



Tecnológico Nacional de México

**Centro Nacional de Investigación
y Desarrollo Tecnológico**

Tesis de Doctorado

**Método para la reducción del espacio de búsqueda de
alineamientos de genes V en vertebrados**

presentada por

MCC. Juan Antonio Miguel Ruiz

como requisito para la obtención del grado de
Doctor en Ciencias de la Computación

Director de tesis

Dr. Javier Ortiz Hernández

Codirector de tesis

Dr. Jesús Martínez Barnetche

Cuernavaca, Morelos, México. Agosto de 2026.

 cenidet [®] Centro Nacional de Investigación y Desarrollo Tecnológico	ACEPTACIÓN DE IMPRESIÓN DEL DOCUMENTO DE TESIS DOCTORAL	Código: CENIDET-AC-006-D20
		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 1

Cuernavaca, Mor., a 03 de agosto de 2026

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
P R E S E N T E

AT'n: DRA. ANDREA MAGADÁN SALAZAR
PRESIDENTA DEL CLAUSTRO DOCTORAL DEL
DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Los abajo firmantes, miembros del Comité Tutorial del estudiante **M.C. JUAN ANTONIO MIGUEL RUIZ** manifiestan que después de haber revisado el documento de tesis titulado **"Método para la reducción del espacio de búsqueda de alineamientos de genes V en vertebrados"**, realizado bajo la dirección del **Dr. Javier Ortiz Hernández** y Codirigida por, el **Dr. Jesús Martínez Barnetche**, el trabajo se **ACEPTA** para proceder a su impresión.

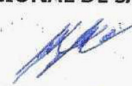
ATENTAMENTE

Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"


DR. JAVIER ORTIZ HERNÁNDEZ
TECNM/CENIDET


DR. JESÚS MARTÍNEZ BARNETCHE
INSTITUTO NACIONAL DE SALUD PÚBLICA


DR. DANTE MÚJICA VARGAS
TECNM/CENIDET


DR. HUGO ESTRADA ESQUIVEL
TECNM/CENIDET


DRA. MARÍA YASMÍN HERNÁNDEZ PÉREZ
TECNM/CENIDET


DR. JUAN MAURICIO TÉLLEZ SOSA
INSTITUTO NACIONAL DE SALUD PÚBLICA

c.c.p: C Verónica Sotelo Boyas/ Jefa del Departamento de Servicios Escolares
c.c.p: C Née Alejandro Castro Sánchez / Jefe del Departamento de Ciencias Computacionales
c.c.p: Expediente

CENIDET-AC-006-D20

Rev. 0



Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 04/agosto/2026

Oficio No. SAC/226/2026

Asunto: Autorización de impresión de tesis

JUAN ANTONIO MIGUEL RUIZ
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Método para la reducción del espacio de búsqueda de alineamientos de genes V en vertebrados"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®

"Conocimiento y Tecnología al servicio de México"

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



Interior Internado Palmira S/N, Col. Palmira,
C.P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4106,
e-mail: acad_cenidet@tecnm.mx | cenidet.tecnm.mx

Dedicatoria

A mis padres y a mis amigos.

Por darme un hogar —y una segunda familia— donde cada sueño encontró espacio para crecer y cada meta se volvió alcanzable. Por enseñarme, con el ejemplo silencioso de cada día, que el esfuerzo constante siempre encuentra su recompensa y que ninguna caída es definitiva mientras exista la voluntad de levantarse.

Gracias por acompañarme en los desvelos, por sostenerme en los momentos de duda y por celebrar conmigo cada pequeño avance como si fuera el más grande de los triunfos. Todo lo que soy y lo que he logrado lleva su huella.

Este trabajo, fruto de años de dedicación, es también suyo.

Agradecimientos

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCYT), hoy Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), por la beca otorgada, que me permitió dedicarme plenamente al desarrollo de este doctorado y cumplir cada uno de los objetivos trazados en el camino.

Al Tecnológico Nacional de México, campus CENIDET, por brindarme sus instalaciones y el entorno académico que hizo posible mi formación profesional, personal y educativa a lo largo de estos años.

A cada uno de los profesores que trazaron el camino para completar este trabajo. Gracias a su enseñanza he podido resolver los problemas que se presentaron a lo largo del desarrollo de la investigación.

A mis queridos amigos, a quienes he tenido el placer de conocer por más de ocho años. Gracias a ustedes logré completar lo que jamás habría alcanzado solo; gracias al gran equipo que formamos, y que seguiremos siendo, pude llegar a esta meta de mi vida.

A la maestra Yanelin Abarca Ponce, por su compañía y apoyo incondicional durante esta etapa, por su paciencia en los momentos de mayor exigencia y por recordarme, cuando más lo necesitaba, la razón para seguir adelante.

Al doctor Javier Ortiz Hernández, por cada consejo, por cada enseñanza y por su dirección a lo largo de esta tesis. Ha sido un profesor brillante y un amigo incondicional, de quien no he dejado de aprender y de quien seguiré aprendiendo. Gracias por creer en este trabajo desde el primer día, por su paciencia en los momentos difíciles y por impulsarme siempre a seguir preparándome y a no conformarme.

Al doctor Jesús Martínez Barnette, por su codirección y por su acompañamiento constante durante toda esta investigación. Gracias a usted tuve la oportunidad de unir dos ciencias tan distintas que jamás imaginé que llegaría a combinar. Le agradezco de manera infinita su paciencia, su generosidad al compartir su conocimiento y la dedicación con la que guio este trabajo. Sin su visión, este puente entre la biología y el cómputo no habría sido posible.

Resumen

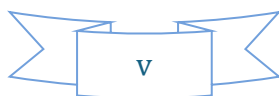
Esta investigación aborda un problema central en la caracterización del genoma: la identificación de genes V de inmunoglobulinas en genomas de vertebrados, una tarea obliga a los expertos a revisar manualmente miles de regiones candidatas producidas por diversas herramientas de alineamiento de secuencias. Este proceso, realizado en el Centro de Investigación sobre Enfermedades Infecciosas (CISEI) del Instituto Nacional de Salud Pública, constituye un cuello de botella que escala con el tamaño del genoma y limita la capacidad de caracterización en organismos poco estudiados como los quirópteros.

Como solución se propone el método *Reduction for the Identification of V-genes by sequence Alignment and Machine learning* (RIVAM), orientado a la reducción del espacio de búsqueda para la identificación de genes V. El método integra principios de diseño derivados del dominio biológico con técnicas de aprendizaje automático no supervisado, siguiendo la metodología de Ciencia del Diseño (DSR) combinada con una metodología de flujo de conocimiento del dominio. RIVAM consolida los resultados de cinco sistemas de alineamiento (*HMMscan*, *Exonerate:est*, *Exonerate:prot*, *Tblastn* y *Tblastx*) en una representación vectorial unificada basada en matrices de adyacencia, sobre la cual aplica reducción de dimensionalidad y agrupamiento para aislar el subconjunto de regiones con mayor densidad de genes V reales.

Al evaluar el método se utilizaron doce genomas de quirópteros anotados por expertos. Se probaron tres configuraciones distintas y sus resultados se compararon con las pruebas de Friedman y Nemenyi. Los resultados indicaron que UMAP+GMM fue la mejor configuración, con diferencias estadísticamente significativas y mejoras de hasta 37 puntos porcentuales en algunos casos.

La validación externa confirmó la capacidad de generalización del método sobre tres genomas con características deliberadamente distintas: *Plecotus auritus* (91.06% de cobertura sobre 850 genes de referencia), *Ambystoma mexicanum* (55.17% en un anfibio filogenéticamente distante) y *Molossus nigricans* (un quiróptero sin anotación experta). La comparación con *IgDetective*, método de referencia del estado del arte, mostró que RIVAM lo supera en tres de cinco genomas evaluados y, de forma destacada, identifica sistemáticamente el primer exón (SP_exon), ausente en todas las predicciones de *IgDetective*.

RIVAM constituye un aporte metodológico al problema de reducción del espacio de búsqueda, demostrando que el aprendizaje no supervisado permite acotar grandes volúmenes de soluciones sin sustituir las herramientas disponibles ni sacrificar la cobertura de los genes V en el proceso de anotación y caracterización del genoma.



Abstract

This research addresses a central problem in genome characterization: the identification of immunoglobulin V-genes in vertebrate genomes, a task that requires experts to manually review thousands of candidate regions produced by various sequence alignment tools. This process, carried out at the Centro de Investigación sobre Enfermedades Infecciosas (CISEI) of the Instituto Nacional de Salud Pública, represents a bottleneck that scales with genome size and limits the characterization capacity for understudied organisms such as bats.

As a solution, the Reduction for the Identification of V-genes by sequence Alignment and Machine Learning (RIVAM) method is proposed, aimed at reducing the search space for V-gene identification. The method integrates design principles derived from the biological domain with unsupervised machine learning techniques, following the Design Science Research (DSR) methodology combined with domain knowledge flow methodology. RIVAM consolidates the results of five alignment systems (HMMscan, Exonerate:est, Exonerate:prot, Tblastn, and Tblastx) into a unified vector representation based on adjacency matrices, upon which dimensionality reduction and clustering are applied to isolate the subset of regions with the highest density of true V-genes.

Method evaluation was conducted using twelve expert-annotated bat genomes. Three distinct configurations were tested, and their results were compared using the Friedman and Nemenyi statistical tests. Results indicated that UMAP+GMM was the best-performing configuration, with statistically significant differences and improvements of up to 37 percentage points in specific cases.

External validation confirmed the generalization capacity of the method across three genomes with deliberately distinct characteristics: *Plecotus auritus* (91.06% coverage over 850 reference genes), *Ambystoma mexicanum* (55.17% in a phylogenetically distant amphibian), and *Molossus nigricans* (a bat species without expert annotation). Comparison with IgDetective, the state-of-the-art reference method, showed that RIVAM outperforms it in three of five evaluated genomes and, notably, systematically identifies the signal peptide exon (SP_exon), absent in all IgDetective predictions.

RIVAM represents a methodological contribution to the search space reduction problem, demonstrating that unsupervised machine learning can effectively narrow down large volumes of candidate solutions without replacing available tools or compromising V-gene coverage in the genome annotation and characterization process.

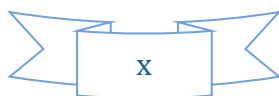
Tabla de contenido

Dedicatoria	iii
Agradecimientos	iv
Resumen	v
Abstract.....	vi
Capítulo 1. Introducción	1
1.1 Antecedentes	2
1.2 Contexto del proceso de caracterización del genoma.....	4
1.3 Planteamiento del problema	6
1.4 Justificación.....	7
1.5 Objetivo de la tesis.....	8
1.6 Objetivos específicos	8
1.7 Alcances y limitaciones.....	8
1.8 Metodología de la investigación.....	10
1.8.1 Metodología para la Ciencia del diseño	11
1.8.2 Metodología de solución propuesta.....	12
1.8.3 Mapeo de capítulos a ciclos	14
1.9 Estructura de la tesis	14
Capítulo 2. Marco conceptual	16
2.1 Conceptos biológicos fundamentales.....	17
2.1.1 Del genoma a las proteínas.....	17
2.1.2 Anticuerpos b.....	17
2.1.3 Recombinación V(D)J.....	18
2.1.4 Estructura del gen del segmento v	19
2.1.5 Genes funcionales y no funcionales	21
2.2 Datos, formatos y sistemas para el análisis del genoma.....	21
2.2.1 Archivos fasta.....	22
2.2.2 Alineamiento de secuencias	22
2.2.3 Archivos GFF	25
2.2.4 Integrative genomics viewer.....	26

2.3	Conceptos computacionales clave	28
2.3.1	Reducción de un espacio de búsqueda	28
2.3.2	Algoritmo mezclas gaussianas	30
2.3.3	Análisis de componentes principales	31
2.3.4	Uniform manifold approximation and projection	32
2.3.5	Particle swarm optimization.....	33
Capítulo 3.	Análisis del proceso de identificación y anotación de genes v	34
3.1	Descripción del proceso de anotación en el cisei	35
3.1.1	Etapas y reglas de identificación	37
3.2	Limitaciones del proceso actual.....	39
3.2.1	Volumen excesivo del espacio de búsqueda.....	39
3.2.2	Costo temporal de la revisión manual.....	40
3.2.3	Susceptibilidad al error en la revisión manual	41
3.2.4	Ausencia de un flujo de trabajo sistematizado	41
3.3	Limitaciones de las herramientas disponibles.....	42
3.3.1	IgDetective en el contexto de quirópteros	42
3.4	Oportunidades de intervención metodológica	43
3.4.1	Punto de intervención en el proceso	44
3.4.2	Fundamentación del método propuesto	44
Capítulo 4.	Revisión sistemática.....	46
4.1	Antecedentes dentro de cenidet	47
4.2	Aplicación de la revisión sistemática	48
4.2.1	Trabajos sobre el alineamiento de secuencias.....	51
4.2.2	Trabajos sobre la identificación de genes.....	52
4.3	Brechas identificadas en la literatura	55
4.3.1	Ausencia de métodos específicos para la reducción del espacio de búsqueda en genes V.....	55
4.3.2	Falta de métodos adaptables a especies poco estudiadas	56
4.3.3	Ausencia de sistematización del conocimiento experto.....	57
Capítulo 5.	Fundamentos conceptuales del método RIVAM	59
5.1	Principios de diseño.....	60

5.1.1	Base teórica de otros métodos de ciencia de datos.....	60
5.1.2	Reducción de redundancia	62
5.1.3	Preservación de coherencia biológica	64
5.1.4	Reusabilidad.....	65
5.1.5	Interoperabilidad	66
5.2	Justificación de las decisiones de diseño	67
5.2.1	¿Por qué reducción de dimensionalidad?	67
5.2.2	¿Por qué agrupamiento no supervisado?	69
5.2.3	¿Por qué explorar múltiples variantes?	71
5.2.4	¿Por qué PSO para optimización?	72
5.3	Componentes metodológicos	73
5.3.1	Comprensión y representación de datos.....	74
5.3.2	Integración de fuentes de información	76
5.3.3	Estandarización de representaciones	77
5.3.4	Reducción de dimensionalidad	78
5.3.5	Modelado y selección.....	79
5.3.6	Representación final e interoperabilidad	80
5.4	Iteración y refinamiento.....	81
5.4.1	Decisiones iniciales de diseño	81
5.4.2	Refinamiento a partir del conocimiento del dominio	82
5.4.3	Diseño conceptual consolidado	83
Capítulo 6.	Desarrollo iterativo del método RIVAM	85
6.1	Desarrollo iterativo del método RIVAM	86
6.1.1	Primera iteración: unificación de datos y primera propuesta de agrupamiento.....	86
6.1.2	Segunda iteración: incorporación de métricas y comparación de algoritmos	105
6.1.3	Tercera iteración: normalizaciones y unificación por grupos.....	117
6.1.4	Cuarta iteración: incorporación de reducción de dimensionalidad	124
6.1.5	Consolidación de las tres variantes finales.....	135
6.2	Especificaciones técnicas.....	137

6.2.1	Entornos de ejecución	137
6.2.2	Tecnologías y bibliotecas.....	138
6.2.3	Tiempos de ejecución	139
Capítulo 7.	Diseño del experimento.....	140
7.1	Datos experimentales.....	141
7.1.1	Genomas de evaluación.....	141
7.1.2	Genomas de validación externa	142
7.1.3	Criterios de selección.....	144
7.2	Proceso de evaluación.....	145
7.2.1	Protocolo experimental	145
7.2.2	Configuración del experimento.....	147
7.3	Análisis estadístico.....	148
7.3.1	Estadística descriptiva.....	148
7.3.2	Evaluación de normalidad	149
7.3.3	Prueba de friedman	149
7.3.4	Análisis post hoc de Nemenyi.....	151
7.4	Comparación con método de referencia	151
7.4.1	Descripción de IgDetective	152
7.4.2	Criterios de comparación.....	153
7.4.3	Genomas seleccionados para la comparación.....	154
Capítulo 8.	Resultados y discusión	156
8.1	Resultados de la evaluación comparativa.....	157
8.1.1	Estadística descriptiva y análisis visual.....	157
8.1.2	Evaluación de normalidad	161
8.1.3	Prueba de friedman	163
8.1.4	Análisis post hoc de Nemenyi.....	164
8.1.5	Selección de la variante óptima	166
8.2	Resultados de la validación externa	168
8.2.1	Validación con <i>Plecotus Auritus</i>	168
8.2.2	<i>Ambystoma Mexicanum</i>	170



8.2.3	Molossus Nigricans.....	171
8.3	Comparación con IgDetective.....	173
8.3.1	Resultados comparativos por genoma.....	173
8.3.2	Análisis de precisión estructural en IGV	175
8.3.3	Discusión de diferencias.....	180
8.4	Flujo de procesamiento del método propuesto	182
8.5	Discusión.....	184
8.5.1	Interpretación de los resultados en el contexto del problema	184
8.5.2	Limitaciones identificadas.....	185
8.5.3	Verificación de los principios de diseño del método	187
Capítulo 9.	Conclusiones	189
9.1	Trabajo futuro	191
Referencias		193
Anexo A	resúmenes del estado del arte.....	198
	Resúmenes del estado del arte para el alineamiento de secuencias	198
	Resúmenes del estado del arte del proceso de anotación del genoma e identificación de genes.....	200
Anexo B	resultado de comparativa de algoritmos.....	212
Anexo C	producción científica asociada a la investigación	216

Figura 1.1	Etapas de la caracterización del genoma adaptado de la tesis (Chavez Serret, 2025)	5
Figura 1.2	Metodología de la investigación adaptada de (Hevner et al., 2010; Peffers et al., 2020).....	10
Figura 1.3	Metodología de la investigación.....	12
Figura 2.1	Representación de receptor de células B (BCR) obtenida de (Elizabeth Ernestina Godoy Lozano, 2015)	18
Figura 2.2	Proceso de generación de linfocito B inmaduro (IgM) obtenida de (Godoy Lozano, 2015).....	19
Figura 2.3	Estructura del gen V adaptada de (Lefranc & Lefranc, 2001)	20
Figura 2.4	Comportamiento del alineamiento local vs global adaptada de (Ganesh et al., 2018)	23
Figura 2.5	Vista general del espacio de búsqueda de genes V las herramientas de alineamiento en IGV	27

Figura 2.6 Ejemplo de alineamientos redundantes y superpuestos sobre una región candidata de genes V	27
Figura 2.7 Comparación de alineamientos parcialmente coincidentes sobre una región candidata a gen V	28
Figura 3.1 Modelo BPMN del proceso de identificación y anotación de genes V, obtenida de (Chavez Serret, 2025).....	36
Figura 3.2 Subproceso de identificación de las partes del gen V, obtenida de (Chavez Serret, 2025).....	38
Figura 4.1 Diagrama PRISMA del proceso de búsqueda y selección de literatura	50
Figura 5.1 Diagrama del proceso CRISP-DM, adaptado de (Chapman et al., 2000)	61
Figura 5.2 Diagrama de la Metodología Fundamental para la Ciencia de Datos, obtenida de (Rollins, 2015).....	62
Figura 5.3 Representación del gen V y distancias validadas por RIVAM.....	65
Figura 5.4 Diagrama conceptual de los componentes metodológicos del método RIVAM	74
Figura 6.1 Etapas desarrolladas en el refinamiento del método	86
Figura 6.2 Gráfico QQnormal y de densidad de las posiciones de inicio	88
Figura 6.3 Proceso de unificación de datos y sus cinco etapas	89
Figura 6.4 Subproceso de importación de datos para la unificación de datos.....	91
Figura 6.5 Identificación del archivo con menor redundancia.....	92
Figura 6.6 Extracción y preparación de datos	92
Figura 6.7 Construcción y validación de registros de genes V	94
Figura 6.8 Construcción y validación de registros de genes V	95
Figura 6.9 Grafo del gen V con sus posiciones como nodos y distancias como aristas.	97
Figura 6.10 Gráficas de densidad del conjunto de datos: (A) escala original, (B) datos escalados, (C) datos escalados y normalizados.....	100
Figura 6.11 Gráficos de dispersión: (A) datos sin normalizar, (B) datos normalizados	100
Figura 6.12 Gráfico de dispersión del primer agrupamiento con K-Means (20 grupos)	102
Figura 6.13 Resultados en IGV: vista general de genes identificados	103
Figura 6.14 Resultados en IGV: comparación con referencia (gen 1)	104
Figura 6.15 Resultados en IGV: comparación con referencia (gen 2)	104
Figura 6.16 Métodos para definir el número de grupos	112
Figura 6.17 Gráficos de dispersión de Spectral clustering y K-Means	113
Figura 6.18 Gráficos de dispersión de Fuzzy C-Means y Gaussian Mixtures	114
Figura 6.19 Gráficos de dispersión del conjunto EPNI con MaxAbsScaler y segundo cuartil.....	123
Figura 6.20 Cuarta iteración del método RIVAM	125

