SBARS

Spectral-**B**ased **A**pproach for **R**epeats **S**earch

User manual

Authors:
Maxim Pyatkov
Anton Pankratov

http://mpyatkov.github.com/sbars/

November 8, 2013

Contents

1 Introduction

The SBARS is fast and efficient tool for identifying dispersed (direct, inverted) and tandem DNA repeats. The main feature of the program is rapid detection of repetitive patterns (repeats) of different types(direct, inverted, etc) in the long sequences. We propose a novel approach based on spectral methods. The general idea of this approach is that the periodical structures are recognized not within the nucleotide sequence directly but within function derived from this sequence.

SBARS is developed in C++, GUI is based on the library QT.

2 The basis of the method

The method is based on the analysis of the function obtained from the nucleotide sequence. The algorithm is split into the following stages (Fig. 1):

- Stage 1. Presentation of the nucleotide sequence in a set of discrete functions.
- Stage 2. Conversion of analog functions into a spectral representation.
- Stage 3. Comparison of the vectors of expansion.
- Stage 4. Display and analysis of the spectral homology matrix.

We assume that scheme (Fig. 1) is clear enough to understand. Mathematical foundation and an implementation details of the algorithm you can see in paper [Pankratov and Pyatkov *et al*, 2012]

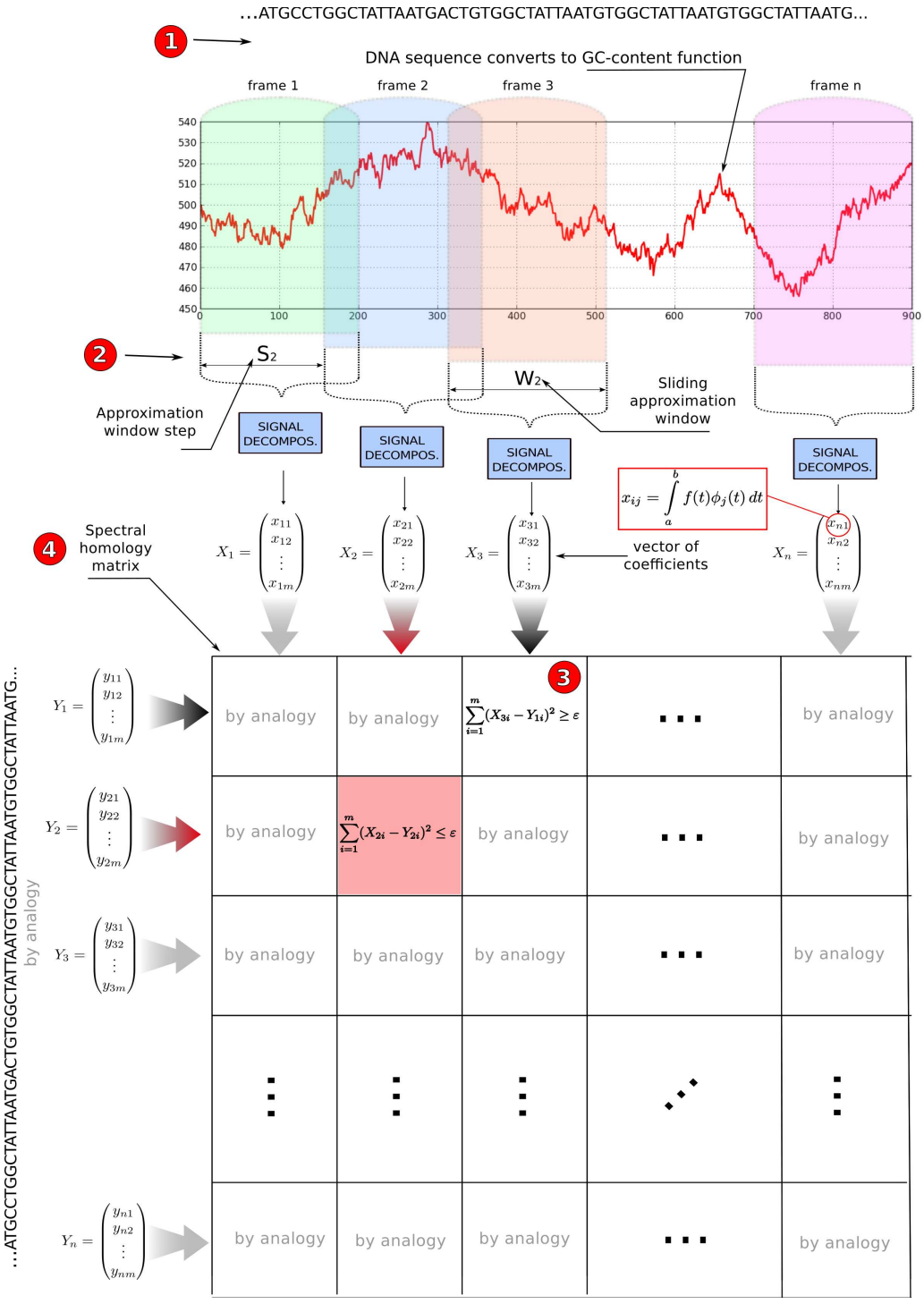


Figure 1: Approximate scheme of the algorithm. Red circles are start points of corresponding stages of algorithm(1 – getting analog function from nucleotide sequence, 2 – spectal representations of analog functions, 3 – comparison of appropriate decomposition vectors, 4 – making the matrix of spectral homology)

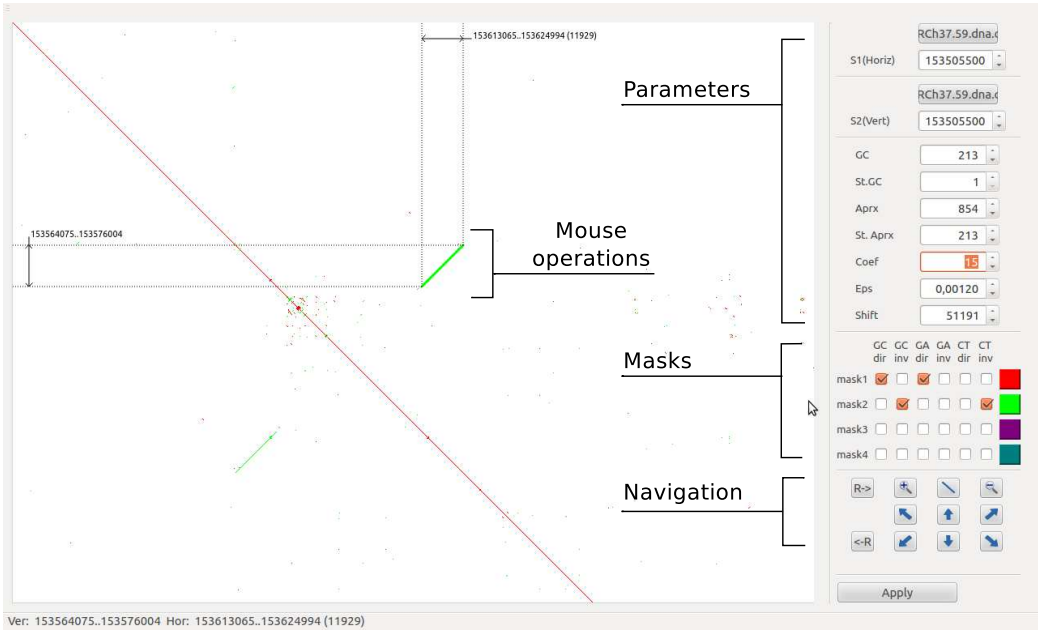


Figure 2: Main window of SBARS

3 Interface and options

The SBARS is a complete implementation of the algorithm given above with graphic user interface (GUI) (Fig. 2). Interface has additional useful functions such as detection of coordinates by the click of the mouse, zoom the selected area, and so on. Next, we will describe the main features and capabilities of the program.

