

Pontificia Universidad Católica Del Ecuador
Facultad De Ingeniería
Maestría en Biología Computacional

Plan De Titulación marzo 2024:

**Metodología Para Una Herramienta De Minería De Datos Para La Evaluación De
La Novedad En Secuencias Genómicas**

Autor: BQF. Gabriela Valarezo Álvarez

Contenido

1. Datos Generales	3
1.1. Título del proyecto	3
1.2. Autor	3
2. Resumen Ejecutivo	3
3. Descripción detallada de la propuesta	4
3.1. Justificación	4
3.2. Planteamiento del problema	5
3.3. Contextualización del tema u objeto	5
3.4. Objetivos	6
3.4.1. Objetivo general	6
3.4.2. Objetivos específicos	6
3.5. Marco Teórico y Conceptual del plan	6
3.5.1. Antecedentes o marco referencial	6
3.5.2. Marco teórico	7
3.5.3. Marco conceptual	10
3.6. Metodología y Técnicas	10
3.6.1. Revisión del Estado del Arte	10
3.6.2. Determinación de Requerimientos Técnicos y Funcionales	11
3.6.3. Propuesta de Metodología para el Desarrollo	11
3.7. Esquema básico de contenidos	11
3.8. Limitaciones del estudio	12
3.9. Resultados esperados	12
4. Cronograma	13
5. Bibliografía	13

1. Datos Generales

1.1. Título del proyecto

Metodología para una Herramienta de Minería de Datos para la Evaluación de la Novedad en Secuencias Genómicas

1.2. Autor

BQF. Gabriela Valentina Valarezo Álvarez

2. Resumen Ejecutivo

En el campo de la biología computacional, la evaluación de la novedad de secuencias genómicas es un proceso crítico para determinar su potencial patentabilidad. Actualmente, aunque existen herramientas que facilitan este trabajo, la adaptación a los requerimientos locales es un proceso laborioso y requiere una comparación manual exhaustiva con las secuencias ya patentadas y publicadas en bases de datos genómicas. Es necesario adaptar la herramienta a las necesidades específicas de la región. Esto incluye enfocarse en las secuencias genómicas relevantes para la biodiversidad local y los desafíos específicos de la biotecnología en Ecuador. Además, incorporar el idioma español y las normativas locales de propiedad intelectual, facilitando su uso por profesionales ecuatorianos

Considerando los nuevos enfoques de investigación aplicada que se están desarrollando en el país, la propuesta metodológica de una herramienta de minería de datos que evalúe automáticamente la novedad de secuencias genómicas comparándolas con las secuencias existentes patentadas y publicadas ofrecerá múltiples beneficios. Facilitará la identificación rápida y precisa de secuencias potencialmente patentables, reducirá significativamente el tiempo y los costos asociados con la investigación preliminar de patentes, al mismo tiempo que mejorará la competitividad de los investigadores y empresas en el ámbito de la biotecnología en Ecuador.

3. Descripción detallada de la propuesta

3.1. Justificación

Las invenciones que utilizan recursos genéticos (RG) representan una categoría única dentro de la protección de la propiedad intelectual, en la que convergen elementos de la biotecnología moderna, el fitomejoramiento, y la transformación digital. La creciente utilización de los RG en la innovación de las ciencias de la vida resalta la necesidad de un enfoque transversal que contemple la interacción entre la propiedad intelectual y los RG. Aunque los RG en su estado natural no son patentables, las invenciones derivadas de estos sí pueden obtener protección legal, lo que destaca su relevancia para la innovación tecnológica y científica ¹.

Ecuador reconoce la importancia de la propiedad intelectual en el acceso y distribución de beneficios derivados de los RG, facilitando que personas obtengan reconocimiento de derechos por desarrollar procesos o productos innovadores a partir de estos recursos, con potencial para ser producidos y comercializados a escala industrial ².

Una breve búsqueda en el estado del arte sobre publicaciones realizadas en Ecuador que mencionen las palabras genoma, o secuencia genética o biotecnologías y sus variantes muestra 762 resultados, cifra modesta en comparación con países industrializados. Sin embargo, este volumen refleja una inversión significativa de tiempo y recursos en investigaciones relevantes.

