Hi-C map color panel. Different gradient color maps are available to customize the Hi-C heatmap display

- **Chromosome thickness (Fig.14)**: This panel is used to control the thickness of the chromosomes. The range of thickness is from 0.05 to 5. The default value is 1.

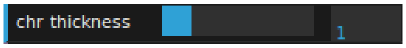
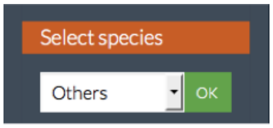

chr thickness 1

Figure 14: Chromosome thickness panel. The thickness of the chromosomes can be changed by moving the sliding button.

5.3 Select Species

HiC-3DViewer also enables users to select corresponding species and show their cytobands. When the user chooses regions of the same chromosome from the Hi-C contact map, these regions will also be highlighted in cytobands. All cytobands are downloaded from http://genome.ucsc.edu/. Note that only several species have detailed cytobands data (including human, drosophila, mouse and rat). Other species just have one long band for each chromosome.


Select species

Others

Figure 15: Select species. Cytobands of different species can be displayed, users can select the suitable one. If the selected cytoband is different from the displayed species genome a while panel will be displayed.

5.4 Loading custom 3D model

HiC-3DViewer also enables users to load their own 3D model, with their own Hi-C data, to display and analyze the 3D genome of different species (Fig.16).

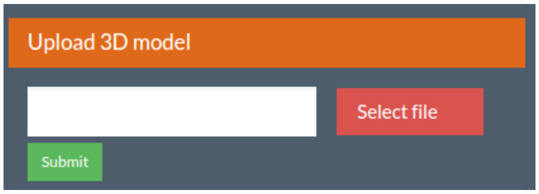


Figure 16: Upload 3D model option. Users can use this option to upload their own 3D models predicted using different methods not available in HiC3D-Viewer.

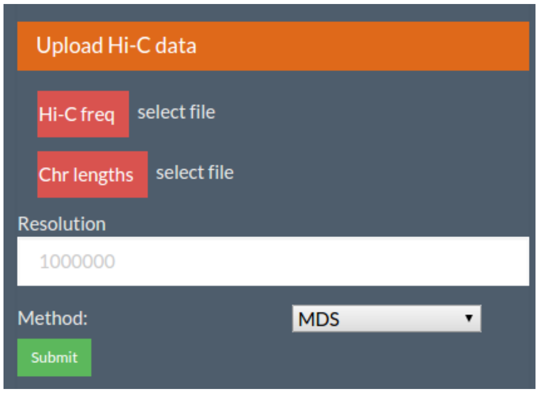
Only tabular files are supported in this version, where the first column indicates the chromosome, the second one indicates the locus and the following three columns correspond to the 3D position as shown below:

```
chrom locus  X3D_x  X3D_y  X3D_z
    1     1 79.452 56.972 90.239
    1   192 80.132 57.616 90.590
    1   383 80.812 58.260 90.941
    1   574 81.492 58.904 91.291
```

Once the model is uploaded it should be displayed on the list of available genomes (Fig.3)

5.5 Model Prediction

This option enables the user to build a 3D model given a Hi-C frequency matrix (Fig.17).



Resolution: 1000000

Method: MDS

Figure 17: 3D prediction panel. This panel used to upload Hi-C frequency map for 3D model prediction

5.5.1 Supported data formats

Three types of information should be provided :

- **Hi-C frequency matrix**: The Hi-C matrix can be uploaded using different formats. For smaller genomes, you can directly upload a tabulated txt file that contains the interaction frequency between each two bins. An example of the 4×4 Hi-C matrix is shown here:

```
     1    2   3   4
1    0 1119 242 272
2 1119    0 910 245
3  242  910   0 234
4  272  245 234   0
```

However, with smaller resolutions, the Hi-C matrix tends to be sparse and the usage of a tabulated Hi-C tends to generate larger files. Therefore, *HiC-3DViewer* enables users the upload his data using more compressed formats that represents only non-zero values. Three types of tabluated data are supported:

  - **3-column format:**: In this format the Hi-C contact map is represented as 3 column file, where the first two columns indicate the id of the interacting bins and the third one indicates the interaction frequency. The chromosome positions will be determined according to the `lengths` file and the `resolution`. An example is shown below:

    ```
    1 2 1119
    1 3  242
    1 4  272
    1 5   71
    ```

  - **5-column format:** In this format, the first 4 columns indicate the chromosome number and the bin id of the interacting regions and the last column indicates the interaction frequency. An example is shown below:

    ```
    1 1 1 2 1119
    1 1 1 3  242
    1 1 1 4  272
    1 1 1 5   71
    ```

  - **7-column format:** In this format, the first 6 columns indicate the chromosome number and the start and stop gnomic positions of the interacting regions. The last column indicates their interaction frequency. An example is shown below:

    ```
    1 10000 20000 1 20000 30000 1119
    1 10000 20000 1 30000 40000  242
    1 10000 20000 1 40000 50000  272
    1 10000 20000 1 50000 60000   71
    ```

- **Chromosome lengths**: To know the number of bins occupied by each chromosome, the user should upload a text file that indicates the size of each chromosome (in bp). The uploaded file should be a two-column text file where each line contains the name and the length of each chromosome. The chromosome names should begin with 'chr', for example, the name of chromosome 1 should be 'chr1'.

