# SIGI-CRF-AUTO tool manual

02/15/2015

## Contents

# 1 Introduction

SIGI-CRF is a tool for prediction of horizontal gene transfer (HGT) events within a sequence given raw DNA as input. SIGI-CRF does not need any further knowledge of the sequence and can be run even if annotation has not been done yet.

Sigi-CFR has two central parameters that control the prediction to some degree, and it may be of interest to the biologist to see differences between a larger number of parameter variations graphically.

This is where the add-on tool SIGI-CRF-AUTO comes in: SIGI-CRF-AUTO allows to run SIGI-CRF with different parameters and to create a graphic showing the combined results of the different SIGI-CRF-runs. You can think of SIGI-CRF-AUTO as a robot that starts various runs of SIGI-CRF one after the other. As soon as one run finishes, the next run starts. This continues until all parameter combinations in question have been run. After all runs of SIGI-CRF have finished their results are collected and used to build the result graphic.

For a sample graphic created with SIGI-CRF-AUTO tool see figure 1.

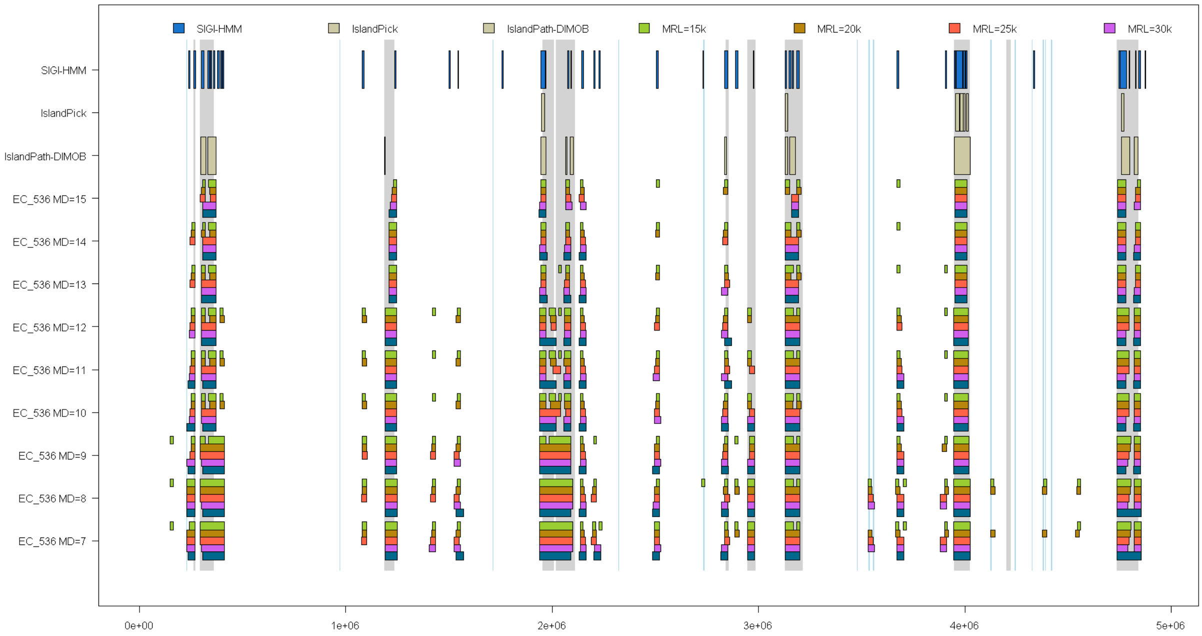


Figure 1: Sample plot generated by SIGI-CRF-AUTOTool

The two central parameters of SIGI-CRF are 'Minimum Dissimilarity' (MD) and 'Minimum Region Length' (MRL). A basic explanation for these two parameters follows:

- **'Minimum Region Length' (MRL)** SIGI-CRF is segmenting input sequences into segments sized 5k each. A Conditional Random Field (CRF) is used to reasonably smooth the prediction and to prevent too short HGT events originating from irregular DNA regions. A minimum length of detected islands (as well as of 'normal' regions between two islands) is effectivly enforced by the CRF. The key point of using a CRF is that it performs a global optimization, even if rivaling features are detected next to each other.

  So if regions smaller than the given parameter length consisting of only single or few segments do show one or more strong HGT-like features then the CRF has two choices: The first alternative is to value the HGT-like features as non-substantially and thus to regard the region as species-specific. The other alternative is to extend the region with some strong HGT-like features in either direction to have it last at least the length given by the MRL parameter. Which of the two options is chosen by the CRF depends on both the 'strengths' of the HGT-like features and the features of the surrounding segments.

- **'Minimum Dissimilarity' (MD)** Basically SIGI-CRF compares each segment to all entries of SIGI-CRF's built-in reference profiles database. Because organisms that are near relatives often have similar genetic signature special care must be taken to eliminate near neighbours from the reference profile database before the CRF does the HGT prediction. The approach taken here is that all reference profiles are compared to the profile of the input sequence and profiles too similar are not used in the comparison.

  Technically this parameter is some $L_1$-Distance over profile vectors containing some nucleotid markov probabilities.

  As a first guidance for parameter values: Different organisms of the same species have mostly values less than 3, different organisms of the same genus have mostly values less than 6, and 9 out of 10 organisms of the same class but of different order have a value of less than 17[1]

The main idea of the SIGI-CRF-AUTO tool is to allow to specify ranges for these two SIGI-CRF parameters and build graphics files that concentrate all individual SIGI-CRF results into one overview picture.

## 2 Starting SIGI-CRF-AUTO

SIGI-CRF-AUTO is packed as a Java jar file named *sigi-crf-auto.jar*. With a Java runtime version 1.6 or higher installed on the system this file can be directly started, for example with Windows a simple double-click in the Windows Explorer does start the program.

