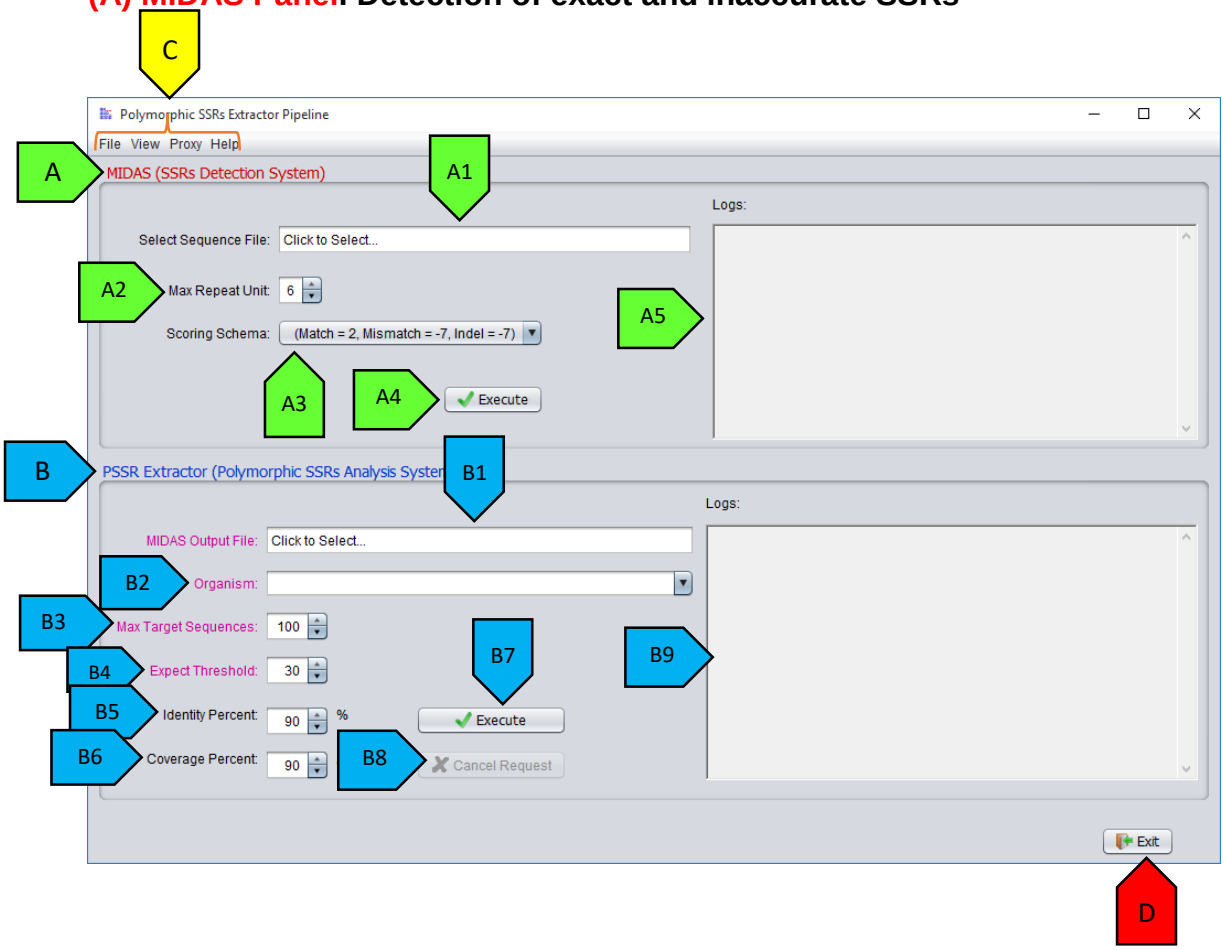
Polymorphic SSRs Extractor Pipeline

Introducing components/functions in graphic interface.

Background

Polymorphic SSRs Extractor Pipeline is a software that allows the fast and reliable identification of polymorphic SSRs loci in genomic sequences. Global processing is done by concatenation of MIDAS [1] and PSSR-Extractor [2] programs, the latter making a remote invocation to NCBI BLAST [3].

(A) MIDAS Panel. Detection of exact and inaccurate SSRs



(A1) Select a file (genome in FASTA or GBFF format, both single or multi-locus).

(A2) Sets the maximum size of the repeat unit to be scanned in the genome (1-6).