3.1 Parameters

SBARS accept as input two files in the FASTA format or files without headers in format like a simple string of nucleotides. Input files can be of any length, the program does not keep them in RAM, reading required fragment by request from HDD. The program cycle starts with load the files by using two buttons in the right top corner. The first button loads a nucleotide sequence that corresponds to the horizontal side of the matrix, the second button – the vertical side of the matrix, respectively. Parameters **S1(Horiz)** and **S2(Vert)** allow you to specify the initial coordinates for each of the se-

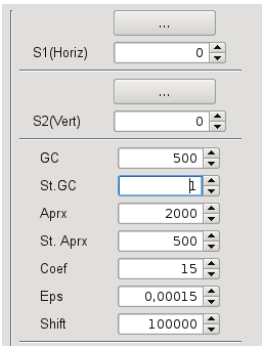


Figure 3: Parameters

quences. **GC** – is a size of sliding window that we use
on the first stage to obtain GC-curve from the nucleotide sequence. **St.GC** – step of GC sliding window. In common case this steps are equal **1** but sometimes we need to explore large sequence, for example, a part of chromosome. In this case this option gives you greater speed of reading files with slightly reduced quality of recognition. But you should use it carefully to gain correct results. The **Aprx** and **St.Aprx** are another sliding window and step. On the scheme (Fig. 1) **Aprx** correspond to the W_2 – size of approximation window and the **St.Aprx** to the S_2 – step of approximation window. The size of decomposition vectors you can set with the option **Coef**. The threshold value ε is a basic parameter at the third stage of the algorithm. It varies in the range from **0** to **1**. Changing this value allows you to change the intensity of the color matrix. The last option **Shift** will be described in the paragraph devoted to navigation.

Selecting options, you should adhere to the following set of rules, to simplify your life:

- One point on the matrix corresponds to the comparison of two vectors, which, in turn, correspond to the comparison of the two curves caught in the approximation window. The size of the approximation window (**Aprx**) is the minimum size of repeat that can be displayed on a matrix.
- Parameters **GC** and **St.Aprx** according to our estimates should be 4-10 times less than parameter **Aprx**.
- If after enlargement of the matrix fragment everything is white, increasing ε will allow to increase the intensity of the matrix.

3.2 Masks

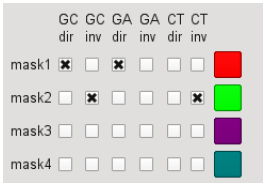
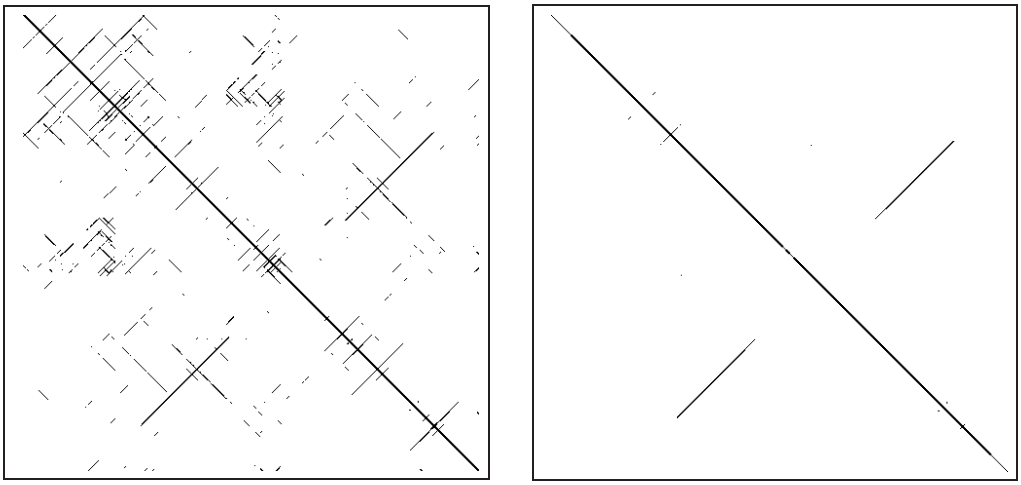