## 3 Fast Lane

If you simply want to have the graphic, then basically all you need to do is:

- Go to tab 'System settings' and select both your base dir of Colombo/SIGI-CRF and the Java executable from your Java dir. Set the Xmx parameter to 1000 to allow SIGI-CRF to use up to 1 Gig of memory. Switch back to 'Parameters' tab.
- For the 'Minimum Distance (MD)' accept the values From=10, To=25 and Step=1
- For the 'Minimum region length (MRL)' accept the values From=15000, To=30000 and Step=5000
- Select an input sequence file with either .embl, .gbk or .fasta (single contig only) file format
- Select both 'Run SIGI-CRF (multi)' and 'Generate plot files' and click 'Start'

[1] All these guidance values are estimated from experiments using a complety copy of all GenBank completely sequenced bacterial genomes from April 2010, with only few seuqnces removed to due sequences either shorter than 500kb or no taxonomic information available.

- Depending on the settings and your machine performance the runtime may be up to few hours. Wait until progress bar and message area indicate that the job is finished.

- Find your .R graphics script file in the sequence directory. To view the graphic run this file with R (See section 7 for details)

All settings are described below in more detail.

# 4 Reading the generated graphic

## 4.1 Composition

The basic composition of a SIGI-CRF-AUTO graphics is as follows:

- The sequence spans the graphic from left to right. I.e. at the left side of the picture is the start of the sequence, at the right side of the picture is the end of the sequence.

- Each row of the graphic corresponds to a particular *Minimum Dissimilarity (MD)* value. All different SIGI-CRF predictions that are done with this particular MD value share the same row (one will have more than one prediction per row only if different Minimum region length (MRL) values are used). The sequence's basenname and the row's MD value is printed as a label left to the row. See figure 1 for a graphic where MD values ranging from 7 to 15 are drawn to 9 distinct rows.

- Plotting of different MRL values often happens to be in the same row but still all of them will be visible. SIGI-CRF predictions that share the same row but have different MRL values differ in two properties:

  1. Color. Each MRL value has a distinct color. A color legend at the top tells which color is which MRL.

  2. Bar height. Each MRL value corresponds to a particular bar height. The larger the MRL value, the smaller the bar height.

  Predictions with smaller MRL length will be plotted first and then overdrawn with predictions from larger MRL values. Because predictions with larger MRL values have smaller bar heights but will be plottet last, all predictions will be visible at the same time

  In figure 1 five different MRL values ranging from 15000 to 35000 have been predicted by SIGI-CRF-AUTO.

- To learn about integrated other reference data see section 9.

## 4.2 Interpretation

### 4.2.1 Degrees of Dissimilarity

Figure 2 shows an example area where a sequence shows two HGT-like events. The first one is both smaller in length and weaker in dissimilarity. The first signal is predicted for MD values up to 9. But with higher MD values SIGI-CRF can no longer detect this region as HGT-like.

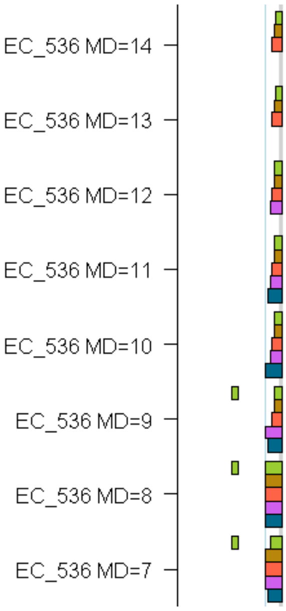


Figure 2: Different dissimilarities

What does this mean? Let's recall section 1. A lower MD value generally means that more species that are taxonomically near relatives are considered as possible donors for HGT events. Higher MD values generally mean that SIGI-CRF will exclude more genomic reference profiles from using them for prediction of HGT-events. So if the first signal is still detected when SIGI-CRF was run with a MD value of 9, but no longer with a MD value of 10, then this suggests that this region has a genomic signature profile similar to a relative of distance of 9.

In contrast let's look at the second signal. The second signal will still be predicted as HGT-like even with MD-values as high as 14. This means that this region is much more dissimilar to the sequence's backbone as the first.

#### 4.2.2 Uniformness of predicted regions

Figure 3 shows an example where implications of different MRL values are apparant.

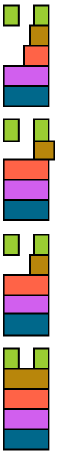


Figure 3: MRL

For SIGI-CRF runs with a shorter *Minimum region length (MRL)* value there are two signals nearby that are predicted as HGT-like (plotted in green). However with larger MRL values the region between them will be bridged (plotter in blue for example).

This may be or not may not be correct in a particular case. Often, genomic islands will consist of not-too-small clusters of HGT genes, and mobile elements may integrate at a whole. So if some strong signals are found nearby it seems plausible that they belong to the same island, even if some intermediate segment shows only a weak signal.

Please note in this example also, that with lower MD value (the lower shapes in the example) a gap may be bridged, that will not be bridged with higher MD value (higher shape): The prediction plotted in red will bridge the gap only in the three lower shapes of the example. Also the prediction plotted in brown (it's MRL value is smaller than the MRL value of the prediction in red. . . ) will bridge a gap only in the lowest shape, i.e. for the lowest MD value. This example may be interpreted in a way that there are two strong signals nearby, with a weaker signal in the middle. By using lower values for MD SIGI-CRF has more near relatives available as comparison profiles, so it will detect more signals that only deviate slightly from the genomic signature of the sequencte tested. Of course a low MD value will also raise the chance of false positiv predictions.

### 4.2.3 Graphic significance

At first sight the picture should give a good overview of the interesting regions of a sequence. Then it should guide the user to get a better understanding of the parameters.

