The SciRoKo 3.1 Manual

1. General Information

SciRoKo is free for use. If you intend to use SciRoKo commercially or you suffer from a surplus budget, a small donation of e.g.: 100$ or 100€ will not be rejected. All donated funds will be used for surviving a monetary ungrateful PhD-thesis or buying new hardware. Please use the following account:
IBAN: AT733439000000012427
BIC (SWIFT): RZOOAT2L390

Please cite:
SciRoKo: A new tool for whole genome microsatellite search and investigation
Robert Kofler; Christian Schlotterer; Tamas Lelley
Bioinformatics 2007; doi: 10.1093/bioinformatics/btm157

SciRoKo is entirely written in C# and primarily designed for the use with Windows. Due to the tireless efforts of the people from the Mono-Project, SciRoKo also works in Linux, MacOS X, Free BSD, UNIX, and Solaris. See the Mono-Project: http://www.mono-project.com/Main_Page

SciRoKo 3.1 is the first SciRoKo-release employing modern software design patterns such as MVC (model-view-control), Template-method, Strategy and Abstract Factory. This should ensure easy maintainability and bug-fixing. Since SciRoKo 3.1 is probably the last major release containing new features our focus is shifting from new features towards ease of maintenance.

2. Installation

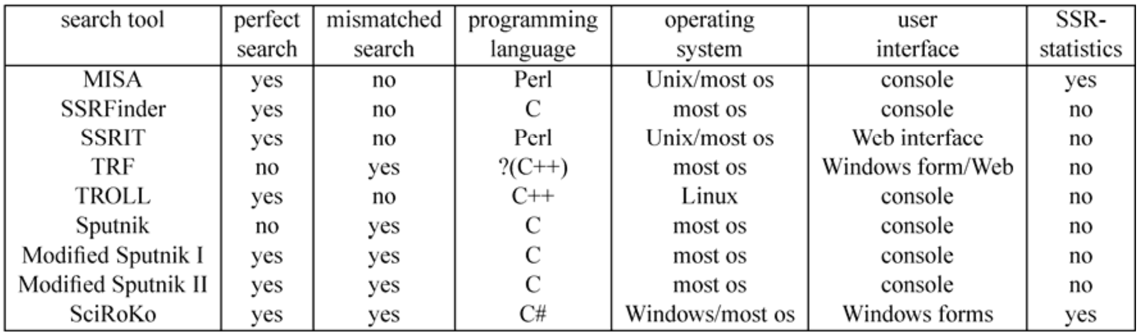
1. Download the SciRoKo.zip from www.kofler.or.at/Bioinformatics/
   The SciRoKo executable is platform independent and can be used with all operating systems.
2. Unpack the archive in a folder of your choice
3. Make sure that the Microsoft .net Framework 2.0 is installed, if not ask your computer administrator or download it yourself from the vendors homepage.
   Alternatively the mono-project can be used, this approach is necessary for operating systems other than Microsoft Windows.
4. Just double click on SciRoKo.exe

Links:
-Mono-Project software: http://www.mono-project.com/Downloads
-Microsoft .net Framework 2.0:
http://www.microsoft.com/downloads/details.aspx?FamilyID=0856EACB-4362-4B0D-8EDD-AAB15C5E04F5&displaylang=en
-General information about .net and C#: http://en.wikipedia.org/wiki/.net_framework

3. Introduction

Several tools for the analysis of microsatellites already exist. (See Table 1)

Table 1. Common programs and tools for SSR search: summary of the features and properties



search tool	perfect search	mismatched search	programming language	operating system	user interface	SSR-statistics
MISA	yes	no	Perl	Unix/most os	console	yes
SSRFinder	yes	no	C	most os	console	no
SSRIT	yes	no	Perl	Unix/most os	Web interface	no
TRF	no	yes	?(C++)	most os	Windows form/Web	no
TROLL	yes	no	C++	Linux	console	no
Sputnik	no	yes	C	most os	console	no
Modified Sputnik I	yes	yes	C	most os	console	no
Modified Sputnik II	yes	yes	C	most os	console	no
SciRoKo	yes	yes	C#	Windows/most os	Windows forms	yes

In our opinion this tools have the following drawbacks: (i) except SSRIT and TRF this tools are not user friendly because they require console input or Unix/Linux as operating systems;
(ii) many of them allow only search for perfect SSRs, although an imperfection within an SSR does not necessarily mean the end of this SSR. It may continue beyond this imperfection, moreover, the imperfection may be caused by sequencing error;
(iii) Sputnik, Modified Sputnik I-II and TRF are very slow ;
(iv) Sputnik, Modified Sputnik I-II do not allow SSR search for hexanucleotide motifs (TRF only with difficulties); (v) TRF allows parameter adjustment only within a limited range and produces a cornucopia of inconvenient output files;
(vi) Finally, none of the listed tools allows detailed subsequent statistical analysis of the search results.

3.1 How the SciRoKo SSR-search algorithm works in detail:

The SciRoKo SSR-search module offers four search modes, two for perfect and two for mismatched SSR search. In the two perfect SSR search modes each nucleotide at position i is tested for identity with the nucleotide at position i+t, where t is the motif length (1-6). Upon identity i is increased i=i+1 until no further identity can be found. If this SSR meets the minimum requirements as specified in ``Minimum repeats" or ``Minimum total length", depending on which one of the two perfect search modes is used, the microsatellite is directly reported into the output file. The score of perfect microsatellites equals their total length.

In the two mismatched SSR search modes, perfect SSRs (SSR-seeds) act as origin for subsequent 5' and 3' extensions. The minimum requirements for the SSR-seeds can be set as low as 2 repeats or 3 nucleotides. Whether a microsatellite is finally reported into the output file depends only on the achieved score. The SciRoKo scores are calculated according to the first two equations:

1. Score_Fixed_Penalty = hits – mmP * mm
2. Score_Variable_Penalty = hits – mm * (m_L * mmP)
3. Score_Sputnki = hits - m_L – mmP * mm

The parameters are: hits (matches with the virtual perfect microsatellite; see below), number of mismatches (mm), mismatch penalty (mmP) and the length of the SSR motif (m_L). Equation 1 is used in the ``Fixed mismatch penalty" mode, equation 2 in the ``Variable mismatch penalty" mode and equation 3 is used in the Sputnik family SSR search tools.
If the score of an imperfect SSR achieves the ``Required score", the SSR is reported into the output file.

The process of SSR-extension used in the two mismatched SSR search modes progresses in major loops and mismatch permutations (See Figure 1)

Initially the SSR-seed is set as highscore SSR. The highscore SSR acts as origin for the 5' and 3' major loops. Within a major loop a number of mismatch permutations (see below) is created. The mismatch permutation achieving the highest score is called permutation highscore. If the permutation highscore is equal or higher than the highscore, the permutation highscore is set as the new highscore and acts as origin of the next major loop. The SSR-seed is at first 5' extended with major loops until the permutation highscore is lower than the highscore and then 3'. Three types of mismatches can be found in an SSR: deletions, base substitutions and insertions. Within a major loop a recursion creates for a given ``Max mismatches at once'' (mmao) all possible combinations of mismatches, allowing for perfect microsatellite stretches between the mismatches: $3^{mmao}+3^{mmao-1}+\ldots+3^{1}$

The recursion is aborted premature if the end of the file or the end of a previous SSR has been reached. Branchings within the recursion only occur at mismatch sites. For the mismatched SSR-search SciRoKo creates a virtual perfect microsatellite (vpm) from the SSR-seed motif, starting at the first position of the SSR-seed (Figure 1). The vpm continues indefinitely in the 5' and 3' direction and acts as template for comparisons with the DNA sequence. Initially the position pointer moves one position from the SSR-start or SSR-end to the 5' or 3' direction for 5' or 3' extension respectively.