La expansión continua de la innovación biotecnológica subraya la necesidad de identificar y patentar nuevas secuencias genómicas como una pieza clave para el progreso científico y comercial. Ecuador, con su rica biodiversidad, presenta un potencial considerable para innovar en áreas como la medicina, la agricultura y la industria. Sin embargo, la falta de herramientas automatizadas adaptadas a las necesidades locales retrasa la identificación y evaluación de la novedad de estas secuencias ³.

Por lo tanto, la propuesta metodológica de una herramienta que permita realizar esta evaluación de manera eficiente, tomando en cuenta tanto la biodiversidad única del país como las normativas locales de propiedad intelectual, es esencial para impulsar la competitividad de los investigadores y empresas ecuatorianas en el ámbito internacional.

3.2. Planteamiento del problema

Una revisión en el estado de la técnica, utilizando los mismos parámetros detallados anteriormente, mostraron que actualmente no existe ningún registro de patente de inventores o aplicantes ecuatorianos. En contraste, se obtuvieron 1409 resultados de patentes protegidas en el territorio nacional. Esto nos indica que el Ecuador es una jurisdicción atractiva para la protección intelectual en estas temáticas, sin embargo, las invenciones son de origen extranjero.

Con estos antecedentes, cabe destacar que, este vacío en la innovación biotecnológica en Ecuador no es producto de falta de investigación, ya que el estado del arte cuenta con resultados contrastantes. Además, tampoco es causado por motivos de trabas legales y burocráticas, ya que observamos que invenciones extranjeras han logrado completar el proceso de protección intelectual en nuestro país. Por lo tanto, el desfase entre estado del arte y el estado de la técnica debe encontrarse en el proceso de gestión del conocimiento y transferencia tecnológica.

Si bien existen plataformas para la comparación entre secuencias propuestas y secuencias patentadas, el proceso de análisis de la novedad de secuencias genómicas en Ecuador es laborioso y propenso a errores, debido a que estas herramientas no están adaptadas a las particularidades locales. La relativamente reciente aplicación del formato ST.26 como un estándar que establece los requisitos para presentar listados de secuencias en las solicitudes de patentes, incrementa un requerimiento adicional que debe considerarse en este proceso ⁴.

Esta situación limita la capacidad de los investigadores para identificar rápidamente secuencias potencialmente patentables, retrasando el avance en áreas clave de la biotecnología. Además, la falta de integración con las normativas locales y el idioma español dificulta el acceso y uso de estas herramientas por parte de la comunidad científica ecuatoriana.

3.3. Contextualización del tema u objeto

La biodiversidad ecuatoriana es una fuente rica de secuencias genómicas que podrían tener un alto valor en biotecnología. Sin embargo, la identificación y evaluación de la novedad de estas secuencias son procesos que requieren herramientas avanzadas de minería de datos, capaces de manejar grandes volúmenes de información y realizar comparaciones precisas con secuencias existentes. En el contexto ecuatoriano, es fundamental desarrollar una herramienta que no solo se ajuste a las necesidades técnicas,

sino que también esté alineada con las regulaciones locales y las particularidades lingüísticas, para facilitar su adopción y uso efectivo por parte de investigadores y empresas ^{2,3}.

3.4. Objetivos

3.4.1. Objetivo general

Establecer una metodología para el desarrollo de una herramienta automatizada de minería de datos que evalúe la viabilidad de protección de una secuencia genética al determinar las diferencias entre la secuencia del recurso natural no patentable y la secuencia propuesta, además determinar la novedad realizando comparaciones con las secuencias ya patentadas y publicadas en bases de datos genómicas, con el fin de facilitar la determinación de su potencial patentabilidad

