Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Reporte final (100%) del Semestre Sabático

No. de autorización SS-2-006/2024

Desarrollo de modelos y métodos computacionales para la anotación genómica del locus de inmunoglobulinas en vertebrados

a) Programa de Investigación Científica, Tecnológica o Educativa.
A.1 Proyectos de Investigación Científica

Dr. Javier Ortiz Hernández

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Morelos, 11/marzo/2025

REPORTE FINAL SEMESTRE SABÁTICO

I.Identificación del Proyecto

Instituto o Centro: TecNM/Centro Nacional de Investigación y Desarrollo Tecnológico
Nombre del/de la docente: Javier Ortiz Hernández
Título del proyecto: Desarrollo de modelos y métodos computacionales para la anotación genómica del locus de inmunoglobulinas en vertebrados
Tipo de investigación: Científica
Duración del proyecto: 1 semestre
Fecha de inicio del proyecto: 2 septiembre 2024
Fecha de término del proyecto: 1 de marzo 2025

II.Resultados

1. Resumen del proyecto

a) Introducción

En el Centro de Investigación Sobre Enfermedades Infecciosas (CISEI) del Instituto Nacional de Salud Pública (INSP), biólogos moleculares llevan a cabo el proceso de anotación de los genes de los segmentos V(D)J de anticuerpos de diversas especies. La anotación consiste en a)identificar la ubicación de los genes y en b)identificar su estructura y funcionalidad biológica. Por razones de alcance derivados de la complejidad del problema, en este proyecto únicamente se abordará la anotación del segmento V, en el entendido que la anotación requerida para los segmentos D y J es muy similar, y parte de los resultados obtenidos del segmento V podrían ser extra polados para realizar la anotación de los segmento D y J. El proceso que realizan comienza con la búsqueda de la ubicación de los genes que integran el segmento V en el genoma, para lo cual se emplean técnicas de alineamiento de secuencias que les proporcionan a los biólogos moleculares aproximaciones de la ubicación de los diferentes genes V. La anotación se realiza con el apoyo de diversas herrramientas informáticas y de visualizadores del genoma. Sin embargo, el enorme espacio de búsqueda que se presenta para la anotación de un solo gen demanda una gran cantidad de recursos y de tiempo. Además del enorme espacio de búsqueda, otro desafío importante es conocer la

estructura del gen. Los sistemas de alineamiento no son capaces de identificar simultáneamente toda la estructura del gen V. Esto implica que los expertos deben enfrentarse a la tarea de identificar diferentes partes del gen por separado, añadiendo dificultades al proceso de anotación.

En el ámbito computacional, el enfoque para abordar el problema del proceso de anotación se centra en dos aspectos fundamentales: la integración de os datos de entrada para realizar el análisis y la reducción del espacio de búsqueda. La integración de los datos de entrada es crucial debido al principio del alineamiento de secuencias, donde cada sistema detecta una sección específica del gen V. El desafío computacional radica en cómo representar los datos del alineamiento de secuencias para obtener un único conjunto de datos que represente de manera completa todos los genes V analizar. En este proyecto se busca Implementar un modelo de inteligencia artificial para reducir el espacio de búsqueda de los alineamientos de secuencias maximizando la identificación de genes V.

b) Objetivos:

Objetivo general

Desarrollar modelos y métodos computacionales para la anotación genómica del locus de inmunoglobulinas en vertebrados

Objetivos específicos

-Desarrollar la integración de un único conjunto de datos con las posiciones de inicio y fin de los alineamientos de secuencias.de los genes que integran el segmento V a partir de diversas fuentes de datos disponibles.

-Implementar un modelo de inteligencia artificial para reducir el espacio de búsqueda de los alineamientos de secuencia con el uso del conjunto de datos del objetivo anterior.

c) Metas:

Meta 1. Integración de un único conjunto de datos con las posiciones de inicio y fin de los alineamientos de secuencias de los genes que integran el segmento V a partir de diversas fuentes de datos. Cantidad: 1

Meta 2. Modelo de inteligencia artificial para reducir el espacio de búsqueda de los alineamientos de secuencias. Cantidad: 1

Meta 3. Redacción y sometimiento de dos articulos en revista indizada en el área. Cantidad: 2

Meta 4: Formación de recursos humanos a nivel doctorado en desarrollo: Cantidad: 2

d) Desarrollo y resultados del proyecto

Metas 1 y 2

1 Introducción

Un problema de optimización busca maximizar o minimizar una o varias funciones numéricas de una o más variables independientes o dependientes de ciertas restricciones (Bazaraa et al., 2006). En esencia, un problema de optimización busca identificar la o las mejores soluciones entre un conjunto de opciones factibles que puede abarcar la minimización o maximización de una función. Este enfoque puede aplicarse en diversas areas, como problemas computacionales, biológicos, de telecomunicaciones o financieros. Sin embargo, el amplio espacio de soluciones en muchos problemas puede implicar mayores esfuerzos para identificar la mejor solución (Dagdia & Mirchev, 2020).
En el presente estudio, el desafío de optimización se aborda en el contexto de la caracterización del genoma, un proceso que consiste en identificar y asignar información a secuencias de caracteres que representan a un gen. El proceso cuenta con dos etapas clave: (a) la identificación del gen, donde se determinan las posiciones de inicio y fin de un gen; y (b) la anotación o clasificación del gen, que consiste en la asignación de información relevante. Una identificación precisa facilita significativamente la etapa de anotación, lo que resalta su importancia para los especialistas (Bergman, 2007; Megrian, 2014).

3 En el Centro de Investigación Sobre Enfermedades Infecciosas (CISEI) del Instituto Nacional de Salud Pública (INSP), ubicado en Cuernavaca, México, se realiza actualmente la caracterización de los genes Variable, de Diversidad, y de Unión por sus siglas en ingles genes V, D y J. Estos genes desempeñan un papel crucial en la formación de anticuerpos B, que detectan y neutralizan agentes externos dañinos, 32 como virus y bacterias (Kindt et al., 2007; Pieper et al., 2013). Esta investigación se enfoca exclusivamente en los genes V debido a su alta frecuencia en los genomas de vertebrados y su compleja estructura.

Durante la identificación de genes, los expertos del CISEI utilizan diversos sistemas de alineamiento de secuencias para estimar las posiciones de inicio y fin de los genes y sus secciones, pero la imprecisión de estos alineamientos impacta al espacio de soluciones, debido a que dificultan la identificación del gen. Además, los sistemas actuales no pueden detectar toda la estructura del gen V simultáneamente, lo que obliga a analizar sus secciones por separado y por ende aumentar la complejidad del proceso.

La reducción del espacio de soluciones minimizara la cantidad de alineamientos hasta obtener una única propuesta cercana a la realidad, permitiendo a los especialistas optimizar su tiempo y enfocarse en la anotación de genes en lugar de validar manualmente cada posición. Los principales retos de esta investigación involucran:

- Integración de datos: los alineamientos provienen de diferentes fuentes, lo que requiere unificar la información para representar de manera completa los genes V.
- Reducción del espacio de soluciones: el gran volumen de las propuestas para ubicación del gen V dificulta la etapa de la identificación.

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

El siguiente estudio propone un enfoque novedoso que combina la integración de datos con una reducción efectiva del espacio de soluciones. El método propuesto se evalúa comparándolo con las referencias manuales generadas por los expertos del CISEI, utilizando métricas para medir el error y su eficiencia. Este método no solo busca mejorar la precisión en la etapa de identificación, sino que también marca un punto de referencia para una anotación más eficiente, contribuyendo al conocimiento y estudio de los genomas de vertebrados.

1.1 Estructura del gen V y proceso de identificación

La identificación del gen V es una actividad que consiste en buscar sus secciones y determinar las posiciones de inicio y fin. Una correcta identificación mejora la etapa de anotación y por ende la representación de la funcionalidad del gen (Bergman, 2007; Megrian, 2014). Un método común para identificar genes es la homología (Mount, 2004; Olivieri & Gambón-Deza, 2019), que utiliza secuencias validadas como referencia para encontrar genes similares en un genoma. Aunque en la presente investigación no se entrara en detalle en la caracterización del genoma los siguientes estudios lo abordan a fondo (Ejigu & Jung, 2020; Miguel-Ruiz et al., 2024; Serret et al., 2023).

Las secuencias validadas se utilizan en sistemas de alineación para obtener los datos que son necesarios en la identificación de los genes V (Amin et al., 2018; Ejigu & Jung, 2020; Mount, 2004). Los sistemas de alineamiento más comunes son Blast, Hmmer y Exonerate. Estos sistemas son eficaces para obtener alineamientos de secuencias, pero cada uno de estos usa un principio diferente y se enfoca en secciones específicas del gen (EMBL-EBI, 2022). Uno de los principales inconvenientes de estos sistemas es su baja precisión para identificar de forma precisa la totalidad del gen. Aunque estos sistemas ofrecen una aproximación a las ubicaciones de las secciones del gen V en el genoma, no proporcionan una ubicación exacta (Mount, 2004).

Cada una de las secciones del gen V cuenta con características particulares como: (a) **Péptido señal o SP** es una secuencia corta de aminoácidos que indica dónde comienza el gen V, este cuenta con una longitud entre 40 a 46 caracteres; (b) **Exon** es la región donde se contiene toda la información del gen V es el bloque que da nombre al gen V y cuenta con una longitud entre 300 a 350 caracteres; y (c) **la señal de recombinación o RSS** que es la encargada de la unión entre dos genes, la longitud de la RSS puede ser de 38 a 40 caracteres (Lefranc & Lefranc, 2001). La Figura 1 presenta la estructura del gen V.

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

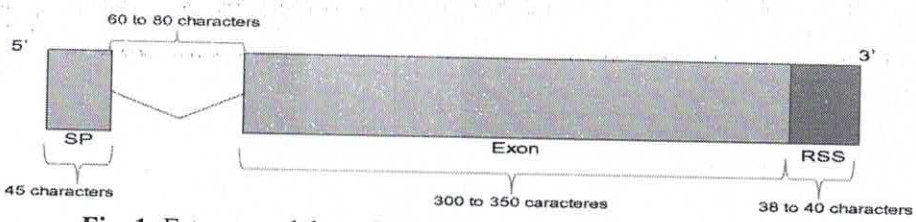

Fig. 1. Estructura del gen V obtenida de (Lefranc & Lefranc, 2001)

Los genes V pueden encontrarse en dos direcciones dentro del genoma: de izquierda a derecha, iniciando en SP y finalizando en RSS, Figura 1, o de derecha a izquierda, manteniendo el mismo orden estructural Figura 2. La caracterización del genoma se realiza mediante el sistema IGV (Integrative Genomics Viewer), herramienta utilizada por los expertos para analizar alineamientos. La Figura 2 muestra un gen completo, donde la referencia experta aparece en rojo junto con los alineamientos colapsados, lo que facilita su interpretación, aunque sin mostrar las posiciones exactas del gen.

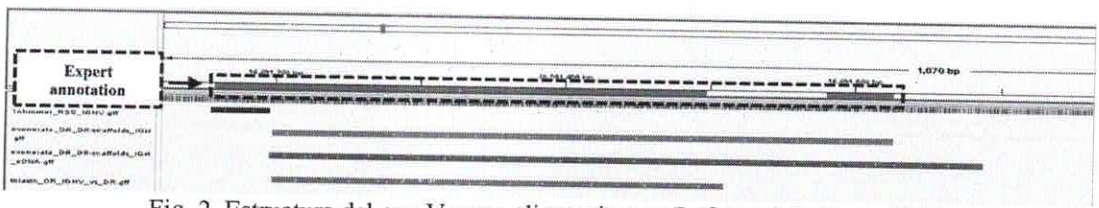

Fig. 2. Estructura del gen V y sus alineamientos (Lefranc & Lefranc, 2001)

La versión expandida de los datos se presenta en la Figura 3, mostrando con mayor detalle los alineamientos generados por una herramienta. La cantidad de alineamientos puede variar entre 50 y 200 por archivo, puesto que, al considerar los demás archivos, complica significativamente el proceso de identificación. Aunque la Figura 3 proporciona más información sobre la ubicación del gen, ninguno de los alineamientos coincide exactamente con la referencia, lo que obliga a los expertos a usar su conocimiento para eliminar redundancias y asignar correctamente las posiciones del gen. Pero este proceso es laborioso e implica un gran esfuerzo.

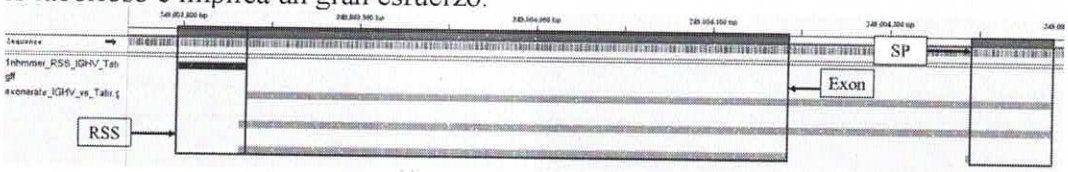

Fig. 3. Sistema IGV de (Lefranc & Lefranc, 2001)

Como se mencionó al inicio de la investigación, cada sistema de alineamiento genera archivos con información clave para la identificación del gen V, debido a que cada uno detecta una sección

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

específica. Hmmerscan identifica la señal de recombinación (RSS) a nivel de nucleótidos; Exonerate:protein2genome detecta el exón y el péptido señal (SP) a nivel de aminoácidos; Exonerate:est2genome identifica el exón y el péptido señal (SP) a nivel de nucleótidos; Tblastx encuentra el exón y el péptido señal (SP) a nivel de aminoácidos; y Tblastn identifica el exón y el péptido señal (SP) a nivel de nucleótidos. Cada sistema opera a un nivel específico, utilizando caracteres particulares que influyen en el procesamiento y análisis de los datos para la identificación del gen V.

1.2 Estado del arte

La revisión de la literatura tuvo el objetivo de identificar trabajos similares donde busquen abordar la identificación y anotación de genes. La pregunta de investigación planteada fue ¿Cómo se ha abordado la sistematización del proceso de caracterización de genes y sus principales técnicas? En esta revisión se buscó identificar sistemas, arquitecturas, métodos, algoritmos y técnica para la sistematización del proceso de caracterización de genes, por ello se hizo uso de la metodología de elementos de informe referidos (PRISMA). La metodología PRISMA originalmente desarrollada para meta-análisis de ensayos clínicos permite su adaptación para la revisión de otros tipos de estudios (Urrútia & Bonfill, 2010). El proceso determinó una cadena de búsqueda que permitió esclarecer cómo se ha abordado la sistematización del proceso de caracterización de genes. La cadena de búsqueda generada para este estudio consistió en: ("gene annotation" OR "gene identification") AND ("machine learning" OR "computational modeling") AND ("V(D)J" OR "antibodies" OR "V genes") and "vertebrates".