Subsequently, the position pointer compares each nucleotide in the DNA sequence with the corresponding nucleotide in the vpm. Each of the three mismatch types has an own distinct pattern (Figure 1). In a major loop all possible mismatch permutations, i.e. combination of mismatch patterns, are tested and the mismatch permutation achieving the highest score is set as the permutation highscore.

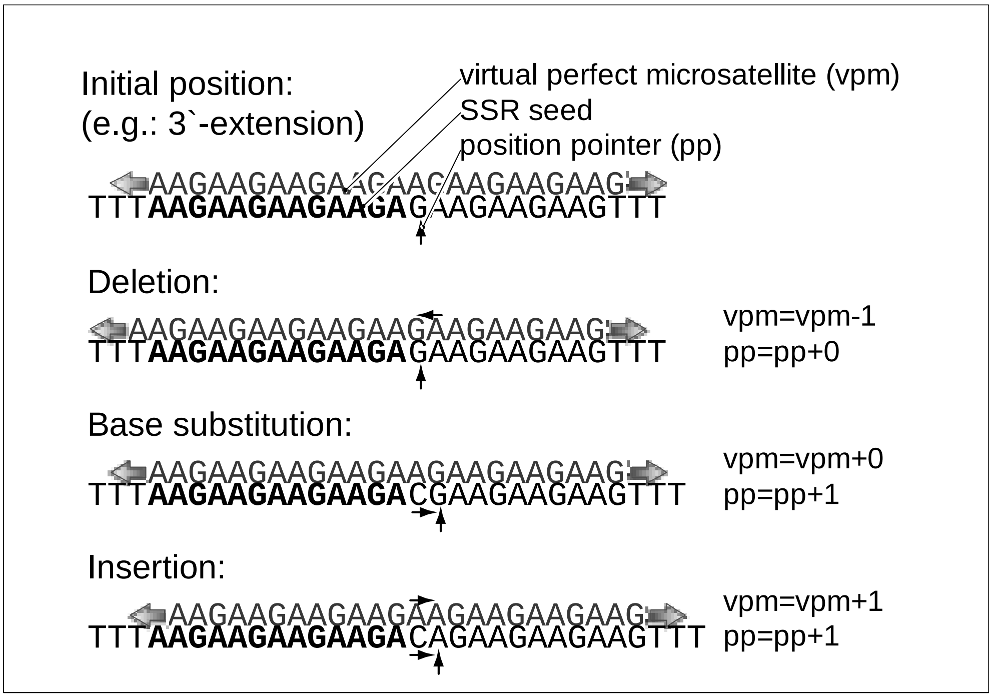
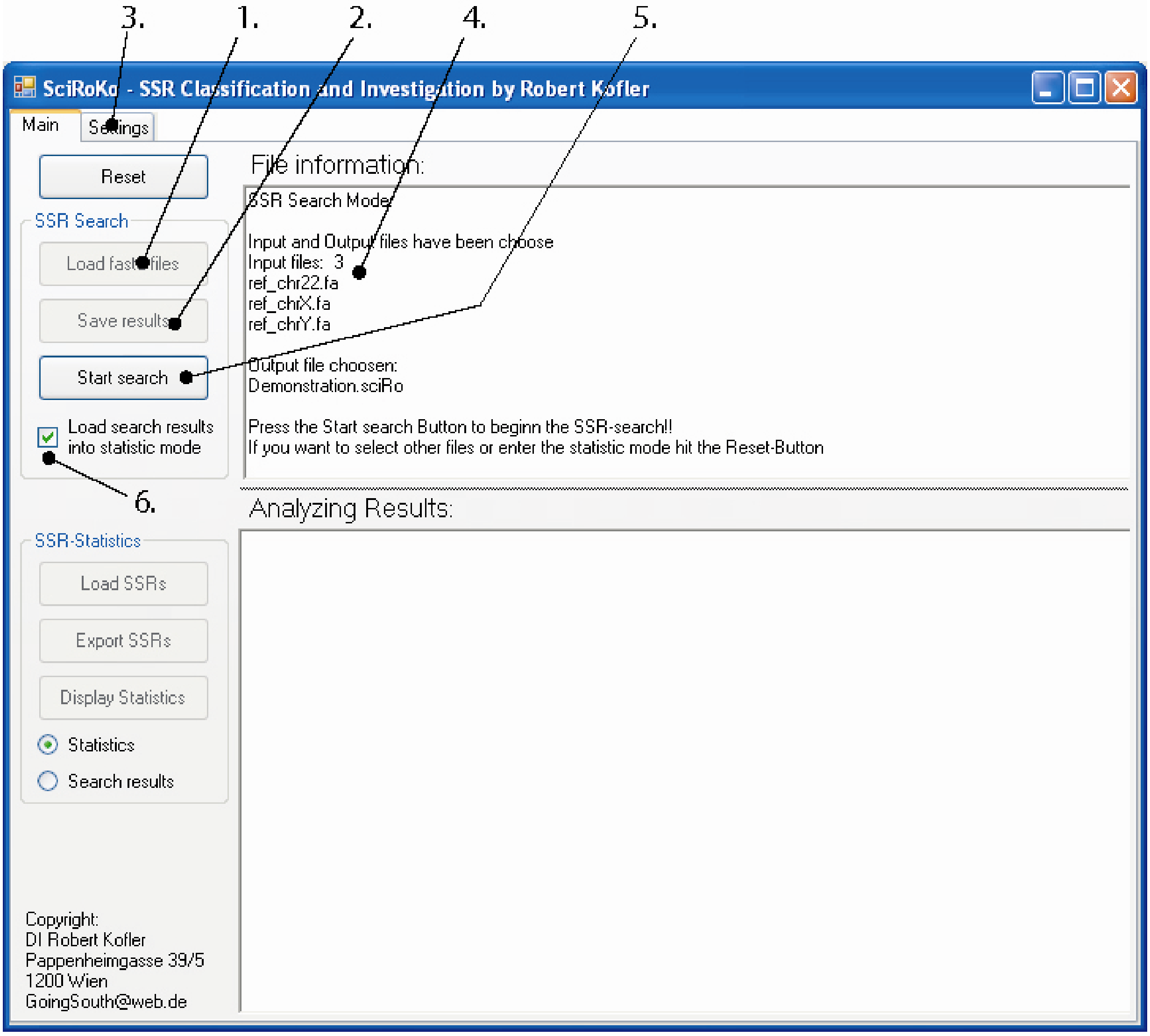


Fig. 1: Pattern of mismatches at a recursion branching site during 3'-extension. For the identification of a deletion the virtual perfect microsatellite (vpm) is moved one bp to the 5'-direction. For a base substitution the position pointer is moved one bp to the 3'-direction and for a insertion the vpm and the position pointer are moved one bp to the 3'-direction.

4. The SciRoKo SSR-search Module

The SciRoKo main menu. To perform SSR search follow the steps in numerical order.