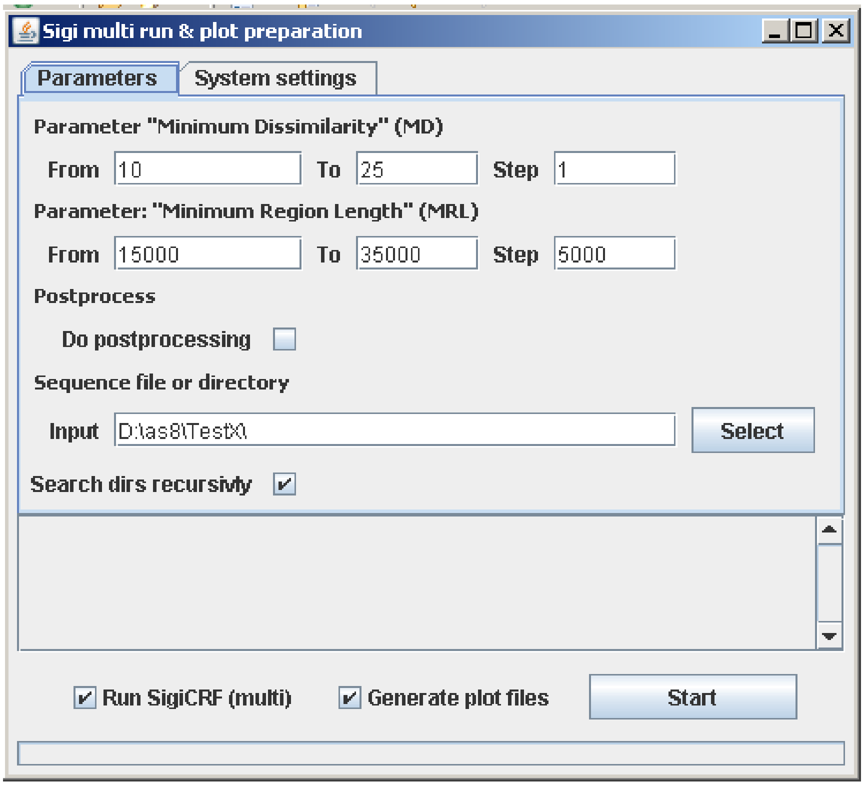
Besides the two basic patterns explained above, there are more deductions to be drawn by the interested user.

## 5 Basic appearance

After starting SIGI-CRF-AUTO the program presents it's user interface, which is designed in a dialog based manner. A couple of parameter input fields are arranged on two tab pages. While the first tab page contains the parameters the user may want to change from start to start, the second tab page basically contains system settings that rarely change over time.

The area below the parameter input fields is used for message output and starting the task. Two kinds of messages are output to the message area: First, all user entries will be checked for plausibility. Any issues will be indicated in the message area. Second, some messages will indicate the progress of the main task.

For screenshots of the dialog based user interface with either of the two tabs selected see figures 4 and 5.



Parameter "Minimum Dissimilarity" (MD)

From 10 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 35000 Step 5000

Postprocess

Do postprocessing □

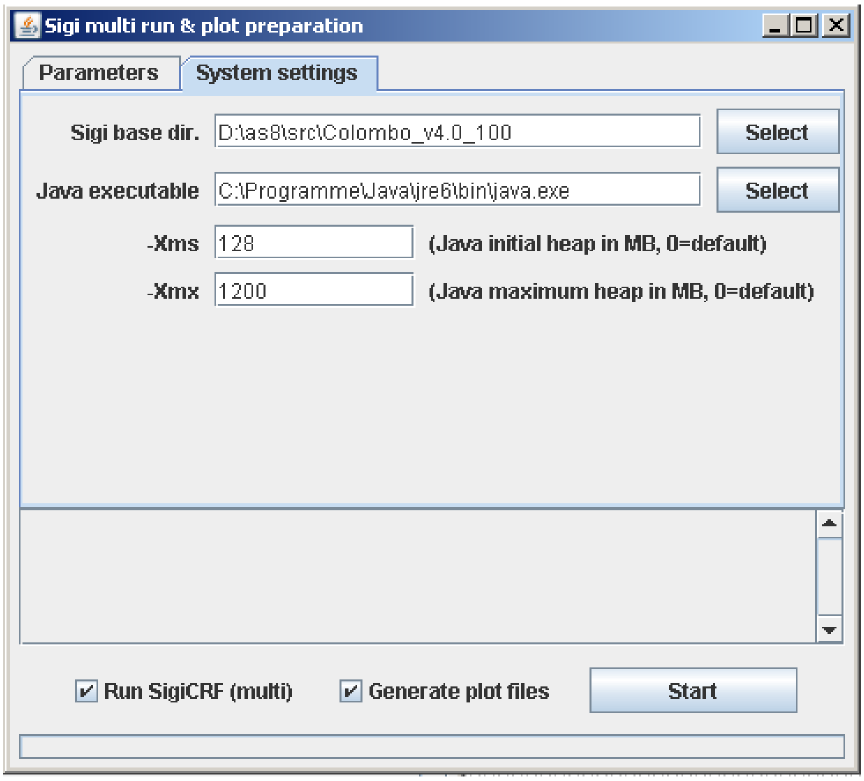
Sequence file or directory

Input D:\as8\TestX\

Search dirs recursivly ☑

☑ Run SigiCRF (multi) ☑ Generate plot files

Figure 4: User interface with tab Parameters selected



Sigi multi run & plot preparation

Parameters | System settings

Sigi base dir. D:\as8\src\Colombo_v4.0_100 Select

Java executable C:\Programme\Java\jre6\bin\java.exe Select

-Xms 128 (Java initial heap in MB, 0=default)

-Xmx 1200 (Java maximum heap in MB, 0=default)

☑ Run SigiCRF (multi) ☑ Generate plot files Start

Figure 5: User interface with tab System Settings selected

# 6 User interface areas and input parameters

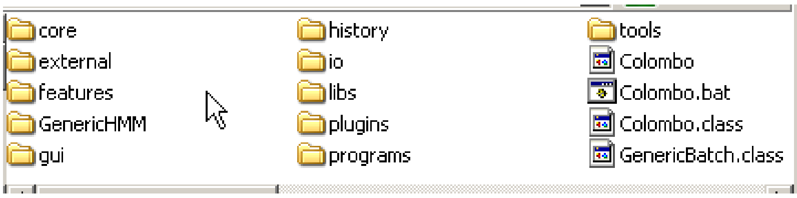
## 6.1 System settings

All system settings are seperated on the second tab page. System settings needs to be initially set, but rarely need to be changed afterwards. In figure 6 the System Settings tab is highlighted green.

### 6.1.1 System setting: Sigi base dir.

To select, find the Colombo/SIGI-CRF install directory on the machine, and select this directory.