En la revisión se buscaron artículos indexados a revistas y artículos presentados en diversos congresos. En la revisión se hizo uso de bases de datos como PubMed, MDPI, NCBI, Springerlink y Science Direct en conjunto con el buscador Google académico. En la revisión sólo se incluyeron artículos escritos en inglés y que fueran publicados a partir del 2019 al 2024. El resultado de la búsqueda para la cadena de búsqueda fue de un total de 160 artículos. De los 160 artículos 9 fueron de relevancia para este estudio.

Tabla 1. Artículos relevantes identificados

N.	Breve descripción	Método, técnica o algoritmo	Referencia
1	GeMoMa sistema web basado en homología con el uso de archivos GFF y Fasta.	Se basa en homología con archivos GFF	(Keilwagen et al., 2019)
2	VgeneFinder sistema basado en el principio de homología con secuencias de referencia	Se basa en el uso de motivos y la secuencia del genoma	(Olivieri & Gambón-Deza, 2019)
3	DeepGSR sistema basado en aprendizaje profundo para la detección de señales y patrones genómicos	Se basa en una red neuronal convolucional.	(Kalkatawi et al., 2019)
4	Helixer es un sistema basado en aprendizaje profundo	Se basa en BLSTM para anotar secuencias con Keras	(Stiehler et al., 2020)

5	GOODORFS es un sistema basado en el método ab initio	Se basado en Kmeans	(McNair et al., 2021)
6	RAPID es un sistema web que realiza el análisis de anticuerpos y anotación de clonotipos	No se especifica.	(Y. Zhang et al., 2021)
7	TOGA es un pipeline de alineamiento de secuencias	Se basa en la homología y aprendizaje automático.	(Zhang et al., 2021)
8	IGDetective es un sistema basado en homología	El sistema se basa en teoría de grafos y homología.	(Sirupurapu et al., 2022)
9	Mgcod sistema basado en homología para identificar zonas de un gen.	Se basa en un pipeline.	(Pfennig et al., 2022)

La revisión de la literatura evidencia la importancia de la identificación de genes V, destacando trabajos como los de Olivieri & Gambón-Deza (2019) y Sirupurapu et al. (2022a). Aunque existen pocos estudios específicos, la mayoría de los sistemas de anotación genética dependen de una identificación previa, creando una gran área de oportunidad. Las técnicas más utilizadas para la anotación de genes incluyen métodos de homología y enfoques ab initio, donde modelos como LSTM, redes neuronales convolucionales y aprendizaje profundo han demostrado ser eficaces (Kalkatawi, Magana-Mora, et al., 2019). Sin embargo, muchos estudios aún dependen del alineamiento de secuencias, utilizando estos datos como referencia.

Un punto clave es la aplicación de IGDetective en la identificación y anotación de genes, un claro ejemplo de esto fue 2024, en el trabajo de (Zhu et al., 2024) donde se sigue utilizando para anotar genes. lo que confirma la relevancia de estos avances en la caracterización genómica. Esto refuerza la necesidad de optimizar los procesos de identificación de genes.

2 Método y datos para la reducción del espacio de soluciones en alineamientos de secuencias para genes V

La literatura científica incluye numerosas propuestas orientadas a la sistematización de los procesos de caracterización de genes V, como los trabajos de Olivieri & Gambón-Deza (2019) y Sirupurapu et al. (2022). Sin embargo, persisten desafíos significativos relacionados con los alineamientos de secuencias por su baja precisión para predecir la ubicación exacta de los genes, lo que dificulta la identificación. Por lo tanto, se propone un método basado en distancias para reducir el espacio de soluciones de los alineamientos de genes V. Este enfoque aprovecha la estructura del gen V y las distancias definidas entre sus tres secciones fundamentales para obtener un nuevo conjunto de datos utilizado en la reducción del espacio de soluciones.

2.1 El método

Se organiza en tres etapas principales que abordan desde el preprocesamiento de datos hasta el posprocesamiento. En la Figura 4 se presenta una visión general del método y a continuación, se describen sus etapas:

Preprocesamiento: Busca asegura la calidad de los datos disponibles en formato GFF, consta de las siguientes actividades:

- Unificación de datos: se integran varias fuentes de alineamientos (e.g., Hmmer, Exonerate, Tblastn) en un único conjunto de datos representativo. Esto asegura que cada registro incluya todas las secciones del gen V.
- Normalización de datos y división de datos: los datos se ajustan utilizando técnicas como Normalizer, para mejorar el el desempeño de los algoritmos de agrupamiento. También son divididos los datos en entrenamiento y pruebas.

Procesamiento: cuenta con dos actividades centradas en el uso de algoritmos de agrupamiento y la unificación de los datos de los grupos:

- Entrenamiento y uso de un modelo de agrupamiento: en esta actividad se implementa un modelo basado en el algoritmo de Gaussian Mixture.
- Unificación de grupos: Se consolidan los registros dentro de cada grupo, generando un único registro por gen. Esto reduce el espacio de soluciones y permite una medición justa de los resultados.

Posprocesamiento: consiste en dos actividades que permiten evaluar los resultados de cada grupo y la transformación de los datos en un formato usable por los expertos.

- Medición del error y selección del mejor grupo: se comparan los resultados del procesamiento con la referencia de los expertos, utilizando métricas como MAE, MSE y RMSE.
- Transformación al formato GFF: El mejor grupo obtenido se transforma al formato GFF, asegurando compatibilidad con herramientas de análisis genómico como IGV.

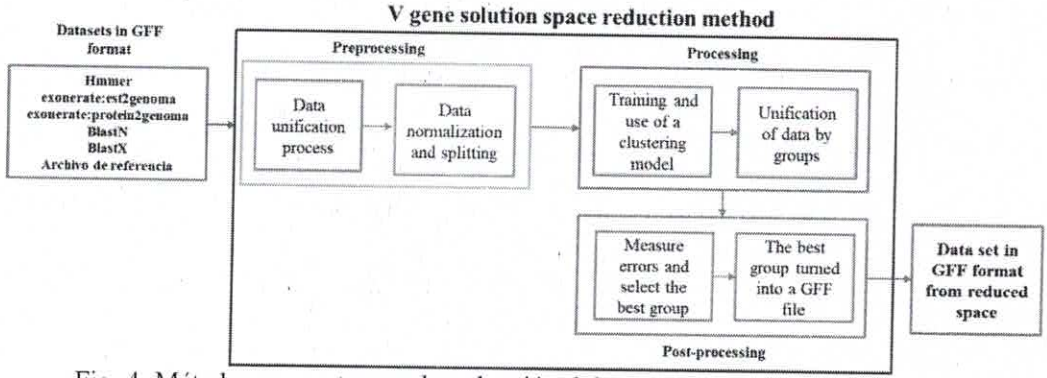

Fig. 4. Método propuesto para la reducción del espacio de búsqueda de genes V

El método propuesto permite una reducción significativa del espacio de soluciones en la identificación de genes V, logrando un desempeño eficiente tanto en genomas grandes como pequeños, con tiempos de procesamiento inferiores a 60 minutos. En las siguientes secciones del documento se abordarán aspectos clave como: (a) conjuntos de datos en formato GFF, (b) las actividades asociadas al preprocesamiento, (c) las actividades del procesamiento de los datos, y (d) los resultados con la etapa del posprocesamiento.

2.2 Conjuntos de datos en formato GFF

Los archivos en formato de características generales (**GFF**) son ampliamente utilizados en proyectos biológicos y bioinformáticos debido a su capacidad para organizar y dar sentido a fragmentos de un genoma. Este formato de texto plano consta de nueve columnas delimitadas por tabulaciones, lo que permite una representación estructurada y detallada de las características genómicas (EMBL-EBI, 2022; García Simón, 2018). En el contexto de esta investigación, los archivos GFF son comparables a conjuntos de datos utilizados en ciencia de datos, debido a que ambos contienen atributos asociados a características específicas.

Los datos utilizados en este estudio provienen de diferentes sistemas especializados en alineamiento de secuencias, que identifican zonas de interés en un genoma y en específico de los genes V. Los sistemas que son utilizados por los expertos del CICEI son Blast, Hmmer y Exonerate, que emplean distintos principios para la ejecución de los alineamientos (EMBL-EBI, 2022). A pesar de su efectividad, estos sistemas presentan limitaciones importantes, como su baja precisión para determinar la ubicación de la estructura completa de un gen y además de la falta de presión en la ubicación exacta (Mount, 2004).

Cada sistema detecta una o varias secciones del gen, pero no en su totalidad, por lo cual se deben combinar múltiples resultados de diferentes sistemas, esta codependencia entre diferentes sistemas permite a los expertos realizar el proceso de caracterización. Un ejemplo de esto es Exonerate:protein2genome y Exonerate:est2genome, que identifican el exón y el péptido señal (SP), cada uno trabaja a niveles diferentes del genoma y se complementan entre sí. Debido a que si protein2genome no identifica el SP est2genome puede completar la identificación. Estos sistemas generan datos clave en el formato GFF, que representan las ubicaciones donde se produjeron alineamientos de secuencias. Este formato proporciona un marco estructurado para analizar características genéticas, facilitando su integración y análisis. En la Tabla 2 se presenta un fragmento de un archivo GFF utilizado en esta investigación. Este ejemplo muestra cómo se organizan los datos en las columnas clave.

Tabla 2. fragmento de un archivo GFF utilizado en esta investigación

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	exonerate:est2genome	gene	16489607	16489647	NA	-	NA	NA
CM040297.1	exonerate:est2genome	gene	16489648	16489956	1239	-	NA	NA
CM040297.1	exonerate:est2genome	gene	16489648	16490157	2011	-	NA	NA

Los archivos GFF contienen nueve atributos que permiten la identificación de información clave. En esta investigación, las columnas Start, End y Strand son especialmente relevantes, debido a que describen con precisión la estructura del gen V. A continuación, se presentan los nueve atributos principales:

1. SeqID: Identificador único de la secuencia, como el cromosoma o scaffold donde se encuentra la característica.
2. Fuente (Source): Herramienta o método que generó el alineamiento, por ejemplo, Hmmer, Exonerate o Blast.
3. Tipo (Type): Tipo de característica alineada, como Gen, Exon, CDS o mRNA.
4. Inicio (Start): Posición inicial de la característica en la secuencia.
5. Fin (End): Posición final de la característica en la secuencia.
6. Puntuación (Score): Valor de confianza asociado a la anotación. Puede ser un número o un punto (.) si no está disponible.
7. Cadena (Strand): Dirección de la característica en el ADN, indicada por + o -.
8. Fase (Phase): Especifica cómo deben leerse los codones. Los valores posibles son 0, 1 o 2.
9. Atributos (Attributes): Información adicional sobre la característica, como identificadores únicos, nombres de genes o notas.

En el contexto de la caracterización de genomas, los murciélagos (quirópteros) han sido ampliamente estudiados debido a su diversidad genética y adaptaciones únicas. Específicamente, en el CISEI, se ha trabajado con especies como Tadarida brasiliensis, un murciélago nativo del oeste y sur de los Estados Unidos, México, Centroamérica y otras regiones (Webster et al., 2024). Esta especie destaca por su genoma extenso, que contiene más de 800 genes V, mientras que uno solo de sus archivos GFF supera los 167,165 registros, representando una cifra intermedia para los datos utilizados en la caracterización del genómica. El ejemplar utilizado en este estudio tiene como referencia del cromosoma CM061257.1, lo que permite un análisis detallado de su organización genética y su referencia en el banco de datos del NCBI. En el análisis de los datos, se realizó un estudio de normalidad y densidad utilizando el atributo de Inicio del archivo Tblastn, el cual contiene 167,165 registros. Los gráficos correspondientes se presentan en la Figura 5. Una característica destacable de estos datos es su escala, que puede abarcar valores en miles o incluso millones, lo que explica la falta de normalidad observada en el QQ plot. En cuanto a las gráficas de densidad, estas reflejan la magnitud de los datos y confirman la amplitud de su

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

distribución, proporcionando una visión más clara sobre la dispersión de los valores analizados. Esto se repite para el atributo de fin.

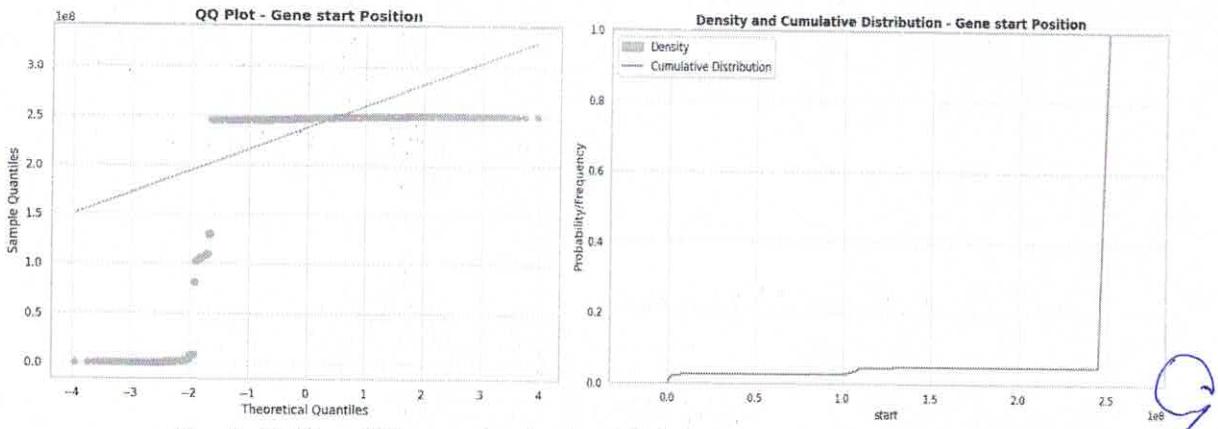

Fig. 5. Gráfico QQnormal y de densidad de las posiciones de inicio

Cuando se realiza un análisis detallado de las características de los archivos GFF, se justifica la selección de tres columnas clave: Start, End y Strand. Estas columnas concentran la información más relevante para describir la estructura y las características genómicas, como se observa en la Tabla 3. Aunque las posiciones de inicio y fin presentan una escala que puede superar los millones, esta granularidad es esencial para identificar cada una de las secciones del gen V. Un desafío importante al trabajar con estos archivos es que cada uno identifica únicamente una sección del gen, lo que hace ineficiente cualquier intento de análisis independiente. Por esta razón, es fundamental unificar la información dispersa en registros que integren las tres secciones del gen V.