SSR Search Mode

Input and Output files have been choose
Input files: 3
ref_chr22.fa
ref_chrX.fa
ref_chrY.fa

Output file choosen:
Demonstration.sciRo

Press the Start search Button to beginn the SSR-search!!
If you want to select other files or enter the statistic mode hit the Reset-Button

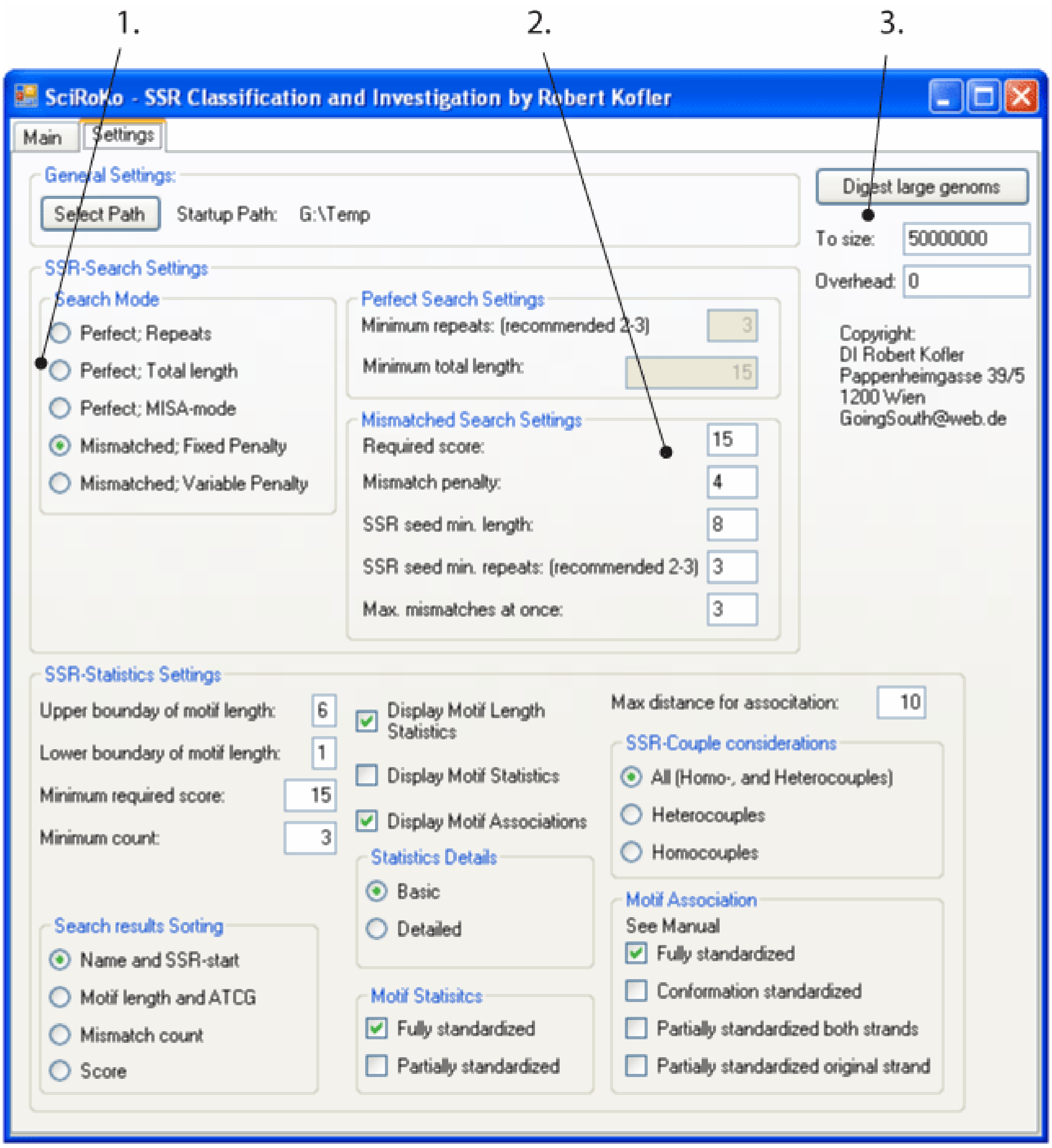
Copyright:
DI Robert Kofler
Pappenheimgasse 39/5
1200 Wien
GoingSouth@web.de

Fig. 2.: SciRoKo 2.1 main menu

1. Choose an input file. SciRoKo accepts all fasta files with the extensions *.fa ; *.fasta; *.txt ;
   A single fasta file might contain multiple fasta sequences separated through the character '>'. Additionaly SciRoKo accepts multiple fasta files at once. For instance, the whole rice genome can be investigated at once, with each chromosome representing a single file or with all chromosomes copied into a single file.
2. Choose the output file for the SSR search. Two file types are supported as output formats. The SciRoKo format (*.sciRo) and the Tab delimited format (*.td). If exporting of the SSR-search results into the sputnik-family file formats is required, the SSR-search results have to be loaded into the SSR-statistics module prior to exporting (See Chapter 5).
3. Adjust the settings used for SSR-search (see below)
4. Make sure the appropriate files have been chosen for input and output. Press the Reset-Button to choose different files.

5. Press the Start button to start SSR search
6. When this box is checked the SSR-search results are directly loaded into the SSR-statistics module. Even when checked, the SSR-search results are first reported into the chosen output file.

4.1 Adjust the SSR-search settings:


1.
2.
3.
SciRoKo - SSR Classification and Investigation by Robert Kofler
Main
Settings
Select Path
Digest large genoms

General Settings:

Startup Path: G:\Temp

To size: 50000000

Overhead: 0

Copyright:
DI Robert Kofler
Pappenheimgasse 39/5
1200 Wien
GoingSouth@web.de

SSR-Search Settings

Search Mode

○ Perfect; Repeats
○ Perfect; Total length
○ Perfect; MISA-mode
● Mismatched; Fixed Penalty
○ Mismatched; Variable Penalty

Perfect Search Settings

Minimum repeats: (recommended 2-3) 3

Minimum total length: 15

Mismatched Search Settings

Required score: 15

Mismatch penalty: 4

SSR seed min. length: 8

SSR seed min. repeats: (recommended 2-3) 3

Max. mismatches at once: 3

SSR-Statistics Settings

Upper boundary of motif length: 6

Lower boundary of motif length: 1

Minimum required score: 15

Minimum count: 3

☑ Display Motif Length Statistics
☐ Display Motif Statistics
☑ Display Motif Associations

Max distance for association: 10

SSR-Couple considerations

● All (Homo-, and Heterocouples)
○ Heterocouples
○ Homocouples

Statistics Details

● Basic
○ Detailed

Search results Sorting

● Name and SSR-start
○ Motif length and ATCG
○ Mismatch count
○ Score

Motif Statisitcs

☑ Fully standardized
☐ Partially standardized

Motif Association

See Manual

☑ Fully standardized
☐ Conformation standardized
☐ Partially standardized both strands
☐ Partially standardized original strand