3.4.2. Objetivos específicos

- Realizar un análisis exhaustivo del estado del arte de las metodologías existentes en minería de datos y bioinformática, enfocadas en la evaluación de la novedad y viabilidad de protección de secuencias genéticas, identificando las técnicas más avanzadas y relevantes aplicables al contexto de recursos naturales no patentables.
- Determinar los requerimientos técnicos y funcionales necesarios para el desarrollo de una herramienta automatizada, que integre capacidades de comparación y análisis de secuencias genéticas con bases de datos públicas y privadas, asegurando su alineación con los estándares de patentabilidad y protección de la propiedad intelectual.
- Diseñar y proponer una metodología robusta para la creación de la herramienta de minería de datos, incluyendo la especificación de algoritmos, modelos de datos y procesos de validación, de manera que sea viable su implementación por parte de desarrolladores de software especializados en el área.

3.5. Marco Teórico y Conceptual del plan

3.5.1. Antecedentes o marco referencial

El Estado ecuatoriano ejerce soberanía sobre la biodiversidad y los recursos genéticos, y su administración y gestión se realizan con responsabilidad intergeneracional. La conservación de la biodiversidad y su patrimonio genético es de interés público. Además, las comunidades locales tienen derecho a participar

en los beneficios derivados del acceso a los recursos genéticos y a recibir indemnizaciones por los perjuicios sociales, culturales y ambientales que puedan causarles ².

Para el análisis de la novedad de una secuencia genética se requieren mecanismos especializados o de pago, los cuales generalmente los disponen consultoras legales dedicadas a la evaluación de esta novedad, quienes tienen altas tarifas para el proceso. De forma alternativa, el análisis de la pertinencia en propiedad intelectual de una determinada secuencia se realiza de forma manual, involucrando tiempo y recursos.

Existen herramientas online que permiten la comparación entre secuencias, facilitando el proceso de evaluación sea con finalidad académica o innovadora. Sin embargo, la mayoría de estas herramientas han sido desarrolladas en contextos internacionales, sin considerar las necesidades específicas de países con alta biodiversidad, como Ecuador. Además, el primer paso para verificar esta viabilidad es determinar la diferencia entre el recurso genético original y la secuencia propuesta. El proyecto pretende abordar esta brecha que genera atrasos o desviaciones en la innovación biotecnológica en Ecuador ^{2,4}.

3.5.2. Marco teórico

3.5.2.1. Recursos Genéticos y Propiedad Intelectual

Los recursos genéticos se refieren a cualquier material de origen vegetal, animal, microbiano o de otro tipo que contenga unidades funcionales de la herencia y que tenga valor real o potencial. En el contexto de la propiedad intelectual (PI), los RG son importantes porque pueden ser la base para innovaciones en campos como la biotecnología, la agricultura y la medicina ¹.

Los recursos genéticos en su estado natural, como secuencias de ADN o proteínas encontradas en la naturaleza, se consideran descubrimientos. No son patentables porque no son invenciones creadas por el ser humano, sino que ya existen en la naturaleza. Si alguien utiliza estos recursos genéticos para crear algo nuevo y útil, como un medicamento o una técnica de ingeniería genética, esa invención sí puede ser patentable ¹. La clave es que la invención debe ser nueva, no obvia y tener una aplicación industrial; para proteger los RG y los datos relacionados con ellos bajo la propiedad intelectual, es crucial entender los diferentes tipos de derechos de PI que pueden aplicarse:

- Patentes: Para que una invención basada en RG sea patentable, debe ser nueva, no obvia y tener aplicabilidad industrial. Las secuencias de nucleótidos o aminoácidos pueden ser patentables si

cumplen estos criterios y proporcionan una solución práctica a un problema específico en su campo tecnológico¹.

- Derechos de Autor: Protegen las expresiones originales de ideas, como textos científicos, gráficos, diagramas y software. En el contexto de RG esto puede incluir mapas genómicos y anotaciones de datos de secuencias¹.

3.5.2.2 Novedad en Secuencias

La novedad es un requisito clave para la protección de secuencias genéticas bajo patentes. Una secuencia debe ser nueva, es decir, no debe haber sido divulgada públicamente antes de la fecha de presentación de la solicitud de patente. Esto incluye secuencias descritas en publicaciones científicas, bases de datos públicas o cualquier otra forma de divulgación pública. Para ser patentables, las secuencias de nucleótidos y/o aminoácidos deben ser nuevas, no obvias y aplicables industrialmente. Las secuencias que simplemente describen una estructura sin proporcionar una solución práctica no son patentables¹.

Además, los listados de secuencias que describen secuencias de nucleótidos o aminoácidos no son patentables por sí mismos, pero su divulgación puede hacer que las secuencias descritas sean parte del estado de la técnica, afectando su novedad¹.

3.5.2.3. Normativa regulatoria de recursos genéticos en Ecuador

El Reglamento al Régimen Común Sobre Acceso a los Recursos Genéticos en Ecuador, define a los RG como todo material de naturaleza biológica que contiene información genética de valor o utilidad real o potencial. Esto incluye organismos vivos y sus partes, así como cualquier componente biótico que contenga información genética².

El acceso a los recursos genéticos implica la obtención y utilización de estos recursos, ya sea en condiciones ex situ (fuera de su entorno natural) o in situ (en su entorno natural). Este acceso puede tener fines de investigación, prospección biológica, conservación, aplicación industrial o aprovechamiento comercial².

El acceso a los recursos genéticos está regulado por la Autoridad Ambiental Nacional, que es responsable de la conservación y uso sostenible de estos recursos. Cualquier actividad relacionada con el acceso, uso, manejo y aplicación tecnológica de los recursos genéticos está sujeta a monitoreo obligatorio. Para acceder a los recursos genéticos, es necesario firmar un contrato de acceso con la Autoridad Ambiental Nacional.²

3.5.2.4. Propiedad intelectual de recursos genéticos en Ecuador

En Ecuador, no se reconoce ningún derecho de propiedad intelectual sobre productos derivados o sintetizados obtenidos a partir del conocimiento colectivo asociado a la biodiversidad nacional. Se prohíbe toda forma de apropiación de conocimientos colectivos en el ámbito de las ciencias, tecnologías y saberes ancestrales^{2,4}.

La propiedad intelectual sobre los recursos genéticos y sus productos derivados está estrictamente regulada. El Instituto Ecuatoriano de Propiedad Intelectual coordina las acciones para determinar la existencia de un componente intangible asociado a los recursos genéticos. Antes de otorgar cualquier derecho de propiedad intelectual, se debe presentar el número de registro del contrato de acceso y una copia del mismo. Esto asegura que los productos o procesos cuya protección se solicita hayan sido obtenidos de manera legal y ética. Además, los beneficios derivados de la utilización de los recursos genéticos deben ser mutuamente acordados entre el Estado y el interesado. Estos beneficios pueden ser monetarios o no monetarios, como la transferencia tecnológica y regalías².

3.5.2.5. Herramientas Bioinformáticas en Propiedad Intelectual

Los mecanismos bioinformáticos en propiedad intelectual se refieren a la aplicación de herramientas y técnicas de bioinformática para proteger y gestionar los derechos de propiedad intelectual (PI) relacionados con datos biológicos y genéticos. La bioinformática, que combina biología, informática y tecnología de la información, juega un papel crucial en la identificación, análisis y almacenamiento de información genética y biomolecular^{1,5}.

Uno de los principales desafíos en la protección de la PI en bioinformática es la complejidad y la naturaleza multidisciplinaria de los datos biológicos. Los algoritmos y herramientas bioinformáticas permiten la identificación de secuencias genéticas únicas y la predicción de funciones biológicas, lo cual es fundamental para patentar secuencias genéticas, proteínas y otros productos biotecnológicos. Además, la bioinformática facilita la gestión de grandes volúmenes de datos biológicos, lo que es crucial para la vigilancia y el cumplimiento de los derechos de PI^{1,5}.

3.5.2.5. Minería de Datos en Recursos Genéticos

La minería de datos en recursos genéticos se refiere al uso de técnicas avanzadas de análisis de datos para extraer información valiosa de grandes conjuntos de datos genéticos. Esta disciplina, también conocida como data mining, es fundamental para la investigación genética, la biotecnología y la conservación de la

biodiversidad. La minería de datos permite identificar patrones, relaciones y tendencias en los datos genéticos que pueden ser utilizados para desarrollar nuevas terapias, mejorar cultivos y conservar especies en peligro de extinción⁵.

3.5.3. Marco conceptual

En el contexto del desarrollo de una herramienta automatizada de minería de datos para la evaluación de la viabilidad de protección de secuencias genéticas, la selección de metodologías bioinformáticas adecuadas es crucial para garantizar un análisis preciso y exhaustivo. La viabilidad de patentabilidad de una secuencia genética depende en gran medida de su novedad y su diferencia con secuencias preexistentes, tanto en recursos naturales no patentables como en bases de datos de secuencias ya patentadas. A continuación, se detallan las metodologías tentativas que podrían emplearse en este proyecto, junto con su relevancia y aplicación en el proceso de desarrollo. El proceso de evaluación de la patentabilidad de una secuencia genética se puede estructurar en varios pasos clave:

- **Alineamiento de Secuencias:** Las secuencias genéticas propuestas se alinean contra bases de datos públicas y privadas para identificar similitudes y diferencias con secuencias ya existentes.
- **Comparación con Patentes Existentes:** Se cruzan los resultados obtenidos con bases de datos de patentes para determinar si la secuencia cumple con los requisitos de novedad y no obviedad necesarios para su protección.
- **Generación de Reportes:** Los resultados del análisis se sintetizan en reportes automáticos que incluyen visualizaciones y métricas clave, facilitando la toma de decisiones por parte de los evaluadores de patentes.

3.6. Metodología y Técnicas

3.6.1. Revisión del Estado del Arte

Con la finalidad de identificar y analizar las metodologías existentes en el campo de la minería de datos y bioinformática aplicadas a la evaluación de la novedad y patentabilidad de secuencias genéticas. Se propone una búsqueda bibliográfica en bases de datos científicas, así después evaluar las metodologías encontradas, centrándose en su aplicabilidad para la comparación de secuencias genéticas y la determinación de novedad.

3.6.2. Determinación de Requerimientos Técnicos y Funcionales

Es necesario establecer los requerimientos necesarios para el desarrollo de la herramienta automatizada, considerando tanto aspectos técnicos como funcionales. En primera instancia se debe establecer las funcionalidades clave de la herramienta, como la capacidad de realizar alineamientos de secuencias, comparación con bases de datos de patentes y generación de reportes.

Como parte de los requerimientos técnicos se busca determinar las especificaciones técnicas necesarias, incluyendo el entorno de desarrollo, lenguajes de programación y paquetes bioinformáticos. De forma paralela, se debe evaluar la viabilidad técnica y económica de implementar estos requerimientos, considerando recursos disponibles y limitaciones.

3.6.3. Propuesta de Metodología para el Desarrollo

El diseño de una metodología robusta para guiar el desarrollo de la herramienta, desde la concepción hasta la implementación, empieza con la esquematización del flujo de trabajo que la herramienta seguirá para procesar y analizar las secuencias genéticas, incluyendo la integración de diversas herramientas bioinformáticas.

Además, con la fundamentación adecuada, continuar con la propuesta de las herramientas y paquetes específicos que serán utilizados en cada etapa del proceso, asegurando su compatibilidad y efectividad. Así mismo, proponer un conjunto de criterios y pruebas para validar la efectividad y precisión de la herramienta en la determinación de la novedad de secuencias.

Con esta validación, finalmente se elaborará un informe detallado que documente la metodología final propuesta, incluyendo las especificaciones técnicas, flujos de trabajo, herramientas utilizadas y resultados de las pruebas de validación.

3.7. Esquema básico de contenidos

Índice de contenidos	
1. Introducción	Problema Justificación Objetivos
2. Marco Teórico	<ul style="list-style-type: none"> • Recursos Genéticos y Propiedad Intelectual • Novedad en Secuencias

	<ul style="list-style-type: none"> • Propiedad intelectual de recursos genéticos en Ecuador • Herramientas Bioinformáticas en Propiedad Intelectual • Minería de Datos en Recursos Genéticos
3. Metodología	<ul style="list-style-type: none"> • Estrategias de búsqueda • Bases de datos
4. Resultados y discusión	<ul style="list-style-type: none"> • Revisión comparativa de estrategias bioinformáticas • Propuesta de modelo de arquitectura • Propuesta de código
5. Conclusión	
6. Bibliografía	

3.8. Limitaciones del estudio

La principal limitación técnica de esta propuesta de herramienta es la dependencia de la calidad y disponibilidad de las bases de datos genómicas y de patentes, ya que los resultados pueden verse afectados si estas contienen información incompleta o desactualizada. Adicionalmente, la herramienta se basa en algoritmos de alineamiento de secuencias que, aunque eficaces, pueden presentar dificultades para identificar variaciones estructurales complejas. Otra limitación técnica es la capacidad de procesamiento computacional necesaria para manejar grandes volúmenes de datos genómicos, lo cual puede requerir infraestructura robusta y costosa, algo que podría no estar disponible para el desarrollo final.

Una limitación importante radica en la complejidad de las regulaciones sobre propiedad intelectual en relación con los recursos genéticos, especialmente en un contexto local como Ecuador, donde las normativas pueden ser ambiguas o no estar completamente alineadas con las internacionales. Además, la herramienta debe cumplir con las normas de confidencialidad y seguridad de datos, lo que agrega una capa de complejidad en su desarrollo y uso.

3.9. Resultados esperados

Se espera que como resultado de este trabajo se realice una comparativa exhaustiva entre las diferentes estrategias y herramientas bioinformáticas existentes para la evaluación de la novedad en secuencias genómicas. Esta comparativa permitirá identificar las mejores prácticas y adaptarlas al contexto local. Adicionalmente, se desarrollará una arquitectura de software que será la base para el futuro desarrollo de la herramienta automatizada de minería de datos. Esta arquitectura estará diseñada para ser escalable y

adaptable, lo que facilitará la implementación de futuras mejoras en función de las necesidades de los usuarios.

4. Cronograma

Actividad	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Revisión del estado del arte																								
Análisis de herramientas existentes																								
Comparativa de estrategias existentes																								
Diseño y validación de la arquitectura																								
Definición de algoritmos principales																								
Desarrollo de módulos de código																								
Pruebas, revisión y ajustes																								
Redacción de documento final																								

5. Bibliografía

1. OMPI. (2022). Intellectual Property Guide for Genetic Resources and Genetic Sequence Data. https://www.wipo.int/export/sites/www/tk/en/docs/ip_gr_grdataguidefin.pdf
2. REGLAMENTO AL REGIMEN COMUN SOBRE ACCESO A LOS RECURSOS GENETICOS, 553 Registro Oficial Suplemento 553 11 (2011). www.lexis.com.ec
3. Gomez-Hinostroza, E. S., Gurdo, N., Alvan Vargas, M. V. G., Nikel, P. I., Guazzaroni, M. E., Guaman, L. P., Castillo Cornejo, D. J., Platero, R., & Barba-Ostria, C. (2023). Current landscape and future directions of synthetic biology in South America. In *Frontiers in Bioengineering and Biotechnology* (Vol. 11). Frontiers Media S.A. <https://doi.org/10.3389/fbioe.2023.1069628>
4. Colmenarez, Y. C., Smith, D., Walsh, G. C., France, A., Corniani, N., & Vásquez, C. (2023). Regulatory Frameworks for the Access and Use of Genetic Resources in Latin America. *Neotropical Entomology*, 52(2), 333–344. <https://doi.org/10.1007/s13744-022-01017-x>
5. Singh, G. B. (2015). *Fundamentals of Bioinformatics and Computational Biology* (1°). Springer. <https://doi.org/10.1007/978-3-319-11403-3>