Tabla 3. Estadísticas básicas del conjunto de exonerate:est2genome

Column	Min	Max	Mean	Median	NA_Count
seqid	13	CM061257.1	0	13	-
start	15537	250241728	239268703.757382	247595660	0
end	15746	250242045	239269002.931899	247595962	0
strand	1	-	0	1	-

2.3 Etapa de preprocesamiento de datos

El análisis de los datos en formato GFF muestra que su estructura dispersa dificulta la representación precisa del gen V, lo que limita la aplicación de algoritmos de aprendizaje automático. Cada registro captura sólo una parte de la información, impidiendo una identificación completa del gen. Para resolver

este problema, se ha desarrollado un proceso avanzado de preprocesamiento y unificación, que integra los datos fragmentados en nuevos registros completos, mejorando su representación y análisis. Este enfoque se basa en las distancias relativas entre las secciones del gen V, permitiendo omitir temporalmente la secuencia de caracteres y centrarse en la información estructural de los archivos GFF. A continuación, se presenta el proceso para la unificación de datos.

2.3.1 Unificación de datos

En esta investigación, es fundamental implementar una técnica de preprocesamiento que se enfoque en la unificación e integración de datos de diferentes fuentes en registros completos y precisos. Este proceso no sólo permite representar la estructura del gen V, sino que también incorpora las distancias relativas entre sus secciones (SP, Exon y RSS), reduciendo la dependencia de posiciones absolutas y mejorando la flexibilidad y precisión del análisis, incluso cuando las herramientas varían en sus valores de inicio y fin.Este enfoque optimiza el tiempo y los recursos en la identificación de genes V al consolidar datos dispersos y reducir el espacio de soluciones, haciendo el proceso de anotación más eficiente. La integración sistemática de estos datos no solo mejora la identificación de genes V, sino que también crea un marco reutilizable para consolidar datos en otros genomas. En la siguiente sección, se describen las cinco actividades del proceso de unificación:

1. Importación de datos GFF: Consiste en la recuperación y adaptación de los archivos GFF, transformándolos en un formato tabular que facilite su manipulación y análisis posterior.
2. Identificación del archivo con menor redundancia: Se comparan todos los archivos disponibles para determinar cuál tiene la menor cantidad de registros redundantes. Este archivo se selecciona como punto de partida, ya que representa la información más específica y confiable para la construcción de registros de genes V.
3. Extracción y preparación de datos: Con base en las posiciones del archivo con menor redundancia, se extraen los datos faltantes de los otros archivos no utilizados. Esto permite mejorar la representación del gen V.
4. Construcción y validación de registros de genes V: Se utiliza la estructura del gen V para garantizar que los registros generados no excedan distancias biológicamente coherentes entre puntos.
5. Construcción de matrices de adyacencia y vectorización: En esta etapa, se genera una matriz de adyacencia por cada gen para representar los datos de cada posible gen. Esta matriz se utiliza para derivar nuevas características basadas en distancias, que son posteriormente vectorizadas para facilitar su uso algoritmos de aprendizaje máquina.

Estas cinco actividades se articulan para garantizar que los datos procesados sean de alta calidad y estén listos para su análisis posterior. La combinación de estas actividades no solo mejora la precisión de la

identificación de genes V, sino que también simplifica el manejo de datos complejos, promoviendo un flujo de trabajo más eficiente y confiable. En la Figura 6 se presenta la estructura general del proceso y sus cinco etapas denominado proceso principal.

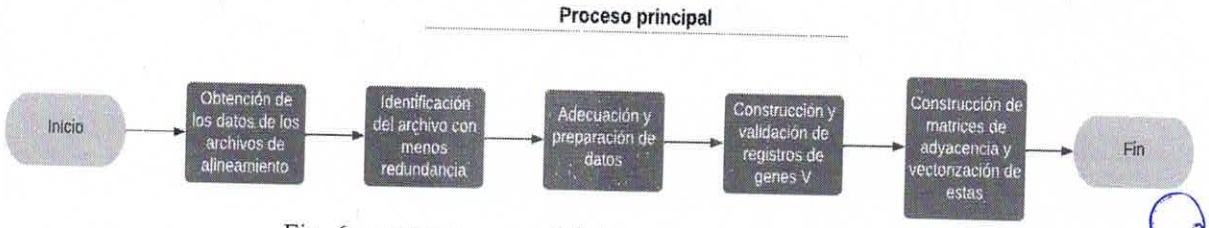

Fig. 6. estructura general del proceso y sus cinco etapas

El primer subproceso consiste en la identificación del archivo con menor redundancia, asegurando que el conjunto seleccionado contenga solo contenga información única. Para ello, se calcula la redundancia de cada conjunto GFF, verificando si al menos una de las columnas Inicio o Fin contiene valores únicos y si estos representan la totalidad de los registros dentro del archivo analizado. Una vez identificada la opción adecuada, el conjunto con menor redundancia es asignado a la variable DUMrev o DUMforw (Data Unique Matrix), dependiendo de la dirección del alineamiento. En esta etapa, se toma en cuenta si el alineamiento es revers o forward. Finalmente, las matrices DUMrev y DUMforw son retornadas para su uso en las siguientes etapas del proceso de unificación. El subproceso se presenta en la Figura 7.

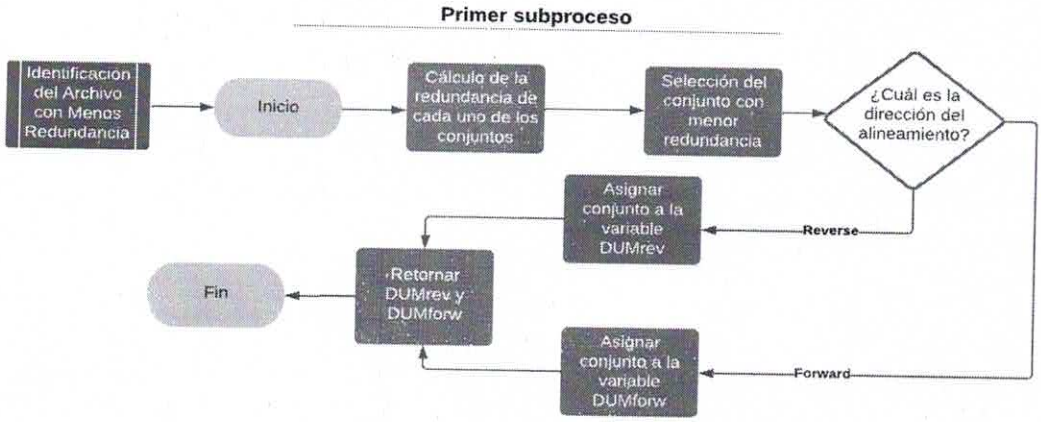

Fig. 7. primer subproceso consiste en la identificación del archivo con menor redundancia

El segundo subproceso se centra en la extracción y preparación de datos, asegurando que los registros sean estructurados correctamente. Por ello, se extraen las posiciones de inicio de las matrices DUMrev y

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

DUMforw en pares, utilizándose como rangos. Con estos rangos, se recuperan los datos de los demás archivos GFF que coincidan dentro de los intervalos establecidos por estos rangos. Posteriormente, los registros extraídos son asignados a listas según su dirección: LISTMrev o LISTMforw. Estas listas almacenan los registros de manera estructurada para su posterior análisis. El proceso continúa iterativamente hasta que ya no existan datos pendientes en DUMrev o DUMforw. El proceso se muestra en la Figura 8.

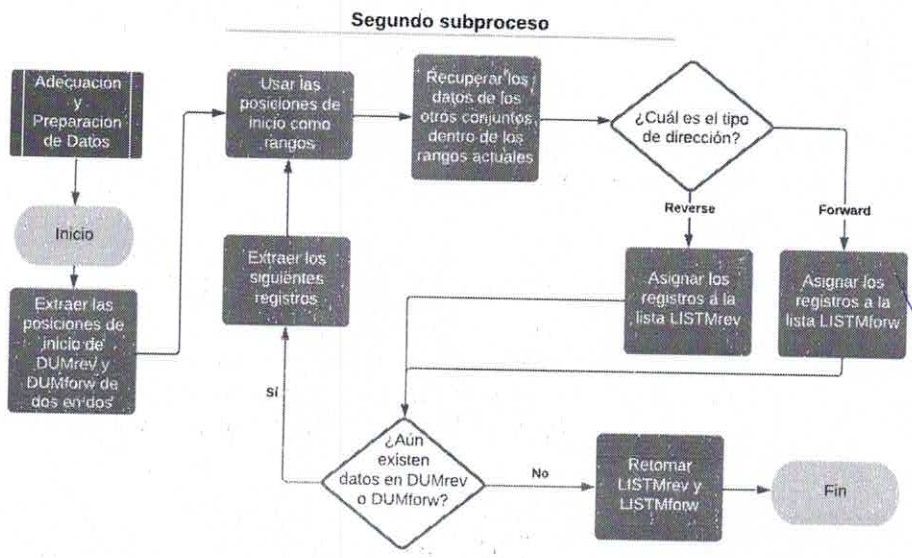

Fig. 8. segundo subproceso se centra en la extracción y preparación de datos

El tercer subproceso se aborda en la Figura 9, centrado en la construcción y validación de registros de genes V. El objetivo es asegurar que los registros generados cumplan con distancias biológicamente coherentes entre las secciones, consolidando la información de los subprocesos previos. El subproceso comienza con la extracción de un registro de DUMrev o DUMforw y se asigna a la variable RSS. El siguiente paso es procesar los datos de LISTMrev o LISTMforw, según el DUM del que se haya extraído para la RSS. Las siguientes asignaciones siguen la estructura de: el primer registro de LISTM (dependiendo de la dirección) se asigna a Exon, mientras que el siguiente se asocia a SP.

La construcción de registros se basa en una estructura genérica del gen V, asegurando que las tres secciones estén correctamente representadas y que cada registro sea biológicamente válido. Para mantener esta coherencia, se establecen cuatro reglas fundamentales que delimitan la estructura del gen:

1. Distancia entre RSS y Exon: la diferencia absoluta entre la posición de inicio de la RSS y el inicio del Exon debe ser menor o igual a 15 y mayor a -1 carácter. Esto garantiza que las secciones estén dentro de un rango aceptable sin solapamientos ni separaciones excesivas. Si no se cumple, se extrae el siguiente registro como Exon y se reevalúa.
2. Distancia dentro del Exon: la diferencia absoluta entre el inicio y el final del Exon debe estar entre 280 y 350 caracteres. La distancia se evalúa en valores absolutos, debido a que los alineamientos tipo Reverse aparecen en orden creciente y los Forward en decreciente. Si la condición no se cumple, se analiza el siguiente registro y se vuelve a iniciar todo el proceso.
3. Distancia dentro del SP: la diferencia absoluta entre el inicio y el final del SP debe ser mayor a 45 caracteres, pero no superar los 50. Si esta condición no se cumple, se extrae el registro como SP y se reinicia todo el proceso de evaluación.
4. Distancia entre SP y Exon: la distancia absoluta entre el inicio del SP y el final del Exon debe estar en un rango de 60 a 150 caracteres. Este criterio asegura una separación biológicamente coherente entre estas regiones del gen. Si la condición no se cumple, se extrae otro registro como posible SP y se reinicia todo el proceso.

Si todos los criterios de validación se cumplen, los registros de RSS, SP y Exon se almacenan en la matriz Mat_Gen según la dirección del alineamiento y se almacenan estas matrices en ListMat_Genrev o ListMat_Genforw dependiendo de la dirección. Una vez que no quedan más datos en DUMrev o DUMforw, se retornan las listas, que contienen los registros finales listos para su análisis en la siguiente fase del método de unificación. Este subproceso es fundamental para estructurar los datos de manera coherente, asegurando que las secciones del gen V sean representadas con precisión biológica.

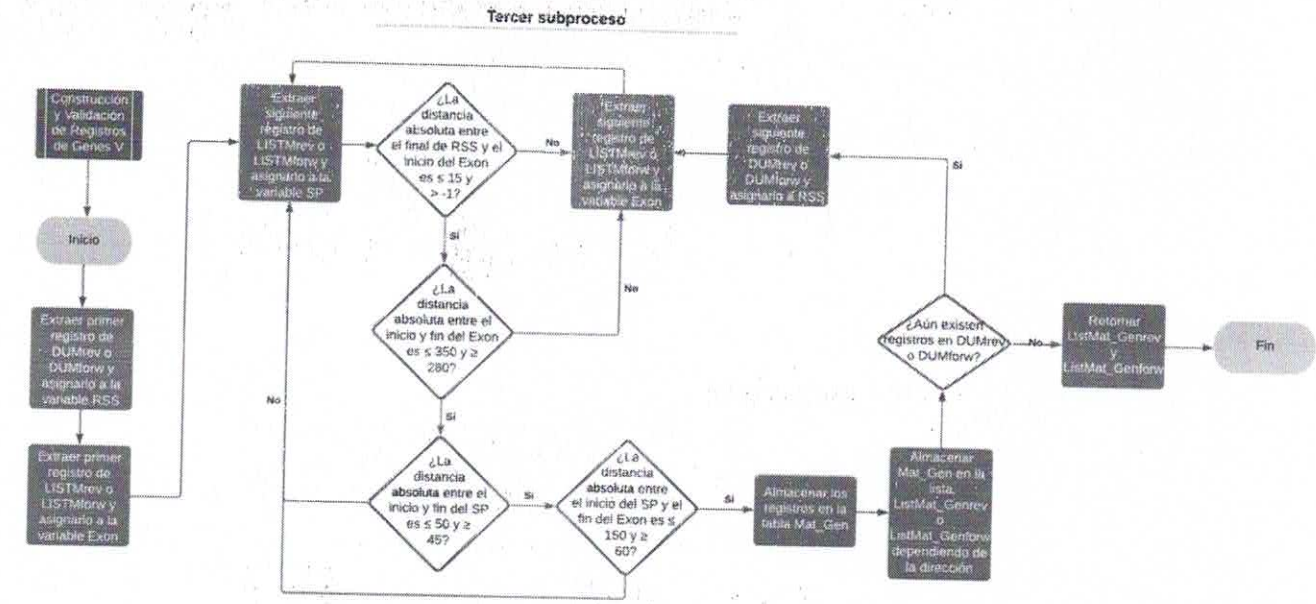

Fig. 9. Diagrama del proceso general

La Figura 10 representa el cuarto subproceso, enfocado en la construcción de matrices de adyacencia y su vectorización, una fase crítica para transformar los registros del gen V en una representación matricial y vectorial. Este proceso optimiza el análisis en etapas posteriores, asegurando una relación precisa entre las secciones del gen. El procedimiento inicia con la combinación de las listas ListMat_Genrev y ListMat_Genforw, generando una nueva lista denominada ListMat, que contiene todos los registros unificados. A partir de ListMat, se extrae el primer registro y se almacena en MatrizGen, estableciendo la base para la construcción de relaciones entre los datos.