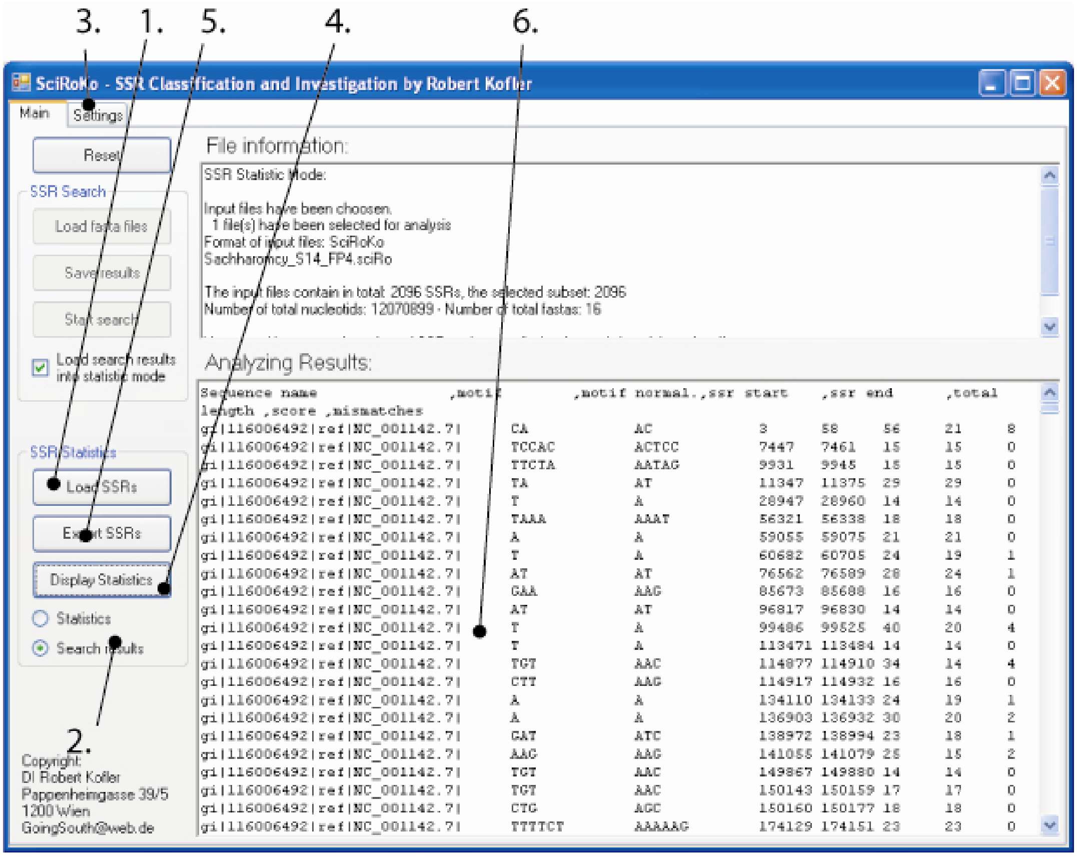
1. First choose the SSR-search mode. SciRoKo offers three modes for perfect SSR search, one for SSR-search according the total length of a microsatellite and two for perfect SSR-search according the number of repeats. The “Perfect; MISA-mode” requires an input string of the form *mono-di-tri-tetra-penta-hexa*. For instance, the input string 12-6-7-5-5-5 states that a trinucleotide microsatellite has to have at least 7 repeats.
Additionally SciRoKo provides two mismatched SSR search modes, one using a fixed mismatch penalty and one using a variable mismatch penalty.

2. Adjust the settings used for SSR search such as total length or required repeat number in the perfect SSR search modes.
   When using the mismatched SSR search modes adjust the mismatch penalty, the required score, the requirements for the SSR-seed and the maximum number of mismatches allowed in a row (max. mismatches at once is equivalent to the depth of the recursion). An SSR-seed is each perfect microsatellite meeting the specified requirements like minimum repeats or minimum total length. The lower boundary for the SSR-seed settings is: 2 repeats and a minimum length of 3.
3. Large genomes, like the human, have chromosomes larger than 200 Mbp. Unfortunately 200 Mbp are to large for ad hoc analysis with SciRoKo 3.1, SciRoKo accepts fasta files with a size up to 50 Mbp. It is therefore necessary to digest large genomes into smaller chunks of the chosen size. We recommend using a chunk size of 50 Mbp with no overhead specified.
   Once pre-treated SciRoKo analysis the whole human genome in 460 seconds.

   The pre-treated chromosome chunk files are stored in a subfolder, file names and file number are kept identical.

5. The SciRoKo SSR-Statistics Module

5.1 Display the SSR-search results:



```
SSR Statistic Mode:

Input files have been choosen.
  1 file(s) have been selected for analysis
Format of input files: SciRoKo
Sachharomcy_S14_FP4.sciRo

The input files contain in total: 2096 SSRs, the selected subset: 2096
Number of total nucleotids: 12070899 - Number of total fastas: 16
```

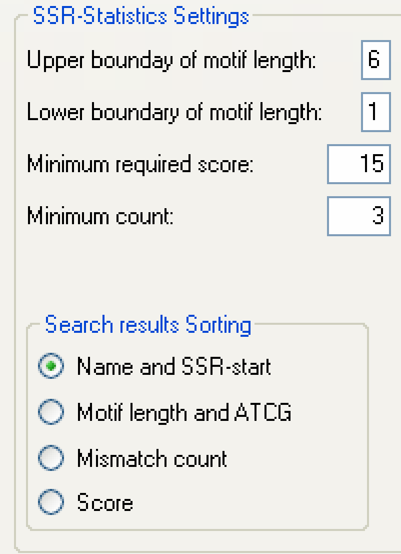
```
Sequence name                   ,motif          ,motif normal.,ssr start    ,ssr end       ,total
length ,score ,mismatches
gi|116006492|ref|NC_001142.7|      CA           AC            3      58     56      21      8
gi|116006492|ref|NC_001142.7|      TCCAC        ACTCC         7447   7461   15      15      0
gi|116006492|ref|NC_001142.7|      TTCTA        AATAG         9931   9945   15      15      0
gi|116006492|ref|NC_001142.7|      TA           AT            11347  11375  29      29      0
gi|116006492|ref|NC_001142.7|      T            A             28947  28960  14      14      0
gi|116006492|ref|NC_001142.7|      TAAA         AAAT          56321  56338  18      18      0
gi|116006492|ref|NC_001142.7|      A            A             59055  59075  21      21      0
gi|116006492|ref|NC_001142.7|      T            A             60682  60705  24      19      1
gi|116006492|ref|NC_001142.7|      AT           AT            76562  76589  28      24      1
gi|116006492|ref|NC_001142.7|      GAA          AAG           85673  85688  16      16      0
gi|116006492|ref|NC_001142.7|      AT           AT            96817  96830  14      14      0
gi|116006492|ref|NC_001142.7|      T            A             99486  99525  40      20      4
gi|116006492|ref|NC_001142.7|      T            A             113471 113484 14      14      0
gi|116006492|ref|NC_001142.7|      TGT          AAC           114877 114910 34      14      4
gi|116006492|ref|NC_001142.7|      CTT          AAG           114917 114932 16      16      0
gi|116006492|ref|NC_001142.7|      A            A             134110 134133 24      19      1
gi|116006492|ref|NC_001142.7|      A            A             136903 136932 30      20      2
gi|116006492|ref|NC_001142.7|      GAT          ATC           138972 138994 23      18      1
gi|116006492|ref|NC_001142.7|      AAG          AAG           141055 141079 25      15      2
gi|116006492|ref|NC_001142.7|      TGT          AAC           149867 149880 14      14      0
gi|116006492|ref|NC_001142.7|      TGT          AAC           150143 150159 17      17      0
gi|116006492|ref|NC_001142.7|      CTG          AGC           150160 150177 18      18      0
gi|116006492|ref|NC_001142.7|      TTTTCT       AAAAAG        174129 174151 23      23      0
```

Copyright:
DI Robert Kofler
Pappenheimgasse 39/5
1200 Wien
GoingSouth@web.de

1. Choose the input files for the SSR-statistics module. Although multiple input files can be selected it is not recommended, because this might cause problems with the total number of nucleotides or files
   Important note: Alter the file-extensions of the Sputnik, Modified Sputnik I-II files prior to use with the SciRoKo SSR-statistics module. The following list indicates the required file extensions.

