Artículo original

Aplicación Web para el Minado In Silico de Loci Polimórficos de Microsatélites

Web Application for In Silico Mining of Microsatellite Polymorphic Loci

Carlos M. Martínez Ortiz[1*] 0000-0002-1852-3905
Alejandro Rivero Bandinez[1] 0000-0003-2396-346X

[1] Universidad de Ciencias Médicas de La Habana, ICPB "Victoria de Girón", Departamento de Bioquímica, La Habana, Cuba.

Autor para la correspondencia: cmmoprogram@gmail.com; Tel: +5353443074

RESUMEN

Esta nota de aplicación presenta una nueva plataforma web diseñada para la minería in silico de loci de microsatélites polimórficos (SSR), ofreciendo una mayor eficiencia en comparación con métodos tradicionales. La herramienta integra MISA, BLAST y el script PSSR-Extractor en un flujo de trabajo continuo, permitiendo la detección y análisis automatizado de SSRs a partir de secuencias genómicas. A diferencia de otras herramientas que dependen de bases de datos preexistentes, esta plataforma procesa secuencias específicas de interés, utilizando BLAST para buscar secuencias similares y poder evaluar el polimorfismo.

Validada en 23 genomas de Mycobacterium tuberculosis, el sistema identificó más de 4400 loci SSR y extrajo 414 *loci* polimórficos no redundantes. Una comparación con herramientas similares, como PSSRdt, muestra las ventajas de la plataforma en términos de flexibilidad de entrada, informes de polimorfismo y facilidad de uso. Esta aplicación web acelera la minería de SSR y proporciona valiosas perspectivas sobre la diversidad genética, convirtiéndose en un recurso poderoso para la investigación en genética de poblaciones, biología evolutiva y epidemiología.

Palabras clave: SSR; microsatélites; loci polimórficos; biología computacional.

ABSTRACT

This application note introduces a novel web platform designed for in silico mining of polymorphic microsatellite loci (SSRs), offering improved efficiency over traditional methods. The tool integrates MISA, BLAST, and the PSSR-Extractor script into a seamless

workflow, enabling the automated detection and analysis of SSRs from genomic sequences. It supports both standalone and web-based use, with compatibility for various input formats (FASTA, GBBF). Unlike other tools that rely on pre-existing databases, this platform processes specific sequences of interest, using BLAST to assess polymorphism. Validated on 23 Mycobacterium tuberculosis genomes, the system identified over 4,400 SSR loci and extracted 414 non-redundant polymorphic loci. A comparison with similar tools, such as PSSRdt, shows the platform's advantages in input flexibility, polymorphism reporting, and ease of use. This web application accelerates SSR mining and provides valuable insights into genetic diversity, making it a powerful resource for research in population genetics, evolutionary biology, and epidemiology.
Keywords: SSR; microsatellite; polymorphic loci; computational biology.

Recibido: 30/10/2024
Aprobado: 13/11/2024

Introducción

Los microsatélites, o repeticiones de secuencias simples (SSR), son secuencias de ADN cortas y repetitivas distribuidas en genomas de organismos procariotas y eucariotas. Estas secuencias, consistentes en motivos repetidos de 1 a 6 pares de bases, son altamente polimórficas, lo que las convierte en valiosos marcadores genéticos para una variedad de aplicaciones, incluyendo estudios evolutivos, genética de poblaciones y seguimiento epidemiológico (Ellegren, 2004). La naturaleza polimórfica de los SSR surge de sus altas tasas de mutación, especialmente en el número de unidades repetidas, lo que genera una diversidad genética significativa incluso dentro de las especies.

La minería in silico de SSRs polimórficos consta de dos etapas generales: primero, la detección de SSRs, para la cual se han desarrollado una amplia gama de aplicaciones; y segundo, la determinación de si estos SSRs muestran polimorfismos en el número de copias. Esto requiere comparar sus secuencias con repositorios de especies relacionadas o secuencias de la misma especie. Para la validación experimental de los polimorfismos, se diseñan cebadores que flanquean las regiones SSR, utilizando software como Primer-BLAST (Ye et al., 2012). [(2)] Los cebadores se sintetizan y amplifican mediante PCR, y los resultados se analizan a través de electroforesis o secuenciación.