La matriz de adyacencia (AdMatriz) se construye para representar todas las distancias posibles entre los datos del registro actual y en específico de sus campos inicio y fin. Para evitar redundancias, se establece que todos los valores por debajo de la diagonal sean cero, asegurando una estructura coherente. Luego, los registros se vectorizan para su almacenamiento. Las posiciones de inicio y fin se almacenan en el arreglo Gen, mientras que la matriz de adyacencia se vectoriza y se guarda en Distans. Estas dos listas se combinan para formar el arreglo Totalgen, facilitando su manipulación y estructuración en un conjunto de datos ordenado.

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

En la siguiente etapa, se determina la dirección de los registros a partir de MatrizGen, clasificándolos como Forward o Reverse. Dependiendo de su dirección, se agrega un nuevo registro a Totalgen, asignando valores 0 ó 1 según corresponda. Todos los registros generados para Totalgen se almacenan en MatFinal, que será el nuevo conjunto de datos, debido a que contendrá la estructura del gen, su dirección y sus distancias. El subproceso se repite hasta que no existan más registros en ListMat.

El proceso de unificación es esencial para generar registros completos y estructurados que representan de manera integral un gen V. Cada registro debe abarcar toda la estructura del gen, evitando fragmentaciones parciales. La construcción de registros se fundamenta en la formación de matrices de adyacencia y su vectorización, procesos que se detallarán en la sección 2.2.1.1, debido a que dependen de la aplicación de distancias y métodos específicos de estructuración de datos. El método de unificación de datos es altamente adaptable a cualquier genoma que contenga genes V y utilice alineamientos de secuencias para su identificación. Su versatilidad permite reutilizar el método en distintos estudios genómicos, asegurando precisión, eficiencia y aplicabilidad en la caracterización genética.

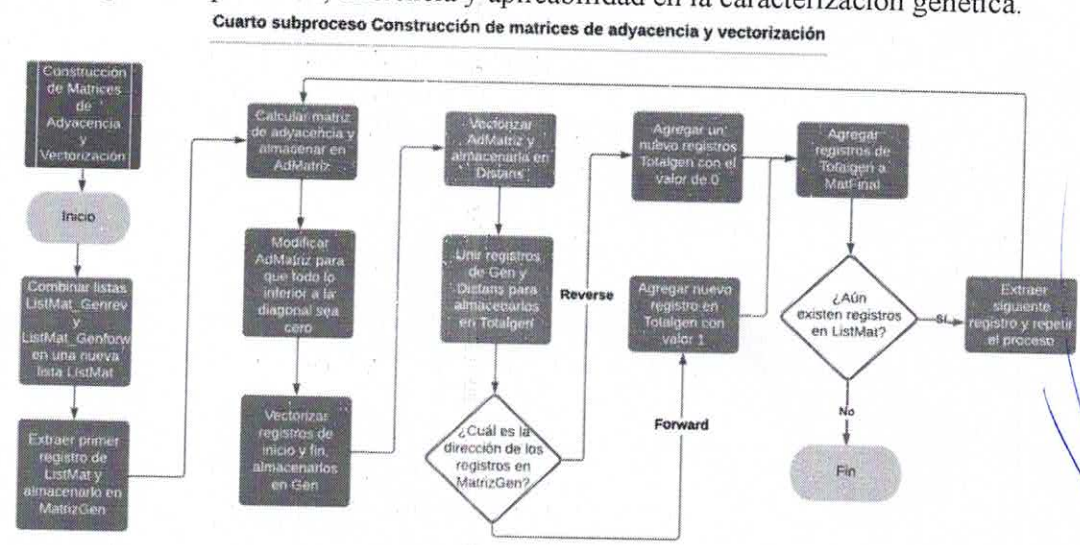

Fig. 10. Diagrama de la construcción y validación de registros de genes V

2.3.1.1 Formación de matrices de adyacencia y su vectorización

Uno de los pasos fundamentales del método de unificación de datos es la formación de registros funcionales a partir de matrices estructuralmente válidas. La sección se centrará en la construcción de estos registros utilizando una matriz de adyacencia, con el objetivo de garantizar que cada registro dentro

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

del conjunto represente distancias validadas entre las secciones del gen V. La formación de matrices depende de los registros generados por los alineamientos de secuencias, organizados en tablas que agrupan tres registros para representar las secciones de un gen. El ejemplo que se presenta se basa en los datos del individuo Tadarida brasiliensis, el cual cuenta con seis archivos de alineamientos de secuencias: (a) Hmmerscan para RSS, (b) Exonerate:protein2genome, (c) Exonerate:est2genome, (d) Tblastx, (e) Tblastn y (f) Hmmerscan para SP.

La Tabla 4 muestra un ejemplo de los registros contenidos en ListMat_Genforw o ListMat_Genrev que se almacenan en MatrizGen. Cada uno de estos registros representa un gen V, con sus respectivas secciones y direcciones asociadas. Sin embargo, aunque la columna Feature especifica la sección del gen que se ha identificado, no siempre se puede tomar como criterio absoluto en la formación de registros. En este caso el primer registro representa la RSS, el segundo el Exon y el tercero el SP, completando así a un gen.

Tabla 4. Contenido del archivo CSV

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	hmmerscan	RSS	3712219	3712180	NA	-	NA	NA
CM040297.1	exonerate:est2genome	exon	3712181	3711873	NA	-	NA	NA
CM040297.1	exonerate:est2genome	gene	3711792	3711746	NA	+	NA	NA

El primer paso en la construcción de registros es la asignación de datos a la variable Gen, utilizando la información de la Tabla 4. En este proceso, se extraen únicamente las posiciones absolutas en el genoma y se vectorizan. La variable Gen define las posiciones que se utilizaran para la matriz de adyacencia. Cada uno de los registros representa una de las secciones del gen V, los primeros 3 registros representan el inicio de la RSS, el Exon y el SP mientras que los últimos tres son las posiciones de fin de estos, a continuación, se presenta el registro:

3712219	3712180	3712181	3711873	3711792	3711746

El cálculo de la distancia se basa en la sustracción progresiva de cada posición en Gen con respecto a todas las demás, repitiendo este procedimiento para cada registro. La Figura 11 presenta este proceso. El resultado es una matriz de adyacencia que representa las distancias absolutas entre las secciones del gen V. Cada posición se compara tanto con otras como consigo misma, asegurando que todas las relaciones espaciales sean correctamente evaluadas. Para evitar valores negativos y garantizar la coherencia estructural, el cálculo se realiza en términos absolutos, manteniendo la precisión en la representación del gen.

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

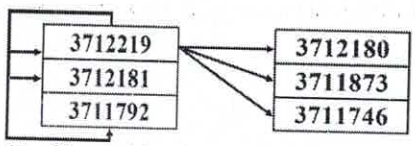

Fig. 11. Obtención de distancias entre puntos

La Tabla 5 presenta la nueva matriz de adyacencia en el cual se tienen los valores absolutos entre las diferentes distancias, Es relevante mencionar que en una primera instancia esta matriz presenta una forma de espejo. Debido a que el espejo se produce en todos los valores por debajo de la diagonal, por lo cual se decide convertir todo lo que esté por debajo de la diagonal en cero. La nueva matriz podría ser utilizada en su forma actual pero no sería del todo apropiada para un algoritmo de aprendizaje automático, por ende, se debe vectorizar para generar un solo registro. Este proceso se repite para cada uno de los registros.

Tabla 5. nueva matriz de adyacencia

	3712219	3712180	3712181	3711873	3711792	3711746
3712219	0	39	38	346	427	473
3712180	0	0	1	307	388	434
3712181	0	0	0	308	389	435
3711873	0	0	0	0	81	127
3711792	0	0	0	0	0	46
3711746	0	0	0	0	0	0

Con la matriz de adyacencia generada, se procede a vectorizar los valores de las distancias y agregarlos al vector Totalgen. Posteriormente, se verifica la dirección del gen V utilizando el primer registro de la variable MatrizGen, lo que permite incorporar esta información dentro del vector. Finalmente, la variable Totalgen se integra en la Tabla MatFinal, donde se almacenan todos los registros de los genes V. El nuevo conjunto de datos se compone de 43 características de las cuales las primeras siete son las posiciones de los alineamientos en el genoma. Estas primeras características se mantienen para preservar la relación entre los registros y su ubicación en el genoma. A continuación, se presentan los siete primeros atributos.

Atributos del 1 al 7	Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP	Dirección

La Tabla 6 presenta los 36 atributos que representan las relaciones entre cada una de las secciones del gen V y las distancias entre ellas. Como se mencionó previamente, todos los valores situados debajo de la diagonal de la matriz de adyacencia se convierten en ceros. Con la integración de estos datos, se obtiene el conjunto final, listo para su uso en algún algoritmo de aprendizaje automático.

Table 6. Atributos representan las relaciones entre cada una de las secciones del gen V

Atributos del 8 al 13	Atributos del 14 al 19	Atributos del 20 al 25	Atributos del 26 al 31	Atributos del 32 al 37	Atributos del 38 al 43
InicioRSS-InicioRSS	Fin_RSS-InicioRSS	Inicio_Exon-Inicio_RSS	Fin_exon-Inicio_RSS	Inicio_SP-Inicio_RSS	Fin_SP-Inicio_RSS
InicioRSS-FinRSS	Fin_RSS-Fin_RSS	Inicio_Exon-Fin_RSS	Fin_exon-Fin_RSS	Inicio_SP-Fin_RSS	Fin_SP-Fin_RSS
InicioRSS-Inicio_Exon	**Fin_RSS-Inicio_Exon**	Inicio_Exon-Inicio_Exon	Fin_exon-Inicio_Exon	Inicio_SP-Inicio_Exon	Fin_SP-Inicio_Exon
InicioRSS-Fin_Exon	**Fin_RSS-Fin_exon**	**Inicio_Exon-Fin_Exon**	Fin_exon-Fin_Exon	Inicio_SP-Fin_Exon	Fin_SP-Fin_Exon
InicioRSS-Inicio_SP	**Fin_RSS-Inicio_SP**	**Inicio_Exon-Inicio_SP**	**Fin_exon-Inicio_SP**	Inicio_SP-Inicio_SP	Fin_SP-Inicio_SP
InicioRSS-Fin_SP	**Fin_RSS-Fin_SP**	**Inicio_Exon-Fin_SP**	**Fin_exon-Fin_SP**	**Inicio_SP-Fin_SP**	Fin_SP-Fin_SP

Es fundamental comprender tanto la estructura de los datos originales como la del nuevo conjunto, debido a que estos siguen un enfoque de aprendizaje no supervisado. Para evitar la pérdida de información relevante, se debe conservar la trazabilidad de los registros durante el proceso de transformación y unificación. Además, la selección de algoritmos adecuados es crucial para procesar los datos de manera óptima y garantizar una correcta interpretación de las relaciones espaciales entre las secciones del gen V. Por ello, la siguiente actividad consiste en realizar un filtrado de columnas y la normalización de los datos.

2.3.2 Normalización de datos

La evaluación de la efectividad del proceso de unificación de datos se basó en su uso en los conjuntos de datos GFF de la especie Tadarida brasiliensis. Esta especie fue seleccionada para pruebas, debido a la disponibilidad y cantidad de datos de alineamientos presentes en su genoma. El conjunto analizado consta de 1,006,797 registros, distribuidos entre sus seis archivos GFF, representando la magnitud de datos con la que los expertos deben trabajar diariamente en la caracterización del genoma y la identificación del gen V. El total de registros posteriores a la aplicación del proceso de unificación es de 46,787 registros. El porcentaje en la reducción del espacio de búsqueda se calcula mediante la siguiente fórmula: **((Total inicial−Total final) /Total inicial)×100**. Esta fórmula permite obtener el primer porcentaje sobre la reducción en el espacio de soluciones, proporcionando una medida cuantitativa del proceso de unificación de datos. En este caso, la reducción alcanzada es del 95.35%, lo que representa una disminución significativa en la cantidad de registros a analizar.

Uno de los primeros puntos que son abordados sobre el conjunto de datos son las columnas compuestas completamente por ceros. La Figura 12 presenta una fracción del conjunto de datos resultante del proceso. Ejemplos de columnas sin relevancia son: a) InicioRSS-InicioRSS o b) InicioRSS-FinRSS, estos atributos se conforman completamente de cero por lo cual son filtradas y eliminados. Posterior a este filtrado, el conjunto de datos queda reducido a 22 columnas, las cuales contienen información sobre las distancias entre secciones del gen V y sus posiciones originales en el genoma. Finalmente se divide el conjunto y se extraen las primeras siete columnas y solo se harán uso de 15 atributos, las distancias.

direccion	Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP	InicioRSS-Ini	InicioRSS-fin	InicioRSS-Ini	InicioRSS-Fir	InicioRSS-Ini	InicioRSS-Fir	Fin_RSS-Inici	Fin_RSS-Fin_	Fin_RSS-Inici
0	15507	15543	15931	15542	15857	15975	0	35	36	350	424	468	0	0	1
0	15507	15543	15929	15542	15854	15970	0	35	36	347	422	463	0	0	1
0	15507	15543	15929	15542	15851	15970	0	35	36	344	422	463	0	0	1
0	15507	15543	15928	15542	15851	15975	0	35	36	344	421	468	0	0	1
0	15507	15544	15927	15542	15846	15968	0	35	37	339	420	461	0	0	2
0	15507	15544	15927	15542	15849	15968	0	35	37	342	420	461	0	0	2
0	15507	15544	15928	15542	15846	15975	0	35	37	339	421	468	0	0	2
0	15507	15544	15928	15542	15849	15975	0	35	37	342	421	468	0	0	2

Fig. 12 fracción del conjunto de datos resultante del proceso

Al iniciar el análisis de los datos, el primer paso es verificar su normalidad. Los gráficos de distribución en la Figura 13 representan exclusivamente las distancias entre las secciones del gen y en particular, la relación entre el inicio de la RSS y el fin del Exon. En la gráfica sin normalizar se observa que las distancias oscilan entre 300 a 38 caracteres, un rango común para la distancia entre las dos secciones del gen.