  It is better that the number of chromosomes in the chromosome lengths file is the same as the number of the chromosomes encoded in the Hi-C heatmap. For example if the uploaded Hi-C matrix represents only the interactions in one chromosome, then only one line is needed in the chromosomes-length file, if not, *HiC-3DViewer* will correct it automatically.

  An example of the chromosomes-length file is presented bellow:

```
 240000
 820000
 320000
1540000
```

- **Resolution**: This information is used by *HiC-3DViewer* to know the number of bins each chromosome occupies in the Hi-C matrix.
  For example, from the chromosomes-length file we know that chromosome 1 has a length of 249,250,621bp, so if the Hi-C heatmap was constructed using a resolution of 1,000,000bp, then chromosome 1 should occupy: 249250621/1000000 $\approx$ 250 bins.
- **The 3D model prediction method**: The user should also specify the 3D model prediction algorithm to use. Actually, the 4 algorithms provided by *Pastis* are used for predictions. For more information on the algorithms please check the algorithms explanation in the next page.

5.5.2 3D model prediction algorithms used in HiC3D-Viewer

HiC-3DViewer is based on the 3D model prediction algorithm developed by N.Varoquaux.*et al*[1] available in the *pastis* package which implements four 3D model prediction algorithms:

- Two multidimensional scalling (MDS) based algorithms : metric and non-metric MDS respectivelly .
- Two statistical models that assume that the counts between two loci follow a Poisson distribution and that their intensity decreases with increasing genomic distance between the loci.

In brief, the incorporated methods are based on the following assumptions:

- **Metric-MDS** : It is based on the classical MDS algorithm [2]. Given a Hi-C frequency matrix F, for each pair of interacting bins (i, j) the algorithm calculates a physical estimated distance $\delta_{ij} = (\frac{1}{f_{ij}})^{\alpha}$. The algorithm then, tries to find the best spatial 3D positioning that can generate the physical estimated distances δ_{ij}, by optimizing the following equation:

$$\min_{x} \sum_{(i,j)\in\mathcal{D}} (\parallel x_i - x_j \parallel - \delta_{ij})^2 \tag{1}$$

- **Non-Metric MDS (NMDS)** : To avoid the strong assumptions made about the distance matrix in the metric MDS model, the NMDS method doesn't restrict itself to the fitting of the distance-matrix. However, it is based on the observation that if two points i and j have a higher Hi-C contact frequency than another two points k and l, then i and j should be closer in the 3D space than k and l. In other words, if C is the Hi-C contact matrix we want to find $X \in R^3$ such as:

$$c_{ij} \geq c_{kl} \Leftrightarrow \parallel x_i - x_j \parallel_2 \leq \parallel x_k - x_l \parallel_2 \tag{2}$$

- **Poisson Model 1 (PM1)** : In this model each contact frequency c_{ij} is modeled as an independent Poisson random variable with $\lambda = \beta d_{ij}(X)^{\alpha}$ were $d_{ij}(X)$ is the euclidean distance between i and j. $\beta > 0$ and $\alpha < 0$ represent the parameters of the model.

The likelihood of having a certain 3D positioning X can be then calculated as:

$$\ell(X, \alpha, \beta) = \prod_{i,j} \frac{(\beta d_{ij}^{\alpha})^{c_{ij}}}{c_{ij}!} \exp(-\beta d_{ij}^{\alpha}) \tag{3}$$

The PM1 algorithm tries to find the best β value that maximizes the log-likelihood given that α is fixed.

$$\max_{\alpha,\beta,X} \mathcal{L}(X, \alpha, \beta) = \sum_{i \le j \le n} c_{ij} \alpha \log(d_{ij}) + c_{ij} \log(\beta) + \beta d_{ij}^{\alpha} \tag{4}$$

- **PM2** : The PM2 algorithm uses a non-parametric approach to estimate α and β in equation (4).

5.6 Interactive Hi-C map

In many cases the investigator would like to make a correspondence between the Hi-C contact matrix and the 3D model, for example to check the spatial proximity between two genomic regions. Thus, in *HiC-3DViewer* we enable the user to interactively highlight genomic regions in the 3D model by using 2D Hi-C frequency matrix (Fig.18). Each region is associated with a different color to enable visualization and intuitive manipulation.