Tradicionalmente, la detección de SSR y el análisis de polimorfismo requerían técnicas experimentales laboriosas, como la electroforesis en gel o la secuenciación capilar. Estos métodos, aunque efectivos, son lentos, consumen muchos recursos y no son escalables para grandes conjuntos de datos. Para abordar estas limitaciones, se han desarrollado enfoques computacionales que permiten la identificación y el análisis rápidos de SSR en todo el

genoma. Entre estas herramientas, MISA (MIcroSAtellite identification tool, por sus siglas en inglés) es un software ampliamente utilizado que ha simplificado enormemente la detección de SSR en secuencias genómicas (Beier et al., 2017).[3] Además, la integración de herramientas como BLAST (Basic Local Alignment Search Tool, por sus siglas en inglés), que identifica secuencias homólogas comparando datos de entrada con bases de datos de nucleótidos o proteínas (Altschul et al., 1990),[4] y scripts personalizados, como el PSSR-Extractor, puede proporcionar métodos eficientes in silico para descubrir loci SSR polimórficos en múltiples genomas. Los avances recientes en las herramientas de detección de SSR han optimizado tanto la ve*loci*dad como la precisión de la minería in silico de SSR, especialmente en estudios genómicos a gran escala (Shilpa y Gopalakrishna, 2021). [6]

Pocos intentos se han hecho para identificar automáticamente microsatélites polimórficos explotando la redundancia en sus secuencias. Hasta ahora, existen pocas herramientas para la genotipificación de microsatélites capaces de gestionar grandes volúmenes de secuencias (Cantarella et al., 2015).[7] Ejemplos de software diseñados para este propósito incluyen PolySSR (Tang et al., 2008), PSR (Cantarella et al., 2015), PolyMorphPredict (Das et al., 2019) y PSSRdt (Tian et al., 2019).[8-11] La mayoría de estas herramientas minan datos de secuencias de bases de datos de etiquetas de secuencias expresadas (EST), es decir, regiones transcritas del genoma en humanos u organismos modelo, y emplean pipelines que combinan scripts personalizados con software de terceros con uso generalizado.

Estas implementaciones difieren en los formatos y tamaños de las secuencias procesadas, los métodos de detección de SSR (si están incluidos), y el tipo de SSR (perfectos o aproximados). También pueden variar en cuanto al diseño de cebadores, la dependencia de la plataforma computacional y si son aplicaciones de escritorio o basadas en la web. Además, la combinación compleja de scripts personalizados y software de uso generalizado, junto con representaciones simplificadas de las etapas involucradas sin especificar las reglas exactas a seguir en cada paso, hace que estos algoritmos sean metodológicamente opacos y difíciles de reproducir o mejorar para terceros.

La aplicación web que presentamos aprovecha estas herramientas para automatizar todo el proceso de minería de SSR, ofreciendo una interfaz fácil de usar que elimina la necesidad de flujos de trabajo manuales complejos. Esta plataforma no solo acelera la detección de SSR, sino que también facilita la identificación de polimorfismos en los genomas, proporcionando así valiosas perspectivas sobre la diversidad genética. El enfoque se basa en trabajos previos de Martínez Ortiz y Rivero Bandinez (Martínez y Rivero, 2019), quienes describieron una metodología fundamental para la minería *in silico* de SSR. En este trabajo, ampliamos su marco introduciendo un sistema integrado y completo diseñado para manejar grandes conjuntos de datos genómicos con mayor precisión y eficiencia.

Materiales y Métodos

La aplicación web consiste en varios módulos integrados diseñados para agilizar el proceso de identificación de microsatélites y sus polimorfismos en los genomas.