Sin embargo, los datos aún no presentan una escala uniforme ni una aproximación clara a la normalidad, lo que resalta la necesidad de aplicar un proceso de normalización. Para este estudio, se utilizó la función normalize de scikit-learn, basada en la norma L2 o euclidiana (scikit-learn developers, n.d.), ajustando los valores a un vector de longitud 1. La segunda gráfica de la Figura 13 muestra los datos tras la normalización, evidenciando una distribución más uniforme y una mejor adaptación a los algoritmos de procesamiento, lo que optimiza su rendimiento en las siguientes etapas del análisis.

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

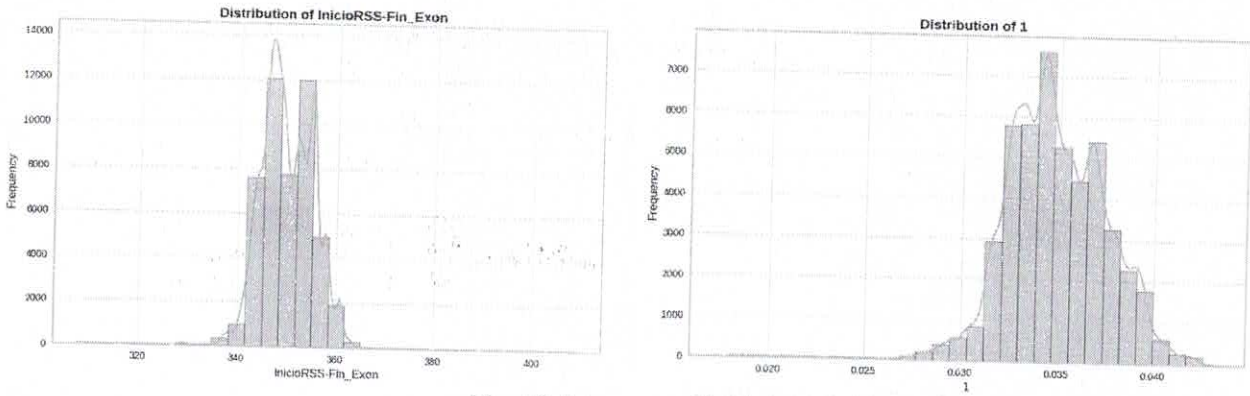

Fig. 13 datos normalizados

Con la normalización de los datos, se procedió con el desarrollo de dos nuevos gráficos que representan la dispersión de los datos. La Figura 15 muestra estos gráficos, donde se observa el estado de los datos antes y después de la normalización. Es importante destacar la nueva distribución posterior a la normalización. El gráfico de la derecha presenta los datos originales y sin procesar mientras que el gráfico de la izquierda presenta los datos después de la normalización. Esta normalización permitirá un correcto procesamiento por parte de los algoritmos a utilizar, permitiendo identificar relaciones y patrones desconocidos.

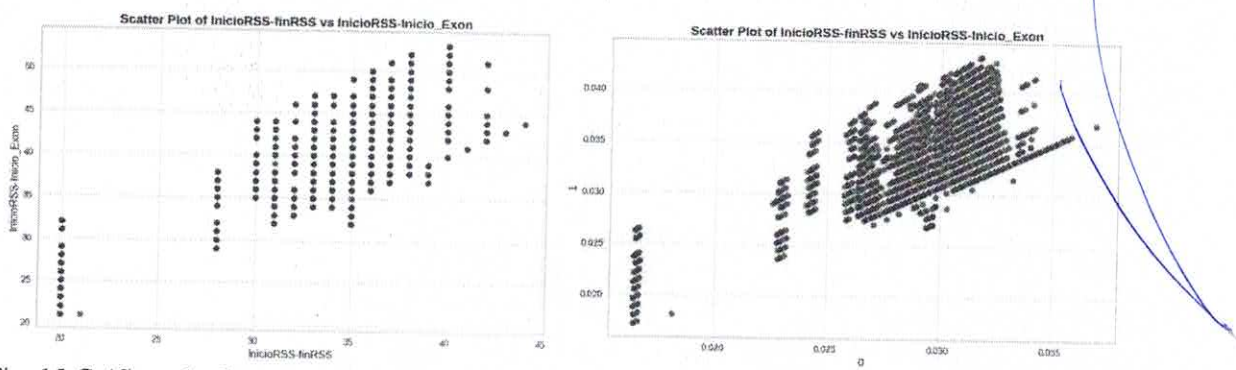

Fig. 15 Gráficos de dispersión de la primera y segunda iteración del algoritmo de unificación de datos

Finalmente cabe mencionar que el proceso de normalización hace uso de los datos del alineamiento y de datos de referencia tratados por el proceso de unificación. La Razón para utilizar todos los datos al mismo tiempo es utilizar todos los datos posibles y verificar el efecto de los datos de referencia. Pero estos datos de referencia son extraídos para el procesamiento de datos para no afectar la implementación del modelo.

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

El algoritmo solo hará uso de los 46,787 registros para su implementación, obviando un total de 918 registros de referencia que serán usados para la validación del mejor grupo.

2.4 Etapa de procesamiento de datos

La etapa de procesamiento consiste en la aplicación de un algoritmo de aprendizaje automático adecuado para el tipo de datos utilizados. En este caso, el objetivo es la reducción del espacio de búsqueda y la identificación de genes V, lo que implica trabajar con un conjunto de datos no clasificado, debido a que no se dispone de características que permitan diferenciar claramente entre clases definidas. Debido a esto, el problema recae dentro del ámbito del aprendizaje no supervisado, lo que requiere la selección de un algoritmo de agrupamiento capaz de identificar registros con mayor similitud entre sí.

El algoritmo seleccionado es Mezclas Gaussianas (Gaussian Mixture Models, GMM) debido a su capacidad para modelar distribuciones de datos complejas y detectar patrones ocultos sin necesidad de etiquetas predefinidas. A diferencia de otros métodos de agrupamiento como K-Means, GMM permite modelar datos que no necesariamente siguen una estructura esférica, lo que lo hace más adecuado para conjuntos de datos con variabilidad en sus distribuciones (Oscar Contreras Carrasco, 2024; scikit-learn, n.d.).

Con el proceso de agrupamiento definido y el algoritmo seleccionado, es necesario determinar el número óptimo de grupos. Aunque esta decisión puede realizar se de forma arbitraría, en el presente estudio se ha decidido hacer uso de uno los enfoques más utilizados para la definición de grupos el cual es el método del codo (Elbow Method), el cual identifica el punto donde la disminución de la inercia (variabilidad dentro de los grupos) (Shi et al., 2021).

Finalmente, con la definición del número de grupos se procede con el último punto del procesamiento de datos la unificación por grupos la cual consiste en unificar los datos por cada uno de los grupos, esto se debe a que en cada grupo se habrán encontrado registros similares por lo cual se busca minimizar la cantidad de resultados de cada grupo. Esta reducción final es necesaria debido a que se provee un solo resultado por cada uno de los genes V de cada grupo. A continuación, se presentará el procesamiento de datos realizados para el conjunto de datos Tadarida brasiliensis.

2.4.1 Selección de grupos y aplicación del algoritmo Mezclas Gaussianas

El primer paso en esta sección fue la selección del algoritmo de agrupamiento más adecuado para la naturaleza de los datos. Como se ha mencionado anteriormente, debido a la estructura de los datos, es necesario utilizar una técnica de aprendizaje no supervisado, en particular un algoritmo de agrupamiento. Aunque existen múltiples enfoques para la segmentación de datos, en este estudio se ha optado por utilizar el modelo de Mezclas Gaussianas (GMM, Gaussian Mixture Model). Este algoritmo asume que los datos provienen de múltiples distribuciones normales superpuestas, lo que permite modelar la incertidumbre en la asignación de los puntos a los grupos. A diferencia de algoritmos como K-Means, que asigna cada punto estrictamente a un único grupo, GMM permite que los puntos tengan una probabilidad de pertenencia a distintos grupos, lo que lo hace más adecuado para datos con estructuras no esféricas (Oscar Contreras Carrasco, 2024; scikit-learn, n.d.).

El análisis de la Figura 15 representa la dispersión de los datos, permitiendo identificar tres características clave en su distribución:

1. Estructura no esférica:
   - En la gráfica, los puntos no forman grupos bien delimitados con forma circular, lo que sugiere que un algoritmo como K-Means no generaría resultados óptimos.
   - GMM modela cada grupo como una distribución gaussiana con su propia media y covarianza, lo que permite capturar formas elípticas y distribuciones más complejas.
2. Superposición de grupos potenciales:
   - La gráfica muestra regiones con alta densidad de puntos, lo que indica la presencia de subgrupos dentro de los datos.
   - GMM, al asignar probabilidades en lugar de etiquetas fijas, permite una clasificación más flexible y adaptable a datos que no tienen límites definidos entre grupos.
3. Heterogeneidad en la dispersión de los puntos:
   - Algunos conjuntos de puntos presentan una mayor dispersión en comparación con otros, lo que sugiere que los grupos pueden tener varianzas diferentes.
   - Mientras que K-Means asume que todos los grupos tienen varianzas iguales, GMM permite que cada distribución gaussiana tenga covarianzas diferentes, lo que lo hace más adecuado para datos con estructuras irregulares.

Con base en estas observaciones, GMM se definió como el algoritmo seleccionado para este estudio, debido a que permite capturar complejidades en la distribución de los datos y generar agrupamientos más representativos de la estructura subyacente en el conjunto analizado. La necesidad de obtener grupos con similitudes claras en sus datos es uno de los principales motivos para su selección.

Con el preprocesamiento de los datos y la definición del algoritmo de agrupamiento se procedió a definir el número óptimo de grupos. Por ello, se decidió utilizar el método del codo, el cual permite determinar el punto en el que aumentar el número de grupos deja de proporcionar mejoras significativas en la compactación de los datos. La Figura 16 muestra la relación entre el número de grupos (k) y la inercia, observándose una reducción drástica de la inercia entre k = 2 y k = 4. Sin embargo, a partir de k = 4, la disminución en la inercia se vuelve menos pronunciada, lo que indica que añadir más grupos no mejora significativamente la calidad del agrupamiento. En este estudio se harán uso de 4 grupos.

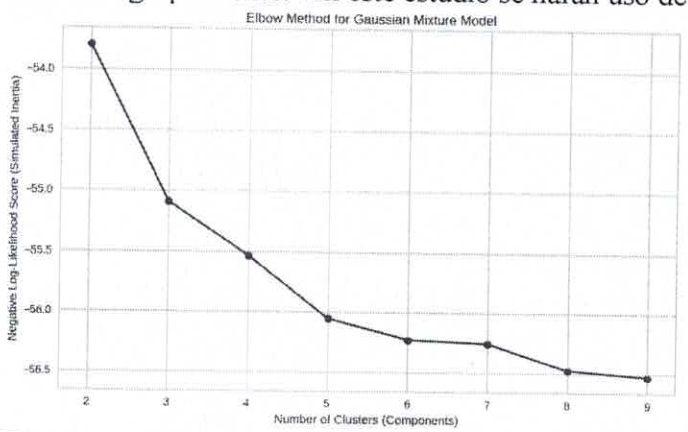

Fig. 16 la relación entre el número de grupos (k) y la inercia

Con la definición del número de grupos y el algoritmo a usar se procedió con el entrenamiento y agrupamiento de los datos, Los gráficos de dispersión con los datos agrupados se presentan en la Figura 17. En estos gráficos, los cuatro grupos se encuentran en colores diferentes además se hizo uso de una técnica de reducción de dimensiones, específicamente PCA (Principal Component Analysis) con el objetivo de mejorar las gráficas para los lectores. Cabe mencionar que en los gráficos se presenta un nuevo grupo con la etiqueta 10, este grupo representa los datos de referencia, los cuales no fueron utilizados en el entrenamiento del algoritmo. El objetivo de graficar estos datos es evaluar su proximidad a los otros grupos y analizar si presentan alguna relevancia dentro de la segmentación realizada.

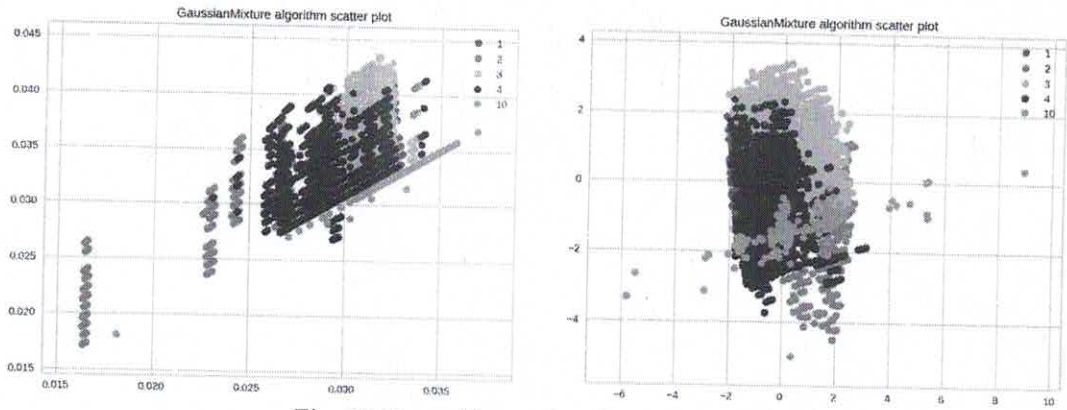

Fig. 17 dispersión con los datos agrupados

Finalmente, el análisis de los grupos indica que: a) el grupo uno ha agrupado **8219**, b) el grupo dos ha agrupado **1117,** c) el grupo tres ha agrupado **33968** y d) el grupo cuatro ha agrupado un total de **3482**. Cada uno de estos grupos contiene una cantidad N de genes V, pero uno de los problemas existentes en estos grupos es que aún existe repetibilidad por lo cual realizar un proceso de unificación de datos resolverá este problema.