Figura 6.21 Gráficos de dispersión de la prueba de optimización del conjunto de datos EPNI.....	130
Figura 6.22 Gráficos de dispersión del Modelo PCA-GMM sobre EPNI.....	132
Figura 6.23 Gráficos de dispersión del Modelo UMAP-GMM sobre EPNI.....	134
Figura 6.24 Diagrama consolidado de las tres variantes finales del método RIVAM...	135
Figura 7.1 Diagrama del protocolo experimental del método RIVAM	146
Figura 8.1 Boxplot del porcentaje de identificación por variante.....	159
Figura 8.2 Boxplot de RMSE por variante.....	160
Figura 8.3 Boxplot del número de genes identificados por variante	161
Figura 8.4 QQ-plots de RMSE por variante	162
Figura 8.5 QQ-plots de porcentaje de identificación por variante.....	162
Figura 8.6 Diagrama de diferencia crítica para RMSE	165
Figura 8.7 Diagrama de diferencia crítica para porcentaje de identificación.....	166
Figura 8.8 Proyección UMAP con grupos GMM sobre el genoma de <i>Plecotus auritus</i> .	169
Figura 8.9 Proyección UMAP con grupos GMM sobre el genoma de <i>Molossus nigricans</i>	172
Figura 8.10 Vista general del cromosoma CM050515.1 de <i>Antrozous pallidus</i>	176
Figura 8.11 Gen 73 de <i>Antrozous pallidus</i> en orientación inversa.....	176
Figura 8.12 Gen 72 de <i>Antrozous pallidus</i> en orientación forward	176
Figura 8.13 Vista general del cromosoma CM063093.1 de <i>Hipposideros larvatus</i>	177
Figura 8.14 Gen 004 de <i>Hipposideros larvatus</i> en orientación forward	177
Figura 8.15 Gen 104 de <i>Hipposideros larvatus</i> en orientación forward	177
Figura 8.16 Vista expandida del cromosoma OY734022.1 de <i>Plecotus auritus</i>	178
Figura 8.17 Gen 005 funcional de <i>Plecotus auritus</i> en orientación inversa.....	178
Figura 8.18 Gen 021 no funcional de <i>Plecotus auritus</i> en orientación inversa	178
Figura 8.19 Vista general del cromosoma CM078066.1 de <i>Molossus nigricans</i>	179
Figura 8.20 Gen 1430 de <i>Molossus nigricans</i> , identificado únicamente por el método propuesto.....	179
Figura 8.21 Gen 1429 de <i>Molossus nigricans</i> en orientación forward.....	179
Figura 8.22 Gen 0067 de <i>Molossus nigricans</i> en orientación forward.....	180
Figura 8.23 Flujo de procesamiento del método propuesto	183
Tabla 4.1 comparativa de trabajos sobre alineamiento de secuencias	51
Tabla 4.2 comparativa de trabajos sobre anotación del genoma e identificación de genes	53
Tabla 5.1 Ejemplo de registros redundantes de dos genes V.....	63
Tabla 5.2 Comparativa teórica de algoritmos de agrupamiento.....	70
Tabla 6.1 Fragmento de datos de un archivo GFF.....	87
Tabla 6.2 Estadísticas básicas del conjunto de exonerate:est2genome.....	89
Tabla 6.3 Cinco etapas del proceso de unificación de datos.....	90

Tabla 6.4 Reglas de validación biológica para la construcción de registros de genes V93	
Tabla 6.5 Ejemplo de registros para un gen V con todas sus secciones.....	95
Tabla 6.6 Ejemplos de genes V organizados en sus seis secciones.....	96
Tabla 6.7 Ejemplo de matriz de adyacencia.....	97
Tabla 6.8 Atributos resultantes de la unificación de datos.....	98
Tabla 6.9 Estadísticas básicas del conjunto de datos unificado.....	99
Tabla 6.10 Estadísticas básicas del conjunto de datos unificado.....	99
Tabla 6.11 Grupos de interés identificados en el primer agrupamiento con K-Means	102
Tabla 6.12 Ejemplo de registros redundantes de un gen dentro de un grupo.....	107
Tabla 6.13 Resultado de la unificación de datos por grupos.....	108
Tabla 6.14 Ejemplo de comparación entre posiciones identificadas y referencia.....	110
Tabla 6.15 Mejores resultados de los algoritmos para el conjunto de datos con todas sus dimensiones.....	114
Tabla 6.16 Comparativa de técnicas de normalización evaluadas en el método RIVAM	119
Tabla 6.17 Grupos del algoritmo GMM con los datos de EPNI con el segundo cuartil y MaxAbsScaler.....	122
Tabla 6.18 Configuración del algoritmo PSO.....	128
Tabla 6.19 Hiperparámetros óptimos encontrados por PSO para el Modelo GMM.....	129
Tabla 6.20 Resultados comparativos de EPNI antes y después de la optimización con PSO.....	130
Tabla 6.21 Hiperparámetros óptimos encontrados por PSO para el Modelo GMM.....	131
Tabla 6.22 Hiperparámetros óptimos encontrados por PSO para el Modelo PCA.....	131
Tabla 6.23 Resultados del Modelo PCA-GMM sobre EPNI.....	131
Tabla 6.24 Hiperparámetros óptimos de GMM para el Modelo UMAP-GMM.....	133
Tabla 6.25 Hiperparámetros óptimos de UMAP para el Modelo UMAP-GMM.....	133
Tabla 6.26 Resultados del Modelo UMAP-GMM sobre EPNI.....	133
Tabla 6.27 Comparativa de las tres variantes finales del método RIVAM sobre EPNI	136
Tabla 6.28 Especificaciones del entorno Google Colab.....	137
Tabla 6.29 Especificaciones del equipo personal ASUS TUF Gaming F15.....	137
Tabla 6.30 Bibliotecas y versiones utilizadas en el desarrollo del método RIVAM.....	138
Tabla 6.31 Tiempos de ejecución por variante del método RIVAM.....	139
Tabla 7.1 Genomas de evaluación utilizados en el estudio.....	141
Tabla 7.2 Genomas de validación externa.....	143
Tabla 7.3 Genomas seleccionados para la comparación con IgDetective.....	155
Tabla 8.1 Estadísticas descriptivas de las métricas de evaluación por variante.....	157
Tabla 8.2 Resultados detallados por genoma para las tres variantes del método RIVAM	158
Tabla 8.3 Prueba de Shapiro-Wilk para el porcentaje de identificación.....	163

Tabla 8.4 Prueba de Shapiro-Wilk para el RMSE	163
Tabla 8.5 Prueba de Shapiro-Wilk para el número de genes identificados	163
Tabla 8.6 Resultados de la prueba de <i>Friedman</i> por métrica.....	164
Tabla 8.7 Análisis post hoc de <i>Nemenyi</i> para RMSE.....	165
Tabla 8.8 Análisis post hoc de <i>Nemenyi</i> para porcentaje de identificación	165
Tabla 8.9 Resultados de la validación sobre el genoma de <i>Plecotus auritus</i>	169
Tabla 8.10 Resultados por cluster sobre el genoma de <i>Molossus nigricans</i>	173
Tabla 8.11 Comparación del desempeño de RIVAM e IgDetective en la identificación de genes V funcionales.....	174
Tabla de anexos 1 Resultados de la comparación de algoritmos para 3 grupos	212
Tabla de anexos 2 Resultados de la comparación de algoritmos para 4 grupos	213
Tabla de anexos 3 Resultados de la comparación de algoritmos para 5 grupos	213
Tabla de anexos 4 Resultados de la comparación de algoritmos para 6 grupos	214
Tabla de anexos 5 Resultados de la comparación de algoritmos para 7 grupos	215

Capítulo 1. Introducción

Este capítulo establece el nexo entre el entorno del problema real y la investigación propuesta. Se presentan antecedentes relacionados con la reducción del espacio de búsqueda en múltiples dominios, el contexto específico del proceso de caracterización del genoma en el Centro de Investigación Sobre Enfermedades Infecciosas (CISEI), el planteamiento formal del problema, su justificación científica y práctica, los objetivos del proyecto, y la metodología de Ciencia del Diseño que fundamenta esta investigación. La estructura del capítulo refleja esta progresión desde lo general hacia lo específico, sentando las bases para los capítulos subsecuentes.

1.1 Antecedentes

En numerosos campos de la ciencia y la ingeniería, la resolución de problemas complejos requiere explorar espacios de búsqueda extensos, compuestos por múltiples combinaciones de posibles soluciones. Estos espacios pueden estar definidos por parámetros, estructuras, ubicaciones o relaciones que deben evaluarse para determinar cuáles satisfacen un conjunto de condiciones u optimizan una función objetivo. Sin embargo, en contextos reales, este espacio suele ser altamente dimensional, ruidoso, redundante y poco informativo, lo cual representa una barrera significativa para el análisis eficiente de datos y el diseño de soluciones efectivas (Bayas-Jiménez et al., 2021; Bode et al., 2019).

Este problema, ampliamente documentado en áreas como la optimización multiobjetivo, el diseño de redes, la inteligencia artificial o la bioinformática, incrementa el costo computacional y el esfuerzo humano, dificulta la interpretación de resultados y limita la escalabilidad de los métodos tradicionales (Bayas-Jiménez et al., 2021; Bode et al., 2019). Si bien se han desarrollado soluciones robustas como los algoritmos evolutivos, bioinspirados, de agrupamiento o búsqueda heurística para explorar estos espacios de soluciones, la literatura muestra que aún persiste la necesidad de consolidar métodos sistemáticos que sean reutilizables y adaptables. Esta carencia es crítica en dominios como la bioinformática, donde los datos provienen de diversas fuentes heterogéneas que requieren integración y consolidación antes de que cualquier técnica de reducción pueda aplicarse de manera efectiva.

Adicionalmente, muchos de los enfoques existentes se centran en la implementación de algoritmos específicos o métricas puntuales, pero omiten aspectos clave como: a) la unificación de datos provenientes de múltiples fuentes, b) la evaluación integrada del rendimiento y c) la construcción de marcos metodológicos completos y replicables (Bode et al., 2019). La ausencia de un marco integral limita la transferencia de soluciones o generalización de soluciones entre dominios y restringe el impacto de las investigaciones. En este sentido, la necesidad de estructurar un proceso metodológico que combine filtrado, integración, agrupamiento y validación se vuelve fundamental para enfrentar problemas de reducción de espacio de soluciones (Ania & Kuri, 2007).

Este trabajo aborda dicha problemática mediante el diseño de un método integral para la reducción del espacio de búsqueda, cuyo objetivo es transformar grandes volúmenes de soluciones parciales o repetitivas en un conjunto compacto, representativo y útil para la toma de decisiones.

El método propuesto combina técnicas de unificación de datos, agrupamiento no supervisado, reducción de dimensionalidad y evaluación estadística, en un marco estructurado que guía el proceso desde la recopilación hasta la selección óptima de soluciones. Esta aproximación no solo busca resolver un problema técnico, sino

también establecer un modelo conceptual que pueda adaptarse a otros contextos donde la redundancia y el ruido dificultan la identificación de patrones significativos.

Existen diferentes instancias del problema de reducción del espacio de búsqueda. En el caso de esta investigación, se aborda un caso de estudio del Centro de Investigación sobre Enfermedades Infecciosas (CISEI) en el Instituto Nacional de Salud Pública (INSP), ubicado en Cuernavaca, México, donde se realiza la caracterización manual de genes Variables, de Diversidad y de Unión conocidos como genes V, D y J. Estos genes desempeñan un papel crucial en la formación de anticuerpos B que detectan y neutralizan agentes externos como virus y bacterias (Kindt et al., 2007; Pieper et al., 2013).

De los tres tipos de genes que conforman los receptores de inmunoglobulina, los genes V producen la mayor cantidad de regiones candidatas durante el proceso de alineamiento. Esto se debe a su alta variabilidad entre especies, la presencia de diversas copias en el genoma y su mayor longitud en comparación con los genes D y J. Como resultado, herramientas como *HMMER*, *Exonerate* y *BLAST* pueden producir miles de candidatos para un solo genoma, los cuales deben ser revisados manualmente por expertos. Esta situación constituye el principal cuello de botella en el proceso de caracterización del genoma realizado en el CISEI. Por ello, el método propuesto se centra en la reducción del espacio de búsqueda de genes V en genomas de vertebrados.

El método no solo mejora la eficiencia en la identificación de genes V, sino que también abre nuevas oportunidades para sistematizar tareas complejas de caracterización del genoma (Gouret et al., 2005). Al reducir el volumen de candidatos a revisar sin comprometer la cobertura, el método contribuye a establecer un puente entre el conocimiento experto y las capacidades computacionales para abordar problemas prácticos en bioinformática y áreas afines.

En este contexto, el presente trabajo propone el método *Reduction for the Identification of V-genes by sequence Alignment and Machine learning* (RIVAM), un método replicable y adaptable a otros dominios que aborda desafíos como: grandes cantidades de soluciones, fuentes de datos heterogéneas y la necesidad de integrar, filtrar y priorizar información relevante de manera eficiente y confiable. Este método contribuye no solo a mejorar procesos existentes, sino también a facilitar el desarrollo de nuevas herramientas en bioinformática y áreas afines. A futuro, este tipo de métodos puede consolidar una base común para enfrentar la creciente complejidad de los datos científicos, promoviendo la creación de soluciones escalables, sostenibles y orientadas a la práctica interdisciplinaria.

1.2 Contexto del proceso de caracterización del genoma

En el Centro de Investigación sobre Enfermedades Infecciosas (CISEI) del Instituto Nacional de Salud Pública (INSP), se desarrolla el proceso de caracterización del genoma, el cual constituye una actividad esencial para investigaciones relacionadas con evolución molecular, enfermedades genéticas, inmunología y desarrollo de nuevas estrategias terapéuticas (Simarro, 2022). La caracterización del genoma permite identificar, localizar y describir regiones funcionales dentro del ADN, facilitando el estudio de genes específicos y su relación con distintos procesos biológicos.

Este proceso está conformado por una serie de etapas interdependientes que transforman datos biológicos crudos en información estructurada y analizable. Cada etapa recibe insumos que son utilizados por la siguiente, formando un flujo de trabajo continuo orientado a la identificación y anotación de genes dentro del genoma.

1. **Secuenciación del genoma (1.1):** consiste en la obtención del ADN de un organismo, generalmente a partir de tejido biológico, que es procesado mediante tecnologías de secuenciación del genoma. El resultado es una secuencia digital de caracteres (nucleótidos) que representa la totalidad del genoma del organismo (García Simón et al., 2018; Mount, 2004).
2. **Alineamiento de secuencias (1.2):** la secuencia de la etapa anterior se compara con secuencias de referencia previamente caracterizadas, utilizando herramientas como BLAST, Exonerate o HMMER. El alineamiento puede ser *global* (comparación de extremo a extremo) o *local* (identificación de regiones específicas de similitud). Los resultados son aproximaciones de la ubicación de genes en el genoma (García Simón et al., 2018; Universidad de Gent, 2015).
3. **Transformación de alineamientos a formato de posiciones (1.3):** los alineamientos se convierten a un formato estandarizado, comúnmente el General Feature Format (GFF), que organiza la información en columnas como ubicación de la secuencia, herramienta utilizada, tipo de secuencia, posiciones de inicio y fin, dirección y otros metadatos (EMBL-EBI, 2022).
4. **Identificación de genes y regiones asociadas (1.4):** los archivos GFF, junto con el genoma, se visualizan en herramientas especializadas como Integrative Genomics Viewer (IGV). Mediante inspección manual, se determina si las secuencias corresponden a genes funcionales y se valida la precisión de sus posiciones para determinar la correcta ubicación del gen en el genoma.
5. **Anotación de secuencias (1.5):** las regiones identificadas y validadas se enriquecen con información biológica. Este paso incluye la asignación de etiquetas, la documentación de funciones y la integración de los genes identificados en el genoma anotado, lo cual es esencial para la construcción de bases de datos y la identificación de nuevos genes.

Con el propósito de comprender de manera más clara el funcionamiento del proceso de caracterización del genoma e identificar sus principales limitaciones, una de las primeras actividades de esta investigación consistió en modelar el proceso con la

técnica SADT (Structured Analysis and Design Technique) (Marca & McGowan, 1987). Este trabajo fue desarrollado en la tesis titulada *“Modelo conceptual del proceso de anotación del gen V de anticuerpos basado en la Ciencia del Diseño”*.

El modelo desarrollado permitió representar la secuencia lógica de actividades, así como las interacciones existentes entre las diferentes etapas del flujo de trabajo. A través de esta representación fue posible identificar los puntos críticos del proceso, particularmente aquellos donde se concentran actividades manuales, grandes volúmenes de información o tareas repetitivas que afectan la eficiencia general del análisis del proceso de caracterización.

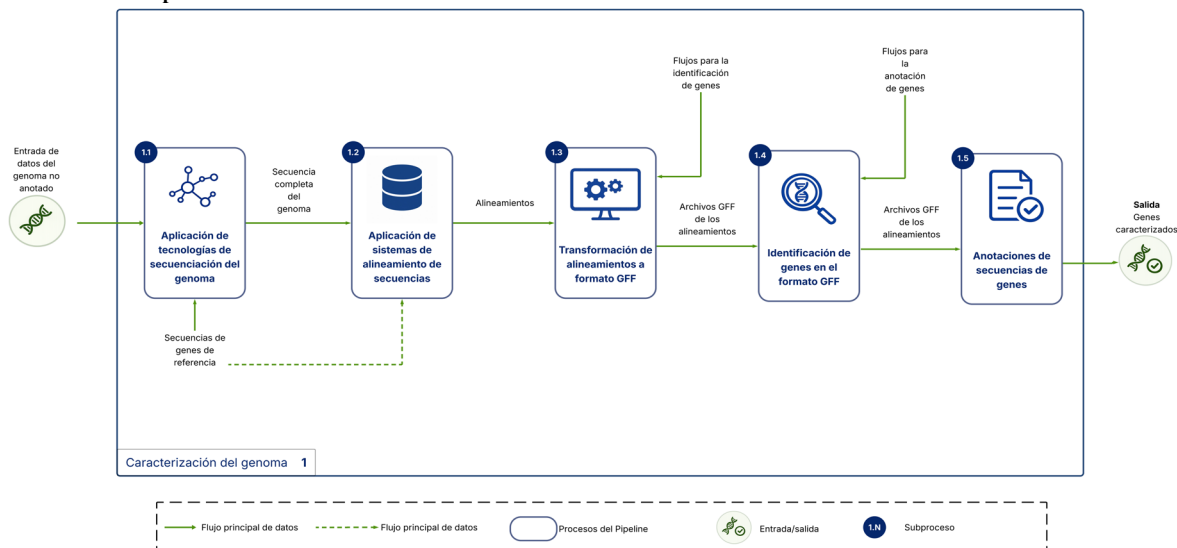


Figura 1.1 Etapas de la caracterización del genoma adaptado de la tesis (Chavez Serret, 2025)

El análisis del diagrama permitió comprender con mayor detalle las actividades realizadas por los expertos durante la caracterización del genoma. Como se observa en el modelo, la actividad 1.4 correspondiente a la identificación de genes en IGV fue identificada como el principal punto de interés para esta investigación.

La razón de enfocar los esfuerzos en esta actividad se debe a que es en esta etapa es donde se presenta el mayor espacio de búsqueda de soluciones, particularmente en el análisis de genes V. La gran cantidad de alineamientos obtenidos dificulta la identificación precisa de regiones de interés de genes V y aumenta considerablemente el tiempo requerido para la validación manual.

En este contexto, la investigación tiene como objetivo reducir la cantidad de alineamientos presentes en los archivos GFF y obtener una salida precisa y consolidada para la actividad 1.4. Con ello se busca proporcionar a los expertos posiciones de genes V más cercanas a la ubicación real dentro del genoma, disminuyendo la carga de análisis manual. Asimismo, se pretende facilitar la etapa 1.5 de anotación de secuencias, permitiendo que los especialistas dediquen más tiempo a la anotación biológica del genoma y menos tiempo a la determinación manual de la ubicación de genes V.