1. Módulo de Detección de SSR
    - Herramienta: MISA (Script en Perl)
    - Proceso: Los usuarios suben secuencias genómicas en formato FASTA, las cuales son procesadas por MISA para detectar SSR con secuencias flanqueantes. La detección se basa en parámetros definidos por el usuario, como la longitud mínima de repetición y tipos específicos de motivos repetitivos. El resultado es un archivo MultiFASTA que marca las regiones de repetición identificadas. Este archivo se utiliza posteriormente para el análisis en el siguiente módulo.
2. Módulo de Detección de Polimorfismo
    - Herramientas: BLAST y PSSR-Extractor (Java)
    - Proceso: El archivo MultiFASTA generado en el paso anterior se usa como entrada para BLAST (acceso remoto), que busca secuencias homólogas dentro de una base de datos especificada por el usuario. Los resultados de BLAST son procesados por el script PSSR-Extractor, que identifica loci polimórficos analizando variaciones en el número de repeticiones a lo largo de diferentes secuencias genómicas. El resultado final es un informe detallado que los usuarios pueden descargar, conteniendo datos sobre la diversidad alélica y otros indicadores clave de polimorfismo.

Interfaz del Usuario y Características

La figura 1 muestra una captura de pantalla de la interfaz del formulario de entrada de la aplicación web. El formulario incluye campos para cargar secuencias genómicas en formato FASTA y establecer parámetros de análisis. Los usuarios pueden especificar parámetros de detección para el proceso de identificación de SSR, como la longitud mínima de la unidad de repetición, el número de repeticiones y los tipos de motivos a detectar (por ejemplo, repeticiones de di-, tri- y tetranucleótidos, etc.). Además, el formulario proporciona campos para ingresar parámetros de búsqueda de BLAST, incluyendo la base de datos objetivo, el umbral de *valor e* y la configuración de alineación para identificar secuencias homólogas. Un botón de "Enviar" inicia el análisis, y una barra de progreso muestra el estado de la tarea en curso.