Sputnik	*.sput
Modified Sputnik I	*.m1sput
Modified Sputnik II	*.m2sput
SSR-Couples	*.ssrCou

2. Select the display search results radio button
3. Adjust the settings:



SSR-Statistics Settings

Upper bounday of motif length: 6

Lower boundary of motif length: 1

Minimum required score: 15

Minimum count: 3

Search results Sorting

● Name and SSR-start

○ Motif length and ATCG

○ Mismatch count

○ Score

A lower and an upper boundary, as well as a minimum required score might be specified. Only microsatellites meeting these criteria will be displayed.

For instance:
Lower boundary: 2 (dinucleotide SSRs)
Upper boundary: 5 (pentanucleotide SSRs)
Minimum required score: 15
-> Only dinocleotide to pentanucleotide microsatellites having a length of at least 15 basepairs will be displayed. (The score is equivalent to the total length of a microsatellite, only imperfect microsatellites have to be even longer)

a. The SSR search results might be sorted according the name (ID) of the fasta sequences and if two SSRs have an equal sequence name according the SSR-start position. This sorting option is also used within the algorithm identifying SSR-Couples;
b. The microsatellites can also be sorted according the motif length and if two SSRs have an equal motif length in descending valuation ATCG. This sorting is used within the two MotifMatrices (see below). Descending valuation ATCG means that the variation of a microsatellite motif having the most A-nucleotides at the beginning is displayed first (e.g.: AAT instead of TAA) then the variation with the most T-nucleotides (e.g.: ATAC instead of ACAT) and so on.

c. The microsatellites might be sorted according the number of mismatches. For instance, this feature allows the identification of the microsatellites containing the most mismatches in the whole human genome.

d. The SSRs might be sorted according their score, allowing identification of the “high scoring” microsatellite.

4. Hit the Display statistics button to display the search results
5. The subset of the microsatellites (within the specified constrains) might also be exported using the selected sort-mode.
6. Hit the Display-Button; Displays the SSR-search results.

5.2 Microsatellite statistics:

SciRoKo 3.1 generates three different statistic outputs; Motif length infos, Motif infos, and Motif association statistics (compound microsatellites).
To allow categorization and statistical analysis of the identified microsatellites, SciRoKo standardizes the microsatellites in two intensities: **Full and partial standardization**.

Only the Motif association statistic requires the full spectrum of the microsatellite motif standardizations used in SciRoKo. The “Motif length info” only requires standardization of the microsatellite motif lengths (e.g.: all trinucleotid microsatellites are grouped into the same category).
During the standardization process microsatellites with similar motifs are grouped together. For instance the microatellite motifs “AG” and “GA” become identical during the process of partial standardization, yielding the partially standardized motif “AG”.
During full standardization, the reverse complements of microsatellite motifs also have to be considered. For instance, full standardization groups the microsatellite motifs “TC”, “CT”, “AG”, and “GA” together into one group (“AG”).

To allow the standardization of the microsatellite motifs SciRoKo contains two hard-coded MotifMatrices which contain each possible microsatellite motifs with each permutation thereof. The MotifMatrices are scanned for the motifs of identified microsatellites and the variation of the microsatellite motif representing the standardization is returned.

A	*T*				
C	*G*				
AT	*TA*				
AC	*CA*	*GT*	*TG*		
AG	*GA*	*CT*	*TC*		
CG	*GC*				
AAT	*ATA*	*TAA*	*ATT*	*TTA*	*TAT*

Fig 3.: Excerpt from the Motif Matrix, fully standardized. Related motifs are arranged in one row. The left motif represents the fully standardized motif.

The MotifMatirx is sorted according to two rules: (i) short motifs are displayed first (mononucleotide motifs first, followed by the dinucleotide motifs and so on) and (ii) motifs are arranged with descending valuation A-T-C-G (Remember: the motifs with the most A-nucleotides first than the T and so on)

This arrangement ensures a high speed of the standardization process, since mononucleotide and dinucleotide motifs are extremely abundant. Therefore an early standardization of the most abundant motifs saves a lot of computation time.
Palindromic microsatellite motifs have also been considered during construction of the MotifMatices (e.g.: CG or ACGT)

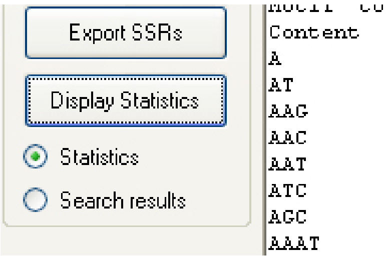
A		
T		
C		
G		
AT	*TA*	
AC	*CA*	
AG	*GA*	
TC	*CT*	
TG	*GT*	
CG	*GC*	
AAT	*ATA*	*TAA*

Fig. 4: Excerpt from the Motif Matrix, partially standardized. The left motifs represent the partially standardized variations.

.

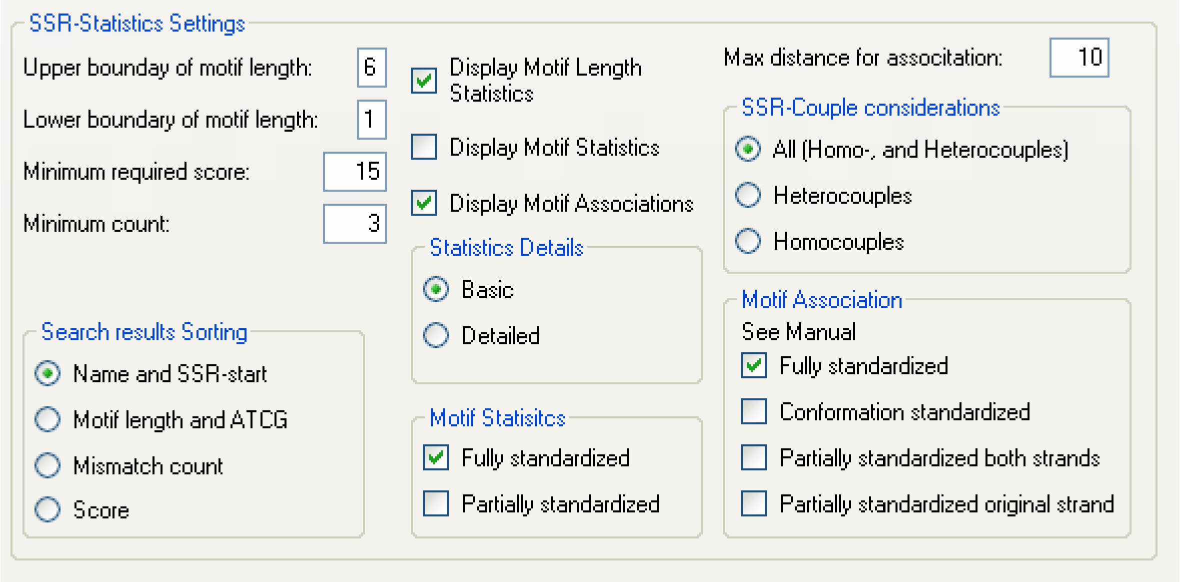
Display microsatellite statistics:

1. First load the file containing the microsatellites into the SciRoKo SSR-statistics module and choose the Statistics Radio Button in the main menu.



```
Motif  Co
Content
A
AT
AAG
AAC
AAT
ATC
AGC
AAAT
```

2. Then choose the settings: Which type of the statistics should be displayed? Display the motif length infos? Display the motif infos? Display the motif association? A frequency threshold might be specified (Minimum count). This feature is especially handy for the motif association statistics. Which detail level is required for the microsatellite statistics?