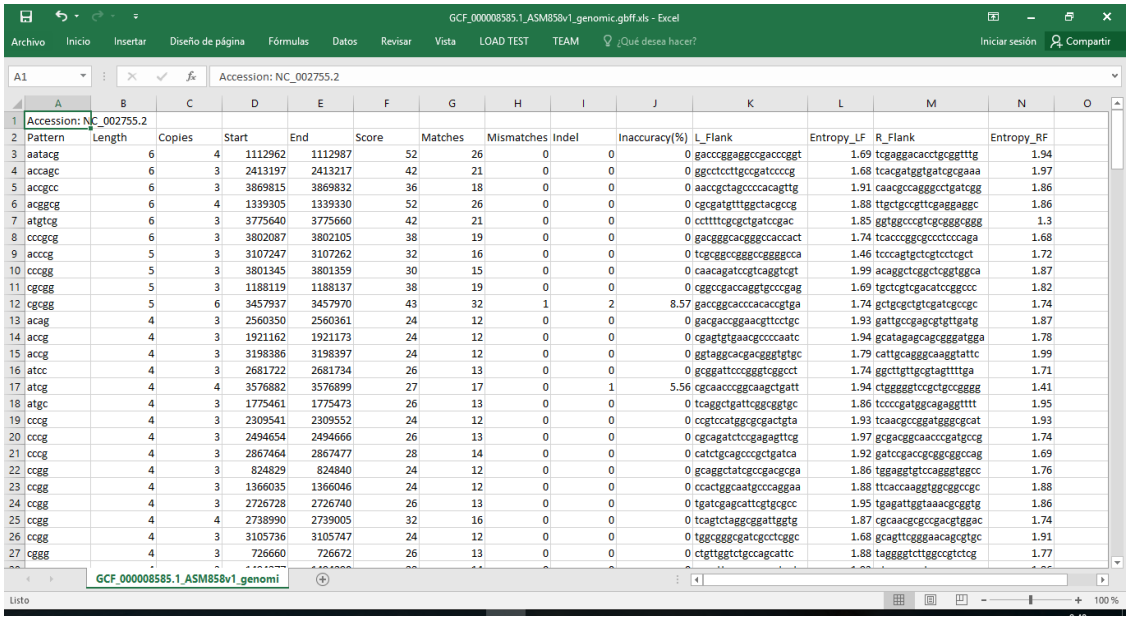
(A3) Alignment parameters scheme for match, mismatch and indel.

(A4) Execute the program according to the previously defined parameters.

(A5) Reports the current status of the application execution.

As outputs MIDAS returns three files with the same name as input file plus the extensions .xls, .dat and .mfaa (.xls can be opened with Excel or similar applications).

.xls file:



Accession: NC_002755.2

Pattern	Length	Copies	Start	End	Score	Matches	Mismatches	Indel	Inaccuracy(%)	L_Flank	Entropy_LF	R_Flank	Entropy_RF
aatacg	6	4	1112962	1112987	52	26	0	0	0	gacccggaggccgacccggt	1.69	tcgaggacacctgcggtttg	1.94
accagc	6	3	2413197	2413217	42	21	0	0	0	ggcctccttgccgatccccg	1.68	tcacgatggtgatcgcgaaa	1.97
accgcc	6	3	3869815	3869832	36	18	0	0	0	aaccgctagccccacagttg	1.91	caacgccagggcctgatcgg	1.86
acggcg	6	4	1339305	1339330	52	26	0	0	0	cgcgatgtttggctacgccg	1.88	ttgctgccgttcgaggaggc	1.86
atgtcg	6	3	3775640	3775660	42	21	0	0	0	ccttttcgcgctgatccgac	1.85	ggtggcccgtcgcgggcggg	1.3
cccgcg	6	3	3802087	3802105	38	19	0	0	0	gacgggcacgggccaccact	1.74	tcaccccggcgccctcccaga	1.68
acccg	5	3	3107247	3107262	32	16	0	0	0	tcgcggcccgggccggggcca	1.46	tcccagtgctcgtcctcgct	1.72
ccccgg	5	3	3801345	3801359	30	15	0	0	0	caacagatccgtcaggtcgt	1.99	acaggctcggctcggtggca	1.87
cgcgg	5	3	1188119	1188137	38	19	0	0	0	cggccgaccaggtgcccgag	1.69	tgctcgtcgacatccggccc	1.82
cgcgg	5	6	3457937	3457970	43	32	1	2	8.57	gaccggcacccacaccgtga	1.74	gctgcgctgtcgatcgccgc	1.74
acag	4	3	2560350	2560361	24	12	0	0	0	gacgaccggaacgttcctgc	1.93	gattgccgagcgtgttgatg	1.87
accg	4	3	1921162	1921173	24	12	0	0	0	cgagtgtgaacgccccaatc	1.94	gcatagagcagcgggatgga	1.78
accg	4	3	3198386	3198397	24	12	0	0	0	ggtaggcacgacgggtgtgc	1.79	cattgcagggcaaggtattc	1.99
atcc	4	3	2681722	2681734	26	13	0	0	0	gcggattcccgggtcggcct	1.74	ggcttgttgcgtagttttga	1.71
atcg	4	4	3576882	3576899	27	17	0	1	5.56	cgcaacccggcaagctgatt	1.94	ctgggggtccgctgccgggg	1.41
atgc	4	3	1775461	1775473	26	13	0	0	0	tcaggctgattcggcggtgc	1.86	tccccgatggcagaggttttt	1.95
cccg	4	3	2309541	2309552	24	12	0	0	0	ccgtccatggcgcgactgta	1.93	tcaacgccggatgggcgcat	1.93
cccg	4	3	2494654	2494666	26	13	0	0	0	cgcagatctccgagagttcg	1.97	gcgacggcaacccgatgccg	1.74
cccg	4	3	2867464	2867477	28	14	0	0	0	catctgcagcccgctgatca	1.92	gatccgaccgcggcggccag	1.69
ccgg	4	3	824829	824840	24	12	0	0	0	gcaggctatcgccgacgcga	1.86	tggaggtgtccagggtggcc	1.76
ccgg	4	3	1366035	1366046	24	12	0	0	0	ccactggcaatgcccaggaa	1.88	ttcaccaaggtggcggccgc	1.88
ccgg	4	3	2726728	2726740	26	13	0	0	0	tgatcgagcattcgtgcgcc	1.95	tgagattggtaaacgcggtg	1.86
ccgg	4	4	2738990	2739005	32	16	0	0	0	tcagtctaggcggattggtg	1.87	cgcaacgcgccgacgtggac	1.74
ccgg	4	3	3105736	3105747	24	12	0	0	0	tggcgggcgatcgcctcggc	1.68	gcagttcgggaacagcgtgc	1.91
cggg	4	3	726660	726672	26	13	0	0	0	ctgttggtctgccagcattc	1.88	taggggtcttggccgtctcg	1.77

Accession: GenBank access code.