1.3 Planteamiento del problema

En múltiples disciplinas científicas y tecnológicas, resolver problemas complejos requiere explorar espacios de búsqueda altamente dimensionales, compuestos por miles o millones de posibles configuraciones. Estos espacios suelen presentar información redundante, ruido o patrones poco informativos, lo cual incrementa los costos computacionales, dificulta la interpretación de resultados y limita la escalabilidad de los métodos convencionales.

Campos como la inteligencia artificial, la optimización multiobjetivo y la minería de datos han desarrollado algoritmos potentes como lo son: evolutivos, bioinspirados y de agrupamiento, para explorar espacios de búsqueda complejos. Sin embargo, estos métodos asumen un espacio de soluciones abstracto o simbólico, no dominios donde la información proviene de fuentes heterogéneas que requieren integración previa.

En el ámbito biológico no existen métodos sistemáticos y reutilizables para reducir ese espacio sin sacrificar la calidad de los resultados. Los enfoques actuales se centran en algoritmos individuales y no integran la unificación de datos, el filtrado por conocimiento del dominio, la selección de representaciones ni la comparación de configuraciones. Como consecuencia, gran parte del proceso sigue siendo manual, lento y propenso a errores.

Este problema adquiere especial relevancia en el contexto de la identificación de genes V en genomas de vertebrados, donde diversas herramientas de alineamiento producen de forma independiente miles de regiones candidatas que cubren distintas secciones del gen. La ausencia de un marco metodológico integral que unifique, filtre y priorice estos resultados impone una carga de revisión manual que escala directamente con el tamaño del genoma, limitando la capacidad de caracterización en organismos poco estudiados como el caso de los quirópteros.

Surge así la necesidad de un método que aborde la reducción del espacio de búsqueda como un proceso de varias etapas: integración de salidas de diversas fuentes, agrupamiento de datos y optimización de configuraciones. Aunque aquí se aplica a la caracterización del genoma, el problema es más amplio dentro de la ciencia de datos, donde explorar grandes espacios con eficiencia y precisión sigue siendo un reto abierto. Los principales retos presentados en esta investigación abordan:

- **Reducción sin pérdida de valor:** desarrollar estrategias que simplifiquen el espacio de búsqueda sin descartar patrones relevantes o configuraciones óptimas.
- **Escalabilidad:** asegurar que el método funcione de manera eficiente al aplicarse a volúmenes crecientes de datos.

- **Estandarización del flujo de trabajo:** estructurar procesos que puedan ser fácilmente reutilizados o adaptados en distintos entornos de análisis.
- **Automatización de decisiones:** reducir la necesidad de intervención humana mediante el uso de criterios de selección automáticos e interpretables.
- **Transferencia entre dominios:** garantizar que los principios del modelo puedan aplicarse en otros contextos científicos o industriales con desafíos similares.

1.4 Justificación

La reducción de los costos de las tecnologías de secuenciación y el crecimiento exponencial de los datos en genómica han convertido la anotación del genoma en un reto central para la biología computacional. En particular, los genes del segmento V(D)J requieren una identificación precisa y eficiente que aún depende, en gran medida, de procesos manuales y fragmentados. Esta situación ha puesto en evidencia la necesidad de desarrollar métodos que no solo automaticen el proceso de anotación, sino que también lo optimicen en términos de tiempo, precisión y escalabilidad.

Una anotación de alta calidad de los genes V permite ampliar el conocimiento sobre la diversidad genética, la evolución del sistema inmunológico y la relación de estos genes con distintas enfermedades. Además, proporciona información más precisa para estudios posteriores, facilitando la identificación y análisis de características genéticas de interés. Esto contribuye al desarrollo de estrategias más efectivas para el diagnóstico, tratamiento y estudio de enfermedades.

Sin embargo, el volumen de datos que generan los alineamientos de secuencias representa un obstáculo. Aquí la reducción del espacio de búsqueda se vuelve una estrategia central: enfoca el análisis en las regiones más informativas y disminuye el volumen de información sin comprometer la calidad de las anotaciones. Con el uso de técnicas de agrupamiento, reducción de dimensionalidad y optimización, esta investigación convierte procesos extensos y costosos en un método eficiente, reutilizable y compatible con la escala actual de los datos biológicos.

En suma, avanzar hacia métodos de reducción del espacio de búsqueda no es solo una cuestión técnica, sino una inversión estratégica: obteniendo conocimiento aplicable a corto plazo en la investigación biomédica, la salud pública y la medicina personalizada.

1.5 Objetivo de la tesis

Desarrollar un método para la identificación de genes V con la reducción del espacio de búsqueda de las posiciones de inicio y fin en alineamientos de secuencias, integrando principios de diseño y técnicas de aprendizaje automático para su implementación y validación.

1.6 Objetivos específicos

1. Definir los principios de diseño del método para la reducción del espacio de búsqueda en genes V, a partir del análisis del proceso de caracterización del genoma, para establecer las bases que guíen su desarrollo e implementación.
2. Construir un conjunto de datos unificado a partir de diversas fuentes de alineamiento de secuencias, con técnicas de integración y normalización que garanticen la representación completa de los genes V, para sustentar el análisis y la evaluación del método.
3. Identificar la variante óptima del método con la comparación sistemática de múltiples variantes, evaluando su precisión en las ubicaciones de inicio y fin de genes V, para seleccionar la que mejor equilibre cobertura y precisión.
4. Comprobar la efectividad del método con su aplicación en genomas anotados manualmente, usando métricas cuantitativas y validación comparativa con expertos, para demostrar su utilidad en escenarios reales.

1.7 Alcances y limitaciones

Esta sección define las fronteras del trabajo. Los alcances describen las capacidades del método RIVAM y su ámbito de validación; las limitaciones reconocen las restricciones que derivan de las decisiones de diseño y de la naturaleza del problema. En conjunto, ambos apartados precisan qué aborda la propuesta, en qué condiciones opera y cuál es su aporte al problema de reducción del espacio de búsqueda.

Alcances

- El método RIVAM aborda específicamente la reducción del espacio de búsqueda para la identificación de genes V de inmunoglobulinas, dejando fuera los segmentos de Diversidad (D) y de Unión (J).
- La validación se realiza principalmente sobre genomas de quirópteros anotados por expertos del CISEI, con una extensión exploratoria a especies filogenéticamente distantes como el anfibio *Ambystoma mexicanum*.
- El método toma como entrada los resultados de cinco sistemas de alineamiento de secuencias específicas (*HMMscan*, *Exonerate:est*, *Exonerate:prot*, *Tblastn* y *Tblastx*), sobre los cuales opera la etapa de reducción.

- RIVAM se fundamenta en técnicas de aprendizaje automático no supervisado (GMM, PCA y UMAP), seleccionadas por la ausencia de etiquetas confiables en genomas de especies no caracterizadas.
- El método se concibe como una etapa previa de reducción que se integra al flujo de trabajo existente con el formato estándar GFF, garantizando su interoperabilidad con las herramientas actuales del proceso de anotación.
- La evaluación de las variantes del método se sustenta en pruebas estadísticas formales (Friedman y Nemenyi) para garantizar la validez de la configuración seleccionada como óptima.

Limitaciones

- El método no realiza la anotación funcional final de los genes V ni sustituye la revisión experta, sino que entrega un espacio de búsqueda reducido para su validación posterior.
- No aborda la clasificación definitiva entre genes funcionales y pseudogenes, tarea que permanece bajo criterio del especialista del dominio.
- No contempla la validación experimental en laboratorio de los genes identificados, limitándose al análisis computacional y a la comparación contra anotaciones de referencia.
- El alcance se restringe al segmento V de las inmunoglobulinas, sin extenderse a otras familias génicas ni a otras regiones del locus de inmunoglobulinas.
- El desempeño del método está condicionado por la calidad y el nivel de ensamblaje de los genomas de entrada, sin incluir etapas de mejora o corrección del ensamblaje.
- La optimización de hiperparámetros se realiza sobre un genoma de referencia específico, sin contemplar una adaptación automática de parámetros para cada genoma analizado.

1.8 Metodología de la investigación

La investigación adopta como eje metodológico la Ciencia del Diseño (Design Science Research, DSR), un enfoque orientado a resolver problemas complejos del mundo real basado en la construcción y validación de artefactos (Hevner et al., 2010; Peffers et al., 2020; Zeledón, 2020). A diferencia de los enfoques descriptivos o explicativos, DSR provee conocimiento a través del diseño, la implementación y la evaluación de soluciones concretas frente a necesidades reales. Aquí, el artefacto es un método computacional para la reducción del espacio de búsqueda, aplicado a la identificación de genes V.

La elección de DSR responde a la naturaleza del problema: no se trata de describir o predecir un fenómeno, sino de diseñar una solución sistemática, evaluarla empíricamente y demostrar su utilidad en un entorno real. Esa orientación hacia artefactos útiles y transferibles es la que lo convierte en el marco más adecuado para investigaciones que buscan impacto tanto científico como práctico.

DSR se complementa con la Metodología Fundamental para la Ciencia de Datos (MFCD), un marco de diez fases orientado a la toma de decisiones basada en datos (Rollins, 2015). Esta combinación asegura que las decisiones metodológicas se fundamenten en los datos y en el conocimiento experto del dominio, y que el método resultante sea a la vez sólido y aplicable en contextos reales. La Figura 1.2 presenta el modelo DSR adoptado en esta investigación, donde se aprecian los tres ciclos interconectados y su relación con el entorno del problema y la base de conocimiento científico.

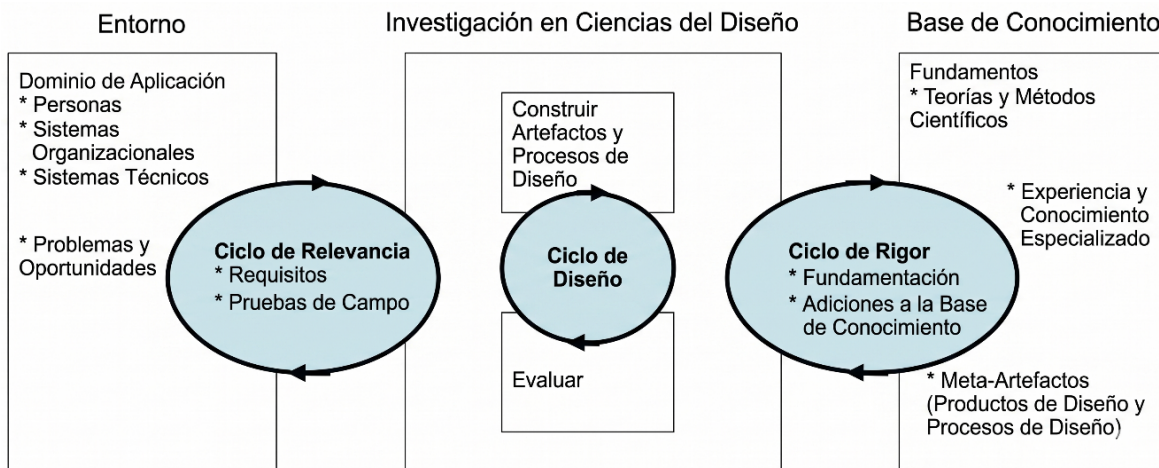


Figura 1.2 Metodología de la investigación adaptada de (Hevner et al., 2010; Peffers et al., 2020)

1.8.1 Metodología para la Ciencia del diseño

El modelo DSR se estructura en tres ciclos interconectados que guían la construcción del conocimiento, la validación empírica y el rigor teórico del proceso investigativo. Estos ciclos no son secuenciales sino iterativos, lo que permite ajustar y refinar el artefacto conforme avanza la investigación. A diferencia de los enfoques de investigación tradicionales, DSR no busca únicamente describir o explicar fenómenos, sino generar conocimiento a través del diseño, construcción y evaluación de soluciones concretas que respondan a necesidades identificadas en contextos reales.

1. **Ciclo de Relevancia:** Establece el puente entre el problema real y el diseño del artefacto. El dominio de aplicación es el análisis del genoma y la anotación de genes V, una tarea central en la bioinformática y el diagnóstico biomédico. A partir de la interacción continua con expertos del CISEI se identificó la necesidad de sistematizar la reducción del espacio de búsqueda que se obtiene de múltiples herramientas de alineamiento. Estas interacciones tradujeron un problema biológico complejo en un reto metodológico concreto y mantuvieron el diseño alineado con las necesidades reales del dominio, evitando que el método quedara como una abstracción técnica desconectada de la práctica.
2. **Ciclo de Rigor:** Asegura la solidez científica del diseño con base en una revisión sistemática de la literatura en bioinformática, inteligencia artificial y ciencia de datos. Se integraron técnicas ampliamente validadas como GMM, PCA, UMAP y PSO. Este análisis identificó las fortalezas y limitaciones de los métodos existentes y extrajo los principios teóricos que fundamentan la propuesta, de modo que el método no surge de forma aislada sino en consonancia con el conocimiento del área. Además, se emplearon recursos estandarizados como scikit-learn y métodos formales de evaluación para validar cada etapa.
3. **Ciclo de Diseño:** Constituye el núcleo operativo de la investigación, donde la propuesta pasa de una concepción teórica a una implementación evaluada en escenarios reales. Se alimenta de los ciclos de relevancia y rigor, y los transforma en un conjunto de fases estructuradas que guían el desarrollo y la validación del método. Integra unificación de datos, reducción de dimensionalidad, agrupamiento y optimización con PSO, evaluados con métricas cuantitativas como el número de genes identificados y el error en la precisión de las ubicaciones. Su naturaleza iterativa incorpora la retroalimentación de expertos y ajusta el método conforme se identifican áreas de mejora.

La articulación de estos tres ciclos garantiza que el método propuesto no solo sea técnicamente válido, sino también relevante para el problema real y fundamentado en el conocimiento científico existente. La forma en que estos tres ciclos se articulan en esta investigación se detalla en la sección 1.8.3, donde se mapea cada ciclo con los capítulos correspondientes de la tesis.

1.8.2 Metodología de solución propuesta

Tomando como base los fundamentos de DSR y los principios de la MFCD, se desarrolló una metodología de solución adaptada específicamente al problema de reducción del espacio de búsqueda de genes V. Esta metodología no replica directamente ninguno de los marcos originales, sino que integra sus elementos más relevantes en un proceso estructurado, iterativo y orientado tanto al rigor científico como a la utilidad práctica (Hevner et al., 2010; Peffers et al., 2020; Rollins, 2015; Sandoval et al., 2019; Zeledón, 2020).

La metodología resultante, presentada en la Figura 1.3, se organiza en torno a los tres ciclos de DSR siendo estos: (a) relevancia, (b) rigor y (c) diseño. Estos se concretan en dos grandes bloques operativos: el diseño del método y su implementación y validación. Cada bloque agrupa un conjunto de fases específicas que traducen los principios metodológicos en acciones concretas y verificables.

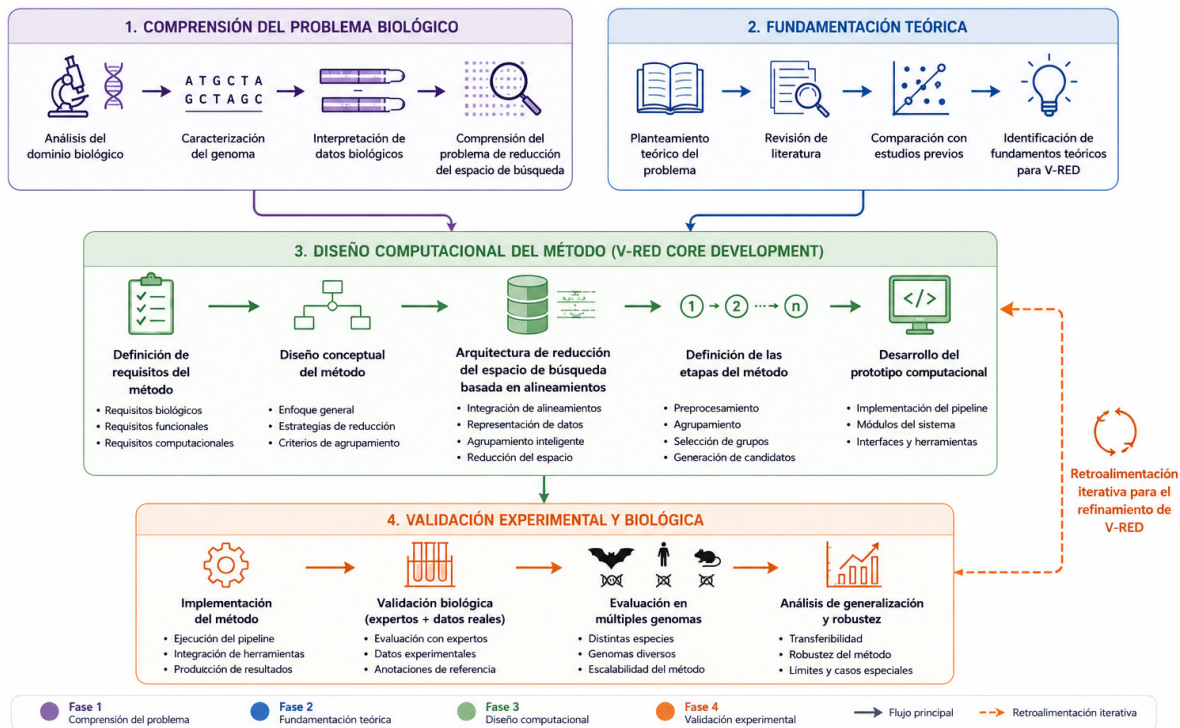


Figura 1.3 Metodología de la investigación

El primer bloque, comprensión del problema biológico, analiza el contexto del genoma asociado a la identificación de genes V. Inicia con el estudio del dominio biológico y la caracterización del genoma, para comprender la organización y complejidad de la información. Después se interpretan los datos biológicos y se analiza el problema de reducción del espacio de búsqueda, identificando las limitaciones de los enfoques tradicionales de detección. Este bloque establece el contexto científico y computacional sobre el que se fundamenta RIVAM.

El segundo bloque, fundamentación teórica, construye el soporte conceptual de la investigación. Desarrolla el planteamiento teórico del problema y una revisión de literatura enfocada en métodos de detección de genes, reducción del espacio de búsqueda y estrategias basadas en alineamientos. Incluye una comparación con estudios previos para identificar fortalezas, limitaciones y áreas de oportunidad en el estado del arte, y establece los fundamentos que sustentan RIVAM integrando conocimientos biológicos, computacionales y metodológicos. Se relaciona con el ciclo de rigor de DSR, debido a que las decisiones metodológicas se apoyan en evidencia científica previa.

El tercer bloque, diseño computacional del método RIVAM, constituye el núcleo de innovación de la investigación y concentra el aporte al campo computacional. Inicia con la definición de los requisitos biológicos, funcionales y computacionales del método. A partir de ellos se diseña la arquitectura de reducción del espacio de búsqueda, cuya decisión central es la representación de los datos heterogéneos de alineamiento basado en matrices de adyacencia, una formulación que transforma un problema biológico disperso en un objeto vectorial tratable por técnicas de aprendizaje automático no supervisado. Sobre esta representación se definen las estrategias de agrupamiento, reducción de dimensionalidad y optimización que estructuran el pipeline. Finalmente, se formalizan las etapas del método y se desarrolla un prototipo funcional que integra las herramientas necesarias para ejecutar la reducción y la detección eficiente de genes V. Este bloque se corresponde con el ciclo de diseño de DSR, pues en él se construye y formaliza el artefacto computacional que representa la contribución central de la tesis.

El cuarto bloque, validación experimental y biológica, evalúa el desempeño, la robustez y la capacidad de generalización de RIVAM. El prototipo se aplica sobre datos reales de alineamiento en formatos GFF y FASTA, y sus resultados se contrastan con anotaciones de referencia y se evalúan con expertos en biología molecular para verificar la coherencia biológica y computacional del método. Después, una evaluación multigenoma analiza su transferibilidad y escalabilidad en distintos organismos. Finalmente, un mecanismo de retroalimentación iterativa refina el método a partir de los resultados experimentales.