Figure 4: Masks

For better recognition of repeats and reduction of noise on spectral matrix we add another one curve which is based on G and A nucleotides. That is, we compare the vectors obtained from the GC-curve, then compare the vectors obtained from the GA-curve and accept the result in case if the first and second comparison satisfies ε. Despite the fact that this action increases the computing time, we are building a better matrix (Fig.5). Using checkboxes you can choose the combinations of curves, thus obtaining different types of repeats. For example, the selected only **GC dir** and **GA dir** chekboxes means that you are searching direct repeats using two curves.



a) b)

Figure 5: Enabling filtration of spectral matrix by addition GA-curve. All calculated with the same parameters a) Only GC-curve b) GC- and GA-curves.

Types of repeats	Combination of checkboxes					
	GCdir	GCinv	GAdir	GAinv	CTdir	CTinv
Direct (only GC)	•					
Direct (GC+GA)	•		•			
Reverse (only GC)		•				
Reverse (GC+GA)		•		•		
Complementary	•				•	
Inverted		•				•

Table 1: The most used types of repeats.

Selected **GC inv** and **CT inv** gives you inverse-complement repeats, because for GA-curve complemented one will be CT-curve. Suffix **inv** in the checkbox names mean *"inverse"* but not *"inverse-complement"*. At the same time you can fill four masks and see four different types of repeats in the matrix, respectively. The most used types of repeats and their checkboxes configuration you can see in Table.1.

3.3 Navigation

To navigate the chromosome, you can use the arrow buttons. The offset value is set by the **Shift** parameter(by default it equal 1/3 of viewing window). The

Figure 6: Navigation

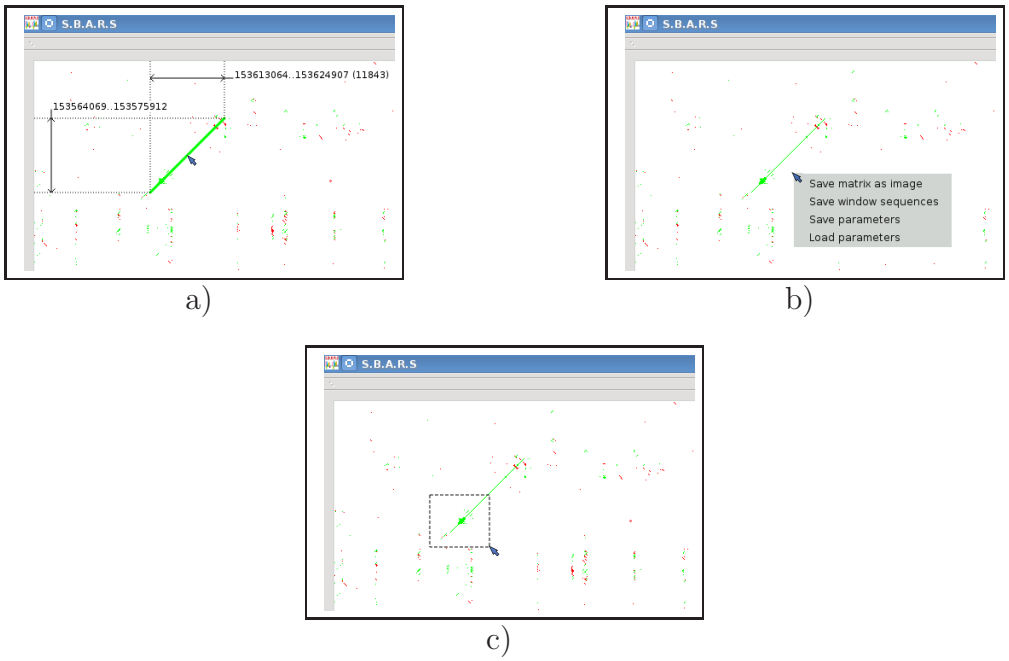
buttons "$R \rightarrow$" and "$\leftarrow R$" visualize repeats found in the matrix.