2.4.2 Proceso de unificación de datos por grupos

La unificación de datos por grupo es el último punto del procesamiento, este permite minimizar la cantidad de resultados dentro de cada agrupamiento. Esta reducción es fundamental para la comparación de la calidad de cada grupo con la referencia. El proceso se fundamenta en que cada grupo identifica una cantidad N de genes V. Sin embargo, estos genes pueden aparecer en varios registros con pequeñas variaciones, generando redundancias dentro de un grupo. Por ejemplo, en la Tabla 7, la posición 15946089 corresponde al inicio del gen siendo el inicio de la RSS, y se repetirán múltiples veces con ligeras variaciones en los registros.

Tabla 7 Posiciones de inicio y fin de los genes

Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP
15946089	15946130	15946516	15946129	15946438	15946566
15946089	15946130	15946516	15946129	15946441	15946566
15946089	15946130	15946516	15946129	15946444	15946566

Dado que esta repetitividad es común en la mayoría de los grupos se ha propuesto resolver este problema, con la implementación de un proceso de unificación de datos por grupo, el cual consiste en calcular el

promedio de cada columna, generando un único registro representativo por gen. El siguiente ejemplo corresponde a los registros de la Tabla 7.

Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP
15946089	15946130	15946519.9	15946129	15946441.4	15946566

Sin embargo, debido a que algunas columnas contienen valores decimales (como Inicio_SP y Fin_Exon), el registro generado no es exacto, lo que impide su correcta interpretación. Para solucionar esto, se busca el registro más cercano inferior dentro del grupo, asegurando que los valores mantengan su coherencia biológica.

Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP
15946089	15946130	15946519	15946129	15946438	15946564

Este proceso de unificación se repite para cada grupo con el objetivo de obtener genes V únicos dentro de cada agrupamiento, permitiendo su evaluación en comparación con el estándar de referencia. Debido a la naturaleza del problema, no es posible aplicar cualquier métrica de evaluación, debido a que se requiere un enfoque específico que garantice una medición precisa y consistente. Para ello, la fase final del método incorpora el posprocesamiento, descrito en la sección de resultados.

3 Posprocesamiento

La última etapa del método corresponde al posprocesamiento, cuyo objetivo es evaluar y seleccionar el mejor grupo dentro de los diferentes grupos generados por el algoritmo GMM. Las métricas utilizadas se basan en la medición del error debido a que permite cuantificar la calidad de los agrupamientos en relación con la referencia. Las métricas utilizadas en esta evaluación incluyen: a) MAE (Mean Absolute Error), b) MSE (Mean Squared Error), c) RMSE (Root Mean Squared Error) y d) Cantidad de genes de cada grupo. Estas métricas permiten medir la precisión del método en la reducción del espacio de soluciones y en la identificación correcta de genes V, garantizando que los registros unificados reflejen con precisión la estructura real del genoma.

Finalmente se aborda la tarea de la obtención del archivo reducido donde se compran las posiciones del genoma del mejor grupo contra los archivos originales para producir un archivo reducido con las mejores posiciones y utilizarlo en la herramienta IGV para demostrar la efectividad del método para la reducción del espacio de soluciones de genes V en genomas de vertebrados

3.1 Evaluación y selección del mejor grupo

En la interpretación de los resultados para la selección del mejor grupo, primero es necesario comprender la relación entre el nuevo conjunto de datos y la referencia utilizada para la validación. Como se mencionó anteriormente, el conjunto de datos unificado contiene 46,787 registros, cada uno representando un posible gen V. Sin embargo, una de las preguntas clave para el método es: ¿Cuántos genes V reales se pueden identificar dentro de estos 46,787 registros? Para responder a esta pregunta, se utiliza una función unique sobre la primera columna del conjunto de datos (Inicio_RSS), lo que permite contabilizar la cantidad de genes V únicos siendo el total de 842 genes V que se pueden identificar. Es fundamental conocer esta cifra, debido a que establece un límite en la cantidad máxima de genes V que pueden ser identificados en el conjunto de datos. A partir de esto, surge la segunda pregunta: ¿Cuántos genes V completos y validados existen en la referencia? La respuesta a esta pregunta es de un total de 917 registros. Estos registros serán utilizados para la evaluación del método.

El proceso de evaluación sigue el principio de medición en regresión, donde se comparan un valor esperado (referencia del experto) y un valor predicho (datos obtenidos del método de reducción de soluciones). La Tabla 8 muestra este proceso de comparación, donde: a) la primera fila representa los valores anotados por los expertos (referencia esperada); b) la segunda fila corresponde a los datos generados por el método de unificación; c) El tercer registro muestra la diferencia entre ambos valores, utilizada para calcular las métricas de error.

Tabla 8 comparación entre esperado y valor predicho

	Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP
Referencia	88965	88926	88539	88930	88615	88495
Método	88967	88929	88536	88929	88621	88490
Diferencia	-2	-3	3	1	-8	5

Los resultados de la evaluación se presentan en la Tabla 9, donde se muestra la calidad de cada uno de los grupos generados. En esta comparación, se evidencia una variabilidad en la cantidad de genes identificados, con un error aproximado de 3 en la mayoría de los casos. Aunque la cantidad total de genes identificados no es de 842, la explicación de esto se debe a que algunos genes presentan alteraciones estructurales que dificultan su correcta agrupación. A pesar de esto, el grupo 3 se destaca como el mejor grupo, debido a que produce el menor error y cuenta con la mayor cantidad de genes V.

Tabla 9 calidad de cada uno de los grupos generados

Grupo	Núm. de datos	Núm. de genes	MAE	MSE	RMSE
1	8219	490	2.820408	22.254422	4.717459
2	1117	21	2.47619	15.793651	3.974123
3	33968	734	1.986376	10.027248	3.166583
4	3482	175	3.333333	40.95619	6.399702

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

La Tabla 9 proporciona una perspectiva global del error, pero las gráficas de distribución de la Figura 17 permiten dimensionar el error y la variabilidad presente en cada sección del gen V. Los gráficos muestran cómo varía el desplazamiento en cada sección del gen V dentro de los grupos analizados. Un ejemplo es la posición de inicio de la RSS, la cual presenta un error cercano a cero en el grupo 3, con la mayoría de los valores desplazándose entre +1 y +4. Esta estabilidad en el error indica que los registros dentro del grupo 3 se ajustan mejor a la referencia del experto en la sección del inicio de la RSS.

En el caso del inicio del Exon se muestra una mayor variabilidad, incluso dentro del grupo 3. En este caso, los valores oscilan entre +1 y +10, con ciertas concentraciones en posiciones como el 3, lo que indica un mayor grado de dispersión. Patrones similares se observan en las demás secciones del gen V, reflejando que, aunque el método logra producir un nivel de error bajo, algunas posiciones presentan mayor variabilidad. Aun así, los resultados establecen un punto de partida sólido para mejorar la precisión del modelo en futuras optimizaciones.

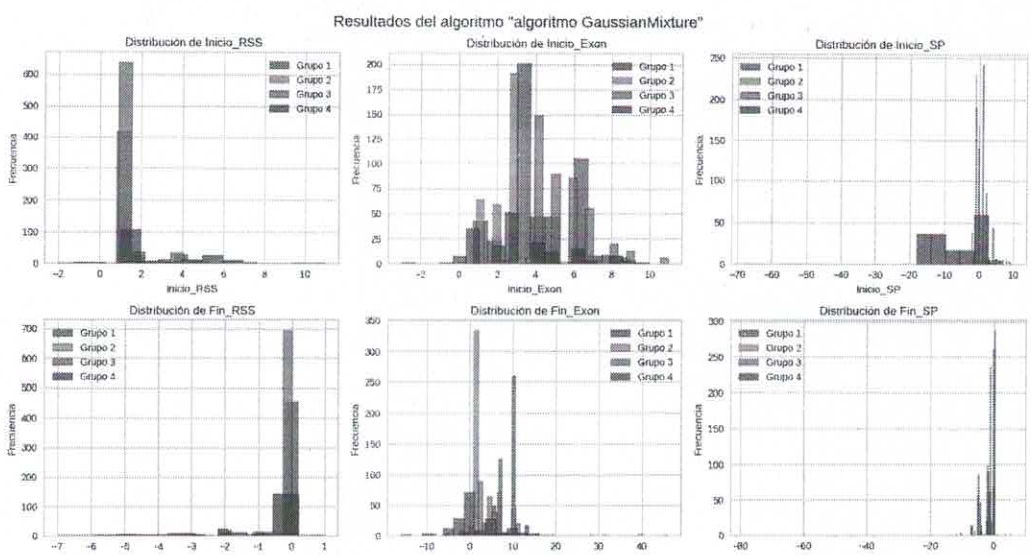

Fig. 17 error y la variabilidad presente en cada sección del gen

3.2 Obtención del archivo reducido y su uso en IGV

Finalmente, el archivo reducido se obtiene a partir del mejor grupo identificado en la sección anterior. En la última fase del método, se deben utilizar las primeras siete columnas del conjunto, donde se registran las posiciones del gen V. A continuación, se presenta un ejemplo de registro perteneciente al

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

Educación
Secretaría de Educación Pública
TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

grupo 3. Como se observa la primera columna indica la dirección, las tres siguientes corresponden a las secciones iniciales del gen, mientras que las tres últimas representan las secciones restantes del gen. Para completar el proceso, es necesario extraer y consolidar las tres primeras columnas en una única columna y repetir el procedimiento con las tres restantes, garantizando así una representación estructurada del gen V.

Dirección	Inicio_RSS	Inicio_Exon	Inicio_SP	Fin_RSS	Fin_Exon	Fin_SP
Reverse	15507	15543	15928	15542	15851	15975

Obtenida una tabla con solo dos registros, se procede a buscar en los archivos originales, registros que generen las mismas posiciones que la tabla por ejemplo Inicio_RSS con 15507 y Fin_RSS con 15542 y se extrae el registro que sea idéntico a estas posiciones. Finalmente, estos resultados se transforman al formato GFF y se unifican en un solo archivo. La Figura 18 muestra un ejemplo de una fracción de este archivo con el presente ejemplo. En este formato, cada tres registros representan un gen, consolidando la información de manera estructurada y facilitando su análisis en herramientas de anotación genómica.

```
1  CM061270.1  hmmerscan   RSS 15507   15542   .   -   .   Query=clustal_IGHV_RSS3
2  CM061270.1  tblastn gene    15543   15845   .   -   .   Query=Arja_IGHV_038
3  CM061270.1  tblastx gene    15929   15970   .   -   .   Query=Rhfe_IGHV_017/1-356
```

Fig. 18 ejemplo de una fracción de archivo

Finalmente, para demostrar la viabilidad del método de manera gráfica, se decidió presentar los resultados en la herramienta utilizada por los expertos. Este enfoque permite mostrar cómo el método mejora la reducción del espacio de búsqueda, facilitando la identificación de genes V en el genoma analizado. En la Figura 19, se presentan los grupos generados por el método, donde se observa la distribución y segmentación de los datos tras la aplicación del proceso. La herramienta de visualización permite apreciar la magnitud del genoma de Tadarida brasiliensis en relación con los genes V, evidenciando la efectividad del método para consolidar información dispersa en alineamientos más precisos.

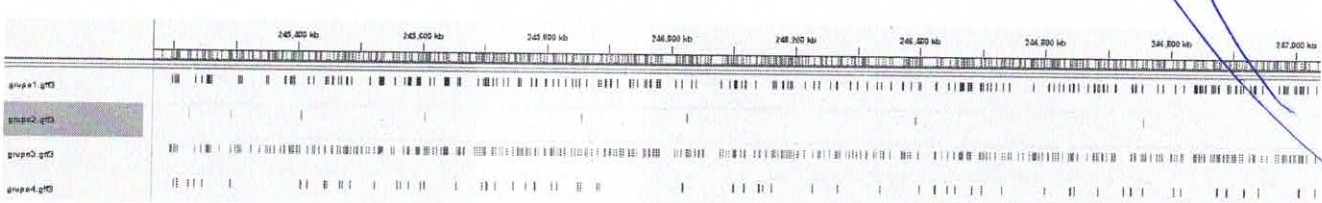

Fig. 19 grupos generados por el método

Educación
Secretaría de Educación Pública
TECNOLÓGICO NACIONAL DE MÉXICO
Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Cuando se amplía el análisis sobre un conjunto de alineamientos, surge un fenómeno clave: algunos grupos identifican genes que otros no. Esto resalta la importancia de comprender que el método no es completamente infalible, debido a que su precisión está influenciada por la naturaleza de los genes V y la variabilidad en sus alineamientos. Si bien ciertos grupos pueden detectar genes que otros no logran identificar, la cuestión fundamental no es solo qué genes se identifican, sino con qué precisión se logran estos resultados. Evaluar este aspecto es esencial para determinar la fiabilidad del método, asegurando que las detecciones sean consistentes y reproducibles dentro del proceso de caracterización genómica. Ver Figura 20.

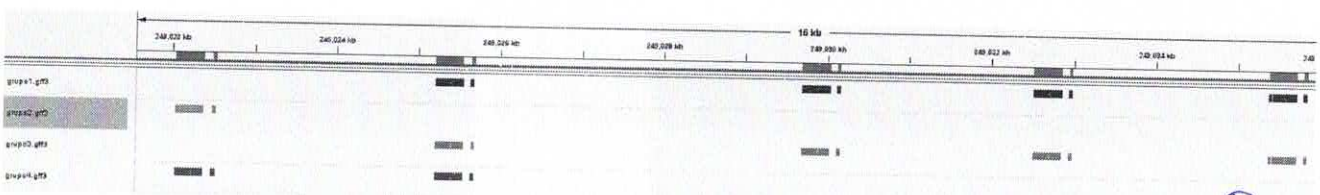

Fig. 20 genes identificados

Al seleccionar un gen se presenta la Figura 21 en la cual muestran las tres secciones del gen en este caso el gen cuenta con una dirección Reverse, en rojo se presenta la referencia del experto. Es relevante mencionar que el grupo tres muestra una gran cercanía a la referencia, pero no es perfecta esto se demuestra por los caracteres faltantes al inicio del gen. Pero lo compensa en gran medida con la exactitud en el SP, la última sección del alineamiento.