En conjunto, estos cuatro bloques conforman el marco metodológico de la presente investigación, integrando componentes biológicos, teóricos, computacionales y experimentales dentro de una estrategia orientada al desarrollo y validación del método *Reduction for the Identification of V-genes by sequence Alignment and Machine learning* (RIVAM). La estructura metodológica definida constituye el eje conductor de la tesis y sirve como base para la organización de los capítulos subsecuentes, tal como se describe en la Sección 1.8.3.

1.8.3 Mapeo de capítulos a ciclos

La estructura de esta tesis refleja directamente los tres ciclos de DSR y los dos bloques de la metodología de solución propuesta, asegurando coherencia entre el marco metodológico y el desarrollo de la investigación. Cada capítulo cumple un rol específico dentro de este esquema, contribuyendo de manera articulada a la construcción, validación y evaluación del método RIVAM.

El Ciclo de Relevancia se centra en los Capítulos 2 y 3. El Capítulo 2 establece los fundamentos teóricos del dominio, proporcionando los conceptos biológicos y computacionales necesarios para comprender el problema. El Capítulo 3 profundiza en el proceso real de caracterización del genoma, identificando los cuellos de botella que justifican la necesidad del método propuesto.

El Ciclo de Rigor se desarrolla en el Capítulo 4, donde se presenta la revisión sistemática de la literatura y el estado del arte global. Este capítulo fundamenta las decisiones técnicas del método, identificando las brechas existentes en la literatura que esta investigación busca atender.

El Ciclo de Diseño abarca los Capítulos 5, 6 y 7. El Capítulo 5 presenta la estructura conceptual del método, el Capítulo 6 detalla su implementación a través del pipeline del método RIVAM y el Capítulo 7 describe el diseño experimental para su validación sistemática en múltiples genomas.

Finalmente, los Capítulos 8 y 9 constituyen la síntesis de la investigación, presentando los resultados, la discusión de hallazgos, las conclusiones y las proyecciones de trabajo futuro, cerrando así el ciclo completo de la investigación bajo el marco de la metodología para la Ciencia del Diseño.

1.9 Estructura de la tesis

El presente documento se organiza en nueve capítulos, cada uno con un rol específico dentro del marco metodológico de la Ciencia del Diseño descrito anteriormente. La progresión de los capítulos refleja el ciclo completo de la investigación.

El Capítulo 1 presenta la introducción de la investigación, estableciendo el contexto del problema, su justificación, los objetivos del proyecto y la metodología que guía el desarrollo de la tesis. Este capítulo sitúa al lector en el dominio del problema y delimita el alcance de la investigación.

El Capítulo 2 desarrolla el marco conceptual, proporcionando los fundamentos teóricos necesarios para comprender el problema desde tres perspectivas complementarias: los conceptos biológicos del dominio de los genes V, los formatos y

datos utilizados en el análisis del genoma, y las técnicas computacionales empleadas en el método propuesto.

El Capítulo 3 describe el proceso real de caracterización llevado a cabo en el CISEI, identificando sus cuatro secciones, los sistemas de alineamiento utilizados y los principales cuellos de botella, con especial énfasis en la actividad de identificación de genes V como punto de intervención de esta investigación. Este capítulo fundamenta la necesidad del método RIVAM a partir del análisis del dominio real.

El Capítulo 4 presenta la revisión sistemática de la literatura y el estado del arte global, analizando los enfoques existentes en caracterización del genoma, reducción de espacios de búsqueda y técnicas de agrupamiento. A partir de este análisis se identifican las tres brechas metodológicas que esta investigación busca atender.

El Capítulo 5 expone los fundamentos conceptuales del método RIVAM, describiendo los cinco principios de diseño que guiaron su desarrollo, el razonamiento metodológico que justifica las decisiones técnicas tomadas y los seis componentes abstractos que conforman su estructura.

El Capítulo 6 detalla el desarrollo iterativo del método RIVAM, describiendo las cuatro iteraciones de construcción y refinamiento que condujeron a las tres variantes finales del método. Incluye la descripción del pipeline completo, las especificaciones técnicas de implementación y los entornos de ejecución utilizados.

El Capítulo 7 presenta el diseño experimental para la validación sistemática del método, describiendo los 12 genomas de evaluación y los 3 genomas de validación externa seleccionados, el protocolo experimental seguido y el análisis estadístico aplicado con las pruebas de *Friedman* y *Nemenyi* para la comparación de las tres variantes.

El Capítulo 8 presenta los resultados cuantitativos y cualitativos obtenidos en la evaluación sistemática y la validación externa del método, así como la discusión de los hallazgos en el contexto del problema planteado, los objetivos establecidos y la literatura revisada. Este capítulo incluye la comparación con IgDetective como método de referencia en el estado del arte.

El Capítulo 9 cierra la investigación con las conclusiones derivadas del proceso de desarrollo y validación del método RIVAM, las limitaciones identificadas durante la investigación y las proyecciones de trabajo futuro, destacando el aporte científico y práctico de la propuesta al campo de la bioinformática computacional.

Capítulo 2. Marco conceptual

Este capítulo presenta los fundamentos teóricos y técnicos que sustentan el desarrollo de la investigación. Se organiza en tres bloques temáticos complementarios: en primer lugar, los conceptos biológicos esenciales para comprender el dominio de los genes V y su papel en el sistema inmune; en segundo lugar, los formatos y tipos de datos utilizados en el análisis del genoma, incluyendo los fundamentos de los alineamientos de secuencias y su papel en la anotación genética; y en tercer lugar, las técnicas computacionales que constituyen la base del método propuesto, entre las que destacan el agrupamiento no supervisado, las técnicas de reducción de dimensionalidad como PCA y UMAP, y los modelos de optimización como PSO.

La selección de estos elementos no es arbitraria, sino que responde a la necesidad de integrar enfoques provenientes del aprendizaje automático, la bioinformática y la optimización computacional en un marco coherente que articule el problema, las herramientas utilizadas y los objetivos de la investigación. La comprensión de estos tres bloques es necesaria para contextualizar las decisiones de diseño presentadas en los capítulos subsecuentes.

2.1 Conceptos biológicos fundamentales

Esta sección introduce los conceptos biológicos esenciales para comprender el dominio en el que se desarrolla esta investigación. Si bien el enfoque de la tesis es computacional, la naturaleza del problema requiere un entendimiento claro de los elementos biológicos involucrados, debido a que las decisiones de diseño del método están directamente influenciadas por las características y restricciones propias del dominio. Los conceptos aquí presentados sientan las bases para comprender qué son los genes V, cómo se estructuran y por qué su identificación representa un reto computacional de alta complejidad.

2.1.1 Del genoma a las proteínas

Todo organismo vivo posee un genoma, compuesto por ácido desoxirribonucleico (ADN), que contiene la información necesaria para su funcionamiento biológico. Esta información está codificada en una secuencia de cuatro nucleótidos —adenina (A), timina (T), guanina (G) y citosina (C)— cuya lectura, en grupos de tres, determina la síntesis de proteínas (Brown, 2008; Virrueta Gordillo, 2012).

Las proteínas son las moléculas que ejecutan la información del genoma en forma de funciones biológicas concretas, y las posiciones de inicio y fin de un gen delimitan la región del ADN que codifica cada una. Entre las proteínas de interés para esta investigación se encuentran los anticuerpos, encargados de la defensa del organismo. El resto del capítulo se centra en los genes que dan origen a estos anticuerpos, en particular el segmento V, cuya identificación constituye el problema abordado.

2.1.2 Anticuerpos b

Los linfocitos B son las células del sistema inmune encargadas de producir anticuerpos, moléculas que detectan y neutralizan agentes externos como virus y bacterias. Cada anticuerpo se compone de cadenas pesadas y ligeras, y cada cadena se divide en dos regiones: la región variable, que reconoce y se une a los antígenos, y la región constante, que determina la clase y función del anticuerpo (Nescakova & Bystricky, 2011; Pieper et al., 2013).

La diversidad de la región variable está determinada por segmentos genéticos específicos: la cadena pesada se codifica por los segmentos V (Variable), D (Diversidad) y J (Unión), mientras que la cadena ligera se compone solo de los segmentos V y J. Esta organización es la que proporciona la enorme diversidad del repertorio inmunológico, capaz de reconocer millones de antígenos distintos (Virrueta Gordillo, 2012). La Figura 2.1 presenta la estructura del receptor de células B (BCR), donde se aprecian estas regiones.

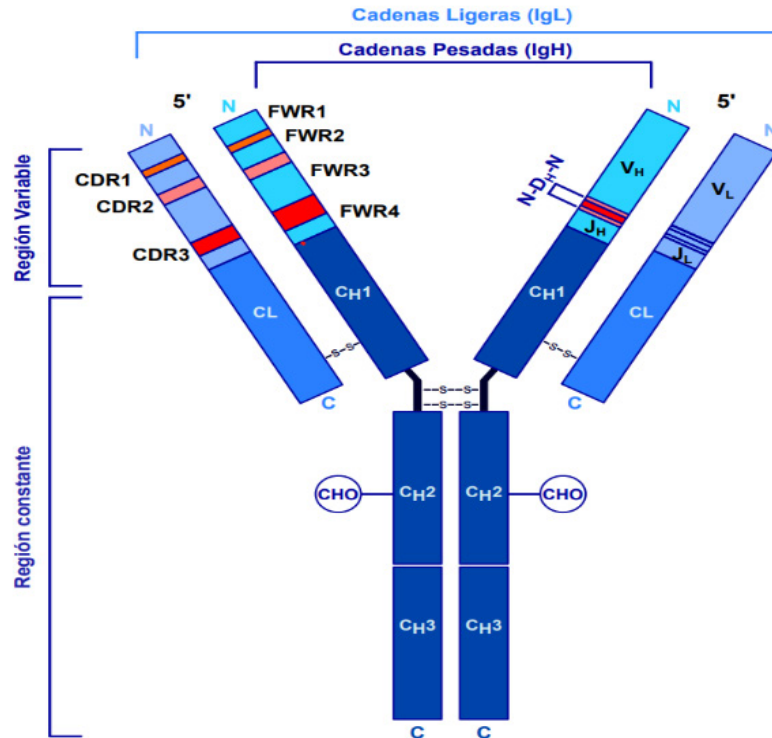


Figura 2.1 Representación de receptor de células B (BCR) obtenida de (Elizabeth Ernestina Godoy Lozano, 2015)

La identificación y caracterización precisa de estos segmentos, en particular el segmento V, representa el problema central que motiva el desarrollo del método propuesto en esta investigación. La complejidad estructural de estos genes, su alta variabilidad y la presencia de múltiples copias en el genoma son precisamente los factores que dificultan su anotación automática y justifican la necesidad de un método sistemático para reducir el espacio de búsqueda asociado a su identificación.

2.1.3 Recombinación V(D)J

El reconocimiento de los antígenos depende de la región variable del anticuerpo, que se forma mediante la recombinación de los segmentos V(D)J en cada cadena. Este proceso es el responsable de la enorme diversidad del sistema inmune, pero tiene una consecuencia directa sobre el problema abordado en esta investigación: el elevado número de genes disponibles para recombinarse implica que múltiples copias coexisten a lo largo del genoma, lo que incrementa significativamente el espacio de búsqueda que debe explorarse durante su identificación (García Merino, 2011; Mikocziova et al., 2021).

Estos segmentos se organizan en el genoma de forma ordenada, con V seguido de D y luego J, estructura que constituye la base de la región variable. La Figura 2.2 muestra esta distribución ordenada. La identificación y clasificación de los segmentos V(D)J es una parte fundamental de la anotación del genoma, debido a que localiza y caracteriza

los genes responsables de la diversidad de los anticuerpos, y resulta esencial para comprender el sistema inmune a nivel molecular (García Merino, 2011; Mikocziova et al., 2021).

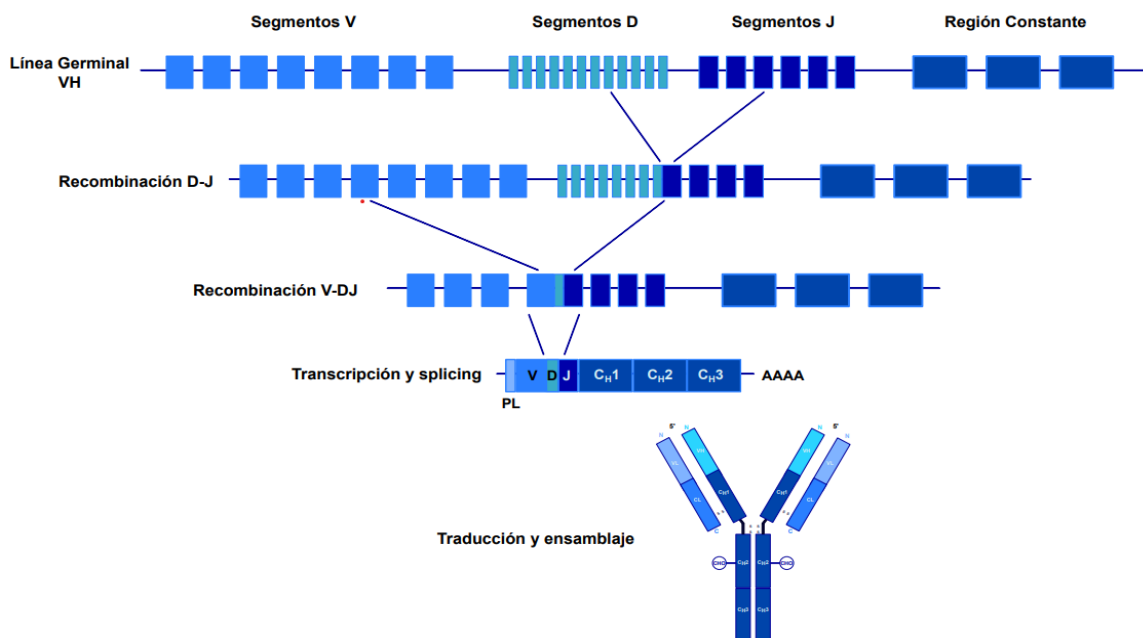


Figura 2.2 Proceso de generación de linfocito B inmaduro (IgM) obtenida de (Godoy Lozano, 2015)

Desde una perspectiva computacional, esta tarea representa un problema de búsqueda en espacios de alta dimensionalidad. Cada especie puede contener numerosas copias de estos genes distribuidas a lo largo del genoma, por lo que herramientas de alineamiento como BLAST, HMMER o Exonerate proporcionan miles de candidatos que deben evaluarse para determinar cuáles son funcionales. Ese volumen de resultados, sumado a la variabilidad entre especies, hace que la identificación manual de genes V sea costosa y propensa a errores, lo que justifica el desarrollo de métodos automatizados como el propuesto en esta investigación.

2.1.4 Estructura del gen del segmento v

El gen del segmento V presenta una estructura interna bien definida, compuesta por tres elementos funcionales que determinan su identificación, localización y papel en la formación de anticuerpos. Comprender esta estructura es fundamental para el método propuesto, ya que las posiciones de inicio y fin de cada uno de estos elementos son precisamente los datos que se buscan identificar y validar durante la anotación.

El primero de estos elementos es el *L-PART1*, el cual codifica para la primera parte del péptido señal o *signal peptide* (SP, por sus siglas en inglés). Esta es una secuencia corta de nucleótidos ubicada al inicio del gen V germinal que codifica los aminoácidos

necesarios para la translocación de la proteína. Su identificación es relevante porque delimita el inicio del gen y lo distingue de las regiones. El segundo elemento es la secuencia de señal de recombinación o *recombination signal sequence* (RSS, por sus siglas en inglés), una secuencia conservada en el extremo 3' que facilita la unión entre segmentos genéticos por medio de la recombinación V(D)J, como ocurre entre el gen V, el segmento D (en cadenas pesadas) y el segmento J durante la maduración del linfocito B. Finalmente, el segundo exón del gen constituye la región central, el cual está compuesto por el *L-PART2* y la región del gen V; es aquí donde se concentra la información funcional (dividida en regiones FR y CDR) que determina la especificidad del anticuerpo y que da nombre al propio segmento.

Además de estos tres elementos, la dirección del gen es otro aspecto crítico que debe considerarse durante su identificación. Convencionalmente, la dirección se expresa en términos de los extremos 5' y 3' de la cadena de ADN. Cuando el gen se lee del extremo 5' al 3', la dirección es de izquierda a derecha; en caso contrario, del 3' al 5', la lectura ocurre de derecha a izquierda (Kindt et al., 2007; Lefranc & Lefranc, 2001; Megrian, 2014). La Figura 2.3 presenta la estructura completa del gen V, indicando la disposición de sus elementos y su dirección de lectura.

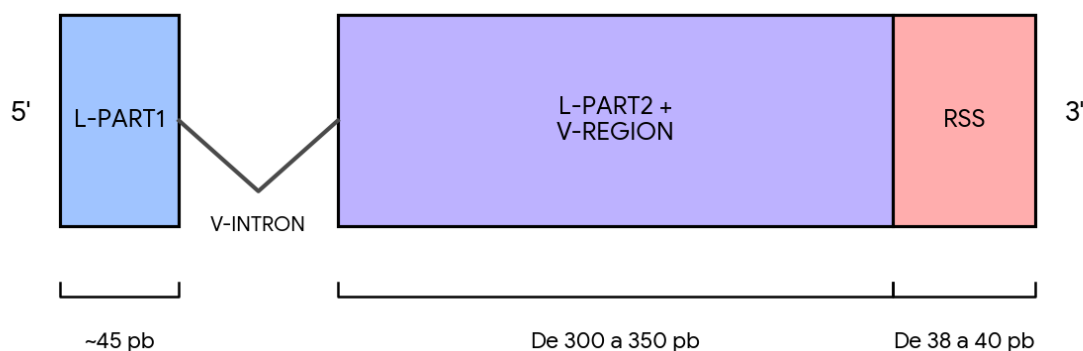


Figura 2.3 Estructura del gen V adaptada de (Lefranc & Lefranc, 2001)

Desde una perspectiva computacional, la correcta identificación de estos tres elementos y la dirección del gen representa uno de los principales desafíos del proceso de anotación. Las herramientas de alineamiento producen candidatos independientes para cada uno de estos componentes sin integrar la información de los tres en un registro unificado, lo que produce un espacio de búsqueda fragmentado y de gran volumen. En genomas de quirópteros como los analizados en esta investigación, este espacio puede superar los 300,000 registros antes de cualquier etapa de filtrado, lo que evidencia la necesidad de un método sistemático de reducción como el propuesto.

2.1.5 Genes funcionales y no funcionales

No todos los genes V presentes en el genoma son capaces de producir una proteína funcional. En función de su integridad estructural y su capacidad de expresión, los genes V se clasifican en dos categorías principales: funcionales y pseudogenes (Lefranc & Lefranc, 2001). Un gen funcional es aquel que conserva intactas todas sus regiones codificantes, incluyendo el péptido señal, el exón y la señal de recombinación, lo que lo capacita para participar activamente en la formación de anticuerpos.

Los pseudogenes son secuencias que, por diversas razones, han perdido su capacidad funcional. Esto puede deberse a mutaciones o truncamientos en su secuencia, o bien a la ausencia de uno o más de sus elementos estructurales, como el péptido señal, el exón o la señal de recombinación. A pesar de no ser funcionales, su presencia en el genoma complica el proceso de identificación, debido a que comparten similitud estructural con los genes funcionales.

En particular, los pseudogenes V representan un componente importante dentro de la anotación del genoma, debido a que forman parte de la diversidad de elementos que deben distinguirse cuidadosamente de los genes V funcionales. Su correcta identificación es clave para evitar errores de anotación y para una interpretación precisa de la organización del repertorio de inmunoglobulinas.

Desde una perspectiva computacional, esta distinción tiene un impacto directo en el problema abordado. Las herramientas de alineamiento no distinguen entre genes funcionales y pseudogenes al producir candidatos, incrementando el espacio de búsqueda con registros que dificultan el proceso de anotación. El método RIVAM busca reducir este espacio priorizando los candidatos con mayor probabilidad de contar con sus tres elementos estructurales, encapsulando tanto genes funcionales como pseudogenes dentro de un conjunto más compacto y representativo.

2.2 Datos, formatos y sistemas para el análisis del genoma

Esta sección describe los principales formatos de datos y procesos utilizados en el análisis del genoma que son relevantes para esta investigación. La comprensión de estos elementos es esencial dado que el método RIVAM opera directamente sobre estos formatos, tanto en la entrada de datos como en la obtención de resultados. Se describen los archivos FASTA y GFF como representaciones estándar de secuencias y anotaciones del genoma, así como el proceso de alineamiento de secuencias como mecanismo principal para la obtención del espacio de búsqueda que el método busca reducir.

2.2.1 Archivos fasta

Los archivos FASTA son archivos de texto plano utilizados en bioinformática para almacenar secuencias biológicas, como la información del ADN de una especie. Cada registro dentro del archivo se compone de dos partes: una línea de cabecera, identificada por el símbolo ">", que contiene el identificador, el nombre y comentarios descriptivos de la secuencia, y las líneas subsecuentes que representan la secuencia biológica propiamente dicha, expresada en nucleótidos o aminoácidos según el tipo de estudio (García Simón et al., 2018; Universidad de Gent, 2015). La simplicidad de este formato lo ha convertido en un estándar adoptado en herramientas bioinformáticas, bases de datos y pipelines de análisis computacional de áreas biológicas.

En el contexto de esta investigación, los archivos FASTA contienen la secuencia completa del genoma de cada especie y constituyen la entrada principal de las herramientas de alineamiento. Estas herramientas comparan la secuencia del genoma contra secuencias de referencia de genes V previamente caracterizados, produciendo como resultado un conjunto de regiones candidatas que son posteriormente procesadas por el método RIVAM.

2.2.2 Alineamiento de secuencias

El alineamiento de secuencias es una técnica fundamental en bioinformática para comparar dos o más secuencias de ADN para identificar regiones de similitud y determinar su grado de correspondencia en el genoma. El proceso busca encontrar la disposición óptima entre las secuencias comparadas, maximizando la similitud e identificando posibles regiones de interés (Chao et al., 2022).

Existen dos enfoques principales para el alineamiento de secuencias. El alineamiento local se centra en identificar regiones específicas de alta similitud entre secuencias, sin considerar la totalidad de estas, siendo útil cuando se buscan patrones concretos dentro de secuencias largas. El alineamiento global, por su parte, busca alinear las secuencias en su totalidad, considerando todas sus posiciones, lo que resulta adecuado cuando se comparan secuencias de longitud y función similares (Chao et al., 2022; Li, 2018). La Figura 2.4 presenta el comportamiento de ambos enfoques y sus diferencias en la identificación de regiones de similitud.

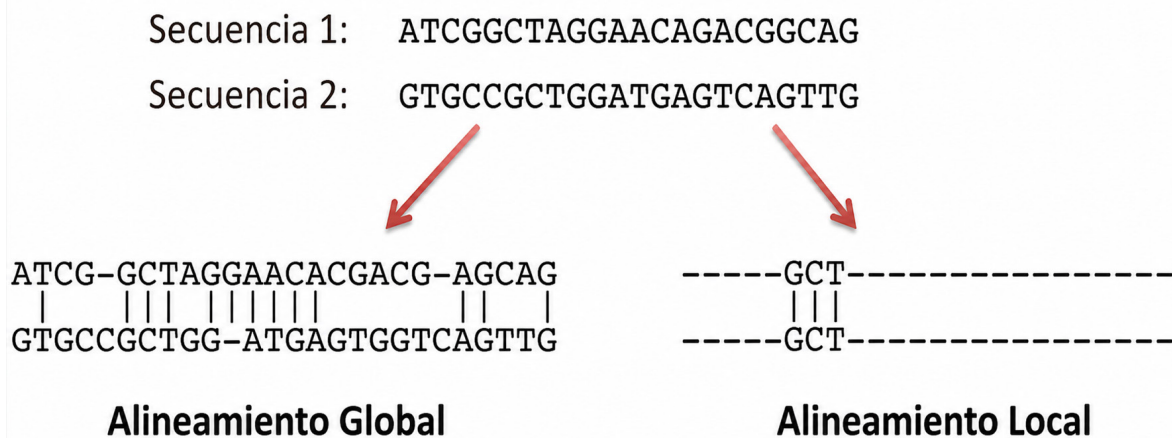


Figura 2.4 Comportamiento del alineamiento local vs global adaptada de (Ganesh et al., 2018)

En el contexto de esta investigación, el alineamiento de secuencias es el mecanismo principal mediante el cual se obtienen los candidatos para la identificación de genes V. Para ello se utilizan tres herramientas especializadas, cada una con características y enfoques distintos:

BLAST (Basic Local Alignment Search Tool) es una de las herramientas más utilizadas en bioinformática para la búsqueda de similitudes entre secuencias biológicas. Su funcionamiento se basa en un enfoque heurístico de alineamiento local que compara de manera eficiente una secuencia de consulta contra una base de datos de referencia (Altschul et al., 1990). BLAST incluye distintas variantes diseñadas para diferentes tipos de secuencias. En el contexto de esta investigación, se utilizan dos: (a) BLASTn, que compara secuencias de nucleótidos contra bases de datos de nucleótidos, y (b) BLASTx, que traduce una secuencia de nucleótidos en sus posibles marcos de lectura proteicos y la compara contra bases de datos de proteínas.

En el proceso de identificación de genes V, estas herramientas comparan secuencias del genoma con referencias de genes V previamente caracterizados. Como resultado, se obtienen alineamientos que indican las regiones de similitud, junto con métricas asociadas como la puntuación del alineamiento y el valor de significancia, lo que facilita evaluar la calidad de las coincidencias detectadas (Altschul et al., 1990).

HMMER (Hidden Markov Models Eddy Research) es una herramienta basada en modelos ocultos de Markov (HMM, por sus siglas en inglés), diseñada para detectar patrones conservados en secuencias biológicas. A diferencia de BLAST, que se fundamenta en la similitud directa entre secuencias, HMMER construye un modelo estadístico a partir de una familia de secuencias relacionadas y lo utiliza para identificar regiones del genoma que se ajustan a ese perfil conservado (Eddy, 2011). Este enfoque reconoce mejor la variabilidad dentro de un mismo conjunto de secuencias sin depender de coincidencias exactas.

En este trabajo se utiliza específicamente *hmmscan*, el cual compara secuencias de consulta contra una base de datos de perfiles HMM. En este caso, dichos perfiles se construyen a partir de genes V previamente caracterizados. El programa produce como salida las regiones del genoma que presentan similitud estadísticamente significativa con estos perfiles, junto con métricas de confianza. Este enfoque es útil para la detección de genes V en genomas poco caracterizados o altamente variables, debido a que los perfiles HMM representan la estructura conservada del gen con mayor sensibilidad que los métodos basados únicamente en similitud directa (Eddy, 2011).

Exonerate es una herramienta de alineamiento diseñada específicamente para la comparación de secuencias del genoma contra secuencias de referencia, con la particularidad de considerar la estructura intrónica y exónica de los genes (Slater & Birney, 2005). A diferencia de BLAST y HMMER, Exonerate no solo identifica regiones de similitud, sino que modela explícitamente la estructura del gen, lo que le ayuda detectar con mayor precisión los límites de los exones y las señales de empalme.

En el contexto de esta investigación se utilizan dos modalidades específicas. La primera es Exonerate:protein2genome:local, que compara secuencias de referencia a nivel de aminoácidos contra el genoma, para identificar las regiones que corresponden a la estructura de una proteína conocida. Esta modalidad es útil para detectar el exón codificante del gen V, debido a que trabaja directamente con la información de la secuencia proteica de referencia. La segunda modalidad es Exonerate:est2genome, que compara secuencias de ADN complementario (cDNA) contra el genoma, siendo útil para identificar la estructura del gen a nivel de nucleótidos, incluyendo el péptido señal y la señal de recombinación.

La combinación de estos dos enfoques ofrece una visión estructural más completa del gen V, complementando los resultados de BLASTn, BLASTx y hmmscan. Mientras BLAST y HMMER aportan información sobre las regiones de similitud y los perfiles conservados, Exonerate añade detalle sobre la organización interna del gen, aportando información clave sobre las posiciones de inicio y fin de cada uno de sus elementos estructurales.

Cada una de estas herramientas produce sus propios resultados de alineamiento en formatos nativos distintos, con diferencias en sensibilidad, especificidad y nivel de detalle de la información reportada. Estos resultados son posteriormente procesados y transformados a formato GFF con el objetivo de estandarizar la información y facilitar su integración en un único conjunto de datos.

La combinación de los resultados de BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y Exonerate:est2genome da como resultado un espacio de búsqueda amplio y heterogéneo. A partir de esta integración, el método

RIVAM reduce dicho espacio de manera sistemática, consolidando la información proveniente de todas las fuentes en un conjunto compacto y representativo de candidatos para la identificación de genes V.

2.2.3 Archivos GFF

Los archivos GFF (General Feature Format) son archivos de texto plano estructurados en nueve columnas delimitadas por tabulaciones, utilizados como formato estándar para representar las características y anotaciones de una secuencia (EMBL-EBI, 2022; García Simón et al., 2018; gmod, 2017). En el proceso de caracterización del genoma del CISEI, los archivos GFF representan las posiciones donde las herramientas de alineamiento detectaron similitud con genes de referencia. Los campos que componen cada registro son:

- **Seqname:** consiste en el nombre del cromosoma o contig.
- **Source:** es el nombre del programa que generó esta característica, o la fuente de datos.
- **Feature:** nombre del tipo de secuencia, por ejemplo, gen, variación o similitud.
- **Start:** posición de inicio de la secuencia que está dado en un dato tipo numérico.
- **End:** posición de fin de la secuencia que está dado en un dato tipo numérico.
- **Score:** es un valor de punto flotante.
- **Strand:** consiste en la dirección de la secuencia y es definida como + (hacia adelante) o - (hacia atrás).
- **Frame:** indica el valor desde donde se leerá la primera letra de la secuencia puede ser desde la primera, segunda o tercera posición.
- **Attribute:** es una lista separada por punto y coma de pares de etiquetas y valores, que proporciona información adicional sobre cada función.

Dado que cada herramienta de alineamiento entrega sus resultados en formatos nativos distintos, es necesario realizar una transformación previa para convertirlos al formato GFF. En el caso de BLASTn y BLASTx, los resultados en formato tabular son procesados para extraer las columnas de posición, puntuación y dirección, y mapearlas a las columnas correspondientes del GFF. Para hmmscan, la tabla de resultados propia de HMMER es procesada para identificar las coordenadas de los dominios detectados y convertirlas al formato estándar. En el caso de Exonerate, tanto la modalidad protein2genome:local como est2genome produce una salida estructurada que incluye información de exones y señales de empalme, la cual es transformada para preservar los límites precisos de cada elemento del gen V en el archivo GFF resultante.

2.2.4 Integrative genomics viewer

El Integrative Genomics Viewer (IGV) es una herramienta de visualización de alto rendimiento desarrollada por el Broad Institute, diseñada para el análisis interactivo de datos del genoma a gran escala. IGV carga y analiza simultáneamente múltiples tipos de archivos, incluyendo secuencias en formato FASTA, alineamientos en formato GFF y otros formatos estándar de la bioinformática, facilitando la inspección visual integrada de la información de un organismo (Robinson et al., 2023).

En el contexto del proceso de identificación y anotación de genes V, IGV desempeña un papel central como interfaz entre los resultados computacionales de las herramientas de alineamiento y el criterio experto del biólogo molecular. A través de su interfaz gráfica, el experto puede analizar de manera simultánea los alineamientos obtenidos por BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y Exonerate:est2genome, superpuestos sobre la secuencia completa del genoma del organismo analizado. Este análisis integrado ayuda al experto identificar las regiones del genoma donde diversas herramientas coinciden en señalar la presencia de un posible gen V, aumentando la confianza en la identificación.

Una de las funcionalidades más relevantes de IGV para este proceso es su capacidad de ampliar regiones específicas del genoma hasta el nivel de nucleótidos y aminoácidos individuales, para que el experto verifique las reglas biológicas de cada sección del gen V directamente sobre la secuencia. Esta funcionalidad es la que hace posible la aplicación de las reglas de identificación descritas en la sección 3.1.1, donde el experto verifica la presencia de patrones específicos como el nonamero ACAAAAACC en la RSS, la Cisteína 104 en el exón V o la metionina inicial en el péptido señal.

La Figura 2.5, 2.6 y 2.7 presentan ejemplos del espacio de búsqueda de genes V en la herramienta Integrative Genomics Viewer (IGV). En estas figuras se observa la superposición de los alineamientos obtenidos por las distintas herramientas de alineamiento sobre regiones específicas del genoma, así como la alta densidad de registros candidatos que deben ser revisados manualmente por el experto durante el proceso de identificación y anotación de genes V.

En la Figura 2.5 se muestra una vista general de una región en el genoma donde se integran simultáneamente los resultados producidos por las diversas herramientas de alineamiento, incluyendo nhmmer, exonerate y BLAST. Cada línea horizontal representa un conjunto de alineamientos proyectados sobre la misma región cromosómica. Puede observarse una elevada concentración de candidatos distribuidos a lo largo de la secuencia, evidenciando la redundancia y complejidad del espacio de búsqueda obtenido por la combinación de múltiples herramientas. Esta situación incrementa

considerablemente la carga de revisión manual requerida por el experto para determinar qué regiones corresponden realmente a genes V funcionales.

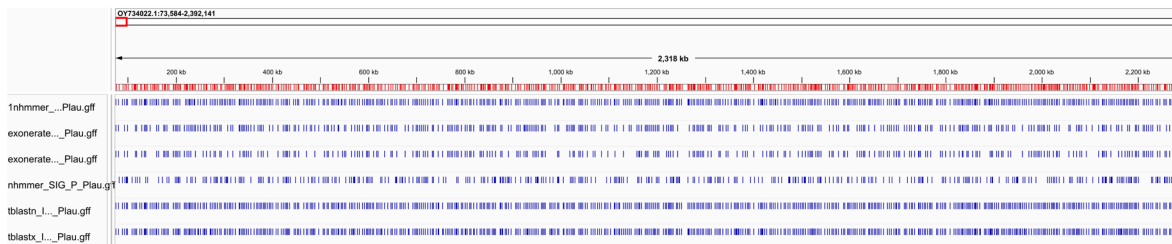


Figura 2.5 Vista general del espacio de búsqueda de genes V las herramientas de alineamiento en IGV

La Figura 2.6 presenta un ejemplo de una región donde una herramienta produjo una gran cantidad de alineamientos altamente redundantes sobre prácticamente el mismo segmento del cromosoma. En esta vista se aprecia cómo múltiples registros se superponen casi completamente, formando una acumulación densa de candidatos alineados sobre una región relativamente pequeña del genoma. Este comportamiento es común en procesos de alineamiento sensibles, donde pequeñas variaciones en las posiciones de inicio y fin producen múltiples registros similares que deben ser posteriormente filtrados y consolidados. La Figura evidencia uno de los principales problemas abordados por esta investigación: la existencia de grandes volúmenes de alineamientos redundantes que dificultan la identificación eficiente de genes V.

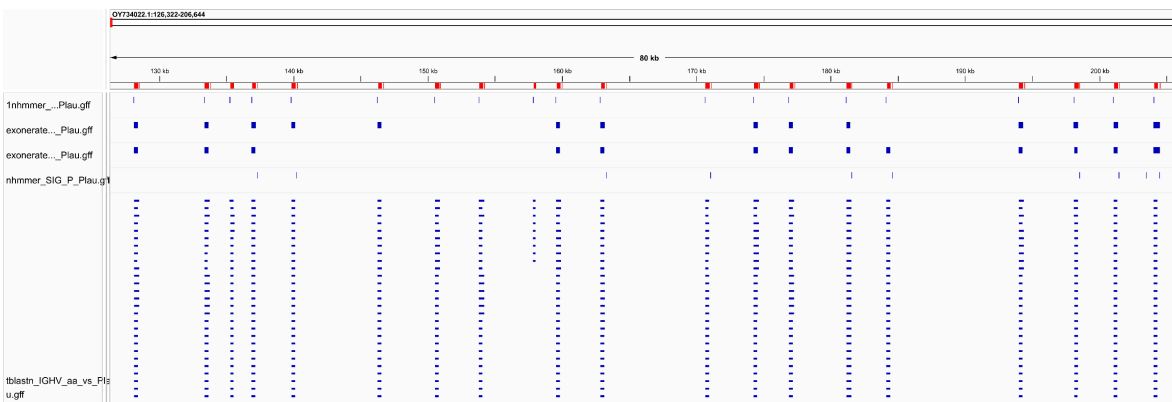


Figura 2.6 Ejemplo de alineamientos redundantes y superpuestos sobre una región candidata de genes V

Finalmente, la Figura 2.7 muestra un caso donde distintas herramientas obtienen alineamientos parcialmente coincidentes sobre una misma región. A diferencia de la figura anterior, aquí los alineamientos presentan diferencias en longitud, posición y cobertura sobre la secuencia, reflejando la variabilidad existente entre algoritmos de alineamiento. Algunas herramientas producen alineamientos más extensos, mientras que otras obtienen regiones más específicas o fragmentadas. Este comportamiento introduce incertidumbre en la delimitación exacta de los genes V y obliga al experto a

comparar manualmente múltiples evidencias para determinar el candidato más representativo.

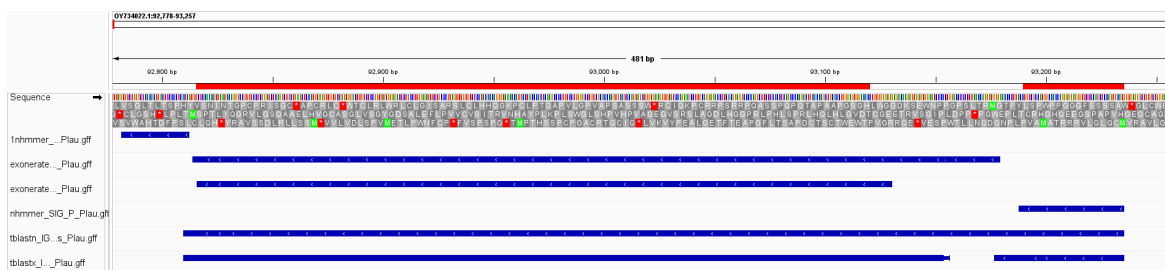


Figura 2.7 Comparación de alineamientos parcialmente coincidentes sobre una región candidata a gen V

En conjunto, estas figuras muestran la magnitud y complejidad del espacio de búsqueda asociado al proceso de identificación de genes V. La elevada cantidad de alineamientos redundantes, la variabilidad entre herramientas y la superposición de candidatos evidencian la necesidad de desarrollar métodos computacionales orientados a reducir el espacio de búsqueda y facilitar la identificación eficiente de genes V en genomas de vertebrados, problemática que motiva el desarrollo del método RIVAM propuesto en esta investigación.

2.3 Conceptos computacionales clave

Esta sección presenta los fundamentos computacionales que sustentan el diseño e implementación del método RIVAM. Los conceptos aquí descritos provienen de tres áreas principales: la ciencia de datos, el aprendizaje automático y la optimización computacional. Su selección responde directamente a las decisiones de diseño del método, donde cada técnica cumple un rol específico dentro del pipeline de reducción del espacio de búsqueda.

Se describen en primer lugar el concepto de reducción del espacio de búsqueda como estrategia central del enfoque, seguido de las técnicas de agrupamiento y reducción de dimensionalidad empleadas, y finalmente el método de optimización utilizado para ajustar los parámetros del modelo.

2.3.1 Reducción de un espacio de búsqueda

En problemas de optimización, aprendizaje automático y análisis de datos, el espacio de búsqueda se define como el conjunto total de posibles soluciones o configuraciones que pueden evaluarse para resolver un problema determinado. Este espacio puede estar conformado por variables continuas, discretas o categóricas, y su tamaño y complejidad aumenta exponencialmente con el número de dimensiones involucradas (Bazaraa et al., 2006).

Formalmente, si un problema de optimización tiene n variables de decisión $x_1, x_2, \dots, x_n$, cada una con un dominio de búsqueda D_i , el espacio total de búsqueda S se puede expresar como (Ec. 2.1):

$$S = D_1 \times D_2 \times \dots \times D_n \quad (\text{Ec. 2.1})$$

Cuando este espacio se vuelve excesivamente amplio o contiene muchas soluciones irrelevantes, la eficiencia de los algoritmos disminuye significativamente, afectando tanto el tiempo de cómputo como la calidad de las soluciones obtenidas (Feng et al., 2022). En este contexto, la reducción del espacio de búsqueda se refiere a la aplicación de técnicas que eliminan o filtran combinaciones no viables, redundantes o poco prometedoras, manteniendo solo aquellas que tienen mayor probabilidad de contener soluciones óptimas o relevantes.