3.4 Mouse interaction

For more friendly using we add some mouse interaction (Fig.7).

Right clicking options:

- **Save matrix as image** – this action allows you save viewed matrix as PNG image
- **Save window sequence** – this action allows you to save the vertical and horizontal sequences, which correspond to the current viewing matrix
- With actions **Save parameters** and **Load parameters** you can save interesting for you parameters as INI files.



a) b)

c)

Figure 7: Different types of mouse interaction. a) Left click on repeat shows its coordinates(in brackets there is length of repeat). b) Right click shows options menu. c) Pressing and moving left mouse button allows select interesting part for zooming.

We should note that in the matrix are considered repetitions longer than 10 pixels.

4 Examples

4.1 Direct and inverted repeats

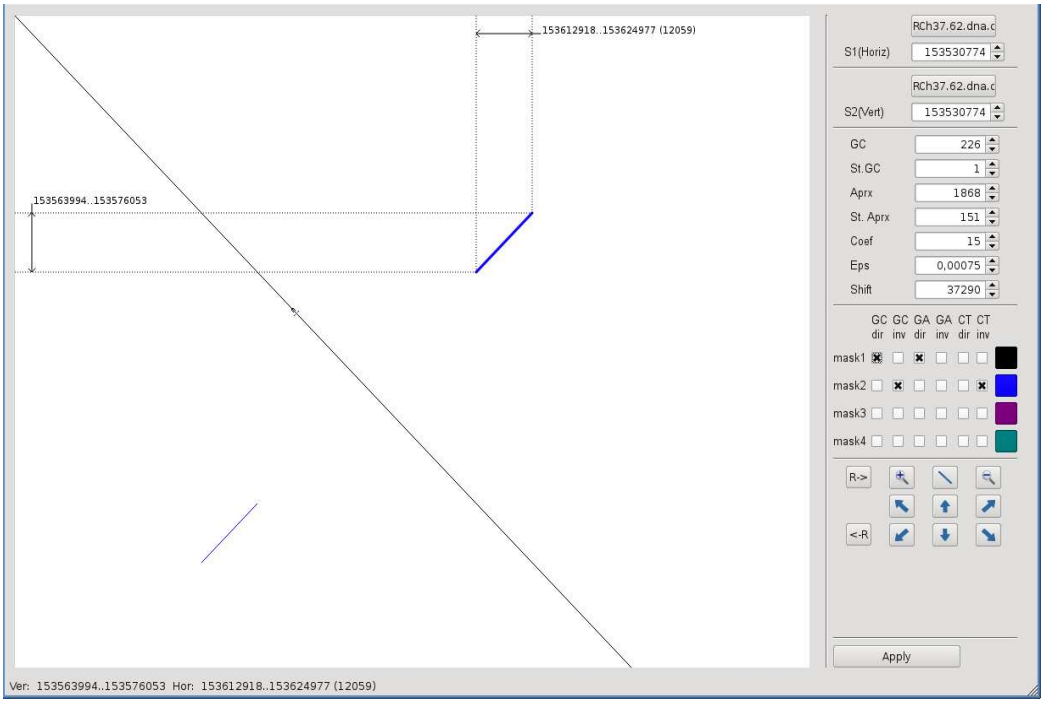
Despite the fact that the role of different types of repeats is not yet fully understood, let us consider the real case, when repeats are the cause of the mutation. In the paper [Small *et al*, 1997] authors shows the fact that the presence of inverted repeats flanking two closely located genes (Emerin and FLNA) results in the inversion of this region.