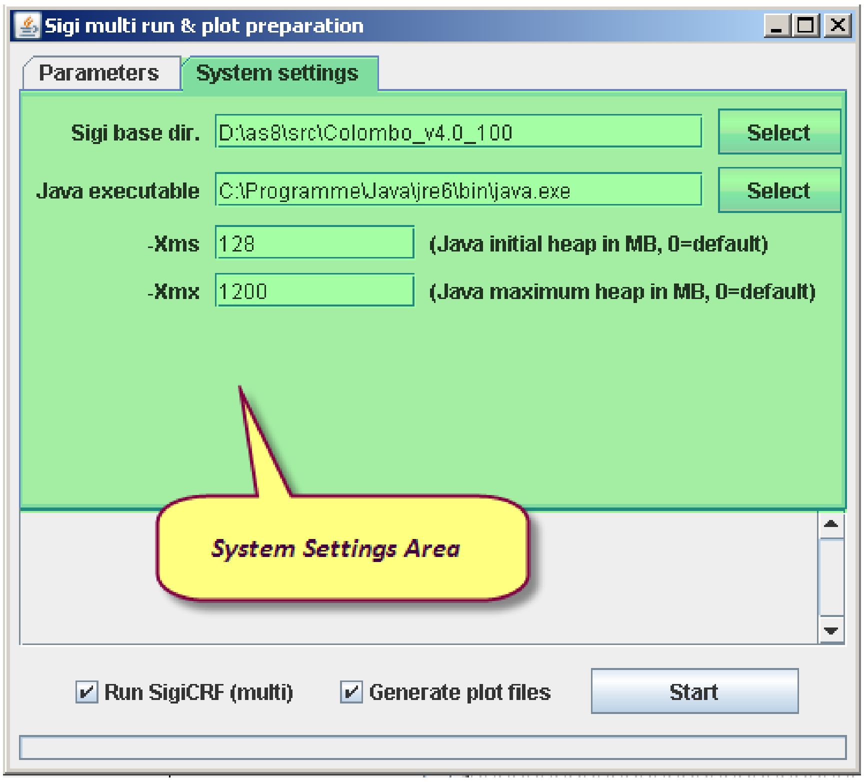
Details: In order to execute SIGI-CRF the SIGI-CRF-AUTO tool needs to know where Colombo (the runtime framework application of SIGI-CRF) has been installed (copied) to. It is required to select the *base directory* of Colomobo, i.e. the directory that contains the Colombo file structure. Within that directory you would find directories like 'libs' or 'plugins' (in which the actual plugin SIGI-CRF will be found).



core
external
features
GenericHMM
gui
history
io
libs
plugins
programs
tools
Colombo
Colombo.bat
Colombo.class
GenericBatch.class

Figure 7: Sample Colombo base directory

SIGI-CRF-AUTO checks some of the files or directories belonging to the Colombo/SIGI-CRF



Sigi base dir. D:\as8\src\Colombo_v4.0_100

Java executable C:\Programme\Java\jre6\bin\java.exe

-Xms 128 (Java initial heap in MB, 0=default)

-Xmx 1200 (Java maximum heap in MB, 0=default)

☑ Run SigiCRF (multi)   ☑ Generate plot files

Figure 6: System settings

distribution. If the directory doesn't contain those expected files or subdirectories, an error message will be shown in the message area. You will also get error messages if no directory is selected or the path does not exist.

Figure 7 shows the contents of a Colombo/SIGI-CRF base directory. You need to select the directory containing such files and directories.

### 6.1.2 System setting: Java executable

To select, find a/the Java VM directory on the machine, go to the bin subdirectory and select the Java application launcher tool called *java* (Unix, Linux, . . . ) or *java.exe* (Windows).

Details: Colombo is stand-alone Java application and will be started as such through an operating system call. SIGI-CRF-AUTO will use the Java application launcher tool for this. For Sun or Oracle Java virtual machines this file is contained in the bin directory of Java VM distributions like Java Runtime Environments (JRE).

A typical filename for Sun/Oracle Java runtime environments on an english Windows installation is `C:\program files\Java\jre6\bin\java.exe`

### 6.1.3 System setting: -Xms - Java initial heap

This parameter allows to adjust the heap size of each started SIGI-CRF run. If this paramter is set to zero, then the Java VM default size will be used, which is just fine for most cases.

Setting this value to a bigger number than the Java default may slightly increase the performance of SIGI-CRF in some rare cases.

The unit of this parameter is Megabytes of memory.

### 6.1.4 System setting: -Xmx - Java maximum heap

If you have more than one but less than four Gigabytes of memory it is recommended to set this parameter to 1024.

Details: This field allows SIGI-CRF to use more memory than the Java VM allows in standard configuration. Depending on the length of the sequence to be analysed as well as on some of the parameters it may be needed to increase this value to a larger value, otherwise SIGI-CRF may abort with a fatal memory error.

The Java VM default value is rather defensivly low. For common Sun/Oracle Java VMs like J2SE 5.0 or 6.0 the default value is smaller than 1/4th of the physical memory or 1GB at most.

The SIGI-CRF-AUTO parameter -Xmx counts in Megabytes of memory. So, to allow SIGI-CRF to use 1 Gigabyte of memory, this parameter needs to set to 1024.

Please note that the program only uses as much memory as it needs at a time. So even when setting this parameter higher, you may not see the larger amount of memory beeing actually consumed.

A value of zero let the Java VM use it's default value.