La reducción puede lograrse mediante diferentes estrategias, tales como:

- **Filtrado basado en restricciones:** se eliminan combinaciones que violan condiciones conocidas del problema.
- **Agrupamiento o clustering:** se consolidan soluciones similares en representaciones más compactas.
- **Proyecciones o transformaciones dimensionales:** como PCA o UMAP, que reestructuran el espacio en dimensiones más significativas.
- **Evaluación heurística temprana:** que descartar soluciones con bajo rendimiento en etapas preliminares (Bode et al., 2019)

El objetivo general es reducir el espacio original S a un subconjunto $S' \subset S$, tal que (Ec. 2.2):

$$|S'| \ll |S| \quad y \quad \text{Óptimo}(S') \approx \text{Óptimo}(S) \quad (\text{Ec. 2.2}):$$

Es decir, se busca que el subespacio reducido S' conserve las mejores soluciones del espacio original S , minimizando el costo de búsqueda sin afectar la calidad de los resultados. La reducción del espacio de búsqueda no solo mejora el rendimiento computacional, sino que también facilita la interpretación y evaluación de las soluciones, en dominios como la bioinformática, el diseño de redes, la ingeniería estructural o la minería de datos complejos (Bode et al., 2019; Cho & Hegde, 2019; Feng et al., 2022).

En el contexto de esta investigación, la reducción del espacio de búsqueda se aplica específicamente al conjunto de registros de alineamiento obtenidos por múltiples herramientas bioinformáticas, donde el espacio original está conformado por miles de posiciones candidatas para genes V. El método RIVAM implementa una combinación de estas estrategias para reducir dicho espacio manteniendo los candidatos con mayor coherencia biológica y precisión en las posiciones de inicio y fin del gen.

2.3.2 Algoritmo mezclas gaussianas

El modelo de mezclas gaussianas (*Gaussian Mixture Model*, GMM) es un algoritmo de aprendizaje no supervisado que asume que los datos se obtienen a partir de una combinación ponderada de varias distribuciones gaussianas. A diferencia de métodos como K-Means, que asignan rígidamente cada punto a un único clúster, GMM asigna una probabilidad de pertenencia a cada componente, lo que resulta en una segmentación más flexible, en escenarios donde los grupos presentan formas no esféricas, varianzas heterogéneas o fronteras difusas (Bishop & Nasrabadi, 2006; Dempster et al., 1977; Oscar Contreras Carrasco, 2024; Pedregosa et al., 2011).

El uso de algoritmos de agrupamiento en la reducción del espacio de búsqueda se justifica teóricamente desde varios enfoques:

- **Reducción de redundancia:** Al identificar elementos similares, se evita evaluar soluciones repetitivas, concentrando el análisis en regiones más informativas del espacio de búsqueda (Jain, 2010). De Amorim y Mirkin en (De Amorim & Mirkin, 2016) propusieron un enfoque basado en clustering para la selección de características redundantes y demuestra que agrupar características similares y seleccionar representantes puede mejorar la eficiencia sin pérdida significativa de información.
- **Segmentación estructurada del espacio:** Los clústeres dividen grandes volúmenes de datos en subconjuntos manejables, optimizando recursos computacionales y facilitando análisis más focalizados. Este enfoque es útil en contextos de datos de alta dimensionalidad, donde la segmentación adecuada puede mejorar significativamente el rendimiento de los algoritmos de aprendizaje automático (Steinbach et al., 2004).
- **Preservación de relaciones semánticas:** en dominios como bioinformática, los elementos agrupados suelen compartir propiedades funcionales o biológicas relevantes (Ishioka, 2000).
- **Compatibilidad con entornos no etiquetados:** al no requerir clases previas, los modelos de clustering son ideales para tareas de exploración, donde no se cuenta con verdad de terreno (Ishioka, 2000).

En este estudio, GMM es seleccionado tras múltiples pruebas preliminares por su capacidad de modelar tres características clave de los datos:

1. **Estructura no esférica:** la dispersión de los puntos no forma agrupaciones circulares, lo que limita el desempeño de algoritmos como K-Means.
2. **Superposición de grupos:** las regiones con alta densidad y superposición requieren modelos probabilísticos para representar adecuadamente la pertenencia.

3. **Varianzas desiguales:** GMM da a cada componente una matriz de covarianza, para adaptarse a estructuras complejas.

La función de densidad de probabilidad de un GMM con K componentes es (Ec. 2.3):

$$p(x) = \sum_{k=1}^K \pi_k \cdot \mathcal{N}(x | \mu_k, \Sigma_k) \quad (\text{Ec. 2.3})$$

donde:

- π_k es el peso (mezcla) del componente k ($\sum \pi_k = 1$),
- μ_k es la media del componente k ,
- Σ_k es su matriz de covarianza.
- $\mathcal{N}(x | \mu_k, \Sigma_k)$ es la función de densidad normal multivariada.

El modelo se entrena mediante el algoritmo **Expectation-Maximization (EM)**, que alterna entre:

- **E-step:** calcular la probabilidad de pertenencia de cada punto a cada componente;
- **M-step:** actualizar los parámetros del modelo en función de las probabilidades obtenidas.

Este enfoque ofrece una representación más fiel de la estructura de los datos y es útil en contextos como la bioinformática, donde los datos presentan alta variabilidad y una estructura compleja. En el contexto de esta investigación, GMM es el algoritmo central del método RIVAM, utilizado para agrupar los registros de alineamiento en función de sus características estructurales y posicionales, para identificar el grupo con mayor concentración de candidatos relevantes para la identificación de genes V.

2.3.3 Análisis de componentes principales

El Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica estadística de reducción de dimensionalidad lineal ampliamente utilizada para transformar conjuntos de datos altamente correlacionados en un nuevo sistema de coordenadas más compacto, conocido como componentes principales. Estos componentes son ortogonales entre sí y están ordenados de acuerdo con la cantidad de varianza que explican del conjunto original (Jolliffe & Cadima, 2016)

El proceso se basa en el cálculo de la matriz de covarianza de los datos y en su posterior descomposición espectral de la cual se obtienen los vectores propios (componentes principales) y los valores propios (varianza explicada por cada componente). Matemáticamente, esto se expresa como (Ec. 2.4):

$$\Sigma v = \lambda v \quad (\text{Ec. 2.4})$$

donde Σ es la matriz de covarianza, λ es un valor propio y v es el vector propio correspondiente.

El resultado es una proyección del conjunto de datos en un espacio de menor dimensión, conservando la mayor cantidad de información posible en los primeros componentes. Esta técnica es útil cuando se busca eliminar redundancias, mejorar la eficiencia computacional y facilitar procesos posteriores como la clasificación o el agrupamiento (Abdi & Williams, 2010).

En este trabajo, PCA se implementa como paso previo al modelo de agrupamiento GMM, con el fin de evaluar si una compresión lineal del conjunto de datos mejora la eficiencia del algoritmo, la definición de los clústeres y la separación entre grupos en contextos del genoma.

2.3.4 Uniform manifold approximation and projection

UMAP (Uniform Manifold Approximation and Projection) es una técnica de reducción de dimensionalidad no lineal que busca preservar tanto la estructura local como la global de los datos. Fue propuesta por (McInnes et al., 2018) y se basa en principios de topología algebraica y teoría de grafos para modelar la conectividad entre los datos. Su objetivo es obtener una representación embebida de menor dimensión que conserve las relaciones topológicas del espacio original.

El algoritmo construye primero un grafo ponderado que representa la proximidad entre puntos en el espacio de alta dimensión. Luego, intenta preservar esa estructura en un espacio de menor dimensión mediante la minimización de una función de pérdida que mide la discrepancia entre la distribución de pares en ambos espacios. Esta función de pérdida se define como una divergencia cruzada o de Kullback-Leibler modificada (Ec. 2.5):

$$C = \sum_{(i,j)} w_{ij} \log\left(\frac{w_{ij}}{p_{ij}}\right) \quad (\text{Ec. 2.5})$$

donde:

- w_{ij} : similitud entre los puntos i y j en el espacio original.
- p_{ij} : similitud proyectada en el espacio embebido.

UMAP ha demostrado ser particularmente efectivo en conjuntos de datos complejos y de alta dimensionalidad, como matrices de expresión, secuencias anotadas o representaciones latentes en bioinformática. Su capacidad para capturar estructuras no lineales lo convierte en una alternativa potente frente a métodos lineales como PCA o t-SNE, cuando se busca conservar relaciones globales y locales simultáneamente (Becht et al., 2019).

En esta investigación, UMAP se utiliza como etapa previa al agrupamiento mediante GMM, con el objetivo de evaluar si su capacidad de preservar estructuras no lineales favorecía una mejor separación de grupos y una mayor cobertura en la identificación de genes V.

2.3.5 Particle swarm optimization

Particle swarm optimization (PSO) u optimización por enjambre de partículas (PSO) es un algoritmo de optimización inspirado en el comportamiento colectivo de enjambres de organismos, como aves o peces. Cada partícula representa una solución candidata y se desplaza en el espacio de búsqueda influenciada tanto por su mejor posición individual como por la mejor posición global encontrada por el grupo.

Durante el proceso de optimización, cada partícula ajusta su trayectoria en función de su propio mejor desempeño (conocido como "best personal") y del mejor desempeño encontrado por el enjambre en su conjunto (denominado "best global"). Así las partículas exploran el espacio de búsqueda de manera colaborativa, balanceando la explotación de zonas prometedoras con la exploración de nuevas áreas para evitar quedarse atrapadas en óptimos locales (Imron & Prasetyo, 2020; Selvaraj & Choi, 2022). Las ecuaciones de actualización de velocidad y posición se muestran en la (Ec. 2.6) y la (Ec. 2.7), respectivamente:

$$v_i(t + 1) = \omega v_i(t) + c_1 r_1 (p_i - x_i(t)) + c_2 r_2 (g - x_i(t)) \quad (\text{Ec. 2.6})$$

$$x_i(t + 1) = x_i(t) + v_i(t + 1) \quad (\text{Ec. 2.7})$$

donde:

- x_i : posición de la partícula i ,
- v_i : velocidad,
- p_i : mejor posición individual,
- g : mejor posición global,
- ω, c_1, c_2 : parámetros de inercia y aceleración,
- r_1, r_2 : números aleatorios en $[0,1]$.

PSO se emplea para optimizar los hiperparámetros del modelo GMM, evaluando iterativamente una función objetivo que considera criterios de calidad del agrupamiento y de identificación de genes V. Su capacidad de adaptación a espacios multidimensionales lo hace ideal para tareas de ajuste fino en modelos de reducción del espacio de búsqueda.

Capítulo 3. Análisis del proceso de identificación y anotación de genes v

Este capítulo describe el proceso real de identificación y anotación de genes V llevado a cabo en el CISEI, a partir del análisis directo con expertos en biología molecular. A diferencia del capítulo anterior, que estableció los fundamentos teóricos del dominio, este capítulo se enfoca en comprender cómo se realiza el proceso en la práctica, identificando sus etapas, las herramientas utilizadas y los principales cuellos de botella que limitan su eficiencia. Este análisis constituye la base del Ciclo de Relevancia de la metodología DSR y justifica la necesidad del método RIVAM propuesto en esta investigación.

3.1 Descripción del proceso de anotación en el cisei

Con el objetivo de comprender el proceso de identificación y anotación de genes V tal como se lleva a cabo en la práctica, se tomó como referencia el trabajo desarrollado por Chavez Serret (2025) en su tesis de maestría titulada "Modelo conceptual del proceso de anotación del gen V de anticuerpos basado en la Ciencia del Diseño", realizada en el CISEI del INSP. En dicho trabajo se modeló el proceso mediante la técnica BPMN (Object Management Group, s/f), permitiendo representar de manera formal la secuencia de actividades, sus interacciones y los puntos críticos del flujo de trabajo.

El modelo fue desarrollado bajo el enfoque de la Ciencia del Diseño (Hevner et al., 2010; Peffers et al., 2020), lo que implicó un proceso iterativo de adquisición de conocimiento, representación y validación. Durante el ciclo de relevancia, se estableció un diálogo continuo con expertos en biología molecular del CISEI, quienes compartieron su conocimiento del dominio a través de reuniones y sesiones de trabajo presencial. Adicionalmente, se colaboró directamente en las actividades que desempeñan los biólogos moleculares durante el proceso de anotación, lo que permitió una comprensión profunda y detallada del flujo de trabajo real.

En la investigación desarrollada se evaluaron y compararon cuatro enfoques de modelado conceptual para seleccionar la representación más adecuada: diagrama de clases (Grady Booch & Rumbaugh, 2011), diagrama de actividades (Villa et al., 2010), modelo SADT (Ahmed et al., 2014; Marca & McGowan, 1987) y modelo BPMN (White, 2004).

El modelo BPMN resultante fue validado en tres fases. En primer lugar, se verificó su conformidad con la norma ISO 19510 (Object Management Group, s/f) y mediante el uso del programa Bizagi Modeler. En segundo lugar, fue revisado por expertos en biología molecular del CISEI, quienes confirmaron que representa fielmente el proceso real de anotación.

El proceso de identificación y anotación de genes V comprende cuatro secciones interrelacionadas que parten de los archivos de alineamiento obtenidos previamente hasta la clasificación y etiquetado final de los genes identificados. La Figura 3.1 presenta el modelo BPMN del proceso completo, compuesto por dos participantes: el experto en biología molecular, encargado de ejecutar el proceso e interactuar con el sistema, y el software Integrative Genomics Viewer (IGV), encargado de administrar los archivos de alineamiento y el genoma del individuo analizado.

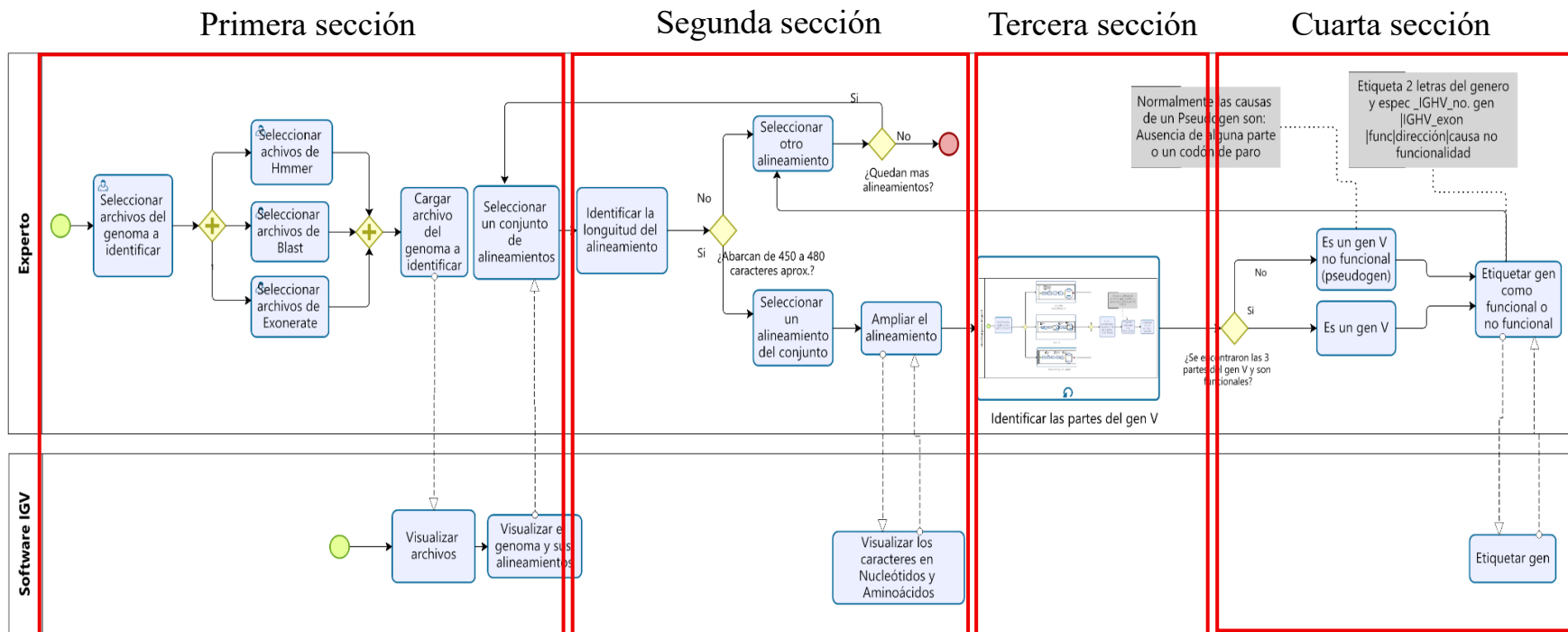


Figura 3.1 Modelo BPMN del proceso de identificación y anotación de genes V, obtenida de (Chavez Serret, 2025)

3.1.1 Etapas y reglas de identificación

El modelo BPMN desarrollado por Chavez Serret (2025) estructura el proceso de identificación y anotación de genes V en cuatro secciones secuenciales. Cada sección representa un conjunto de actividades realizadas por el experto en biología molecular en colaboración con el software IGV, y su correcta ejecución es condición necesaria para avanzar a la siguiente. A continuación, se describe cada una de estas secciones, detallando las actividades, decisiones y reglas que las componen:

Selección y carga de archivos, el proceso inicia con la selección por parte del experto de los archivos de alineamiento obtenidos por las tres herramientas bioinformáticas: Hmmer, BLAST y Exonerate. Cada uno de estos archivos contiene información sobre diferentes secciones del gen V: Hmmer se encarga de identificar la señal de recombinación (RSS), BLAST identifica el péptido señal y Exonerate identifica el exón. Una vez seleccionados los archivos, el sistema IGV los carga y visualiza junto con el genoma del individuo a analizar, mostrando los alineamientos disponibles. El experto procede entonces a seleccionar un conjunto de alineamientos para su análisis.

Análisis de longitud y selección del alineamiento, la segunda sección inicia con la identificación de la longitud del conjunto de alineamientos seleccionado. El experto verifica si el conjunto abarca entre 450 y 480 caracteres aproximadamente, longitud que corresponde a un gen V completo con todas sus partes. Si el conjunto no cumple con esta condición, el experto selecciona otro conjunto de alineamientos; de no existir más conjuntos disponibles, el proceso finaliza. Si el conjunto cumple con la longitud requerida, el experto selecciona un alineamiento específico del conjunto y procede a ampliar la imagen en el IGV. El sistema responde mostrando la secuencia de caracteres en nucleótidos y aminoácidos, lo que permite al experto identificar con mayor precisión las secciones del gen.

Identificación de las partes del gen V, la tercera sección constituye el núcleo del proceso de anotación y el punto donde se concentra el mayor esfuerzo manual del experto. Como se muestra en la Figura 3.2, el subproceso de identificación opera de manera paralela sobre las tres secciones estructurales del gen V, aplicando reglas específicas para cada una. Una vez analizadas las tres secciones, se sitúan las coordenadas de inicio y fin de cada parte identificada y se procede a su etiquetado.

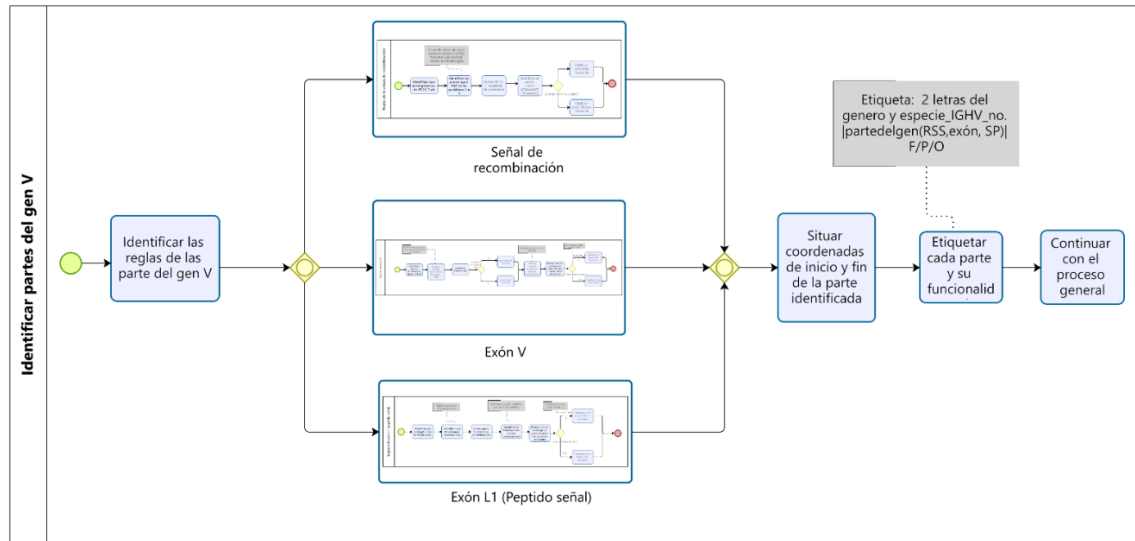


Figura 3.2 Subproceso de identificación de las partes del gen V, obtenida de (Chavez Serret, 2025)

La señal de recombinación (RSS), se identifica una longitud de 39 (± 1) caracteres, se verifica la presencia de un patrón similar a CAC, se cuentan 23 caracteres de nucleótidos y se verifica que finalice con el nonamero ACAAAAACC. Si cumple estas reglas se clasifica como RSS funcional; de lo contrario, como no funcional.