Pattern: Pattern sequence.

Length: Size of the repeating unit.

Copies: Number of copies.

Start: Start position in the genome.

End: End position in the genome.

Score: Alignment score.

Matches: Matching bases.

Mismatches: Mismatched bases.

Indel: Base insertions and deletions.

Inaccuracy: SSR inaccuracy (%).

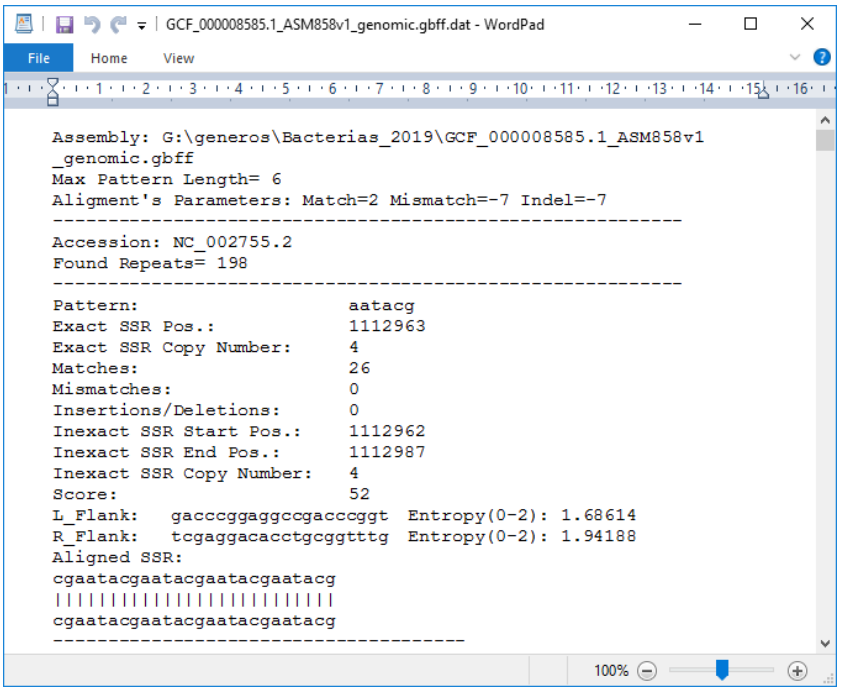
L_Flank: Left flank sequence.

Entropy_LF: Compositional entropy in left flank.

R_Flank: Right flank sequence.

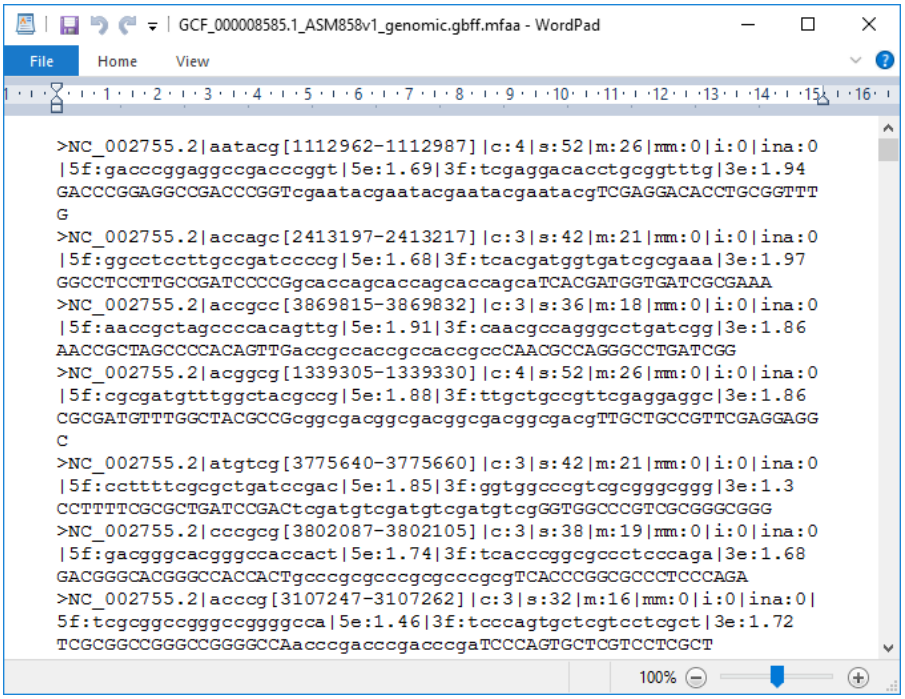
Entropy_RF: Compositional entropy in right flank.

.dat file. It could be visualized in any text editor, presents the data in non-tabular form and shows sequence alignments:



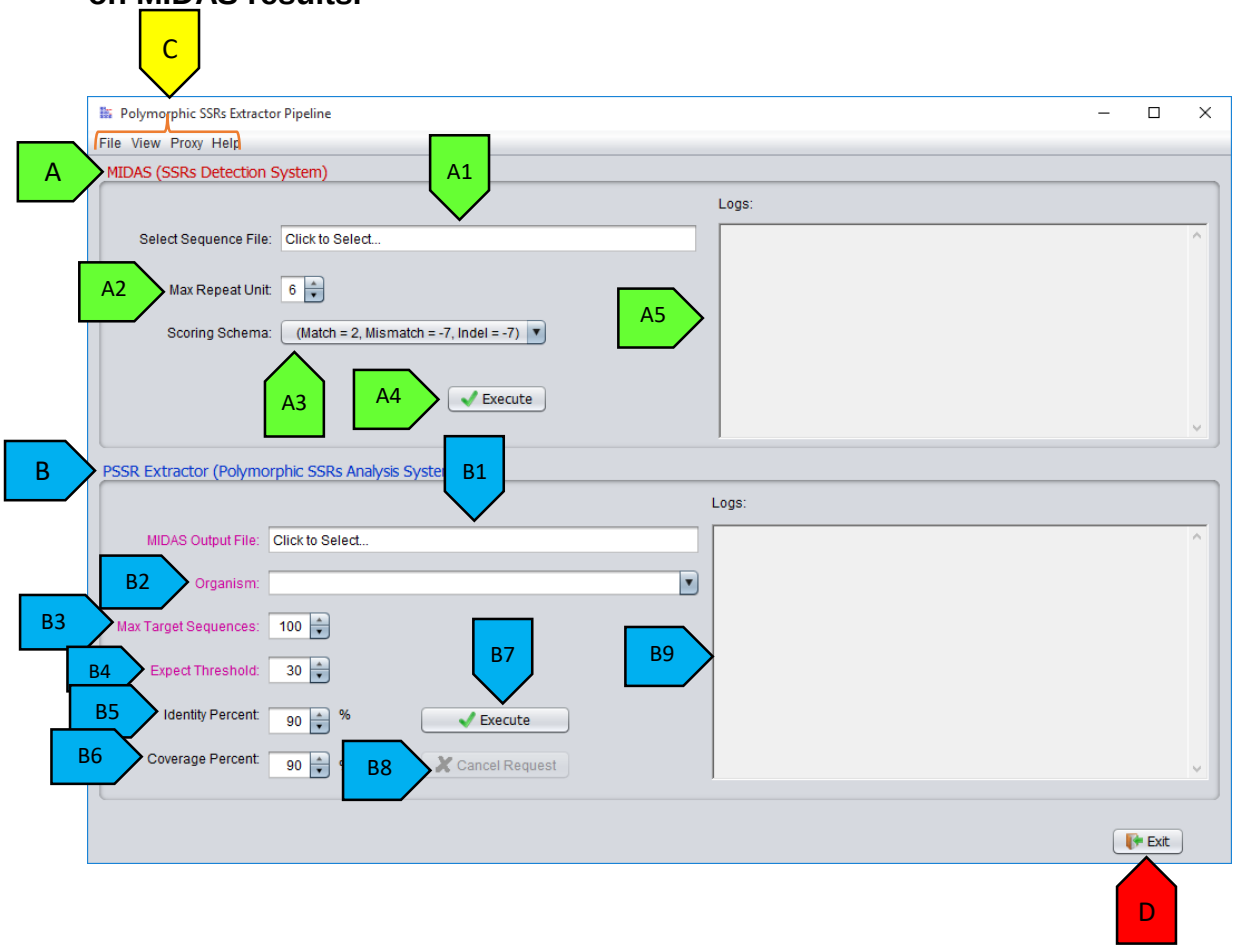
```
Assembly: G:\generos\Bacterias_2019\GCF_000008585.1_ASM858v1
_genomic.gbff
Max Pattern Length= 6
Aligment's Parameters: Match=2 Mismatch=-7 Indel=-7
-------------------------------------------------------
Accession: NC_002755.2
Found Repeats= 198
-------------------------------------------------------
Pattern:                    aatacg
Exact SSR Pos.:             1112963
Exact SSR Copy Number:      4
Matches:                    26
Mismatches:                 0
Insertions/Deletions:       0
Inexact SSR Start Pos.:     1112962
Inexact SSR End Pos.:       1112987
Inexact SSR Copy Number:    4
Score:                      52
L_Flank:   gacccggaggccgacccggt  Entropy(0-2): 1.68614
R_Flank:   tcgaggacacctgcggtttg  Entropy(0-2): 1.94188
Aligned SSR:
cgaatacgaatacgaatacgaatacg
||||||||||||||||||||||||||
cgaatacgaatacgaatacgaatacg
--------------------------------------
```