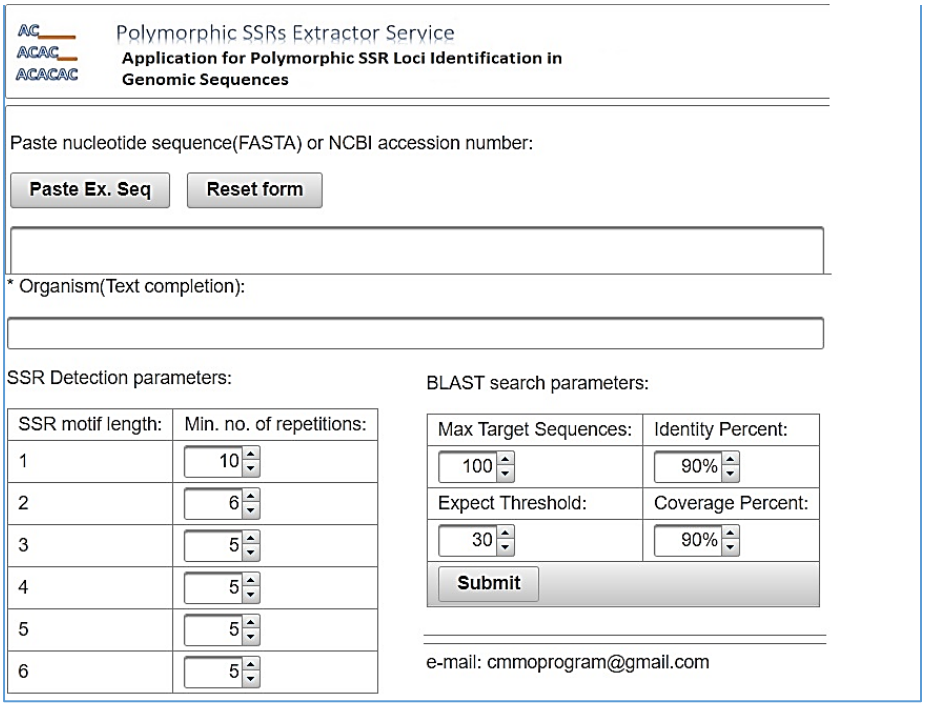
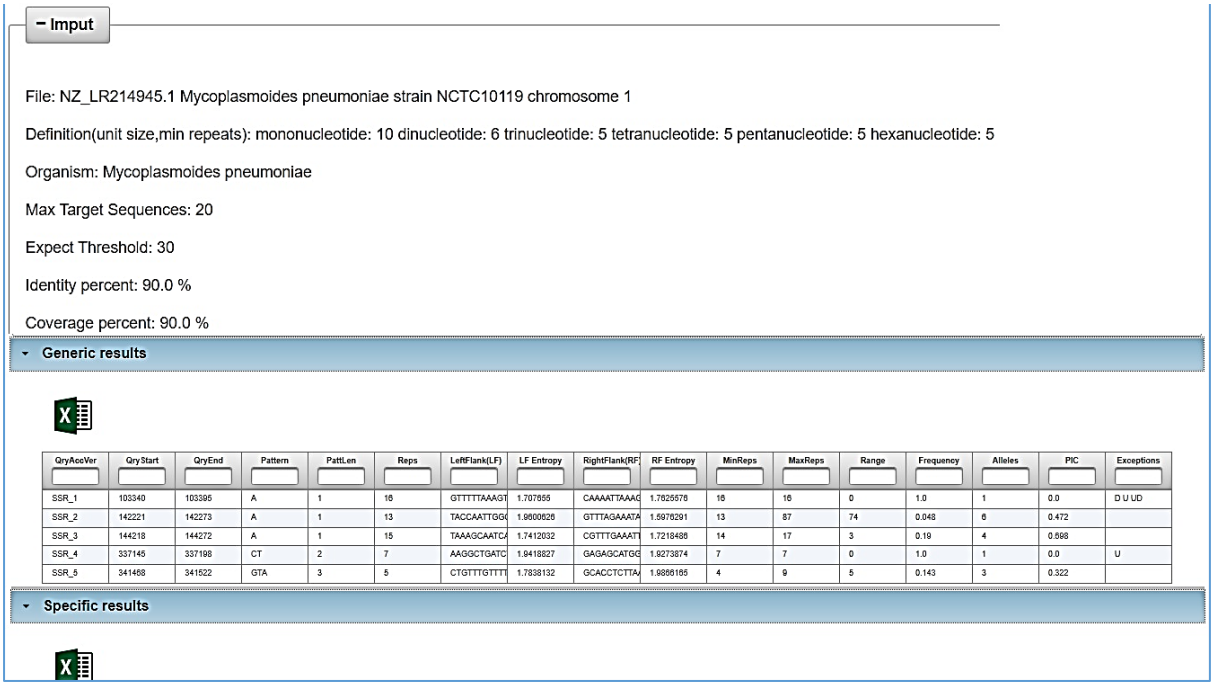


Fig. 1- Interfaz de la aplicación web para la detección de SSR y los parámetros de búsqueda de BLAST.

La figura 2 muestra una captura de pantalla de la interfaz de tablas de resultados de la aplicación web. Las tablas incluyen columnas para el identificador SSR, la posición genómica, la unidad de repetición, el recuento de repeticiones e indicadores de polimorfismo, como la frecuencia alélica y el PIC (Polymorphic Information Content, por sus siglas en inglés). Cada fila representa un locus SSR identificado, y los usuarios pueden ordenar y filtrar los resultados según diferentes criterios. La interfaz también ofrece opciones para descargar los resultados en formato Excel.