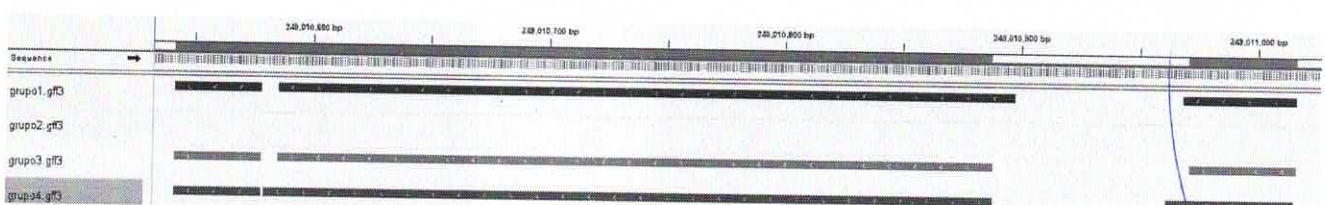

Fig. 21 tres secciones del gen

Finalmente, cuando se realiza un mayor acercamiento al inicio del Exon y al alineamiento del SP se muestra la Figura 20 donde se aprecia la cercanía de cada grupo a la referencia. Como se muestra en grupo 3 y 4 presentan una gran cercanía a la referencia en el caso del inicio del Exon pero en el alineamiento del SP es donde se demuestra la diferencia de cada grupo. En el caso del SP para el grupo 3 es el que presenta una mayor exactitud de todos los grupos, en el caso del grupo 3 solo varia por una posición siendo casi perfecto.

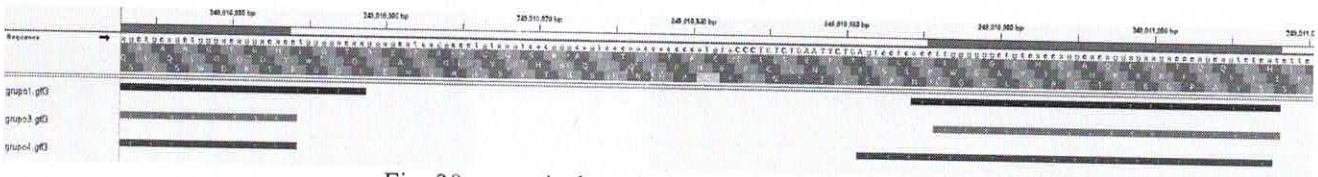

Fig. 20 cercanía de cada grupo a la referencia

4 Discusión

El método propuesto para la unificación de datos y reducción del espacio de soluciones demostró ser una herramienta eficaz para mejorar la identificación de genes V en grandes volúmenes de datos genómicos. La integración de múltiples fuentes de alineamiento permitió una representación más completa del gen, reduciendo la redundancia y minimizando el esfuerzo manual requerido por los expertos. Sin embargo, a pesar de la mejora en la precisión y eficiencia, el método aún presenta desafíos, especialmente en la identificación de secciones específicas del gen, como el Exon, donde se observó una variabilidad mayor en comparación con la referencia del experto.

Uno de los aspectos más relevantes del análisis fue la evaluación del rendimiento del método en términos de tiempo de cómputo. La ejecución en un equipo con especificaciones moderadas (Asus TUF con 16 GB de RAM, disco duro mecánico, Intel Core i7U y una tarjeta gráfica NVIDIA RTX 3070) permitió completar el método en aproximadamente 50 minutos, lo que representa una mejora significativa en comparación con el método manual. No obstante, este tiempo podría optimizarse con el uso de arquitecturas computacionales más avanzadas o mediante la comparación de otros modelos o técnicas de reducción de dimensiones que faciliten el procesamiento de los datos.

Otro punto clave en la discusión es la fiabilidad del modelo de Mezclas Gaussianas (GMM) utilizado para el agrupamiento de los datos. Si bien GMM permitió modelar estructuras complejas y capturar variaciones en la distribución de los alineamientos, la asignación probabilística de los registros a los grupos sugiere que algunos genes podrían haber sido clasificados en múltiples agrupamientos. En futuras investigaciones se explora la combinación de GMM con otros enfoques de aprendizaje automático, como la reducción de dimensiones para variar la estructura y dispersión de los datos. Como también analizar otras técnicas de normalización y unificación de datos por grupos e identificar si estas producen alguna variación para la mejora del método.

Por último, la transformación de los resultados al formato GFF fue un paso crucial para garantizar la aplicabilidad del método en herramientas de análisis genómico como IGV. Sin embargo, se identificó que la precisión de los alineamientos depende en gran medida de la calidad de los datos de entrada. En el futuro, se recomienda una evaluación más extensa con diferentes conjuntos de datos. Validar la

generalización a diferentes genomas permitirá robustecer el método y verificar que realmente el método es aplicable a múltiples especies.

5 Conclusiones

El presente estudio presentó un método innovador para la reducción del espacio de soluciones en la identificación de genes V, logrando una mejora sustancial en la precisión y eficiencia del proceso. La combinación de técnicas de preprocesamiento, normalización y agrupamiento permitió disminuir significativamente la cantidad de registros a analizar, facilitando la identificación de genes completos con una mayor aproximación a la referencia experta. Esto se demuestra en la reducción del espacio de soluciones donde se comenzó con un total de 1,006,797 registros para todos los genes y finalizando con grupos con un máximo de 734 registros uno por gen. Aunque existen pérdidas en los datos, la optimización del método permitirá seguir mejorando el desempeño.

Los resultados indican que el método puede ser aplicado a otros genomas, siempre que los datos de entrada cumplan con los criterios de alineamiento y estructuración definidos. La integración de múltiples fuentes de datos en un único conjunto de referencia mejora la calidad de los registros y reduce la redundancia, optimizando el tiempo y los recursos necesarios para la caracterización genómica.

A pesar de sus ventajas, el método aún presenta oportunidades de mejora, especialmente en la identificación de ciertas secciones del gen, como el SP y el Exon, y en la optimización del tiempo de cómputo. Futuras investigaciones podrían enfocarse en el ajuste de los modelos de agrupamiento, la exploración de nuevas técnicas de reducción de dimensionalidad y la evaluación del impacto de la variabilidad genética en la precisión de los alineamientos.

En conclusión, la metodología propuesta representa un avance significativo en la sistematización del proceso de identificación de genes V. Especialmente al hacer uso de distancias y no de la secuencia del genoma, no se descarta hacer uso de la secuencia debido a que esta permitirá terminar de sistematizar el proceso de caracterización del sistema. Sin embargo, este estudio proporciona una base sólida para futuras investigaciones en la bioinformática, la caracterización de genomas y la optimización de procesos computacionales. Su implementación en herramientas de análisis genómico permitirá mejorar la anotación de genes y facilitar la interpretación de datos en estudios de inmunogénica y evolución molecular.

References

Amin, M. R., Yurovsky, A., Tian, Y., & Skiena, S. (2018). Deepannotator: genome annotation with deep learning. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 254–259.
Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). Nonlinear programming: theory and algorithms. John wiley & sons.

Bergman, N. H. (2007). Comparative genomics.
Dagdia, Z. C., & Mirchev, M. (2020). Chapter 15 - When Evolutionary Computing Meets Astro- and Geoinformatics. In P. Škoda & F. Adam (Eds.), Knowledge Discovery in Big Data from Astronomy and Earth Observation (pp. 283–306). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-12-819154-5.00026-6
Ejigu, G. F., & Jung, J. (2020). Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology, 9(9), 295.
EMBL-EBI. (2022). GFF/GTF File Format - Definition and supported options. https://www.ensembl.org/info/website/upload/gff.html
García Simón, A. (2018). Gestión de datos genómicos basada en Modelos Conceptuales.
Kalkatawi, M., Magana-Mora, A., Jankovic, B., & Bajic, V. B. (2019). DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. Bioinformatics, 35(7), 1125–1132.
Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. Gene Prediction: Methods and Protocols, 161–177.
Kindt, T. J., Goldsby, R. A., & Osborne, B. A. (2007). Inmunología de Kuby. McGraw Hill.
Lefranc, M.-P., & Lefranc, G. (2001). The immunoglobulin factsbook. Academic press.
McNair, K., Ecale Zhou, C. L., Souza, B., Malfatti, S., & Edwards, R. A. (2021). Utilizing amino acid composition and entropy of potential open reading frames to identify protein-coding genes. Microorganisms, 9(1), 129.
Megrian, D. (2014). Identificación de genes de inmunoglobulinas en el genoma bovino.
Miguel-Ruiz, J., Serret, N., Ortiz-Hernandez, J., Barnetche, J. M., & Hernández, Y. (2024). A Design Science Approach to Modeling the V Gene Annotation Process. Programming and Computer Software, 50(8), 829–843. https://doi.org/10.1134/S0361768824700798
Mount, D. W. (2004). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press. https://books.google.com.mx/books?id=bvY21DGa1OwC
Olivieri, D. N., & Gambón-Deza, F. (2019). Iterative Variable Gene Discovery from Whole Genome Sequencing with a Bootstrapped Multiresolution Algorithm. Computational and Mathematical Methods in Medicine, 2019, 3780245. https://doi.org/10.1155/2019/3780245
Oscar Contreras Carrasco. (2024). Gaussian Mixture Model Explained . https://builtin.com/articles/gaussian-mixture-model
Pfennig, A., Lomsadze, A., & Borodovsky, M. (2022). Annotation of Phage Genomes with Multiple Genetic Codes. BioRxiv, 2022–2026.
Pieper, K., Grimbacher, B., & Eibel, H. (2013). B-cell biology and development. Journal of Allergy and Clinical Immunology, 131(4), 959–971.
scikit-learn. (n.d.). Gaussian mixture models. Retrieved February 11, 2025, from https://scikit-learn.org/stable/modules/mixture.html
scikit-learn developers. (n.d.). Normalizer. Retrieved July 30, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html
Serret, N., Ortiz-Hernandez, J., Miguel-Ruiz, J., Barnetche, J. M., & Hernández, Y. (2023). Conceptual modeling of the V gene annotation process in antibodies. 2023 11th International Conference in Software Engineering Research and Innovation (CONISOFT), 256–264.
Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1), 31. https://doi.org/10.1186/s13638-021-01910-w
Sirupurapu, V., Safonova, Y., & Pevzner, P. A. (2022). Gene prediction in the immunoglobulin loci. Genome Research, 32(6), 1152–1169.
Stiehler, F., Steinbom, M., Scholz, S., Dey, D., Weber, A. P. M., & Denton, A. K. (2020). Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. Bioinformatics, 36(22–23), 5291–5298.
Webster, C. F., Smotherman, M., Pippel, M., Brown, T., Winkler, S., Pieri, M., Mai, M., Myers, E. W., Teeling, E. C., & Vernes, S. C. (2024). The genome sequence of Tadarida brasiliensis I. Geoffroy Saint-Hilaire, 1824 [Molossidae; Tadarida]. Wellcome Open Research, 9.
Zhang, Y., Chen, T., Zeng, H., Yang, X., Xu, Q., Zhang, Y., Chen, Y., Wang, M., Zhu, Y., & Lan, C. (2021). RAPID: a rep-seq dataset analysis platform with an integrated antibody database. Frontiers in Immunology, 12, 717496.
Zhu, Y., Watson, C., Safonova, Y., Pennell, M., & Bankevich, A. (2024). Assessing assembly errors in immunoglobulin loci: a comprehensive evaluation of long-read genome assemblies across vertebrates. BioRxiv.

Educación
Secretaría de Educación Pública
TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Meta 3. Redacción y sometimiento de 1 artículo en revista indizada en el área. Cantidad: 1

Miguel-Ruiz, J., Serret, N., Ortiz-Hernandez, J. et al. A Design Science Approach to Modeling the V Gene Annotation Process. Program Comput Soft 50, 829–843 (2024). https://doi.org/10.1134/S0361768824700798
JCR IF 0.7, Q4 (Scopus: Q3).

SPRINGER NATURE Link

Log in

Find a journal Publish with us Track your research Search Cart

Home > Programming and Computer Software > Article

A Design Science Approach to Modeling the V Gene Annotation Process

Published: 12 January 2025

Volume 50, pages 829–843, (2024) Cite this article

Programming and Computer Software

Aims and scope →

Submit manuscript →

Juan Miguel-Ruiz, Noemi Serret, Javier Ortiz-Hernandez, Jesus Martinez Barnetche & Yasmín Hernández

31 Accesses Explore all metrics →

Abstract

Design science is a methodology that addresses complex and evolving problems requiring a multidisciplinary approach. In this research, the methodology of design science is employed to construct a model for the process of V gene annotation in vertebrates. The process entails the identification of genes within a genome, the characterization of their structural elements, and the classification of their associated data. The model standardizes the work of molecular biologists, reducing time and errors. The methodology followed in this project includes rigor, relevance, and design cycles, developed in collaboration with

2025 Año de La Mujer Indígena

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

ISSN 0361-7688, Programming and Computer Software, 2022, Vol. .. No. .., pp. ...

A Design Science Approach to Modeling the V Gene Annotation Process

Juan Miguel-Ruiz[1], Noemi Serret[1,*2], Javier Ortiz-Hernandez[1], Jesus Martinez Barnetche[2], Yasmín Hernández[1]

[1] *Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Mexico.*
e-mail: {d20ce065, m22ce046, javier.oh, yasmin.hp}@cenidet.tecnm.mx
[2] *Instituto Nacional de Salud Pública., Cuernavaca, Mexico.*
e-mail: jmbarnet@insp.mx

Abstract— Design science is a methodology that addresses complex and evolving problems requiring a multidisciplinary approach. In this research, the methodology of design science is employed to construct a model for the process of V gene annotation in vertebrates. The process entails the identification of genes within a genome, the characterization of their structural elements, and the classification of their associated data. The model standardizes the work of molecular biologists, reducing time and errors. The methodology followed in this project includes rigor, relevance, and design cycles, developed in collaboration with experts in molecular biology and information systems. The result of the research is a BPMN model, validated by experts and trainees from the Centro de Investigación Sobre Enfermedades Infecciosas (CISEI) of the Instituto Nacional de Salud Pública (INSP) in Cuernavaca, Mexico.