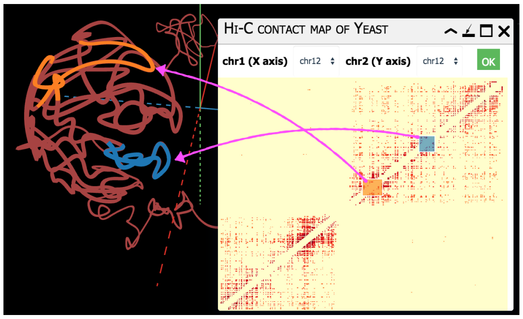


Figure 18: Interactive annotation using 2D Hi-C matrix

If a cytoband species is selected, then a cytoband will be displayed for each annotated region. In case that the selected species is different from the species of the loaded 3D model (for example displaying yeast 3D model while selecting the human cytoband), a white cytoband will be displayed with the annotated regions on it.

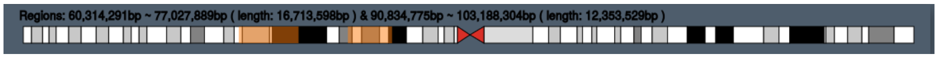
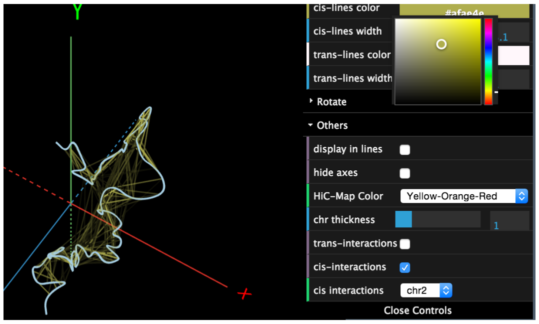

Regions: 60,314,291bp ~ 77,027,889bp (length: 16,713,598bp) & 90,834,775bp ~ 103,188,304bp (length: 12,353,529bp)

Figure 19: Cytoband panel hilights the genomic coordinates of the selected regions on the species genome (here shown in orange). If no cytoband information is available for the species a white band will be displayed instead.

5.7 Inter- and Intra- chromatin interaction

Investigators have also the possibility to visualize inter- and intra- chromatin interactions (Fig.20). The interactions are represented as white lines with different opacity values corresponding to the strength of the interaction. Only trans-interaction lines between visible chromosomes can be shown. Because showing trans-interaction lines will take several seconds, the trans-interaction lines will not update automatically when the user changes the visibility of chromosomes. If user want to update the trans-interaction lines, they may need to hide and show the trans-interactions lines again by using the option in the control panel.



cis–lines color #afae4e
cis–lines width .1
trans–lines color
trans–lines width
‣ Rotate
▾ Others
display in lines □
hide axes □
HiC–Map Color Yellow-Orange-Red
chr thickness 1
trans–interactions □
cis–interactions ☑
cis interactions chr2
Close Controls

Figure 20: The visualization of intra-chromatin interactions (shown in white lines) between genomic loci (on chromosome 2 of the Yeast genome as an example). The transparency of each line is defined by the interaction frequency of the two loci, while the colors on the chromatin correspond to the colors of the domains on 2D Hi-C matrix.

As the display of all the interacting bins in the Hi-C contact matrix will lead to a very crowded display and possibly no useful pattern be observed, we have to restrict the number of displayed interactions. For small datasets we randomly sample 10% of the interactions while for larger datasets (more than 500,000 interactions) we only sample 1%.

The global interaction pattern will not be affected by the sampling process as highly interacting regions will have a higher probability to be sampled.

5.8 Region annotation

In addition to the interactive annotation, the investigator could be also interested in highlighting different genomic regions such as genes or topological domains using a unique color or highlighting regions using gradual colors (for example color regions according to their ChIP-Seq signal value).

For this purpose, users need to upload a 4-columned BED file (as shown bellow) such that the first 3 columns encodes the genomic position and the 4th one indicates its value (for example ChIP-Seq signal value).

```
1 2000 3400 2.9
1 35000 40000 0.5
2 3223 9810 5
2 10293 11400 3
```

6 References

1. Nelle Varoquaux, Ferhat Ay, William Stafford Noble, and Jean-Philippe Vert, A statistical approach for inferring the 3D structure of the genome *Bioinformatics* (2014) 30 (12): i26-i33
2. Kruskal JB, Wish M. Multidimensional Scaling. *Sage University Paper series on Quantitative Application in the Social Science* (1977); 07-011.
3. Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* (1964) ;29:1-27
4. Fudenberg G, Mirny LA. Higher-order chromatin structure: bridging physics and biology. *Curr. Opin. Genet. Dev.* (2012) ;22:115-124.