SSR-Statistics Settings

Upper bounday of motif length: 6

Lower boundary of motif length: 1

Minimum required score: 15

Minimum count: 3

Search results Sorting

● Name and SSR-start
○ Motif length and ATCG
○ Mismatch count
○ Score

☑ Display Motif Length Statistics

☐ Display Motif Statistics

☑ Display Motif Associations

Statistics Details

● Basic
○ Detailed

Motif Statisitcs

☑ Fully standardized
☐ Partially standardized

Max distance for associtation: 10

SSR-Couple considerations

● All (Homo-, and Heterocouples)
○ Heterocouples
○ Homocouples

Motif Association

See Manual

☑ Fully standardized
☐ Conformation standardized
☐ Partially standardized both strands
☐ Partially standardized original strand

3. Press the Display statistics button in the main menu

5.2.1 Motif Length Statistics (mono-, di-, tri- etc nucleotide SSRs)

```
Motif  Counts Average_Length       Average_Mismatches   Counts/Mbp
mononucleotide       456    21,80  0,55   37,78
dinucleotide         326    30,50  1,64   27,01
trinucleotide        509    29,12  1,39   42,17
tetranucleotide      97     20,92  0,53   8,04
pentanucleotide      137    21,57  0,69   11,35
hexanucleotide       205    29,82  1,12   16,98
```

Fig. 5: Excerpt from the Motif Length statistics of Saccharomyzes cereviseae. Detailed information for the frequencies, average mismatches etc are displayed for each motif length category. Basic detail level

The Motif Length Statistics calculate comprehensive statistic information for mono-, di-, tri-, tetra-, penta- und hexanucleotide microsatellites:

BASIC:
Column 1: Motif length category: mononucleotide, dinucleotide etc
Column 2: the total counts
Column 3: the average length of a microsatellite from this length category
Column 4: the average number of mismatches
Column 5: the average counts per million base pairs of a SSR from this length category

DETAILED (additionally):
Column 6: the average GC content. For the average GC content the whole SSR-sequences including the mismatches are considered. Because of mismatches a pure AT-microsatellite can achieve a GC-content of 0,02 = 2%
Column 7: Standard deviation of the microsatellite length. Values have to be treated with care because all microsatellites have a minimum length and the microsatellite length distribution is no Gaussian distribution. Nevertheless it roughly allows estimating which micrsoatellite categories exhibit the most length variations.
Column 8: the number of files / microsatellite; this feature is only interesting for microsatellite enriched librarys or BAC end sequences etc

SciRoKo also allows selecting of a microsatellite subset meeting certain criteria. For instance hexanucleotide microsatellites might be excluded for a better comparison with Modified Sputnik I-II results.

5.2.2 Motif statistics:

Related microsatellite motifs are grouped together and common group specific features are computed. Motif statistics contain two subcategories, fully and partially standardized motif statistics.

```
Motif  Counts Average_Length       Average_Mismatches   Counts/Mbp
A              454  21,82  0,55   37,61
AT             272  25,58  0,81   22,53
AAG            112  26,37  1,22   9,28
AAC            104  28,43  1,34   8,62
```

Fig. 6: Excerpt from the fully standardized motif statistics (Saccharomyzes cerevisaea). Basic detail level

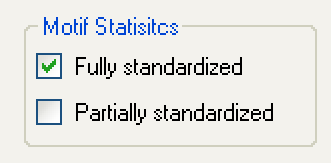
```
Motif  Counts Average_Length       Average_Mismatches   Counts/Mbp
AT             272  25,58  0,81   22,53
A              234  21,55  0,54   19,39
T              220  22,11  0,56   18,23
AAG             58  23,33  0,88   4,80
TTG             55  29,25  1,38   4,56
```

Fig. 7: Excerpt from the partially standardized motif statistics (Saccharomyzes cerevisaea). Basic detail level

The columns of the motif statistics are similar to the motif length statistics:
BASIC:
Column 1: standardized motif (fully or partially)
Column 2: the total counts
Column 3: the average length of microsatellites having the motif in column 1
Column 4: the average number of mismatches
Column 5: the average counts per million base pairs of SSRs having the motif in column 1

DETAILED (additionally):
Column 6: the average GC content.
Column 7: Standard deviation of the microsatellite length. Values have to be treated with care because all microsatellites have a minimum length and the microsatellite length distribution is no Gaussian distribution. Nevertheless it roughly allows estimating which micrsoatellite categories exhibit the most length variation.
Column 8: the number of files / microsatellite;

Settings for motif statistics:

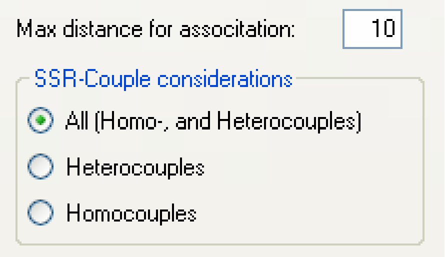

Motif Statisitcs

☑ Fully standardized

□ Partially standardized

5.2.3 Motif association statistics

Motif association statistics are the most complicated and sophisticated part of SciRoKo and a represent a unique feature. Two adjacent or neighboring microsatellites are called SSR-Couple, whereas two microsatellites are neighboring if the distance d between them is smaller as a specified maximum value (max. distance for association)


Max distance for associtation: 10

SSR-Couple considerations

● All (Homo-, and Heterocouples)

○ Heterocouples

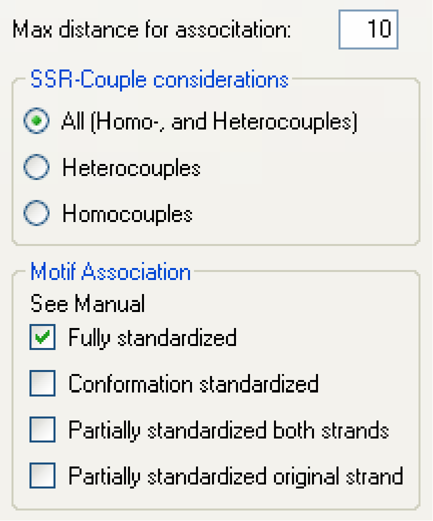
○ Homocouples

To prevent confusion with the motif of a microsatellite we termed the motif of a SSR-Couple as motif association. Attention SciRoKo 3.1 only works with SSR-Couples, a microsatellite cluster consisting of 3 microsatellites will be treated as two SSR-Couples.
SciRoKo further distinguishes Homo- and Heterocouples. If the two microsatellites forming a SSR-Couple have the same partially standardized motif they are referred to as Homocouples (e.g.: (AC)7- (CA)11) if the two microsatellites have different motifs they are referred to as Heterocouples (e.g.: (AG)9-(CT)12).
SciRoKo allows flexible adjustment of the SSR-Couples which should be considered for the motif association statistics.