.mfaa file. This file is the input for PSSR-Extractor program (B). It could be visualized in any text editor and presents the SSRs in multi-fasta format. The repeat region is marked in lowercase and the flanks in uppercase. The header presents information such as the GenBank accession number, the motif, and the positions in the genome:



```
>NC_002755.2|aatacg[1112962-1112987]|c:4|s:52|m:26|mm:0|i:0|ina:0
|5f:gacccggaggccgacccggt|5e:1.69|3f:tcgaggacacctgcggtttg|3e:1.94
GACCCGGAGGCCGACCCGGTcgaatacgaatacgaatacgaatacgTCGAGGACACCTGCGGTTT
G
>NC_002755.2|accagc[2413197-2413217]|c:3|s:42|m:21|mm:0|i:0|ina:0
|5f:ggcctccttgccgatccccg|5e:1.68|3f:tcacgatggtgatcgcgaaa|3e:1.97
GGCCTCCTTGCCGATCCCCGgcaccagcaccagcaccagcaTCACGATGGTGATCGCGAAA
>NC_002755.2|accgcc[3869815-3869832]|c:3|s:36|m:18|mm:0|i:0|ina:0
|5f:aaccgctagccccacagttg|5e:1.91|3f:caacgccagggcctgatcgg|3e:1.86
AACCGCTAGCCCCACAGTTGaccgccaccgccaccgccCAACGCCAGGGCCTGATCGG
>NC_002755.2|acggcg[1339305-1339330]|c:4|s:52|m:26|mm:0|i:0|ina:0
|5f:cgcgatgtttggctacgccg|5e:1.88|3f:ttgctgccgttcgaggaggc|3e:1.86
CGCGATGTTTGGCTACGCCGcggcgacggcgacggcgacggcgacgTTGCTGCCGTTCGAGGAGG
C
>NC_002755.2|atgtcg[3775640-3775660]|c:3|s:42|m:21|mm:0|i:0|ina:0
|5f:ccttttcgcgctgatccgac|5e:1.85|3f:ggtggcccgtcgcgggcggg|3e:1.3
CCTTTTCGCGCTGATCCGACtcgatgtcgatgtcgatgtcgGGTGGCCCGTCGCGGGCGGG
>NC_002755.2|cccgcg[3802087-3802105]|c:3|s:38|m:19|mm:0|i:0|ina:0
|5f:gacgggcacgggccaccact|5e:1.74|3f:tcacccggcgccctcccaga|3e:1.68
GACGGGCACGGGCCACCACTgcccgcgcccgcgcccgcgTCACCCGGCGCCCTCCCAGA
>NC_002755.2|acccg[3107247-3107262]|c:3|s:32|m:16|mm:0|i:0|ina:0|
5f:tcgcggccgggccggggcca|5e:1.46|3f:tcccagtgctcgtcctcgct|3e:1.72
TCGCGGCCGGGCCGGGGCCAacccgacccgacccgaTCCCAGTGCTCGTCCTCGCT
```

(B) PSSR Extractor Panel. Detects polymorphism in microsatellites based on MIDAS results.



(B1) Selects the .mfaa file from MIDAS, which will be the queries for BLAST.

(B2) Selects the organism to which the sequences belong. BLAST will search sequences from the same organism.

(B3) Sets the maximum of similar sequences that BLAST will return.

(B4) Filters sequences with an expected value higher than the selected one.

(B5) Minimum percent identity that the two sequences must have on their flanks.

(B6) Percentage of the number of bases that the two sequences must share on their flanks.

(B7) Execute the program according to the previously defined parameters.

(B8) Cancel the execution of the program. This button is activated after pressing B7.

(B9) Reports the current state of application execution (B).

(D) Exits the program.

BLAST has more parameters than those exposed, but it is necessary for the analysis kept them constant. These parameters and their default value are shown below:

QUERY_BELIEVE_DEFLINE: false.

DATABASE: nr.

LCASE_MASK: true.

FILTER: F.

FORMAT_TYPE: Tabular.

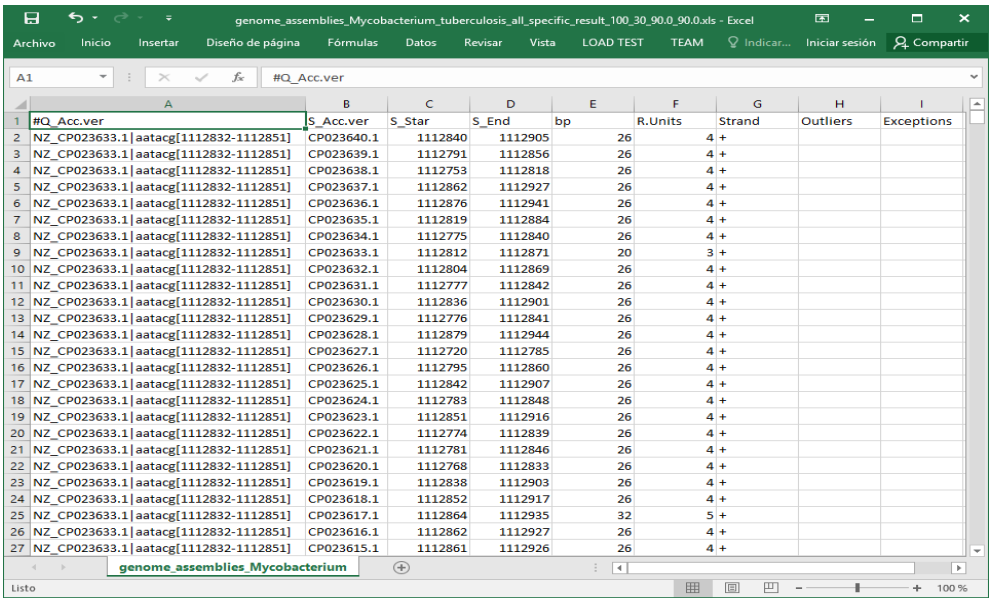
PROGRAM: blastn.

CLIENT: web.

BLAST_PROGRAM: blastn.

PSSR Extractor outputs two .xls files, one detailed and the other generic, whose names have the suffixes _specific_result and _generic_result respectively.

_specific_result.xls. Provides detailed information on each processed subject.



#Q_Acc.ver	S_Acc.ver	S_Star	S_End	bp	R.Units	Strand	Outliers	Exceptions
NZ_CP023633.1|aatacg[1112832-1112851]	CP023640.1	1112840	1112905	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023639.1	1112791	1112856	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023638.1	1112753	1112818	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023637.1	1112862	1112927	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023636.1	1112876	1112941	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023635.1	1112819	1112884	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023634.1	1112775	1112840	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023633.1	1112812	1112871	20	3	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023632.1	1112804	1112869	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023631.1	1112777	1112842	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023630.1	1112836	1112901	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023629.1	1112776	1112841	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023628.1	1112879	1112944	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023627.1	1112720	1112785	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023626.1	1112795	1112860	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023625.1	1112842	1112907	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023624.1	1112783	1112848	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023623.1	1112851	1112916	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023622.1	1112774	1112839	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023621.1	1112781	1112846	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023620.1	1112768	1112833	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023619.1	1112838	1112903	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023618.1	1112852	1112917	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023617.1	1112864	1112935	32	5	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023616.1	1112862	1112927	26	4	+		
NZ_CP023633.1|aatacg[1112832-1112851]	CP023615.1	1112861	1112926	26	4	+		

Q_acc_ver: Access version of the query sequence that was compared with the subject sequences in database.
S_acc_ver: Access version of the subject sequences in database that was compared with query sequence.

S_start: Start position of the subject.

S_end: End position of the subject.

Note: The positions provided by S_start and S_end include the flanks surrounding the microsatellite. If S_start > S_end the subject sequence has a negative direction ("-"), that is, a reverse complement sequence.

bp: Number of nucleotides between two flanks.

R.units: Number of units repeated between two flanks.

Strand: Subject's direction ("+" if the direction of the sequence is from left to right and "-" if the direction of the sequence is from right to left).

Outliers: Represented with "*" and can vary from 1 to 5. It means that the number of repeated units between the flanks of the subjects is doubtful because it is very large, being unlikely a microsatellite between them. The cutoff values established for this exception were mononucleotide: 157 bp, dinucleotide: 364 bp, trinucleotide: 109 bp, tetranucleotide: 45 bp, pentanucleotide: 150 bp, and hexanucleotide: 193 bp. These values were defined after processing SSRs in 200 bacterial genomes, recording their sizes, and establishing the cutoff at 3 times the interquartile range:

- * **bp** > cutoff value.
- ** **bp** ≥ twice the cutoff value.
- *** **bp** ≥ three times the cutoff value.
- **** **bp** ≥ four times the cutoff value.
- ***** **bp** ≥ five times the cutoff value.

Exceptions: Labels that correspond to exceptions:

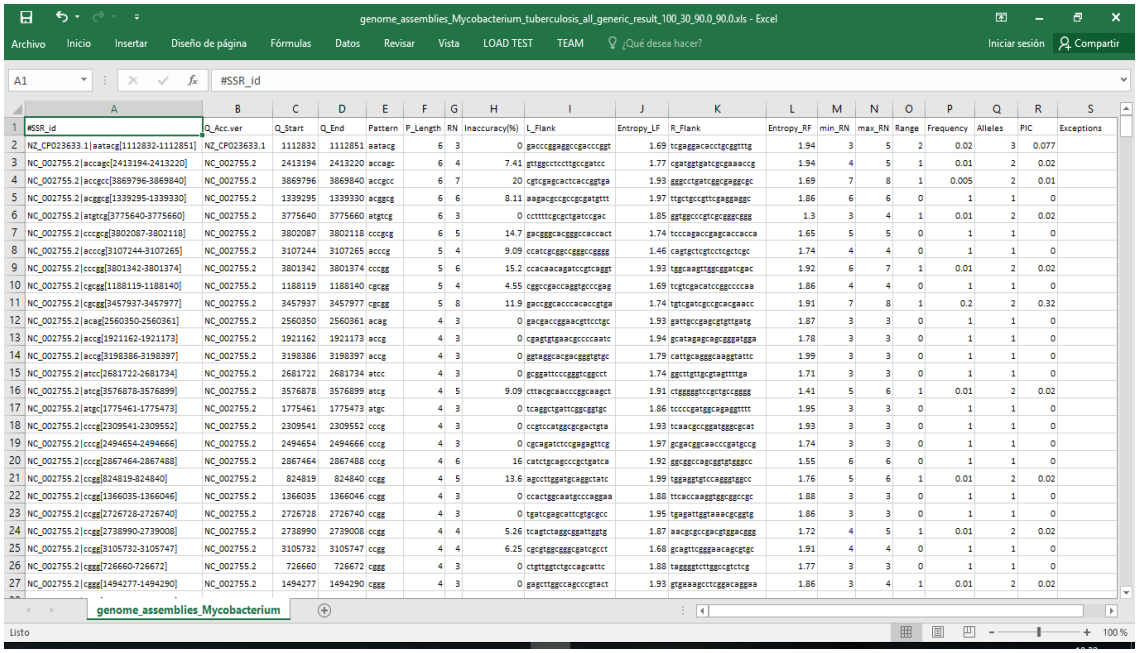
"**D**" (degenerated). Subjects have an Identity Percent <90% and / or a Coverage Percent <90%.

"**NF**" (not found). No subject with similarity was found.

"**O**" (outlier).

"**U**" (unpair). A single edge appears for the same subject sequence.

_generic_result.xls: Provides the information related to the polymorphism for each query, that is, each SSR.



genome_assemblies_Mycobacterium_tuberculosis_all_generic_result_100_30_90.0_90.0.xls - Excel

A1 #SSR_id

#SSR_id	Q_Acc.ver	Q_Start	Q_End	Pattern	P_Length	RN	Inaccuracy(%)	L_Flank	Entropy_LF	R_Flank	Entropy_RF	min_RN	max_RN	Range	Frequency	Alleles	PIC	Exceptions
NZ_CP023633.1|aatacg[1112832-1112851]	NZ_CP023633.1	1112832	1112851	aatacg	6	3	0	gacccggaggccgacccggt	1.69	tcgaggacacctgcggtttg	1.94	3	5	2	0.02	3	0.077	
NC_002755.2|accagc[2413194-2413220]	NC_002755.2	2413194	2413220	accagc	6	4	7.41	gttggcctccttgccgatcc	1.77	cgatggtgatcgcgaaaccg	1.94	4	5	1	0.01	2	0.02	
NC_002755.2|accgcc[3869796-3869840]	NC_002755.2	3869796	3869840	accgcc	6	7	20	cgtcgagcactcaccggtga	1.93	ggccctgatcggcgaggcgc	1.69	7	8	1	0.005	2	0.01	
NC_002755.2|acggcg[1339295-1339330]	NC_002755.2	1339295	1339330	acggcg	6	6	8.11	aagacgccgccgcgatgttt	1.97	ttgctgccgttcgaggaggc	1.86	6	6	0	1	1	0	
NC_002755.2|atgtcg[3775640-3775660]	NC_002755.2	3775640	3775660	atgtcg	6	3	0	ccttttcgcgctgatccgac	1.85	ggtggcccgtcgcggcggg	1.3	3	4	1	0.01	2	0.02	
NC_002755.2|cccgcg[3802087-3802118]	NC_002755.2	3802087	3802118	cccgcg	6	5	14.7	gacgggcacggggccaccact	1.74	tcccagaccgagcaccacca	1.65	5	5	0	1	1	0	
NC_002755.2|acccg[3107244-3107265]	NC_002755.2	3107244	3107265	acccg	5	4	9.09	ccatcgcggccgggccgggg	1.46	cagtgctcgtcctcgctcgc	1.74	4	4	0	1	1	0	
NC_002755.2|cccgg[3801342-3801374]	NC_002755.2	3801342	3801374	cccgg	5	6	15.2	ccacaacagatccgtcaggt	1.93	tggcaagttggcggatcgac	1.92	6	7	1	0.01	2	0.02	
NC_002755.2|cgcgg[1188119-1188140]	NC_002755.2	1188119	1188140	cgcgg	5	4	4.55	cggccgaccaggtgcccgag	1.69	tcgtcgacatccggccccaa	1.86	4	4	0	1	1	0	
NC_002755.2|cgcgg[3457937-3457977]	NC_002755.2	3457937	3457977	cgcgg	5	8	11.9	gaccggcacccacaccgtga	1.74	tgtcgatcgccgcacgaacc	1.91	7	8	1	0.2	2	0.32	
NC_002755.2|acag[2560350-2560361]	NC_002755.2	2560350	2560361	acag	4	3	0	gacgaccggaacgttcctgc	1.93	gattgccgagcgtgttgatg	1.87	3	3	0	1	1	0	
NC_002755.2|accg[1921162-1921173]	NC_002755.2	1921162	1921173	accg	4	3	0	cgagtgtgaacgcccaatc	1.94	gcatagagcagcgggatgga	1.78	3	3	0	1	1	0	
NC_002755.2|accg[3198386-3198397]	NC_002755.2	3198386	3198397	accg	4	3	0	ggtaggcacgacgggtgtgc	1.79	cattgcagggcaaggtattc	1.99	3	3	0	1	1	0	
NC_002755.2|atcc[2681722-2681734]	NC_002755.2	2681722	2681734	atcc	4	3	0	gcggattcccgggtcggcct	1.74	ggcttgttgcgtagttttga	1.71	3	3	0	1	1	0	
NC_002755.2|atcg[3576878-3576899]	NC_002755.2	3576878	3576899	atcg	4	5	9.09	cttacgcaacccggcaagct	1.91	ctgggggtccgctgccgggg	1.41	5	6	1	0.01	2	0.02	
NC_002755.2|atgc[1775461-1775473]	NC_002755.2	1775461	1775473	atgc	4	3	0	tcaggctgattcggcggtgc	1.86	tccccgatggcagaggtttt	1.95	3	3	0	1	1	0	
NC_002755.2|cccg[2309541-2309552]	NC_002755.2	2309541	2309552	cccg	4	3	0	ccgtccatggcgcgactgta	1.93	tcaacgccggatgggcgcat	1.93	3	3	0	1	1	0	
NC_002755.2|cccg[2494654-2494666]	NC_002755.2	2494654	2494666	cccg	4	3	0	cgcagatctccgagagttcg	1.97	gcgacggcaacccgatgccg	1.74	3	3	0	1	1	0	
NC_002755.2|cccg[2867464-2867488]	NC_002755.2	2867464	2867488	cccg	4	6	16	catctgcagcccgctgatca	1.92	ggcggccagcggtgtgggcc	1.55	6	6	0	1	1	0	
NC_002755.2|ccgg[824819-824840]	NC_002755.2	824819	824840	ccgg	4	5	13.6	agccttggatgcaggctatc	1.99	tggaggtgtccagggtggcc	1.76	5	6	1	0.01	2	0.02	
NC_002755.2|ccgg[1366035-1366046]	NC_002755.2	1366035	1366046	ccgg	4	3	0	ccactggcaatgcccaggaa	1.88	ttcaccaaggtggcggccgc	1.88	3	3	0	1	1	0	
NC_002755.2|ccgg[2726728-2726740]	NC_002755.2	2726728	2726740	ccgg	4	3	0	tgatcgagcattcgtgcgcc	1.95	tgagattggtaaacgcggtg	1.86	3	3	0	1	1	0	
NC_002755.2|ccgg[2738990-2739008]	NC_002755.2	2738990	2739008	ccgg	4	4	5.26	tcagtcctaggcggattggtg	1.87	aacgcgccgacgtggacggg	1.72	4	5	1	0.01	2	0.02	
NC_002755.2|ccgg[3105732-3105747]	NC_002755.2	3105732	3105747	ccgg	4	4	6.25	cgcgtggcgggcgatcgcct	1.68	gcagttcgggaacagcgtgc	1.91	4	4	0	1	1	0	
NC_002755.2|cggg[726660-726672]	NC_002755.2	726660	726672	cggg	4	3	0	ctgttggtctgccagcattc	1.88	taggggtcttggccgtctcg	1.77	3	3	0	1	1	0	
NC_002755.2|cggg[1494277-1494290]	NC_002755.2	1494277	1494290	cggg	4	3	0	gagcttggccagcccgtact	1.93	gtgaaagcctcggacaggaa	1.86	3	4	1	0.01	2	0.02	

genome_assemblies_Mycobacterium

SSr_id: GenBank query access code, repeating unit and positions in the genome.

Q_acc.ver: GenBank query access code.

Q_Start: Start position of query in genome.

Q_End: End position of query in genome.

Pattern: Pattern sequence.

P_length: Size of the pattern.

RN: Number of pattern sequences.

Inaccuracy: Inaccuracy of the repeat (%).

L_Flank: Flank sequence at left.

Entropy_LF: Compositional entropy in left flank.

R_Flank: Flank sequence at right.

Entropy_RF: Compositional entropy in right flank.

min_RN: Minimum **RN** in all subjects.

max_RN: Maximum **RN** in all subjects.

Range: Difference between **max_RN** and **min_RN**.

Frequency: Allelic frequency of SSR in query.

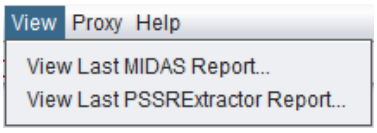
Alleles: Number of alleles.

PIC: Polymorphic Information Content $(1 - \sum_i p_i^2)$, also known as expected average of heterozygosity or Nei genetic diversity. It gives a measure of the probability that a pair of randomly chosen alleles in individuals from the same population are different.

Exceptions: Shows all the labels that correspond to exceptions (in **_specific_result.xls**). If all the subject sequences present exceptions then the labels are placed, otherwise the cell appears in blank.

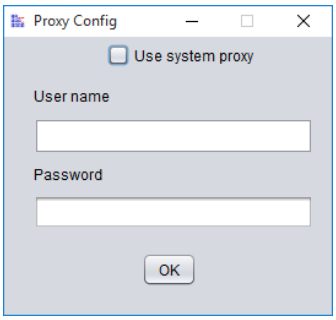
Other options in menu bar.

Report Options in View Menu:



Direct access to the latest MIDAS report.

Direct access to the latest MIDAS report.



It must be checked in case the internet connection is through a proxy. This option uses the proxy settings established by the operating system. If the internet connection through the proxy requires username and password, these fields must be filled in. The "OK" button saves the configuration.

Referencias:

1. Ortíz CMM. MIDAS: Computer application for the identification of exact and inaccurate microsatellites in genomic sequences. Revista Cubana de Informática Médica. 2018;18.
2. Ortíz CMM, Bandinez AR. Methodology for in silico mining of microsatellite polymorphic loci. Revista Cubana de Informática Médica. 2019;19.
3. BLAST Homepage and Selected Search Pages: Introducing the BLAST homepage and form elements/functions of selected search pages2016. Available from: https://blast.ncbi.nlm.nih.gov.
4. Fassler J, Cooper P. BLAST Glossary. Available from: https://blast.ncbi.nlm.nih.gov.
5. QBlast's URL API User Guide. Available from: http://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html.