## 6.2 Parameters area

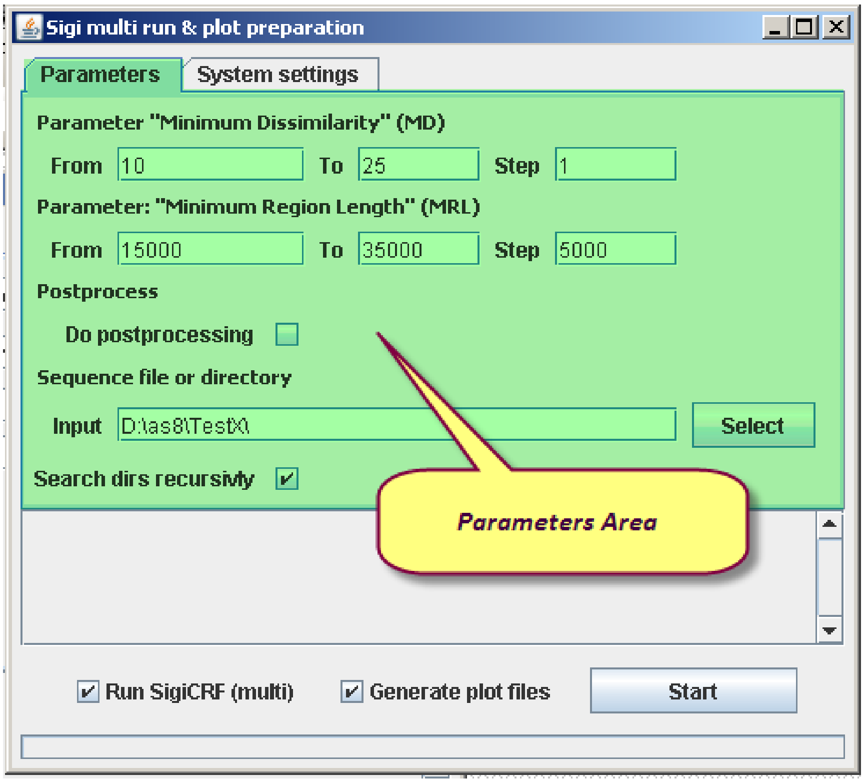
The parameters area contains all settings that are required for a custom graphic generation and can vary from run to run. All these parameters are found on the first tab page. In figure 8 the Parameters tab is highlighted green.

### 6.2.1 Parameters for Minimum Dissimilarity (MD)

As pointed out in section 1 the main purpose of SIGI-CRF-AUTO tool is to run SIGI-CRF with a number of parameter combinations and integrate all the results into one result graphic. *Minimum Dissimilarity (MD)* is one of the two main parameters of SIGI-CRF and MultiPlotTool shall try all requested values for MD. Section 1 also contains a short description for this parameter.

With the Parameters for Minimum Dissimilarity (MD) *From*, *To* and *Step* one controls the range of the MD parameters over all runs of SIGI-CRF. The three subparameter *From*, *To* and *Step* identify the range in a manner of a for-loop. Basically they mean: Run SIGI-CRF once for values between *From* and *To* and in each run instruct SIGI-CRF to use the current value as it's parameter MinimumDissimilarity (MD). The subparameter *Step* determines how far the individual values are apart. If *Step* equals one, then each value between *From* and *To* will be used, if *Step* equals two, then every second value between *From* and *To* will be used, and so forth.

Allowed values range from 1 to 40. Recommended are from 5 or 10 to 25.



Parameter "Minimum Dissimilarity" (MD)

From 10 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 35000 Step 5000

Postprocess

Do postprocessing □

Sequence file or directory

Input D:\as8\TestX\

Search dirs recursivly ☑

☑ Run SigiCRF (multi) ☑ Generate plot files

Figure 8: Parameters area

See following subsection 6.2.2 for relation with Parameters for Minimum Region Length (MRL).

### 6.2.2 Parameters for Minimum Region Length (MRL)

*Minimum Region Length (MRL)* is the second main parameter of SIGI-CRF and MultiPlotTool shall try all requested values for MRL as well. Compare section 1 for a short description of this parameter.

With the Parameters for Minimum Region Length (MRL) *From*, *To* and *Step* one control's the range of the MRL parameters over all runs of SIGI-CRF. The three subparameter *From*, *To* and *Step* identify the range in a manner of a for-loop.

As SIGI-CRF is segmenting input sequences into windows of 5k length, all parameters need to be multiples of 5000. As an example, if you have *From*=15000, *To*=35000 and *Step*=5000 then SIGI-CRF will be run with different values of 15000, 20000, 25000, 30000 and 35000.

Allowed values range from 5000 to 40000. Recommended are values of 15000 and more.

Please note that parameters for Minimum Dissimilarity (MD) and parameters for Minimum Region Length (MRL) form a *two-dimensional* loop, i.e. every possible combination of their values will be tried.

### 6.2.3 Parameter: Do Postprocessing

Optionally SIGI-CRF can try to fine-tune the borders of predicted genomic islands. This is done as a postprocessing step after the main Conditional Random Field algorithmn. Without enabling the postprocessing step, start and length of predicted genomic islands will always be multiples of 5000.

### 6.2.4 Parameters: Input sequence file or directory / Search dirs recursivly

The obvious option for selecting input sequence data is to just select a sequence's file. If selecting a single sequence file, SIGI-CRF-AUTO will of course handle exactly that sequence.

But SIGI-CRF-AUTO has another option for selecting input files: Instead of selecting a single sequence file, it is also possible to select a directory containing one or more sequence files. If more than one sequence file is found in a selected directory, SIGI-CRF-AUTO will handle one after the other.

If the option *Search dirs recursivly* is selected, then SIGI-CRF-AUTO also searches all subdirectories of a directory for sequence files. This option allows to put each sequence file in it's own subdirectory and still be able to process all of them with only one start of SIGI-CRF-AUTO.

See table 1 for a list of known file formats and their accepted file suffixes.

| File format | Filename suffix | Restrictions |
|---|---|---|
| Genbank file | .gbk | |
| EMBL file | .embl | |
| FASTA file | .fasta | No support for multi FASTA |

Table 1: Accepted file formats

All GFF files generated by the runs of SIGI-CRF and the resulting graphic .R file will be written to the same directory as the sequence file and will share the same basename with it. See section 8 for details.

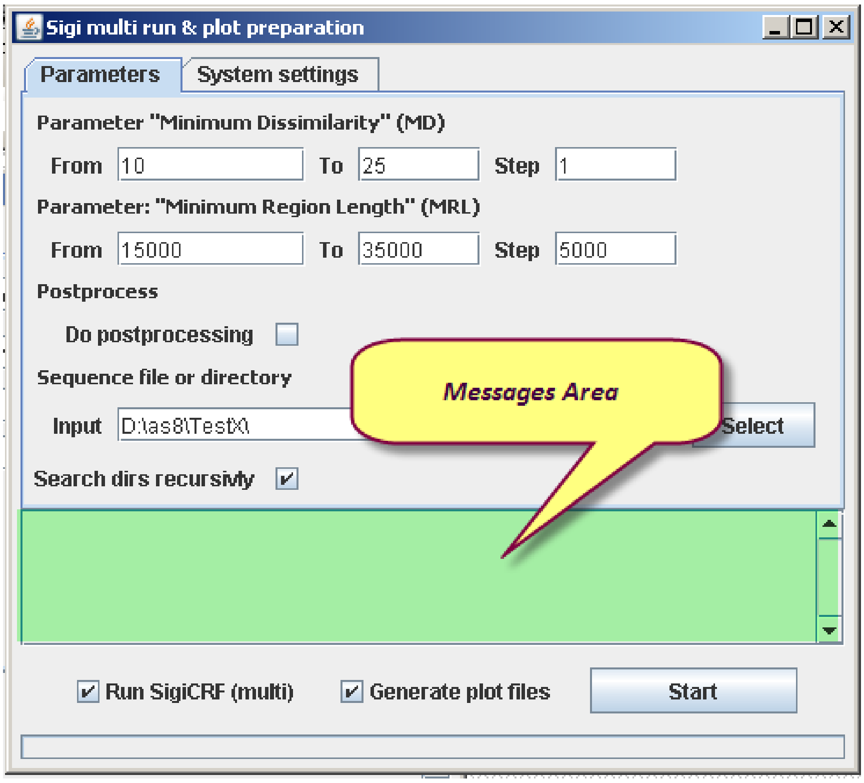
## 6.3 Messages Area and indicating parameter issues

The message area has two purposes - the first is to support entering input parameters by indicating any parameter errors or conflicts.

In figure 9 the Message Area is highlighted green.

Until plausible values are entered into the input fields for parameters or system settings, any detected problems will be written to the message area. All error messages are written in *red*.

Figure 10 shows the Message Area indicating errors or conflicts with parameter settings.



Parameter "Minimum Dissimilarity" (MD)

From 10 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 35000 Step 5000

Postprocess

Do postprocessing □

Sequence file or directory

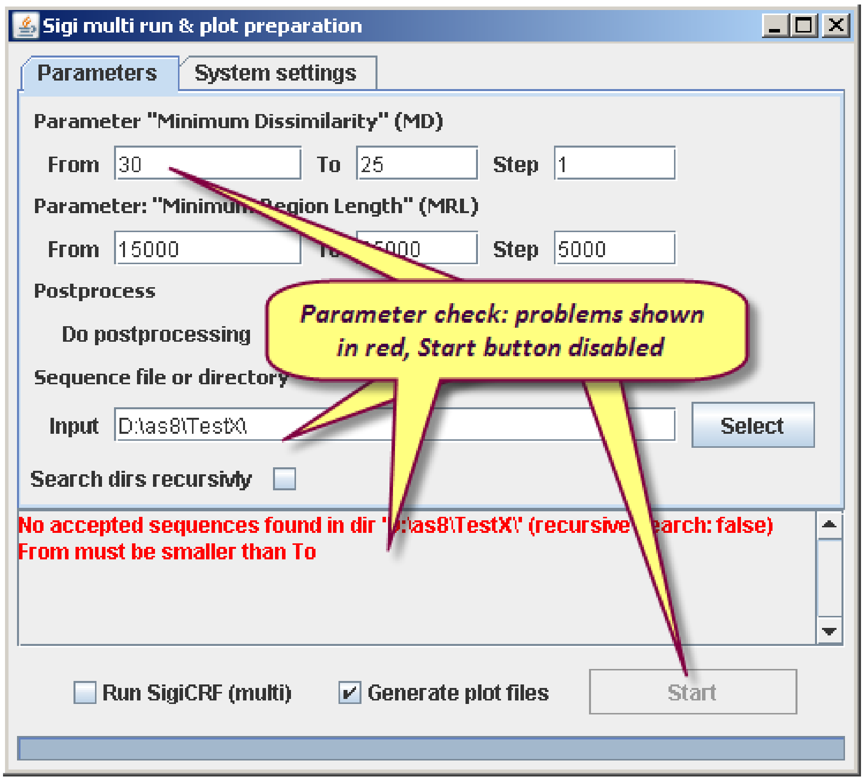
Input D:\as8\TestX\

Search dirs recursivly ☑

☑ Run SigiCRF (multi) ☑ Generate plot files

Messages Area

Figure 9: Message Area



Parameter "Minimum Dissimilarity" (MD)

From 30 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 5000 Step 5000

Postprocess

Do postprocessing

Sequence file or directory

Input D:\as8\TestX\

Search dirs recursivly □

No accepted sequences found in dir ':\as8\TestX\' (recursive earch: false)
From must be smaller than To

□ Run SigiCRF (multi) ☑ Generate plot files

Parameter check: problems shown in red, Start button disabled

Figure 10: Checking parameters

### 6.4 Starting the program and the Start area

The purpose of SIGI-CRF-AUTO tool is to run SIGI-CRF with several variations of parameters and create some .R script files to generate overview graphics.

The straightforward way to start this is to just have both checkboxes *Run SIGI-CRF (multi)* and *Generate plot files* selected and click start. If both tasks are selected then all GFF files will be created and used for the graphic - just as one would normally expect and sufficient for normal use.

See figure 11 for a screenshot of SIGI-CRF-AUTO with the Start Area highlighted green.

But there's also an advanced way of using this tool that some users may find useful: With selecting only one of both checkboxes at a time the Start Area has the option to seperate the runs of SIGI-CRF from the generation of the graphic when starting. And it's fine for graphic generation to use only subsets of the parameter combinations used for creating the SIGI-CRF output GFF files.

While generating the .R files from the SIGI-CRF GFF output files is a quick task, running SIGI-CRF can take hours of time - the time depends to a some degree on the machine performance but especially it depends on the number of parameter combinations run. Therefore SIGI-CRF-AUTO tool explicitly allows to seperate the time-intensive process of running SIGI-CRF from the graphics generation.

With this in mind, some users may find it helpful to be able to first create a big number of SIGI-CRF runs with many different parameter combinations and then create a couple of graphics interactivly, with each graphic using only a subset of the previously created GFF files.

For this approach the Start Area contains not only the Start button, but also two checkboxes to select either or both steps.

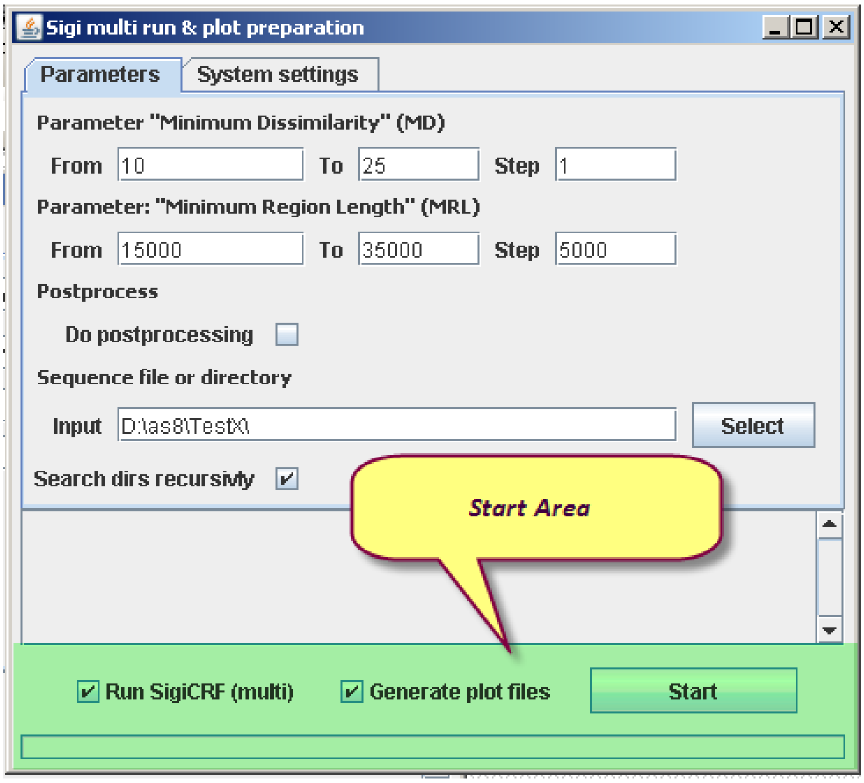
If only the first checkbox 'Run SIGI-CRF (multi)' is selected and 'Start' is beeing clicked, then the program starts with running SIGI-CRF with all selected parameters and generates the resulting GFF files only. (.R file is not generated if second checkbox is not selected).

If only the second checkbox 'Generate plot files' is selected and 'Start' is beeing clicked, then the program expects all GFF files beeing generated beforehand and starts with collecting them to generate the graphic R script file. (SIGI-CRF is not run thereby no GFF files are beeing generated if first checkbox is not selected). Attention: Generating .R script files will not fail, if some of the expected GFF files cannot be found - the task only uses all matching *existing* GFF files. The bottom line to this is that if generation of GFF files is seperated from .R graphics script file generation, then the user needs to pay special attention to enter appropriate parameters to each task and should validate the resulting graphic to meet his expectations.

## 7 Graphics as R script files

The file format that SIGI-CRF-AUTO creates is not directly a graphic file format. To be specific, what the program does create is a script file for the statistical programming environment R. But getting the graphic from the R script is quick and painless.

To cite the R project homepage (http://www.r-project.org/): *R is a free software environment*



Sigi multi run & plot preparation

Parameters | System settings

Parameter "Minimum Dissimilarity" (MD)

From 10 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 35000 Step 5000

Postprocess

Do postprocessing □

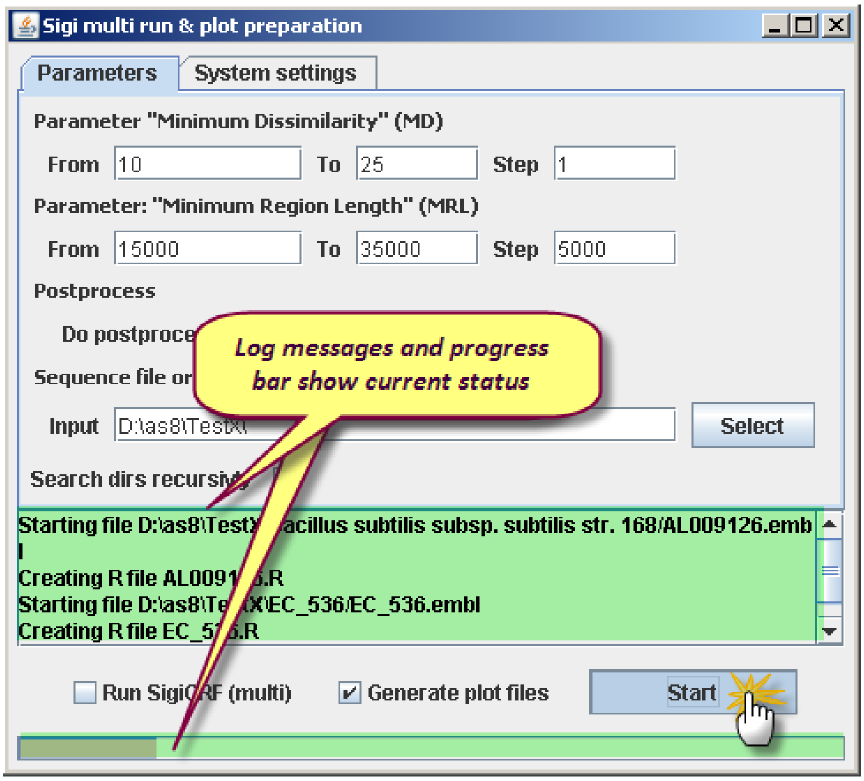
Sequence file or directory

Input D:\as8\TestX\ Select

Search dirs recursivly ☑

☑ Run SigiCRF (multi) ☑ Generate plot files Start

Figure 11: Start area



Sigi multi run & plot preparation

Parameters | System settings

Parameter "Minimum Dissimilarity" (MD)

From 10 To 25 Step 1

Parameter: "Minimum Region Length" (MRL)

From 15000 To 35000 Step 5000

Postprocess

Do postproce

Sequence file or

Input D:\as8\TestX\ Select

Search dirs recursivly

```
Starting file D:\as8\TestX\Bacillus subtilis subsp. subtilis str. 168/AL009126.emb
l
Creating R file AL009126.R
Starting file D:\as8\TestX\EC_536/EC_536.embl
Creating R file EC_536.R
```

□ Run SigiCRF (multi) ☑ Generate plot files Start

Figure 12: Showing processing progress

*for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.*

For Windows R can be installed by means of a simple setup program, which can be easily found on the project homepage.

Once R is installed all that needs to be done to see the graphic is to open the produced script file from the R console or just copy and paste the file’s contents to the R console.

The R graphic can be exported to various common graphic file formats like for example .eps or .png.

# 8 File structure and generated files

The task of running SIGI-CRF with various parameter combinations creates GFF files as intermediate outputs. These GFF files will be collected by the second task of SIGI-CRF-AUTO to produce the final .R file.

The user selects either a sequence file or a (parent) directory containing sequences. The program then either uses the given sequence file or searches the given directory for files. When more than one sequence will be found or used, a top level loop will handle each sequence file seperately.

For each sequence file processed, the directory and names used to write either intermediate GFF or final R files will depend on the sequence file.

All files will be written to the same directoy as the originating sequence file. Further, all intermediate or final output files will use the basename of the sequence file as part of the filename.

For the GFF files generated by the ’Run SIGI-CRF (multi)’ task, the filenames will be build from the sequence file basename, plus some parameter identification and the file suffix ’.gff’. The parameters will be seperated by the tilde sign (‘ ’). For example from the sequence file ’AE017221.embl’, the following or other files could be generated (depending on parameters selected):

AE017221˜mrl35000˜md10˜.gff’
AE017221˜mrl15000˜md10˜.gff’
AE017221˜mrl20000˜md15˜.gff’

The resulting graphic R script file would get the name ’AE017221.R’

# 9 Integrating reference data

## 9.1 Integration of reference GI information (e.g. gold standard)

SIGI-CRF-AUTO tool allows to indicate special reference information within the graphic in a special background manner - pervasive but still low-key. This can be used to compare SigiCRF predictions to known and verified annotations, for example for gold standard GI regions, annotated highly expressed genes etc. Such standard regions will be visualized as solid background bars spanning all prediction rows. See figure 1 for an example showing this feature.

SIGI-CRF-AUTO will allow for up to three special reference data sets to be plotted in different colors. They are refered to as gold standard, silver standard and bronce standard respectivly.

To use this feature create up to three GFF files listing the regions to be indicated. Name them using the basename of the sequence and append each with one of the following three suffixes:

- '~gold.gff'
- '~silver.gff'
- '~bronce.gff'

Put the file or files to the same directory as the sequence.

As an example: For a sequence file 'AE017221.embl' the name for the gold standard file should be 'AE017221~gold.gff'.

For each graphic to be generated SIGI-CRF-AUTO will look for files of such names in the sequence's directory. If one or more of such files can be found then SIGI-CRF-AUTO will read all their entries. From each entry's Start and End position a vertical background bar will be created and drawn all the way from top to bottom, i.e. as background underneath all rows showing prediction results.

'Gold' information will be read and plotted first, so it will build the bottommost graphical elements. 'Bronce' will be read and drawn last of these three, so it will be the topmost of these special standard information graphical elements (but still below are normal predictions).

Please note that SIGI-CRF-AUTO will not apply any row filtering to the GFF file and will ignore any other column contents other than Start and End positions.

## 9.2 Integration of IslandViewer results

What is IslandViewer? To cite it's web site[2]: *An integrated interface for computational identification and visualization of genomic islands. IslandViewer is a computational tool that integrates three different genomic island prediction methods: IslandPick, IslandPath-DIMOB, and SIGI-HMM.*

SIGI-CRF-AUTO tool has also the option to integrate predictions downloaded from the IslandViewer's web site. At the time of writing IslandViewer offers different predictions from different tools for each organism on a dedicated web page. These web pages include not only a circular graphic and a tabular presentation of all predictions but also links to download the predictions. It's exactly the TAB download format that SIGI-CRF-AUTO tool can process.

Please see figure 13 for a partial screenshot of an IslandViewer's organism's web page with an arrow pointing to the TAB file download link.

If you can download a TAB file for the sequence to be analysed with SIGI-CRF-AUTO tool, then save it to the sequence dir and SIGI-CRF-AUTO tool will also automatically integrate these predictions into the graphic.

Such TAB files will use filenames like `islandviewer_NC_005835.1.txt`. SIGI-CRF-AUTO tool

[2] `http://www.pathogenomics.sfu.ca/islandviewer/query.php`

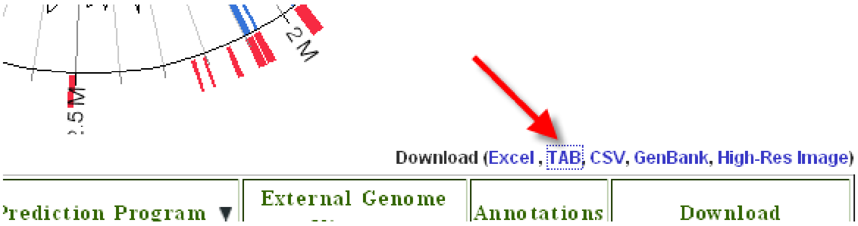


Figure 13: Island Viewer TAB download link

will recognize and use any file whose name starts with 'islandviewer' and ends with '.txt'. Please use at most one such file per directory and have it the correct one for the sequence in that directory.