The last build of the human genome you can get from `www.ensembl.org`. In particular, chromosome X (last build on November,2012) you can take from this url:

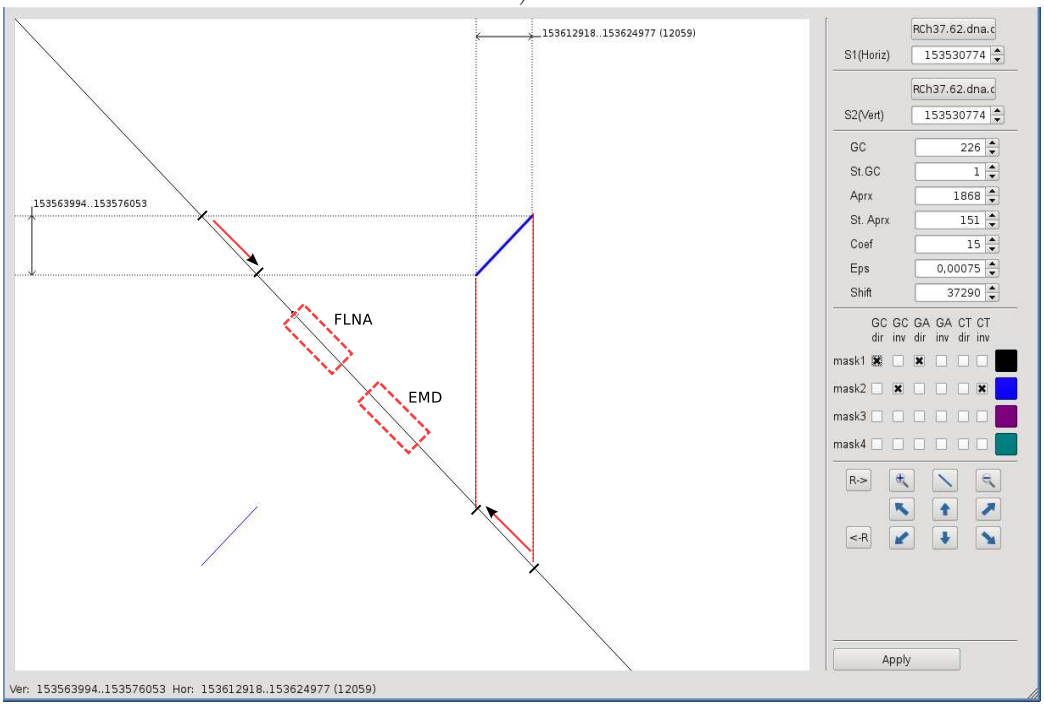
- `ftp://ftp.ensembl.org/pub/release-69/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.69.dna.chromosome.X.fa.gz`

After entering coordinates and parameter selection you will see the following figure (Fig.8).

In the case of direct repeats the pattern would be similar with the only exception that repeats will be parallel to the main diagonal.



a)



b)

Figure 8: Inverted repeats flanking genes. a) Real screenshot b) Add additional explanation

4.2 Tandem repeats

Lets consider an example of long tandem repeat. The repeat named **IMPB_01** [Tetuev and Nazipova, 2010] is a good candiadate for this purpose. In the description it is noted that this region of mouse chromosome 6 located at 114441515-114587670 contains 60 tandem copies of a pattern with average length 2400 bps.