Keywords: design science, process modeling, BPMN, genome annotation

DOI:

1. INTRODUCTION

Design science is concerned with the creation of artifacts that are useful within a specific context. This context includes the circumstances and factors that influence a situation. Context analysis is a crucial element in the design, development, and application of artifacts that are relevant and practical. These artifacts address problems and generate new knowledge, thereby fulfilling the fundamental goal of design science [1], [2], [3].

A properly designed artifact must be practical and relevant to the real world, i.e., it must be relevant to the context in which it will be used. The artifact should be designed to understand, improve or represent a specific problem. By developing a functional and useful artifact, new knowledge is produced, thus fulfilling the fundamental goal of design science [1], [2]. Currently, design science has been used in various research such as in systems development as in [4], [5] and has been addressed in other fields such as: the electrical field, information systems, civil engineering, architecture among others, but currently design science has not been used in the genomic field

This paper employs design science methodology to identify and conceptually model a biological problem. The objective is to represent expert knowledge that is beneficial for comprehending and enhancing genome annotation processes. This paper builds upon and enhances previous research [6].

1.1. *Brief Introduction to Design Science Methodology*

The design science paradigm is a research methodology that focuses on formulating research questions, applying appropriate methodologies to answer them, and defining the

[1*] Corresponding author

La Mujer Indígena

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Meta 4: Formación de recursos humanos a nivel doctorado en desarrollo: Cantidad: 2

1. MC Juan Antonio Miguel Ruiz, Modelo de aprendizaje automático para el análisis de la funcionalidad de segmentos génicos V(D)J en diversas especies de vertebrados, ingreso agosto 2022.
2. Manuel Erazo Valadez, Modelo de aprendizaje automático para el emparejamiento de cadenas pesadas y ligeras de los anticuerpos ANTI-SARS-COV-2, ingreso agosto 2022.

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

EDUCACIÓN

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Avance de Tesis de Doctorado

5to. semestre

Modelo de aprendizaje automático para la identificación de genes V en genomas de vertebrados

Título original: Modelo de aprendizaje automático para el análisis de la funcionalidad de segmentos génicos V(D)J en diversas especies de vertebrados

presentada por

MCC. Juan Antonio Miguel Ruiz

Director de tesis

Dr. Javier Ortiz Hernández

Codirector de tesis

Dr. Jesús Martínez Barnetche

Comité tutorial

Dra. María Yasmín Hernández Pérez

Dr. Dante Mújica Vargas

Dr. Hugo Estrada Esquivel

Dr. Juan Mauricio Téllez Sosa

Cuernavaca, Morelos, México. diciembre de 2024

cenidet

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

cenidet	EVALUACIÓN DE PROYECTO DE INVESTIGACIÓN IV	Código: CENIDET-AC-006-D15
		Revisión: 0
	Referencia a la Norma ISO 9001:2008 7.1, 7.2.1, 7.5.1, 7.6, 8.1, 8.2.4	Página 1 de 3

Cuernavaca, Mor., a 19 de diciembre de 2024

NOMBRE: M.C. JUAN ANTONIO MIGUEL RUIZ

PROGRAMA: DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TITULO DE LA TESIS:
MODELO DE APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE LA FUNCIONALIDAD DE SEGMENTOS GÉNICOS V(D)J EN DIVERSAS ESPECIES DE VERTEBRADOS

En la siguiente tabla se asentarán las calificaciones considerando la información escrita y oral. La escala de calificaciones para el promedio general será de 0 a 100 y la mínima calificación aprobatoria es de 70. Así también para que esta acta tenga validez deberá contener la calificación de al menos tres de los revisores y del director y/o codirector de tesis. En el caso de que algún revisor sea externo o por requerimientos oficiales no esté presente en la evaluación, este podrá evaluar y firmar mediante la digitalización de la presente acta, siendo el primero en asentar calificación y posteriormente los demás miembros podrán evaluar y firmar de forma autógrafa el documento digitalizado.

CRITERIOS	REVISOR 1	REVISOR 2	REVISOR 3	REVISOR 4	DIRECTOR y/o CODIRECTOR
Estructura y claridad del informe. (15%)	14	13	14	14	14
Presentación oral. (15%)	14	14	14	14	14
Evaluar el cumplimiento del 62.5% de avance del proyecto. (50%)	48	17	49	50	50
Nivel de propuesta y descripción detallada de actividades futuras. (20%)	18	17	19	20	20
Suma	94	91	96	98	98
Promedio General					95

DR. JAVIER ORTIZ HERNÁNDEZ
Director de tesis

DR. JESÚS MARTÍNEZ BARNETCHE
Codirector de tesis

DR. DANTE MÚJICA VARGAS
Revisor 1

DR. HUGO ESTRADA ESQUIVEL
Revisor 2

DRA. MARIA YASMIN HERNÁNDEZ PÉREZ
Revisor 3

DR. JUAN MAURICIO TÉLLEZ SOSA
Revisor 4

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA
cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico
TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Documento de avance

Quinto semestre:

Modelo de aprendizaje automático para el emparejamiento de cadenas pesadas y ligeras de los anticuerpos ANTI-SARS-COV-2

Presenta:

Manuel Erazo Valadez

Directora de Tesis:

Dra. María Yasmín Hernández Pérez

Codirectora:

Dra. Elizabeth Ernestina Godoy Lozano

Comité tutorial:

Dr. Javier Ortiz Hernández

Dra. Alicia Martínez Rebollar

Dr. Jonathan Villanueva Tavira

Dr. Juan Mauricio Téllez Sosa

2025 Año de La Mujer Indígena

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

cenidet	EVALUACIÓN DE PROYECTO DE INVESTIGACIÓN IV	Código: CENIDET-AC-006-D15
		Revisión: O
	Referencia a la Norma ISO 9001:2008 7.1, 7.2.1, 7.5.1, 7.6, 8.1, 8.2.4	Página 1 de 3

Cuernavaca, Mor., a 19 de diciembre de 2024

NOMBRE: M.C. MANUEL ERAZO VALADEZ

PROGRAMA: DOCTORADO EN CIENCIAS DE LA COMPUTACIÓN

TITULO DE LA TESIS:
MODELO DE APRENDIZAJE AUTOMÁTICO PARA EL EMPAREJAMIENTO DE CADENAS PESADAS Y LIGERAS DE LOS ANTICUERPOS ANTI-SARS-COV-2

En la siguiente tabla se asentarán las calificaciones considerando la información escrita y oral. La escala de calificaciones para el promedio general será de 0 a 100 y la mínima calificación aprobatoria es de 70. Así también para que esta acta tenga validez deberá contener la calificación de al menos tres de los revisores y del director y/o codirector de tesis. En el caso de que algún revisor sea externo o por requerimientos oficiales no esté presente en la evaluación, este podrá evaluar y firmar mediante la digitalización de la presente acta, siendo el primero en asentar calificación y posteriormente los demás miembros podrán evaluar y firmar de forma autógrafa el documento digitalizado.

CRITERIOS	REVISOR 1	REVISOR 2	REVISOR 3	REVISOR 4	DIRECTOR y/o CODIRECTOR
Estructura y claridad del informe. (15%)	14	13	13	15	14
Presentación oral. (15%)	14	12	13	15	14
Evaluar el cumplimiento del 62.5% de avance del proyecto. (50%)	47	45	50	50	47
Nivel de propuesta y descripción detallada de actividades futuras. (20%)	18	18	18	15	18
Suma	93	88	94	95	93
Promedio General					93

DRA. MARÍA YASMÍN HERNÁNDEZ PÉREZ
Directora de tesis

DRA. ELIZABETH ERNESTINA GODOY LOZANO
Codirectora de tesis

DR. JAVIER ORTIZ HERNÁNDEZ
Revisor 1

DR. JUAN MAURICIO TÉLLEZ SOSA
Revisor 2

DRA. ALICIA MARTÍNEZ REBOLLAR
Revisor 3

DR. JONATHAN VILLANUEVA TAVIRA
Revisor 4

2. Objetivos del proyecto

Objetivo general

Desarrollar modelos y métodos computacionales para la anotación genómica del locus de inmunoglobulinas en vertebrados

Objetivos específicos

Desarrollar la integración de un único conjunto de datos con las posiciones de inicio y fin de los alineamientos de secuencias.de los genes que integran el segmento V a partir de diversas fuentes de datos disponibles.

Implementar un modelo de inteligencia artificial para reducir el espacio de búsqueda de los alineamientos de secuencia con el uso del conjunto de datos del objetivo anterior.

3. Metas: cumplimiento de Metas

Metas cuantificables	Cantidad programada	Cantidad lograda	Observaciones
Incorporación de estudiantes de licenciatura al proyecto (créditos complementarios, servicio social, etc.):			
Estudiantes con residencia concluida			
Tesis en desarrollo de Licenciatura			
Tesis concluida de Licenciatura			
Tesis en desarrollo de Maestría			
Tesis concluida de Maestría			
Tesis en desarrollo de Doctorado	2	2	
Tesis concluida de Doctorado			
Artículos científicos publicados en revistas indizadas enviados	1	1	
Artículos científicos en revistas arbitradas enviados			
Artículos de divulgación enviados			
Artículos científicos en revistas indizadas publicados			
Artículos científicos en revistas arbitradas publicados			
Artículos de divulgación publicados			
Memorias en extenso en congresos			

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

Educación
Secretaría de Educación Pública

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Capítulos de libros enviados para revisión			
Libros enviados para revisión			
Libros editados y publicados			
Registro de Patente (IMPI)			
Registro de Modelo de Utilidad (IMPI)			
Registro de Signos Distintivos (IMPI)			
Registro de Diseño Industrial (IMPI)			
Registro de Esquemas de Trazado de Circuitos Integrados (IMPI):			
Registro de Obra (INDAUTOR):			
Carta de Usuario (Empresa):			

Estudiantes residentes participantes en el Proyecto	
Fecha	Nombre de los estudiantes residentes, nombre del proyecto de residencia y su Programa Educativo

Tesis en desarrollo de Licenciatura	
Fecha	Nombre de la tesis

Tesis concluidas de Licenciatura	
Fecha	Nombre de la tesis

Tesis en desarrollo de Maestría	
Fecha	Nombre de la tesis

Tesis concluidas de Maestría	
Fecha	Nombre de la tesis

Tesis en desarrollo de Doctorado	

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

Educación
Secretaría de Educación Pública
TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Fecha	Nombre de la tesis
26 de noviembre 2024	JUAN ANTONIO MIGUEL RUIZ Modelo de aprendizaje automático para el análisis de la funcionalidad de segmentos génicos V(D)J en diversas especies de vertebrados Doctorado en Ciencias de la Computación
26 de noviembre 2024	MANUEL ERAZO VALADEZ Modelo de aprendizaje automático para el emparejamiento de cadenas pesadas y ligeras de los anticuerpos ANTI-SARS-COV-2 Doctorado en Ciencias de la Computación

Tesis concluidas de Doctorado	
Fecha	Nombre de la tesis

Artículos científicos publicados o enviados en revistas indizadas	
Fecha	Nombre del artículo y acuse de envío

Artículos científicos publicados o enviados enrevistas arbitradas	
Fecha	Nombre del artículo y acuse de envío

Artículos de divulgación enviados	
Fecha	Nombre del artículo y acuse de envío

Memorias en extenso en congresos	
Fecha	Nombre y memoria del Congreso

Capítulo(s) de libros enviados para revisión	
Fecha	Nombre del Libro y del/de los capítulo(s) y acuse de enviado

2025 Año de La Mujer Indígena

Interior Internado Palmira S/N, Col. Palmira,
C. P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 3201,
e-mail: dcc_cenidet@tecnm.mx tecnm.mx | cenidet.tecnm.mx

cenidet

Libro(s) enviados para revisión	
Fecha	Nombre del/de los Libro(s) y acuse de enviado

Libro(s) editado(s) y publicado(s)	
Fecha	Nombre del/de los Libro(s) y Editorial(es)

Registro de Patente (IMPI):	
Fecha	Acuse de solicitud de registro de patente

Registro de Modelo de Utilidad (IMPI)	
Fecha	Acuse de solicitud de registro

Registro de Signos Distintivos (IMPI):	
Fecha	Acuse de solicitud de registro

Registro de Diseño Industrial (IMPI)	
Fecha	Acuse de solicitud de registro

Registro de Esquemas de Trazado de Circuitos Integrados (IMPI):	
Fecha	Acuse de solicitud de registro

Registro de Obra (INDAUTOR):	
Fecha	Acuse de solicitud de registro

Carta de Usuario (Empresa):	
Fecha	Acuse de solicitud de registro

2025 Año de La Mujer Indígena

4. Metodología

1.Comprensión del problema: Realizar reuniones de trabajo con colaboradores del CISEI/INSP para la comprensión del problema desde su punto de vista y a partir de alli establecer pautas para su formulación desde un punto de vista computacional.

2. Revisión de la literatura: Con el apoyo de colaboradores del CISEI/INSP actualizar el marco teórico, integrando nuevos trabajos relacionados con el proceso de anotación genómica en general, y de anotación genómica relacionada con los segmentosV(D)J.

3. Actualización del estado del arte: Con el apoyo de colaboradores del CISEI/INSP actualizar el estado del arte con el que se cuenta actualmente, producto del avance de las tesis doctorales en desarrollo relacionadas con el proyecto.

4. Integración de un conjunto de datos con las posiciones de inicio y fin de los alineamientos de secuencias de los genes que integran el segmento V a partir de las diversas fuentes de datos que se utilizan para realizar la anotación de manera manual .

5. Diseño de un modelo de inteligencia artificial que permita reducir el espacio de búsqueda de los alineamientos de secuencias con el uso del conjunto de datos del objetivo anterior.

5. Difusión

A través del envío para publicación en revista indizada

6. Beneficios y Problemas

En conjunto, estas contribuciones fortalecen la base para un proceso de caracterización de genes más sistemático, eficiente y relevante, con un impacto significativo tanto en la biología molecular como en la bioinformática e informática. Este marco de trabajo sienta las bases para futuros desarrollos y optimizaciones, consolidando un camino hacia una mejor comprensión de los genes V y sus aplicaciones biológicas.

7. Información adicional

Sin información adicional

Subdirector Académico Dr. Carlos Manuel Astorga Zaragoza	Profesor(a) Dr. Javier Ortiz Hernández

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA