Biological Forum – An International Journal **15(12): 177-180(2023)**

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

SSRmine: Python-based Command-line Tool for Precise Genomic SSR Markers' Extraction

Shbana Begam[1], Samarth Godara[2*], Ramcharan Bhattacharya[1], Rajender Parsad[2] and Sudeep Marwaha[2]
[1]ICAR-National Institute for Plant Biotechnology, New Delhi, India.
[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

(Corresponding author: Samarth Godara)*

(Received: 16 October 2023; Revised: 03 November 2023; Accepted: 19 November 2023; Published: 15 December 2023)
(Published by Research Trend)

ABSTRACT: Crop improvement, integral to global food security and environmental sustainability, has been significantly enhanced by molecular markers. Simple Sequence Repeats, or microsatellites, stand out as crucial markers due to their unique tandem repeat structure across genomes. These markers are pivotal in genetic diversity assessment, QTL mapping, and marker-assisted selection in crop breeding programs. Despite existing SSR extraction tools, challenges persist, including the need for language-specific expertise and limitations in handling complex genome data. To address these issues, we introduce SSRmine, a user-friendly Python-based tool designed for efficient SSR extraction from diverse sequence data. SSRmine employs a novel algorithm to analyze sequences systematically, extracting SSR markers of varying lengths. The presented tool's design ensures ease of use, platform independence, and adaptability, catering to researchers with varying computational expertise. The study results show that SSRmine significantly improves accessibility to SSR extraction, offering a robust solution for researchers involved in genetic variation studies. The tool's efficiency and versatility make it a valuable asset in advancing crop improvement strategies through precision breeding.

Keywords: QTL mapping, Molecular marker, Simple Sequence Repeats
Availability and implementation: SSRmine implemented in the Python programming language and available at: https://github.com/ICAR-BIOINFORMATICS/SSRmine

INTRODUCTION

Crop improvement has been a continuous endeavour to meet the ever-growing demands for food security, environmental sustainability, and resilience to changing climates. Traditional breeding methods have played a pivotal role in developing new crop varieties. However, the integration of molecular approaches, specifically molecular markers, has revolutionized the precision and efficiency of crop improvement strategies (Taheri *et al*., 2018). Molecular markers are the specific DNA sequences used to identify genetic variations associated with traits of interest. By leveraging such markers, researchers can expedite the breeding process by directly selecting plants with the desired traits at the molecular level. This not only accelerates the breeding cycle but also enhances precision in trait selection. Simple Sequence Repeats (SSRs), commonly known as microsatellites, are molecular markers commonly used for identifying genetic variations associated with traits of interest. SSRs are short, repetitive DNA sequences (motifs) scattered across the genomes of living organisms (Avise, 2012). Their unique structure, composed of tandemly repeated motifs, makes them invaluable markers for genetic analyses. The variability in the number of repeats within SSRs serves as a genetic fingerprint, offering a glimpse into the diversity and evolution of species (Kumar *et al*., 2022).

The SSRs are basically the repeated motif of one-to-base pairs in the assembled genome/transcriptome sequences. It can be classified into six categories based on the size of the repeat motif-

Table 1: Types of SSR markers.

Sr. No.	Name of SSR Markers	Example
1.	Mono Repeat	Ex. $(A)_n$ where n >= 12
2.	Di Repeat	Ex. $(AT)_n$ where n >= 6
3.	Tri Repeat	Ex. $(ATG)_n$ where n >= 5
4.	Tetra Repeat	Ex. $(ATGC)_n$ where n > = 5
5.	Penta Repeat	Ex. $(ATGGC)_n$ where n >= 5
6.	Hexa Repeat	Ex. $(ATGGCC)_n$ where n >= 5

The significance of SSR markers in agriculture cannot be overstated (Mastromoro *et al*., 2022). These markers play a pivotal role in crop improvement, breeding programs, and the development of resilient and high-yielding varieties. SSRs are utilized for:

- Genetic Diversity Assessment: SSR markers facilitate the characterization of genetic diversity within plant populations, aiding in the selection of diverse and adaptable varieties.
- QTL Mapping: SSRs contribute to the identification of Quantitative Trait Loci (QTLs) associated with desirable traits such as disease resistance, yield, and quality.
- Marker-Assisted Selection: In breeding programs, SSR markers assist in the precise selection of plants with desired traits, expediting the development of improved cultivars.

Currently, there are multiple tools available for extracting SSRs from assembled sequences, such as MISA (Thiel *et al.*, 2003), Kmer-SSR (Pickett *et al.*, 2017), MREPS (Kolpakov *et al.*, 2003), PERF (Avvaru *et al.*, 2018), SSRIT (Temnykh *et al.*, 2001), etc. However, some limitations exist with these tools, such as the requirement for expert knowledge of Perl, Python, or C language to run the system. Additionally, they may not handle complex genome-level data and may not provide amplicon sequence information.

To overcome these limitations, a user-friendly Python tool named SSRmine has been introduced. It facilitates the extraction of SSR markers from any type of provided sequences. It leverages the power of Python to enhance its performance and adaptability. SSRMine is seamlessly integrated with the language, ensuring platform independence and eliminating dependencies on other extraction tools. SSRmine boasts a user-friendly interface, making it accessible to researchers with varying levels of computational expertise. The tool simplifies the SSR marker extraction process, allowing users to navigate effortlessly through their genomic or transcriptomic data.

METHODOLOGY

SSRMine is designed and developed using the proposed algorithm presented in Figure 1, and detailed discussions are presented step by step:

Step 1: SSRmine begins by opening the input file specified in the command line argument and accessing the genomic or transcriptomic data provided for SSR extraction.

Step 2: The tool reads the sequence ID and the corresponding sequence from the input file, preparing to analyze and extract SSR markers.

Step 3: SSRmine initializes the motif length to one and starts the extraction process with the shortest motif.

Step 4: Further, the tool systematically goes through the sequence and extracts the SSRs along with associated information such as start and end positions, repeat motifs, and more.

Step 5: For each extracted SSR, the relevant information is saved in a list and create a comprehensive repository of SSR data.

Step 6: In a process, the motif length is incremented, and the extraction process (step 4) is repeated until the motif length reaches 6, ensuring a thorough analysis of SSRs of varying lengths.

Step 7: Again tool goes back to step 2 to read the sequence from the next ID, repeating the SSR extraction process for all sequences provided in the input file.

Step 8: By utilizing the Pandas library in Python, SSRMine creates a DataFrame to organize and structure the extracted SSR information.

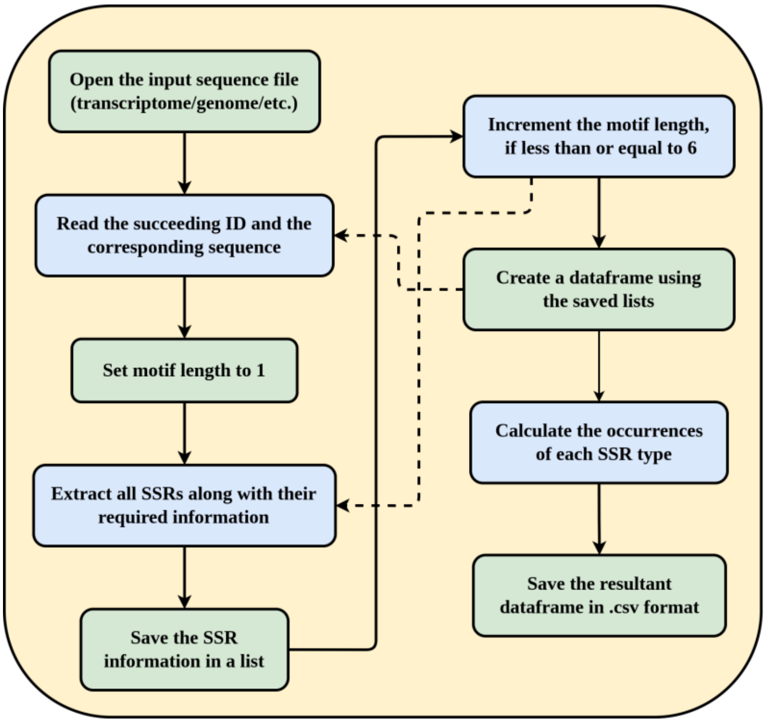
Open the input sequence file (transcriptome/genome/etc.)
Read the succeeding ID and the corresponding sequence
Set motif length to 1
Extract all SSRs along with their required information
Save the SSR information in a list
Increment the motif length, if less than or equal to 6
Create a dataframe using the saved lists
Calculate the occurrences of each SSR type
Save the resultant dataframe in .csv format

Fig. 1. Workflow diagram of the methodology used in SSRmine.

Step 9: Then it calculates the occurrences of each SSR type using a dictionary data structure, providing a count of how frequently each type of SSR appears in the dataset.

Step 10: SSRmine adds a new column to the DataFrame, containing information regarding the number of occurrences for each SSR type in the dataset.

Step 11: The complete DataFrame, now enriched with SSR occurrence information, is saved in a .csv file, providing users with a convenient and portable format for further analysis.

By following these steps, SSRMine systematically processes input assembled sequence data, extracts SSRs of various lengths, and presents the results in an organized and analyzable format. The flexibility of the tool, combined with its user-friendly design, makes it a powerful asset for researchers exploring genetic variation in sequences.

RESULTS AND DISCUSSION

Experimental Data: The validation of SSRmine has been done on 3 different plant genomes like Rice (GCF_001433935.1), Tomato (GCF_000188115.5) and Mango (GCF_011075055.1) collected from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/).

SSRMine is a user-friendly, command-line Python-based tool that makes the SSR extraction process easier for handling complex data. SSRMine provides output in CSV format, and the output file contains the total extracted SSRs along with their respective information.

When the SSRMine tool starts running, it gives the user the choice of whether to perform data preprocessing. After this decision, the tool initiates the actual processing. The output file of SSRMine provides information such as the ID of the SSR marker, the type of SSR marker based on the number of nucleotides, the start and end positions of the SSR marker on the chromosome or contig, and the sequences forming the primer set. Additionally, the output includes information about the occurrence of each SSR, indicating how many times a particular SSR occurred in the entire genome.

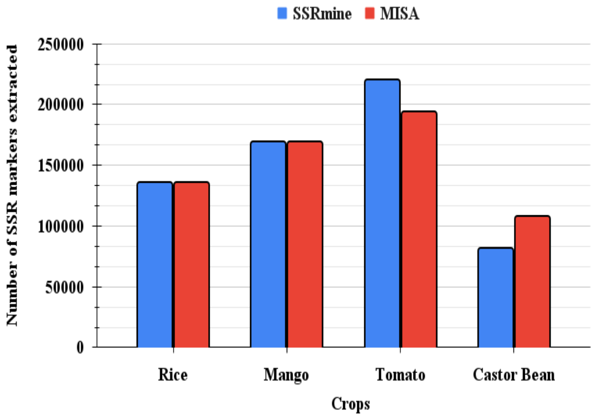

Fig. 2. Statistics of SSR markers obtained from SSRmine and MISA tools.

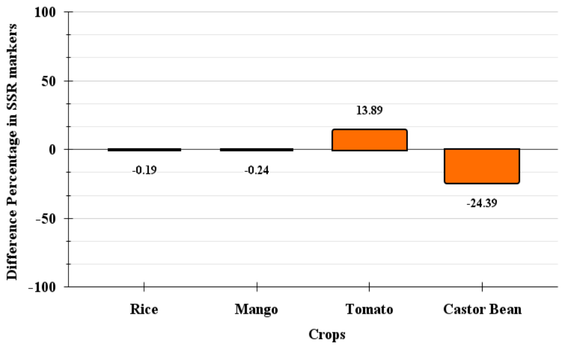

Fig. 3. Statistics regarding the difference (percentage) in the SSR markers obtained from SSRmine and MISA tools.

Fig. 2 illustrates the number of SSR markers extracted from the developed tool SSRmine and the widely used tool MISA. From the graph, it is observed that for the crops of Rice and Mango, both tools extracted a similar number of SSRs (Figure 3 indicates the difference in the percentage). Whereas with the crop of Tomato, SSRmine extracted more number of SSRs than MISA, which was the opposite scenario in the case of Castor Bean. Overall, the results show that the SSRmine successfully extracted approximately all the possible SSRs. Across different crops, SSRMine consistently demonstrates its ability to capture a substantial number of SSRs, showcasing its versatility across diverse genomic landscapes.

Feature Comparison:

Table 2 provides a feature comparison matrix of SSRmine with other existing tools, and it highlights the required programming language to run the tools, input and output formats, as well as supported platforms for each tool. Here, it has been observed that only PERF and SSRMine support all platforms. However, SSRMine exclusively provides output in CSV format, making it more user-friendly and facilitating further data processing.

Table 2: Feature comparison of SSRmine with other existing tools.

Sr. N	Program	Language	Input Format	Output Format	Supported Platform
1.	SSRmine	Python	Fasta	CSV	Linux, MacOSX, Windows
2.	Kmer-SSR	C++	Fasta	Text	Linux
3.	MISA	Perl	Fasta	Text	Linux, Windows
4.	MREPS	C	Fasta	Text	Linux
5.	SSRIT	Perl	Fasta	Text	Linux
6.	PERF	Python	Fasta	Text	Linux, MacOSX, Windows

CONCLUSIONS

In the realm of genomic research, the extraction of SSR is a critical step, providing insights into genetic diversity, evolutionary dynamics, and potential applications in crop improvement. The detailed analysis emphasizes SSRMine's effectiveness in extracting SSR markers, positioning it as a robust tool for researchers exploring genomic repetitive elements in various crops. SSRMine emerges as a potent tool for researchers delving into the intricate world of genomic repetitive elements. Its consistent performance, versatility, and user-friendly design position it as a valuable asset in unravelling genetic mysteries and advancing our understanding of plant genomes. As genomic research continues to evolve, SSRMine stands as a testament to the power of innovative tools in propelling the field forward.

FUTURE SCOPE

The future development and evolution of SSRmine will likely involve a combination of technical refinements, interdisciplinary collaborations, and a keen responsiveness to the evolving needs of researchers in genomics and molecular biology. Further primer design module can be incorporated in this tool.

Acknowledgement. The authors are thankful to ICAR-NIPB, New Delhi and ICAR-IASRI, New Delhi for providing the resources to conduct the study.
Conflict of interest. None.

REFERENCES

Avise, J. C. (2012). *Molecular markers, natural history and evolution*. Springer Science & Business Media.

Avvaru, A. K., Sowpati, D. & Mishra, R. K. (2018). PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences, *Bioinformatics*, *34*(6), 943–948.

Kolpakov, R., Bana, G., & Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic acids research*, *31*(13), 3672-3678.

Kumar, S. J., Susmita, C., Sripathy, K. V., Agarwal, D. K., Pal, G., Singh, A. N., ... & Simal-Gandara, J. (2022). Molecular characterization and genetic diversity studies of Indian soybean (*Glycine max* (L.) Merr.) cultivars using SSR markers. *Molecular Biology Reports*, 1-12.

Lakshmikanth, M. R., Mishra, A., Singh, P., Jagadev, P. N., Pradhan, B., Samal, K.C., Mohanty S. & Verma R.L. (2022). Molecular Screening of Rice (*Oryza sativa* L.) Genotypes for Bacterial Leaf Blight Resistance Genes using Trait-based SNP Markers. *Biological Forum – An International Journal*, *14*(4a), 826-829.

Mastromoro, G., Guadagnolo, D., Khaleghi Hashemian, N., Marchionni, E., Traversa, A., & Pizzuti, A. (2022). Molecular Approaches in Fetal Malformations, Dynamic Anomalies and Soft Markers: Diagnostic Rates and Challenges—Systematic Review of the Literature and Meta-Analysis. *Diagnostics*, *12*(3), 575.

Pickett, B. D., Miller, J. B. & Ridge, P. G. (2017). Kmer-SSR: a fast and exhaustive SSR search algorithm. *Bioinformatics*, *33*(24), 3922-3928.

Taheri, S., Lee Abdullah, T., Yusop, M. R., Hanafi, M. M., Sahebi, M., Azizi, P., & Shamshiri, R. R. (2018). Mining and development of novel SSR markers using next generation sequencing (NGS) data in plants. *Molecules*, *23*(2), 399.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., & McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, *11*(8), 1441-1452.

Thiel, T., Michalek, W., Varshney, R., & Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics*, *106*, 411-422.

How to cite this article: Shbana Begam, Samarth Godara, Ramcharan Bhattacharya, Rajender Parsad and Sudeep Marwaha (2023). SSRmine: Python-based Command-line Tool for Precise Genomic SSR Markers' Extraction. *Biological Forum – An International Journal, 15*(12): 177-180.