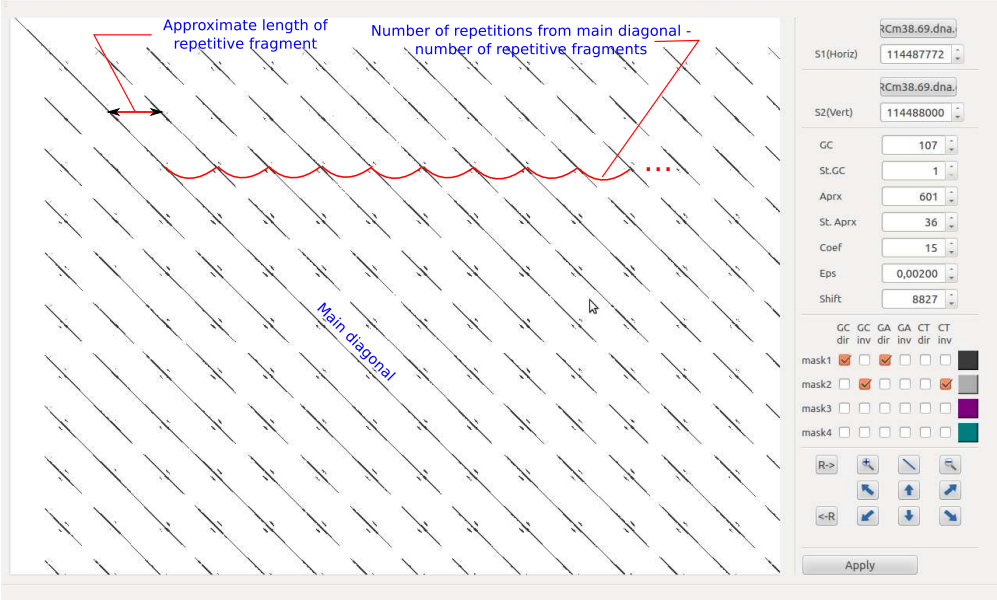
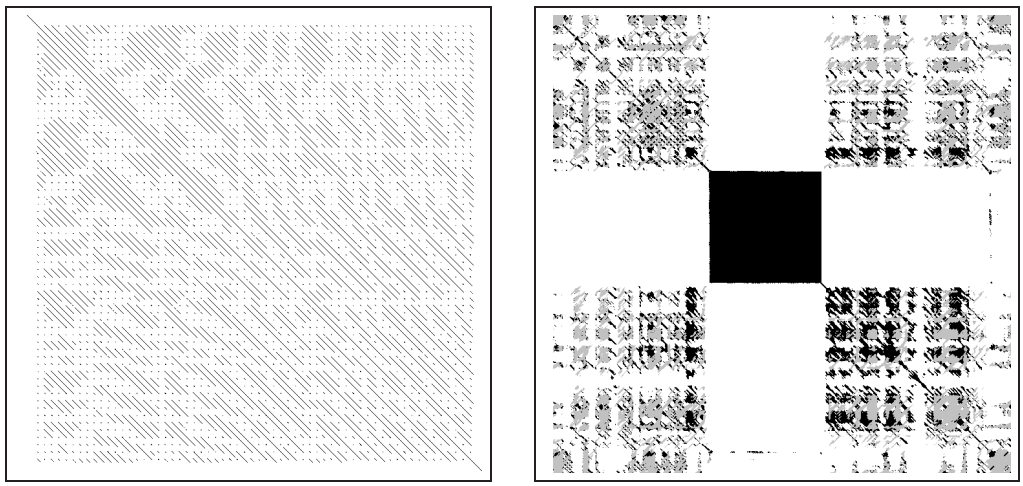


Figure 9: Expalanation of how to explore the tandem repeats(IMPB_01) .

On the picture (Fig.9) you can see how these repeats appear in the program and its explanation. Various options allow you to change scale of viewing area (Fig. 10).



a) b)

Figure 10: Tandem repeat IMPB_01 at different parameters.

References

[Pankratov and Pyatkov *et al*, 2012] Pankratov, A., Pyatkov, M., Tetuev, R., Nazipova, N., Dedus, F. (2012) Search for extended repeats in genomes based on the spectral-analytical method, *Mathematical Biology and Bioinformatics* Vol. 7, **2**, 476–492.

[Small *et al*, 1997] Small K, Iber J, Warren ST. (1997) Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats, *Nature genetics* Vol. 16, **2**, 96–99.

[Tetuev and Nazipova, 2010] Tetuev, R.K., Nazipova, N.N. (2010) Consensus of repeated region of mouse chromosome 6 containing 60 tandem copies of a complex pattern, *Repbase Reports.* Vol. 10, **5**, 776.