QryAccVer	QryStart	QryEnd	Pattern	PattLen	Reps	LeftFlank(LF)	LF Entropy	RightFlank(RF	RF Entropy	MinReps	MaxReps	Range	Frequency	Alleles	PIC	Exceptions
SSR_1	103340	103395	A	1	16	GTTTTTAAAGT	1.707855	CAAAATTAAAC	1.7625576	16	16	0	1.0	1	0.0	D U UD
SSR_2	142221	142273	A	1	13	TACCAATTGG	1.9600626	GTTTAGAAATA	1.5976291	13	87	74	0.048	6	0.472	
SSR_3	144218	144272	A	1	15	TAAAGCAATCA	1.7412032	CGTTTGAAATT	1.7218486	14	17	3	0.19	4	0.598	
SSR_4	337145	337198	CT	2	7	AAGGCTGATC	1.9418827	GAGAGCATGG	1.9273874	7	7	0	1.0	1	0.0	U
SSR_5	341468	341522	GTA	3	5	CTGTTTGTTTT	1.7838132	GCACCTCTTA	1.9886165	4	9	5	0.143	3	0.322	

Fig. 2-Interfaz de resultados que muestra datos de polimorfismo SSR y salidas de análisis.

Resultados y Discusión

La aplicación web fue validada utilizando 23 genomas completos de Mycobacterium tuberculosis de la base de datos NCBI RefSeq, identificando exitosamente 4433 *loci* SSR. De estos, se extrajeron 414 *loci* polimórficos no redundantes. Medidas clave, incluyendo la frecuencia alélica y el contenido de información polimórfica (PIC), confirmaron la precisión de los polimorfismos identificados.

La aplicación web ofrece avances significativos en la detección automatizada de *loci* de microsatélites polimórficos (SSR) mediante métodos computacionales. Al aprovechar herramientas bien establecidas como MISA y BLAST junto con el script PSSR-Extractor, este sistema integra múltiples funciones en una única interfaz fácil de usar y accesible a investigadores en diversos dominios de la genética y bioinformática.

La aplicación tiene aspectos distintivos en comparación con otros métodos *in silico* reportados en la literatura. La **Tabla 1** presenta una comparación entre *Polymorphic SSR Extractor*, la herramienta que presentamos, y PSSRdt (Tian et al., 2019), que fue elegida debido a su lanzamiento reciente y sus principios algorítmicos claramente descritos.

Tabla 1-- Comparación entre PSSRdt y el Polymorphic SSR Extractor.

Criterio	PSSRdt (Tian et al., 2019)	Polymorphic SSR Extractor
Distribución	Aplicación independiente (combinación de scripts)	Servicio web y aplicación independiente (con interfaz gráfica)
Formato de entrada	FASTA	FASTA, GBBF, entrada libre
Base de datos de búsqueda	Transcriptoma (EST-DB), aplicación independiente	Nucleótidos (BLAST), remoto
Detección de SSR	Script basado en MISA	Script basado en MISA
Principio algorítmico	Busca "en entrada de base de datos EST"	Busca "en secuencia problema de entrada"
Reporte de polimorfismo	Número de repeticiones por locus	Número de alelos y PIC

Entre los criterios mostrados en la Tabla 1, resaltamos el principio algorítmico. A diferencia de la mayoría de las otras aplicaciones diseñadas para estos propósitos que toman una base de datos como entrada, típicamente una base de datos EST, la aplicación Polymorphic SSR Extractor utiliza una secuencia específica como entrada. Esta secuencia representa un interés de investigación particular. Este enfoque es más intuitivo y natural en comparación con los métodos anteriores, ya que los investigadores suelen comenzar con una secuencia de interés (un EST, un fragmento de genoma o un genoma completo) y buscan identificar marcadores de microsatélites dentro de ella y determinar cuáles pueden ser polimórficos.

Polymorphic SSR Extractor utiliza este enfoque, realizando la detección de polimorfismos mediante la consulta de una base de datos remota, estandarizada y curada (en este caso, BLASTdb, NCBI Resource Coordinators, 2018). Después del análisis de datos, proporciona resultados sobre los marcadores SSR de la secuencia de entrada y su nivel teórico de polimorfismo (PIC). Esto permite a los investigadores preseleccionar los marcadores más informativos para experimentos in vitro posteriores.

La plataforma aborda limitaciones claves de los métodos tradicionales de detección de SSR, que históricamente han dependido de enfoques experimentales laboriosos. Aunque efectivos, estos métodos no son escalables para conjuntos de datos grandes y a menudo implican flujos de trabajo complejos. El marco computacional que hemos desarrollado no solo automatiza la detección de SSR, sino que también acelera el análisis de polimorfismos en múltiples genomas. Esto es particularmente beneficioso en estudios que requieren la minería de SSR de alto rendimiento, como la genética de poblaciones y la epidemiología.

La validación en genomas de Mycobacterium tuberculosis demuestra la fiabilidad de la plataforma, con la identificación de 4433 SSR y 414 *loci* polimórficos no redundantes. Estos resultados son consistentes con la diversidad genética observada en la especie, ya que la alta tasa de mutación de los SSR contribuye a una variación alélica sustancial. Indicadores clave de polimorfismo, como la frecuencia alélica y el contenido de información polimórfica (PIC), confirmaron la precisión de los resultados, destacando el potencial del sistema para identificar SSR que pueden servir como valiosos marcadores genéticos en diversas especies.

Es una ventaja crítica la capacidad de la aplicación para manejar conjuntos de datos a gran escala sin sacrificar la precisión. En estudios comparativos, encontramos que nuestro enfoque integrado ofrece una eficiencia mejorada en comparación con herramientas autónomas, especialmente en la gestión de grandes volúmenes de datos donde la intervención manual ralentizaría significativamente el proceso.

Una de las fortalezas de la aplicación radica en su accesibilidad. Al proporcionar una interfaz gráfica de usuario, el sistema reduce la necesidad de conocimientos profundos de programación, lo que permite que un rango más amplio de investigadores la utilicen de manera efectiva. Los usuarios pueden personalizar parámetros clave, como la longitud de las unidades de repetición, los umbrales de detección y los criterios de búsqueda de BLAST, permitiendo análisis personalizados según los objetivos específicos del estudio.

Esta personalización es particularmente relevante en estudios evolutivos, donde diferentes organismos pueden exhibir niveles variados de diversidad de SSR. Por ejemplo, al permitir que los usuarios ajusten la sensibilidad de la detección de SSR, la plataforma garantiza que tanto las regiones altamente variables como las conservadas del genoma puedan ser analizadas.

A pesar de estas fortalezas, existen ciertas limitaciones. Actualmente, la aplicación depende del PSSR-Extractor, que, aunque efectivo, podría beneficiarse de una mayor optimización, particularmente en el manejo de estructuras genómicas más complejas donde los polimorfismos SSR pueden estar enmascarados por la homología de secuencia u otras características genómicas. Además, la versión actual solo soporta un rango limitado de bases de datos para la identificación de secuencias homólogas, y la expansión de este rango mejoraría su aplicabilidad a un espectro más amplio de organismos.

El trabajo futuro se centrará en mejorar la eficiencia algorítmica de la línea de detección de SSR, así como en integrar bases de datos genómicas adicionales para ampliar la utilidad de la plataforma. También se planea una validación adicional en genomas más diversos, particularmente de especies eucariotas con estructuras repetitivas más complejas.

Conclusiones

Este estudio presenta una novedosa y eficiente aplicación web para la minería in silico de *loci* SSR polimórficos, ofreciendo mejoras significativas en comparación con métodos tradicionales y herramientas autónomas. Al integrar herramientas establecidas como MISA, BLAST y el script PSSR-Extractor en una plataforma única y fácil de usar, la aplicación automatiza tanto la detección de SSR como el análisis de polimorfismo, reduciendo la necesidad de flujos de trabajo experimentales laboriosos.

Polymorphic SSR Extractor se destaca al permitir a los investigadores analizar secuencias específicas de interés, en lugar de depender de bases de datos completas, y al utilizar bases de datos remotas y estandarizadas, como BLAST, para la detección de polimorfismo. Comparado con otras herramientas recientes como PSSRdt, nuestra plataforma ofrece una mayor compatibilidad de entrada, un informe de polimorfismo más completo (incluyendo recuentos de alelos y PIC), y flexibilidad tanto en el uso autónomo como en la web.

Validada en 23 genomas de Mycobacterium tuberculosis, la aplicación identificó con éxito más de 4400 loci SSR y extrajo 414 loci polimórficos no redundantes con alta precisión. Este

enfoque proporciona a los investigadores una herramienta poderosa y personalizable para la minería de SSR de alto rendimiento, particularmente en los campos de genética de poblaciones, biología evolutiva y epidemiología.

El desarrollo futuro se enfocará en mejorar la eficiencia algorítmica, ampliar la compatibilidad con bases de datos y validar la herramienta en genomas más complejos. Al facilitar tanto la detección de SSR como el análisis de polimorfismo previo a los experimentos, esta plataforma de código abierto tiene el potencial de avanzar la investigación en genómica computacional y análisis de microsatélites.

Disponibilidad e Implementación

La aplicación está disponible en: https://misapssrextractor.sp1.br.saveincloud.net.br. El código fuente y recursos adicionales están disponibles en GitHub: https://github.com/Thebx1994/MISAPolymorphicSSRsExtractorService.git.

Referencias

1. Ellegren, H. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics [Internet]. 2004;5(6):435-45. Disponible en: http://doi.org/10.1038/nrg1348

2. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics [Internet]. 2012;13:134. Disponible en: http://doi.org/10.1186/1471-2105-13-134

3. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics [Internet]. 2017;33(16):2583-5. Disponible en: http://doi.org/ 10.1093/bioinformatics/btx198

4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology [Internet]. 1990;215(3):403-10. Disponible en: http://doi.org/10.1016/S0022-2836(05)80360-2

5. Shilpa M, Gopalakrishna T. Development of a high-throughput *in silico* SSR mining tool for crop genomes. Computational Biology and Chemistry [Internet]. 2021;95:107603. Disponible en: http://doi.org/ 10.1016/j.compbiolchem.2021.107603

6. Cantarella CD, Mirisola M, Sciacca C, Bivona L, Cavallaro A. PSR: A novel tool for functional annotation of the SSRs in expressed sequences of eukaryotes. Computational Biology and Chemistry [Internet]. 2015;56:119-26. Disponible en: http://doi.org/10.1016/j.compbiolchem.2015.04.002

7. Tang Q, Han J, Wu J, Zhang Y. PolySSR: a pipeline poly SSR for automated SSR discovery and marker development. Genomics [Internet]. 2008;92(4):241-7. Disponible en: http://doi.org/ 10.1016/j.ygeno.2008.06.009

8. Das B, Sahoo L, Debata N. PolyMorphPredict: a tool for large-scale SSR polymorphism detection in EST sequences. BMC Genomics [Internet]. 2019;20:245. Disponible en: http://doi.org/10.1186/s12864-019-5585-9

9. Tian H, Zhou R, He H, Hu M, Jiang J. PSSRdt: a tool for polymorphic SSR detection using transcriptome data. Bioinformatics [Internet]. 2019;35(1):158-60. Disponible en: http://doi.org/ 10.1093/bioinformatics/bty617

10. Martínez Ortiz CM, Rivero Bandinez, A. Methodology for in silico mining of microsatellite polymorphic *loci*. Revista Cubana de Informática Médica [Internet]. 2019;11(1):2-17. Disponible en: https://revinformatica.sld.cu/index.php/rcim/article/view/325

11. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research [Internet]. 2018;46(D1):D8-D13. Disponible en: http://doi.org/ 10.1093/nar/gkx1095

Conflicto de interés

Los autores declaran que no existen conflictos de interés.

Declaración de autoría

Carlos M. Martínez Ortiz: Conceptualización, Metodología, Software, Investigación, Redacción, Validación.

Alejandro Rivero Bandinez: Software, Validación, Edición del manuscrito.