For successful compound microsatellite analysis each fasta-identifier (text after the greater than symbol '>') has to be unique.
Since this is always the case with sequences obtained from Genbank we do not expect any complications.
The algorithm used for identification of motif association's first sorts all microsatellites according the fasta identifier and with equal identifier according to the SSR-start position. If the distance between two neighboring SSRs is less or equal to a specified distance the two SSRs are denoted as SSR-Couples. To reiterate the motif of a SSR-Couple is called motif association.

Settings for motif associations: The specified lower and upper boundary also affects the motif association statisitcs, for instance only motif associations between trinucleotide microsatellites might be displayed.

Motif association specific settings:



Max distance for associtation: 10

SSR-Couple considerations

● All (Homo-, and Heterocouples)

○ Heterocouples

○ Homocouples

Motif Association

See Manual

☑ Fully standardized

□ Conformation standardized

□ Partially standardized both strands

□ Partially standardized original strand

The maximum allowed distance between two neighbouring SSRs for a successful annotation as SSR-Couple can be set to a value of choice. The default value is 10. The minimum distance between two adjacent SSR is 1 and not 0.

Displaying motif association statistics

SciRoKo 3.1 additionally allows the import or export of SSR-Couples, which might be analysed detached from the microsatellites.

```
MotifAssociation     Counts AverageDistance      Av.lengthFirst       Av.lengthSecond
       Av.MismatchesFirst    Av.MismatchesSecond  Counts/Mbp
AAC-AGC           10     1,40 35,40  22,30  2,60   0,70     0,83
AAT-AAC            8     1,63 42,38  25,50  2,50   1,13     0,66
AT-AC              4     3,00 29,25  27,25  0,25   0,75     0,33
ATC-ACG            4     1,00 27,00  25,25  1,50   0,50     0,33
AAT-ACT            4     2,00 43,00  25,00  2,75   0,75     0,33
```

Fig. 8: Example of motif association statistic output generated with SciRoKo. Most frequent motif associations of Saccharomyzes cerevisiaea. Basic detail level.

BASIC:
Column 1: standardized motif association (four intensities are possible)
Column 2: total counts, for this category of motif association
Column 3: average distance between the two microsatellites
Column 4: Average length for the first microsatellite motif (eg. AAT for AAT-AAC)
Column 5: Average length for the second microsatellite motif (eg. AAC for AAT-AAC)
Column 6: Average mismatches for the first microsatellite
Column 7: Average mismatches for the second microsatellite
Column 8: Counts per million basepairs

DETAILED (additionally):
Column 9: GC-content for the first microsatellite motif
Column 10: GC-content for the second microsatellite motif
Column 11: standard deviation of the microsatellite length for the first microsatellite motif
Column 12: standard deviation of the microsatellite length for the second microsatellite motif
Column 13: Average number of fasta sequences per identified SSR-Couple.

5.2.4 Standardization intensities for motif association:

Unfortunately the standardization of motif associations is more complicated than the standardization of simple microsatellite motifs, because for two adjacent microsatellites a number of configurations have to be considered additionally. Each of the two microsatellites forming a SSR-Couple can be standardized in two intensities (partially and fully) additionally the conformation and the 5’-3’ arrangement has to be considered. Therefore standardization of SSR-Couple motif associations requires four standardization intensities compared to only two for microsatellite motifs.

```
5′-AGAGAGAGAGAGAGAGAGTGTGTGTGTGTGTGTGTGTGTG-3′
3′-TCTCTCTCTCTCTCTCTCACACACACACACACACACACAC-5′
```

Fig.9: Example of a compound microsatellite. For brevity the upper strand will be abstracted as AG-TG and the lower strand as CA-CT

The compound microsatellite in Figure 9 will act as example to demonstrate the different standardization intensities which might be applied to SSR-Couple motif association.
The upper strand of the microsatellite in Figure 9 might be written as: 5’-(AG)9-(TG)11-3’
Four our purpose this is still to long, we therefore depict the motif association of this SSR-Couples simply as: **AG-TG** (or F:AG-TG -> the F stands for found because this is the actually found motif association without any standardization applied)
The lower strand compound microsatellite will be written as **CA-CT**.
It can easily be seen that AG-TG and CA-CT actually represent the same compound microsatellite, thereforethe question raises: should this two motif associations really be displayed as two different motif associations? That’s when we enter the domain of motif association standardization.

In introducing the motif association statistics we will start with the least standardization intensity moving forward to the most intense standardizations

Partial standardization single strand PSS: (former Type 4 motif association)

PSS motif associations represent the least intensity of standardization used in SciRoKo.
They represent associations of partially standardized microsatellites, the reverse strand compound microsatellite is not considered the 5’-3’ arrangement is considered

Microsatellite 1: 5’-GAGAGAGAGAGAGAGAGACTCTCTCTCTCT-3’
Microsatellite 2: 5’-AGAGAGAGAGAGAGATCTCTCTCTCTCTCT-3’
Microsatellite 3: 5’-GAGAGAGAGAGAGAGAGAGTCTCTCTCTCT-3’

Fig. 10: Examples of different microsatellites which will be grouped into one category in the type 4 motif association standardization intensity, forming the motif association PSS:AG-TC

For PSS motif associations the two microsatellites forming a SSR-Couple are just partially standardized. Figure 10 demonstrates which microsatellites will be grouped together in this

standardization intensity. The two motif associations introduced in Figure 9 AG-TG and CA-CT are not grouped into one category in the PSS motif associations.
Note that the motif associations AG-CT and CT-AG are not grouped into one category, since the PSS motif association still considers the 5’-3’ arrangement of the two microsatellites as to be important for separating the motif association categories..

TGC-TTG	5	1,80	19,60	26,60	0,20	1,40	0,41
ATT-TTG	3	1,00	48,67	31,00	3,00	2,33	0,25
AAC-AGC	2	1,00	51,00	36,00	4,50	2,50	0,17
AT-TG	2	5,00	38,00	24,00	0,00	0,00	0,17
AAC-AAT	2	1,50	27,50	43,00	0,50	3,00	0,17
TCG-ATC	2	1,00	31,00	21,00	0,50	1,00	0,17

Fig. 11: Example of PSS motif associations identified in Saccharomyzes cerevisaea

Partial standardization both strands PSB: (former Type 3 motif associations)

PSB motif associations apply a more vigorous standardization than PSS motif associations but still not as intense as CS (conformation standardization) motif associations.
PSB motif associations represent associations of partially standardized microsatellites, the reverse strand compound microsatellite is considered, the 5’-3’ arrangement is considered. Finally the two motif associations introduced in Figure 9 - AG-TG and CA-CT - will be grouped into one category in the PSB motif associations.

Note that the motif associations AG-CT and CT-AG are still not grouped into one category, since the PSB motif association still considers the 5’-3’ arrangements of the two microsatellites forming the SSR-Couple an important trait for separating the different motif association categories. See also Figure 12.

```
Microsatellite 1:
5’-AGAGAGAGAGAGAGAGAGAGTGTGTGTGTGTGTGTGTGTGTG-3’
3’-TCTCTCTCTCTCTCTCTCTCACACACACACACACACACACAC-5’

Not equal to microsatellite 2 in type 3 motif associations:
5’-TGTGTGTGTGTGTGTGTGTGAGAGAGAGAGAGAGAGAGAGAG-3’
3’-ACACACACACACACACACACTCTCTCTCTCTCTCTCTCTCTC-5’
```

Fig. 12: Example of two compound microsatellite which are **not** grouped together in PSB motif associations

Note: the important difference to PSS motif associations is that PSB motif associations group the reverse complements strand compound microsatellite into the same category!

AAC-AGC	7	1,57	33,57	24,29	2,29	0,86	0,58
AAC-AAT	5	1,20	29,60	46,40	1,60	3,00	0,41
AT-TG	4	3,00	29,25	27,25	0,25	0,75	0,33
ATG-ACG	3	1,00	20,67	26,67	0,67	0,33	0,25
AAT-AAC	3	2,33	35,67	18,67	1,67	0,33	0,25

Fig. 13: Example of PSB motif associations identified in Saccharomyzes cerevisaea

Conformation standardization CS: (former Type 2 motif associations)

CS motif associations are the next standardization intensity only FS (full standardization) motif associations apply a more vigorous standardization intensity. CS motif associations represent associations of partially standardized microsatellites, the reverse strand compound microsatellite is considered the 5’-3’ arrangement is ignored.
Therefore the two microsatellites introduced in Figure 12 - AG-TG and TG-AG - will be grouped into one category since CS motif associations ignore the 5’-3’ arrangement.
Figure 14 demonstrates that CS motif associations still consider the conformation of the compound microsatellites therefore the motif associations AG-TG and AG-AC are not grouped into one category in the CS motif association.

```
Microsatellite 1:
5′-AGAGAGAGAGAGAGAGAGAGTGTGTGTGTGTGTGTGTGTGTG-3′
3′-TCTCTCTCTCTCTCTCTCTCACACACACACACACACACACAC-5′

Not equal to microsatellite 2 in type 2 motif associations:
5′-AGAGAGAGAGAGAGAGAGAGAGAGACACACACACACACACAC-3′
3′-TCTCTCTCTCTCTCTCTCTCTCTCTGTGTGTGTGTGTGTGTG-5′
```

Fig. 14: Example of two compound microsatellite which are **not** grouped together in CS motif associations

Note: the important difference to PSB motif associations is that CS motif associations ignore the 5’-3’ arrangement, grouping motif associations with the same motifs but different 5’-3’ arrangements into the same category!

```
AAC-AGC          10    1,40 35,40  22,30  2,60   0,70    0,83
AAT-AAC           8    1,63 42,38  25,50  2,50   1,13    0,66
AT-AC             4    3,00 29,25  27,25  0,25   0,75    0,33
ATC-TCG           4    1,00 27,00  25,25  1,50   0,50    0,33
AAG-ATG           2    1,00 26,00  28,00  0,00   1,50    0,17
AAG-AGG           2    1,00 25,00  22,00  1,00   0,00    0,17
```

Fig. 15: Example of CS motif associations identified in Saccharomyzes cerevisae!

Full standardization FS: (former Type 1 motif associations)

FS motif associations represent the most intense standardization degree. FS motif association represent associations of fully standardize motifs, the 5’-3’ arrangement is ignored. The reverse strand compound microsatellite is considerd. The two microsatellites introduced in Figure 14- AG-TG and AG-AC – are grouped into one category in the FS motif associations. FS motif associations just represent the fully standardized microsatellite motifs of a SSR-Couple.

```
AAC-AGC          10    1,40 35,40  22,30  2,60   0,70    0,83
AAT-AAC           8    1,63 42,38  25,50  2,50   1,13    0,66
AT-AC             4    3,00 29,25  27,25  0,25   0,75    0,33
ATC-ACG           4    1,00 27,00  25,25  1,50   0,50    0,33
AAT-ACT           4    2,00 43,00  25,00  2,75   0,75    0,33
AAG-ATC           2    1,00 26,00  28,00  0,00   1,50    0,17
```

Fig. 16 Example of FS motif association identified in Saccharoymzes cerevisae.

The motif association standardization pyramid

With each intensification of the standardization intensity, additional SSR-Couples are grouped into the same category, forming a pyramid with FS motif associations at the top and PSS motif associations at the bottom. Figure 17 demonstrates this principle for the most frequent motif association in *Saccharomyzes cerevisae.*
Figure 17 demonstrates which motif association are grouped together with progressing standardization intensities.

PSS:	**PSB:**	**CS:**	**FS:**
TGC-TTG	AAC-AGC	AAC-AGC	AAC-AGC
AAC-AGC			
TTG-TGC	TTG-TGC		
AGC-AAC			
AAC-TGC	TGC-AAC	AAC-TGC	
AGC-TTG			
TGC-AAC	AAC-TGC		
TTG-AGC			

Fig. 17: Standization pyramid for the most frequent motif association of Saccharomyzes cerevisaea

6. Software tools for processing SciRoKo results

Two tools which closely cooperate with SciRoKo exist. SSR-Cluster processes exported SSR-Couple files (*.ssrCou) and Overrepresentation which processes the SciRoKo SSR-statistic output.

SSR-Cluster:

SSR-Cluster is a console program which processes SSR-Couple files. Each microsatellite might be involved in two SSR-Couples, in a 5’ and a 3’ SSR-Couple. Figure 18 demonstrates this relationship.

```
5′-AGAGAGAGAGAGAGAGAGTGTGTGTGTGTGTGTGTGTGACACACACACACACACACAC-3′
3′-TCTCTCTCTCTCTCTCTCACACACACACACACACACACTGTGTGTGTGTGTGTGTGTG-5′
```

Fig.18 : Example of an SSR-Cluster consisting of three microsatellites. SciRoKo identifies two SSR-Couples. The SSR-Couple F:AG-TG and F:TG-AC. The console program SSR-Cluster reduces overlapping SSR-Couples to one SSR-Cluster and allows exporting of the involved micosatellites which in turn might be processed by SciRoKo again

According to our definition a SSR-Cluster is an aggregation of microsatellites containing at least two microsatellites. The console program SSR-Cluster aggregates overlapping SSR-Couples to SSR-Clusters (Figure 18), the smallest possible SSR-Cluster contains only one SSR-Couple. Therefore SSR-Cluster might be imagined as the highest level annotation of microsatellite clustering behaviour.

The console program SSR-Cluster counts SSR-Clusters (Homo- and Heteroclusters) meeting a specified minimum size (default: 2 microsatellites) and allows exporting of the non redundant set of involved microsatellites (nr. HeMS: non redundant Heterocluster forming microsatellite subset)

Overrepresentation:

The console program overrepresentation calculates the expected number of SSR-Couples and the overrepresentations for a given set of “Fully standardized motif associations”.

The input file has to contain most of the SciRoKo SSR-Statistics output.

- The overall microsattellite statistic (average length of a microsatellite, total number of microsatellites, etc)
- The fully standardized motif statistics
- The partially standardized motif statistics
- And the fully standardized motif association statistics

Overreprsentation calculates:

- The expected number of motif associations of a given fully standardized (FS:) motif association category (eg.: FS:AC-AG) based on the frequency of the fully standardized motifs (eg: AC had 5000 counts and AG 7000 counts)
- The overrepresentation of a motif association of a given fully standardized motif association category:
- The overall expected number of SSR-Couples
- The overall expected number of Homocouples based on the frequencies of the partially standardized microsatellite motifs (eg.: AAG had 2000 counts and TTC 3000, they are calculated separately)
- The overall expected number of Heterocouples (Heterocouples= SSR-Couples-Homocouples)