El exón V, se identifica una longitud aproximada de 300 caracteres, se verifica la presencia de una Cisteína 104 y la ausencia de codones de paro (identificados por el símbolo "*"). De manera opcional se puede identificar la señal de splicing al final del exón y la presencia de la señal de recombinación adyacente.

El péptido señal (SP): se identifica una longitud aproximada de 46 caracteres, se verifica la presencia de una metionina (M) o ATG al inicio, se cuentan aproximadamente 15 caracteres y se verifica que finalice con una glicina (G) o GT. No debe contener codones de paro.

Una vez verificadas las reglas de las tres secciones, se determinan y registran las coordenadas de inicio y fin de cada parte en el sistema IGV. Cada parte recibe una etiqueta siguiendo el formato estándar: dos letras del género y especie, seguidas de IGHV, número de gen, parte del gen (RSS, exón o SP) y su funcionalidad, que puede ser funcional (F), pseudogen (P) u ORF abierto (O).

Clasificación y etiquetado del gen V, la cuarta y última sección del proceso consiste en determinar la funcionalidad global del gen V identificado. Si las tres partes del gen cumplen con sus respectivas reglas, el gen se clasifica como funcional; si alguna de las partes no cumple, se clasifica como pseudogen. En ambos casos, el sistema IGV registra la etiqueta correspondiente y el experto continúa con el proceso general, repitiendo el ciclo para el siguiente conjunto de alineamientos disponible hasta agotar todos los candidatos del genoma analizado.

3.2 Limitaciones del proceso actual

El análisis del proceso descrito en la sección anterior, llevado a cabo en conjunto con los expertos del CISEI, permitió identificar una serie de limitaciones críticas que afectan directamente la eficiencia y escalabilidad del proceso de identificación y anotación de genes V. Estas limitaciones se concentran principalmente en las secciones 2, 3 y 4 del proceso, donde el experto debe revisar manualmente cada conjunto de alineamientos, aplicar las reglas de identificación y clasificar cada gen. A continuación, se describen los cuatro cuellos de botella principales identificados.

3.2.1 Volumen excesivo del espacio de búsqueda

El principal cuello de botella en el proceso de identificación es el volumen de registros que el experto debe revisar. Las cinco herramientas de alineamiento empleadas (BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y Exonerate:est2genome) obtienen en conjunto cientos de miles de registros para un solo genoma, cada uno representando una posible ubicación candidata para alguna sección del gen V. Sin embargo, la gran mayoría de estos registros no corresponden a genes V reales, debido a que pueden ser falsos positivos derivados de similitudes superficiales con las secuencias de referencia, porque corresponden a pseudogenes incompletos o porque representan regiones repetitivas del genoma sin relevancia biológica para el estudio.

Por ejemplo, en el caso de *Desmodus rotundus*, los archivos GFF obtenidos por estas herramientas superan los 167,000 registros, de los cuales únicamente 88 corresponden a genes V validados por expertos. Esta desproporción equivale a más de 1,900 registros por cada gen válido. Sin embargo, este no es el caso más extremo. El volumen de registros varía considerablemente entre especies, como se muestra en los genomas analizados en esta investigación:

- *Antrozous pallidus* tiene 163,001 registros con 90 genes de referencia
- *Eptesicus fuscus* tiene 240,377 registros con 168 genes de referencia
- *Nyctalus aviator* tiene 342,239 registros con 300 genes de referencia
- *Pipistrellus nathusii* tiene 308,848 registros con 228 genes de referencia
- *Vespertilio murinus* tiene 421,862 registros con 325 genes de referencia

Estos datos evidencian que el espacio de búsqueda no solo es enorme en términos absolutos, sino que su magnitud varía significativamente entre especies, lo que hace aún más necesario contar con un método sistemático y adaptable de reducción.

La magnitud de este espacio de búsqueda puede entenderse como un problema de alta dimensionalidad, donde el número de candidatos crece de manera no lineal con el tamaño del genoma y la cantidad de herramientas utilizadas. A diferencia de otros

contextos donde existen estrategias consolidadas para reducir este tipo de espacios, en la identificación de genes V no se contaba, hasta el desarrollo de este trabajo, con un método sistemático y reutilizable que realizara esta reducción de forma integral sin comprometer la coherencia biológica de los candidatos seleccionados.

Como consecuencia, los expertos deben enfrentarse directamente a este volumen masivo de información, dedicando una parte considerable de su tiempo a descartar registros irrelevantes en lugar de enfocarse en el análisis detallado de los candidatos más prometedores. Esta situación se vuelve aún más crítica al considerar que el proceso debe repetirse para cada nuevo genoma analizado, sin mecanismos efectivos para reutilizar de forma automatizada el conocimiento obtenido en análisis previos.

3.2.2 Costo temporal de la revisión manual

El proceso de identificación descrito en la sección 3.1.1 requiere que el experto revise cada conjunto de alineamientos de manera individual, verificando manualmente las reglas biológicas de cada sección del gen V. Se trata de un procedimiento inherentemente secuencial: el experto selecciona un conjunto de alineamientos, evalúa su longitud, amplía la región en IGV, identifica las tres secciones del gen aplicando sus respectivas reglas, determina las coordenadas de inicio y fin, y finalmente etiqueta el resultado antes de continuar con el siguiente caso. Cada uno de estos pasos exige atención detallada y conocimiento especializado, lo que dificulta su automatización directa sin una intervención metodológica previa.

Dado el volumen de registros, este proceso puede extenderse hasta un año de trabajo continuo para un solo genoma de tamaño mediano, dependiendo de su complejidad estructural y del número de genes V esperados. En genomas de mayor tamaño o con alta variabilidad, el tiempo requerido puede incrementarse aún más. Esta situación no solo representa una limitación práctica para los laboratorios, sino también un obstáculo científico relevante, debido a que retrasa la disponibilidad de datos necesarios para estudios comparativos, evolutivos o biomédicos.

Además, la lentitud del proceso tiene implicaciones económicas importantes. El tiempo de un experto con formación especializada en inmunogenómica es un recurso limitado y costoso, por lo que dedicarlo a la revisión manual de grandes volúmenes de registros, en su mayoría irrelevantes, resulta ineficiente. En este contexto, la reducción del espacio de búsqueda se plantea como una estrategia clave, debido a que permite concentrar el esfuerzo del experto en un subconjunto de candidatos con mayor probabilidad de corresponder a genes V reales.

3.2.3 Susceptibilidad al error en la revisión manual

La naturaleza manual, repetitiva y prolongada del proceso lo hace susceptible a distintos tipos de error. La fatiga cognitiva acumulada durante la revisión de miles de registros afecta la capacidad del experto para aplicar las reglas biológicas de forma consistente. Esto puede derivar en omisiones de genes válidos, clasificaciones incorrectas entre genes funcionales y pseudogenes, errores en la delimitación de coordenadas de inicio y fin, así como variabilidad en la interpretación de casos ambiguos donde las reglas no ofrecen una respuesta completamente definida.

Además, la necesidad de integrar y comparar visualmente los resultados de diversas herramientas de alineamiento en IGV incrementa la complejidad del análisis. El experto debe considerar simultáneamente los resultados de BLASTn, BLASTx, hmmscan y Exonerate, identificar regiones donde estos se solapan o se complementan, y determinar cuáles corresponden a estructuras completas. Esta integración visual resulta propensa a errores cuando se realiza de forma manual sobre grandes volúmenes de datos.

La magnitud del error humano en procesos de anotación manual ha sido documentada en la literatura bioinformática, donde se observa que la consistencia entre anotaciones realizadas por distintos expertos puede ser menor de lo esperado, particularmente en regiones con alta variabilidad o en situaciones donde las reglas presentan excepciones (Chavez Serret, 2025). Esta variabilidad afecta la reproducibilidad de los resultados, debido a que estudios posteriores dependen de la calidad y consistencia de estas anotaciones.

En este contexto, la reducción del espacio de búsqueda contribuye a mitigar el problema. Al presentar un conjunto más pequeño y representativo de candidatos, disminuye la carga cognitiva del experto, facilita la aplicación consistente de las reglas biológicas y reduce la probabilidad de omitir genes válidos o cometer errores de clasificación.

3.2.4 Ausencia de un flujo de trabajo sistematizado

El proceso actual carece de criterios estandarizados, documentados y reproducibles que guíen las decisiones del experto de manera consistente a lo largo de todo el proceso de anotación. Si bien el modelo BPMN desarrollado por (Chavez Serret, 2025) representa un avance significativo en la formalización del proceso, las reglas de identificación descritas en dicho modelo dependen en gran medida del criterio y la experiencia individual del analista, particularmente en situaciones donde los alineamientos son ambiguos o donde las características del gen no cumplen exactamente con los patrones esperados.

Esta dependencia del conocimiento experto tiene varias consecuencias negativas para la capacidad de escalamiento y adaptación del proceso. En primer lugar, dificulta la incorporación de nuevos investigadores al proceso de anotación, debido a que el conocimiento necesario para tomar decisiones correctas no está completamente externalizado en documentos o herramientas accesibles. En segundo lugar, limita la posibilidad de replicar el proceso en otros laboratorios o instituciones que no cuentan con expertos con el mismo nivel de formación específica en inmunogenómica de quirópteros o vertebrados. En tercer lugar, impide la automatización directa del proceso, debido a que los sistemas computacionales requieren reglas explícitas y formalizadas para operar de manera autónoma.

La falta de sistematización también se manifiesta en la ausencia de un flujo de trabajo unificado que integre los resultados de las múltiples herramientas de alineamiento en un formato común antes de iniciar la revisión manual. Actualmente, el experto debe gestionar manualmente los archivos obtenidos por cada herramienta, lo que introduce una fuente adicional de variabilidad y error en el proceso. La construcción de un pipeline que automatice esta integración y estandarice los datos de entrada para la revisión experta es uno de los aportes fundamentales del método RIVAM propuesto en esta investigación.

3.3 Limitaciones de las herramientas disponibles

Si bien el proceso descrito en las secciones anteriores representa la práctica actual en el CISEI, la comunidad científica ha desarrollado herramientas computacionales que buscan automatizar parcialmente la identificación de genes V en genomas. Sin embargo, estas herramientas presentan limitaciones importantes cuando se aplican a especies con variabilidad genética específica o con genomas poco caracterizados, que es precisamente el tipo de escenarios que aborda esta investigación. En esta sección se analiza la herramienta más relevante en este contexto: IgDetective.

3.3.1 IgDetective en el contexto de quirópteros

IgDetective es un algoritmo computacional desarrollado por Sirupurapu et al. (2022) con el objetivo de predecir genes en los loci de inmunoglobulinas, incluyendo los segmentos V, D y J, en genomas de diferentes especies. El método fue diseñado para operar de manera automatizada sobre secuencias del genoma completas, identificando regiones candidatas con la combinación de alineamientos de secuencias y modelos estadísticos entrenados con datos de referencia de múltiples especies.

Los resultados reportados por sus autores demostraron que IgDetective fue capaz de identificar genes V en diferentes especies de mamíferos, mostrando un desempeño competitivo frente a métodos de anotación manual en los genomas para los cuales fue

entrenado (Sirupurapu et al., 2022). Esto lo convierte en una referencia importante dentro del estado del arte de la identificación automatizada de genes V.

Sin embargo, su aplicación a especies con características particulares, como las analizadas en esta investigación, revela limitaciones que justifican la necesidad del método RIVAM. En primer lugar, especies como los quirópteros presentan una variabilidad en la arquitectura del gen V que difiere significativamente de la observada en los mamíferos para los cuales IgDetective fue originalmente diseñado y validado. Esta variabilidad se manifiesta en diferencias en la longitud de las secciones del gen, en los patrones de las señales de recombinación y en la proporción de pseudogenes respecto a genes funcionales, lo que reduce la capacidad del modelo estadístico de IgDetective para generalizar correctamente a estas especies. Esta limitación no es exclusiva de los quirópteros, sino que se extiende a cualquier especie cuya variabilidad genética difiera de los datos de entrenamiento del modelo.

En segundo lugar, IgDetective no sistematiza el conocimiento experto que los biólogos moleculares han acumulado sobre las reglas específicas de identificación de genes V en distintas especies. Las reglas descritas en la sección 3.1.1, que incluyen longitudes específicas, patrones de nucleótidos y criterios de funcionalidad, representan un conocimiento especializado que no está completamente incorporado en el modelo de IgDetective, el cual fue entrenado principalmente con datos de referencia de especies más ampliamente estudiadas como el ratón y el humano.

En tercer lugar, IgDetective obtiene sus propios candidatos a partir del genoma sin integrarse con los resultados de las herramientas de alineamiento ya utilizadas en el proceso del CISEI. Esto significa que su adopción implicaría reemplazar el pipeline existente en lugar de complementarlo, lo que no es compatible con el enfoque de esta investigación, que busca mejorar el proceso actual sin descartar ni interrumpir el flujo de trabajo establecido por los expertos.

Estas limitaciones no invalidan la utilidad de IgDetective como herramienta general de identificación de genes V, sino que evidencian la necesidad de desarrollar métodos complementarios que sean adaptables a distintas especies y que se integren con el proceso de anotación existente. Es precisamente esta brecha la que el método RIVAM (descrito en detalle en la Sección 1.7.2) busca cubrir.

3.4 Oportunidades de intervención metodológica

El análisis del proceso de identificación y anotación de genes V presentado en este capítulo, junto con las limitaciones identificadas tanto en el proceso manual como en los métodos computacionales existentes, permite delimitar con precisión el punto donde una intervención metodológica puede obtener el mayor impacto. Esta sección

describe dónde y cómo el método RIVAM se inserta en el proceso actual, y por qué su diseño responde de manera directa a los cuellos de botella identificados.

3.4.1 Punto de intervención en el proceso

A partir del análisis del modelo BPMN desarrollado por Chavez Serret (2025), se identificó que la oportunidad de intervención más significativa se encuentra entre la Sección 1 y la Sección 2 del proceso, es decir, después de que los archivos de alineamiento han sido obtenidos y transformados al formato GFF, pero antes de que el experto inicie la revisión manual de los conjuntos de alineamientos. Es en este punto donde el espacio de búsqueda alcanza su máximo volumen y donde una reducción sistemática puede obtener el mayor beneficio en términos de eficiencia y calidad del proceso.

La intervención no busca reemplazar el juicio experto ni eliminar las etapas del proceso actual, sino reducir la cantidad de registros que el experto debe revisar manualmente, entregando un conjunto compacto, representativo y biológicamente coherente de candidatos. De esta manera, el experto puede concentrar su atención y conocimiento especializado en los registros con mayor probabilidad de corresponder a genes V reales, reduciendo significativamente el tiempo de revisión y la probabilidad de error.

Esta ubicación de la intervención es estratégica por varias razones. Primero, opera sobre datos ya obtenidos por el pipeline existente, lo que garantiza compatibilidad con el proceso actual sin requerir cambios en las herramientas utilizadas. Segundo, actúa antes de la etapa más costosa en tiempo y recursos, maximizando el impacto de la reducción. Tercero, su salida es un archivo GFF reducido que puede integrarse directamente en el IGV, respetando el flujo de trabajo establecido por los expertos.

3.4.2 Fundamentación del método propuesto

El método RIVAM fue diseñado específicamente para abordar los cuellos de botella identificados en este capítulo, integrando en un pipeline sistemático y reutilizable las técnicas computacionales necesarias para reducir el espacio de búsqueda de manera eficiente y biológicamente coherente. Su diseño no surge de manera aislada, sino como respuesta directa a las necesidades identificadas durante el análisis del proceso real de anotación en el CISEI, lo que garantiza que cada componente del método tiene una justificación concreta en el dominio del problema.

El problema del enorme espacio de búsqueda se resuelve mediante la unificación de los resultados de las cinco herramientas de alineamiento en un único conjunto estructurado, eliminando redundancias y registros poco informativos. Este proceso de consolidación no descarta información de manera arbitraria, sino que aplica criterios

biológicos y estructurales para preservar los registros con mayor probabilidad de corresponder a regiones relevantes del gen V, lo que permite reducir cientos de miles de registros a un conjunto manejable de candidatos representativos sin comprometer la cobertura del proceso.

La revisión manual prolongada se aborda mediante técnicas de reducción de dimensionalidad y agrupamiento no supervisado, que permiten identificar y priorizar subconjuntos con mayor concentración de genes V reales dentro del espacio de búsqueda reducido. Al organizar los registros en grupos coherentes desde el punto de vista biológico y estadístico, el método permite al experto concentrar su atención en el grupo más prometedor, disminuyendo significativamente la carga de análisis y acelerando el proceso de anotación sin sacrificar la calidad de los resultados.

El alto error humano se mitiga al incorporar criterios estadísticos y biológicos formalizados que guían automáticamente las etapas de filtrado y selección, reduciendo la dependencia del juicio individual en tareas repetitivas y rutinarias. Al sistematizar estas decisiones, el método garantiza que los mismos criterios se apliquen de manera consistente a todos los registros, independientemente del volumen de datos o de la fatiga acumulada durante el proceso de revisión, lo que se traduce en anotaciones más precisas y reproducibles.

En cuanto a la falta de sistematización, RIVAM propone un pipeline reproducible, documentado y transferible que puede aplicarse de forma consistente a distintos genomas y especies, convirtiendo el conocimiento experto acumulado en el CISEI en reglas estructuradas y reutilizables. Esta característica es relevante para la transferencia del proceso a nuevos investigadores o instituciones, ya que reduce la dependencia de la experiencia individual y facilita la escalabilidad del proceso de anotación a un mayor número de genomas.

En conjunto, estas características posicionan a RIVAM no solo como una solución técnica al problema de la reducción del espacio de búsqueda, sino como un puente entre el conocimiento en biología molecular y las técnicas de ciencia de datos, contribuyendo a establecer un proceso de anotación más eficiente, preciso y escalable para la caracterización de genes V en múltiples especies. El detalle de su diseño e implementación se presenta en los Capítulos 5 y 6.

Capítulo 4. Revisión sistemática

La revisión sistemática de la literatura representa un paso clave en la construcción de cualquier propuesta científica, dado que permite identificar los enfoques, técnicas y marcos teóricos utilizados previamente para abordar problemas similares. En el marco de la metodología de la Ciencia del Diseño, este capítulo corresponde al Ciclo de Rigor, cuyo objetivo es garantizar la solidez científica del diseño a partir del conocimiento acumulado en la literatura. El propósito de esta revisión no es únicamente describir los trabajos existentes, sino reconocer sus limitaciones, detectar vacíos metodológicos y justificar la necesidad de una propuesta diferente, integral y adaptable, posicionando esta investigación en el contexto actual del problema.

La revisión se organiza en tres bloques: los antecedentes dentro del CENIDET, el estado del arte global en caracterización del genoma y reducción de espacios de búsqueda, y las brechas identificadas que motivan el desarrollo del método RIVAM.

4.1 Antecedentes dentro de cenidet

En la revisión de trabajos previos desarrollados en el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) se identificaron trabajos relacionados con el análisis de datos y el modelado conceptual del genoma que guardan una relación directa o indirecta con la presente investigación. Si bien ninguno de estos trabajos aborda específicamente el problema de la reducción del espacio de búsqueda en la identificación de genes V , sus aportaciones en términos de metodología, herramientas y enfoques computacionales constituyen antecedentes relevantes que contribuyeron a definir el contexto institucional de esta investigación.

Selección de variantes genéticas asociadas a enfermedades crónico-degenerativas usando técnicas de aprendizaje automático, la tesis de doctorado desarrollada por Carlos Alberto Moncada Vázquez (Moncada Vázquez, 2019) define un procedimiento que permite aplicar técnicas de aprendizaje automático al problema de clasificación de datos masivos en el dominio genómico. La investigación aborda el problema desde cuatro componentes principales: el preprocesamiento de los datos, el análisis de componentes independientes, el entrenamiento de clasificadores y su evaluación. El objetivo fundamental es proveer una herramienta de diagnóstico molecular adecuada para la solución de problemas complejos en el genoma, concretamente en la secuenciación de ADN asociada al diagnóstico temprano de osteosarcoma, basada en modelos predictivos obtenidos por inferencia automática de conocimiento a partir del manejo de grandes volúmenes de datos.

La relevancia de este trabajo para la presente investigación radica en dos aspectos. Primero, demuestra la viabilidad de aplicar técnicas de aprendizaje automático a problemas de alta dimensionalidad en el dominio genómico, sentando un precedente metodológico dentro del CENIDET para el uso de este tipo de enfoques en bioinformática. Segundo, el problema de clasificación de datos masivos que aborda comparte características estructurales con el problema de reducción del espacio de búsqueda que enfrenta esta investigación, particularmente en lo que respecta a la necesidad de manejar grandes volúmenes de datos heterogéneos y de aplicar criterios automáticos de selección y filtrado.

Recuperación y procesamiento de datos genómicos para la carga de información clasificada en una base de datos basada en un modelo conceptual del genoma humano, la tesis de maestría desarrollada por José Luis Rodríguez Pliego (José Luis Rodríguez Pliego, 2009) tuvo como objetivo establecer una metodología para el desarrollo de módulos de carga y actualización de datos relacionados al esquema conceptual del genoma humano. La metodología propuesta consiste en la extracción automática de datos de bases de datos genéticas para cargarlos en una base de datos obtenida a partir del esquema conceptual del genoma humano, haciendo uso de

recursos como el NCBI y desarrollando una herramienta de carga basada en la metodología propuesta.

Este trabajo es relevante para la presente investigación en tanto aborda el problema de la integración y estructuración de datos provenientes de múltiples fuentes, un desafío que también está presente en el método RIVAM durante la etapa de unificación de datos. La necesidad de consolidar información heterogénea en un formato común y estructurado, identificada en este trabajo previo, es análoga a la necesidad de integrar los resultados de las cinco herramientas de alineamiento en un único conjunto de datos que sirva de entrada para el proceso de reducción.

Finalmente, la tesis de maestría de Chavez Serret (2025), ya descrita en detalle en el Capítulo 3, representa el antecedente más directo de esta investigación dentro del CENIDET, al haber modelado formalmente el proceso de anotación de genes V y haber identificado la revisión manual como el principal cuello de botella que motiva el desarrollo del método RIVAM.

4.2 Aplicación de la revisión sistemática

La presente sección tiene como objetivo identificar en la literatura científica trabajos que guarden una relación directa o semejante con la investigación desarrollada, con el fin de contextualizar el método RIVAM dentro del estado del arte global y fundamentar sus decisiones de diseño. Para estructurar la búsqueda, se plantearon dos preguntas de investigación:

- ¿Cómo se ha utilizado el alineamiento de secuencias para la identificación de genes?
- ¿Cómo se ha abordado la sistematización del proceso de anotación de genes y sus principales técnicas?

Estas preguntas dieron origen a dos núcleos temáticos. El primero se enfoca en trabajos que hayan utilizado el enfoque de alineamiento de secuencias para la identificación de genes, incluyendo los principios de funcionamiento del alineamiento y los principales algoritmos utilizados. El segundo núcleo busca esclarecer cómo se ha abordado la sistematización del proceso de anotación de genes, sin limitarse a los genes del segmento V, sino generalizando hacia cualquier tipo de gen. Lo que se espera de este segundo núcleo es identificar sistemas, arquitecturas, métodos, algoritmos y técnicas para la sistematización del proceso de anotación con enfoques de homología o ab initio.

A partir de los núcleos temáticos identificados, se definieron dos cadenas de búsqueda incorporando variaciones terminológicas y sinónimos para mejorar la cobertura y relevancia de la información recuperada:

- "sequence alignment" AND "gene segments" AND "germline" AND "machine learning" AND "immunoglobulin"
- ("gene annotation" OR "gene identification") AND ("machine learning" OR "graph theory" OR "computational modeling") AND ("V(D)J" OR "antibodies" OR "immunoglobulins") AND "vertebrates"

El proceso de búsqueda de la revisión se llevó a cabo entre el 7 de septiembre de 2022 y el 30 de marzo de 2023, consultando bases de datos académicas reconocidas: PubMed, MDPI, NCBI, SpringerLink y ScienceDirect, complementadas con Google Académico.

La selección de estas fuentes se fundamenta en su amplia cobertura de investigaciones en bioinformática y aprendizaje automático, lo que permitió incluir los trabajos más relevantes para esta investigación. Se consideraron únicamente artículos escritos en inglés, publicados entre 2018 y 2023, indexados en revistas científicas o presentados en congresos de reconocido prestigio en el área.

Se excluyeron artículos de opinión, revisiones no sistemáticas y trabajos sin acceso al texto completo. El criterio de fecha de publicación se estableció con el objetivo de capturar el estado más reciente del arte, considerando que el campo de la anotación computacional ha experimentado avances significativos en los últimos años, con la incorporación de técnicas de aprendizaje automático y reducción de dimensionalidad.

La búsqueda inicial arrojó un total de 220 artículos, distribuidos de la siguiente manera: 81 artículos correspondientes a la primera cadena de búsqueda y 139 artículos a la segunda. Dado el volumen de resultados, fue necesario aplicar un proceso de selección sistemático y reproducible que permitiera identificar los trabajos más relevantes para los objetivos de esta investigación.

Por ello se aplicó la metodología PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), un marco reconocido internacionalmente para la realización de revisiones sistemáticas que incorpora criterios conceptuales y metodológicos para garantizar la transparencia y reproducibilidad del proceso de selección (Urrútia & Bonfill, 2010).

Si bien PRISMA fue desarrollada originalmente para metaanálisis de ensayos clínicos, su adaptación a otros tipos de estudios ha sido ampliamente aceptada en la literatura científica, particularmente en áreas donde la sistematización de la evidencia existente es fundamental para fundamentar nuevas propuestas metodológicas.

La Figura 4.1 presenta el diagrama PRISMA del proceso de búsqueda y selección, donde se detallan los criterios de inclusión y exclusión aplicados en cada etapa, así como el número de artículos descartados en cada fase del proceso. Como resultado de este proceso, se seleccionaron 21 artículos para su análisis detallado: 4 provenientes de la primera cadena de búsqueda y 17 de la segunda. Estos trabajos se organizan en dos grupos temáticos que se describen en las secciones siguientes: trabajos sobre alineamiento de secuencias y trabajos sobre identificación y anotación de genes.

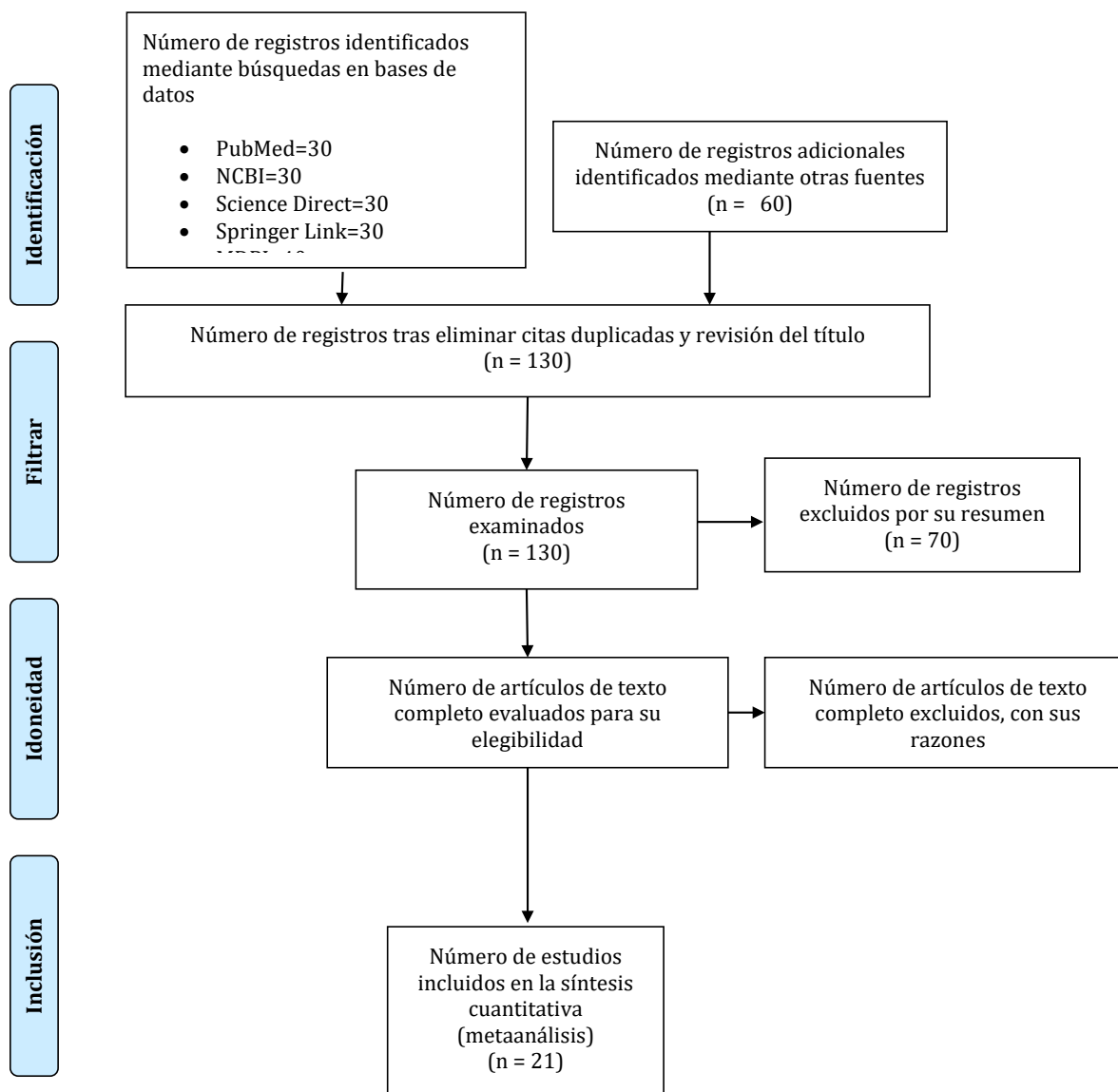


Figura 4.1 Diagrama PRISMA del proceso de búsqueda y selección de literatura

4.2.1 Trabajos sobre el alineamiento de secuencias

Este núcleo temático tiene como objetivo describir trabajos que realicen análisis, investigaciones y recomendaciones sobre el alineamiento de secuencias como técnica fundamental para la identificación de genes. La revisión se centra en identificar los principales enfoques, algoritmos y herramientas utilizados en este campo, con el fin de contextualizar el uso que el método RIVAM hace de estas técnicas en su pipeline de reducción del espacio de búsqueda. La Tabla 4.1 presenta los trabajos seleccionados, organizados en tres aspectos principales: una descripción del trabajo, los métodos y algoritmos utilizados, y la referencia bibliográfica correspondiente.

4.2.1.1 Tabla comparativa de trabajos sobre alineamiento de secuencias

La Tabla 4.1 está dividida en cuatro puntos principales: a) Una breve descripción del trabajo, b) los métodos, algoritmos o sistemas utilizados, c) la referencia del trabajo.

Tabla
Tabla 4.1 comparativa de trabajos sobre alineamiento de secuencias

N.	Descripción	Métodos	Referencia
1	Se realizó una revisión de los diferentes algoritmos y técnicas para el alineamiento de secuencias. Las principales heurísticas usadas son divide y vencerás, programación dinámica acotada y alineación en estrella. Minimap2 es un sistema basado en homología que utiliza semillas de secuencias para analizar el genoma y buscar regiones similares a las referencias. Se evaluó con conjuntos de secuencias de humanos y ratones.	Alineamiento por pares, alineamiento múltiple y modelo oculto de Markov (HMM)	(Chao et al., 2022)
2	El trabajo realiza una revisión sobre el alineamiento de secuencias y el uso del algoritmo genético para abordar el problema. Su método de evaluación se basa en Suma de pares (SOP) y penalización por brecha. COVID-Align es un sistema basado en alineamiento de secuencias múltiples para el análisis de genomas de la cepa hCoV-19 disponibles en GISAID.	Programación dinámica basada en Needleman-Wunsch y alineamiento de secuencia global	(Li, 2018)
3		Alineamiento por pares y múltiple, alineación progresiva y algoritmo genético	(Chowdhury & Garai, 2017)
4		Modelos ocultos de Markov (HMM) y herramienta HMMER	(Lemoine et al., 2021)

4.2.1.2 CONCLUSIONES DE LOS TRABAJOS SOBRE ALINEAMIENTO DE SECUENCIAS

El análisis de los trabajos seleccionados permite identificar patrones y tendencias claras en el uso del alineamiento de secuencias para la identificación de genes. En primer lugar, los dos enfoques principales que predominan en la literatura son el alineamiento por

pares, utilizado cuando se comparan dos secuencias específicas, y el alineamiento múltiple, empleado cuando se busca establecer relaciones entre tres o más secuencias de manera simultánea. Ambos enfoques son complementarios y su elección depende del objetivo específico del análisis y del volumen de datos disponibles.

En segundo lugar, los Modelos Ocultos de Markov (HMM) emergen como el modelo más recurrente en los trabajos analizados, siendo HMMER la herramienta que los implementa con mayor frecuencia. Esta prevalencia se explica por la capacidad de los HMM para capturar patrones estadísticos conservados en familias de secuencias, lo que los hace útiles en contextos donde la variabilidad entre secuencias es alta, como ocurre con los genes V en diferentes especies de vertebrados.

En tercer lugar, ninguno de los trabajos analizados presenta un conjunto de datos estándar o base de datos específica para la evaluación del alineamiento, lo que evidencia una falta de benchmarks comunes en el área. Sin embargo, sí se identifican métricas de evaluación consistentes, siendo la suma de pares (SOP) y la penalización por brecha las más utilizadas para medir la calidad de los alineamientos obtenidos.

La relevancia de este núcleo temático para la presente investigación es directa: el método RIVAM toma como entrada los resultados de cinco herramientas de alineamiento (BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y Exonerate:est2genome), cada una de las cuales implementa alguno de los enfoques identificados en esta revisión. La comprensión de los fundamentos y limitaciones de estas herramientas, documentada en la literatura analizada, fundamenta las decisiones de diseño del pipeline de unificación e integración de datos del método RIVAM.

4.2.2 Trabajos sobre la identificación de genes

Este núcleo temático tiene como objetivo describir trabajos que realicen análisis, investigaciones y propuestas sobre la identificación y anotación de genes, con especial énfasis en los genes de los segmentos V(D)J. La revisión no se limitó exclusivamente a trabajos sobre genes V, sino que buscó identificar sistemas, arquitecturas, métodos, algoritmos y técnicas orientadas a la sistematización del proceso de anotación para cualquier tipo de gen, con el fin de extraer principios y enfoques aplicables al problema abordado en esta investigación.

4.2.2.1 TABLA COMPARATIVA DE TRABAJOS SOBRE EL PROCESO DE ANOTACIÓN DEL GENOMA E IDENTIFICACIÓN DE GENES

La Tabla 4.2 presenta los 17 trabajos seleccionados, organizados en tres aspectos principales: una descripción del trabajo, los métodos y algoritmos utilizados, y la referencia bibliográfica correspondiente.

Tabla 4.2 comparativa de trabajos sobre anotación del genoma e identificación de genes

N.	Descripción	Métodos	Referencia
1	GeMoMa es una herramienta de anotación basada en homología que usa archivos GFF y FASTA.	Disponible en entorno web y local.	(Keilwagen et al., 2019)
2	Geneid permite la predicción de genes basada en homología utilizando archivos GFF y FASTA.	Ejecución desde terminal.	(Alioto et al., 2018)
3	Mgcod identifica regiones específicas dentro de genes mediante enfoques basados en homología.	Pipeline estructurado.	(Pfennig et al., 2022)
4	GOODORFS detecta marcos de lectura abiertos usando métodos <i>ab initio</i> con Python y scikit-learn.	Algoritmo K-Means.	(McNair et al., 2021)
5	TOGA transfiere anotaciones entre genomas combinando homología y aprendizaje automático.	Alineamiento de secuencias + ML.	(Kirilenko et al., 2022)
6	IgDetective predice genes en loci de inmunoglobulinas con variantes específicas del método.	Basado en grafos y homología.	(Sirupurapu et al., 2022)
7	VgeneFinder busca genes V en genomas completos usando secuencias de referencia.	Clasificador Random Forest.	(Olivieri & Gambón-Deza, 2019)
8	DeepGSR detecta señales y patrones basados en aprendizaje profundo.	Red neuronal convolucional (CNN).	(Kalkatawi et al., 2019)
9	Helixer realiza la anotación del genoma combinando aprendizaje profundo y GeenuFF.	BLSTM con TensorFlow/Keras.	(Stiehler et al., 2020)
10	Vgas permite la identificación y anotación de genes en genomas completos.	Alineamiento con BLASTp.	(K.-Y. Zhang et al., 2019)
11	Modelo para clasificar secuencias con múltiples enfoques basados en LSTM.	Redes LSTM (incluyendo bidireccionales).	(Choudhary, 2017)
12	DeepAnnotator analiza genomas completos usando modelos de aprendizaje profundo.	LSTM de doble capa.	(Kalkatawi et al., 2019)
13	Método para identificar patrones y enzimas mediante homología y teoría de redes.	Uso de HMMER en entorno web.	(Sinha et al., 2020)
14	RAPID es una plataforma web para análisis de anticuerpos y anotación de clonotipos.	Herramientas de análisis integradas.	(Y. Zhang et al., 2021)
15	Estudio comparativo de herramientas para análisis de receptores T y B.	Integración con VDJtools y otras herramientas.	(Arunkumar & Zielinski, 2021)
16	Yclon identifica secuencias de anticuerpos similares mediante homología.	Clustering jerárquico con densidad.	(Gervásio et al., 2022)
17	Comparación de métodos para clasificar retrotransposones usando homología y ML.	SVM, Random Forest y HMM.	(Orozco-Arias et al., 2019)

4.2.2.2 CONCLUSIONES DE LOS TRABAJOS SOBRE IDENTIFICACIÓN Y FUNCIONALIDAD EN GENES

El análisis de los 17 trabajos seleccionados permite identificar tendencias, fortalezas y áreas de oportunidad en el estado del arte de la identificación y anotación de genes, con implicaciones directas para el diseño del método RIVAM.

En primer lugar, los trabajos analizados se agrupan en dos grandes enfoques metodológicos: los métodos basados en homología y los métodos *ab initio*. Los métodos basados en homología, como GeMoMa, Geneid, TOGA, IgDetective y VgeneFinder, comparan secuencias del genoma analizado contra secuencias de referencia previamente caracterizadas, identificando regiones con similitud estructural o funcional. Estos métodos son efectivos cuando se dispone de buenas referencias para la especie analizada, pero su desempeño disminuye en especies poco estudiadas o con alta variabilidad genética respecto a las referencias disponibles. Los métodos *ab initio*, como GOODORFS y los basados en aprendizaje profundo, identifican genes a partir de patrones estadísticos aprendidos de los datos, sin depender directamente de secuencias de referencia. Si bien estos métodos son más generalizables, requieren grandes volúmenes de datos de entrenamiento y son más difíciles de interpretar desde una perspectiva biológica.

En segundo lugar, el aprendizaje automático y el aprendizaje profundo emergen como tendencias crecientes en el campo de la anotación del genoma. Modelos basados en LSTM, redes neuronales convolucionales y BLSTM han demostrado capacidad para capturar patrones complejos en secuencias, superando en algunos casos a los métodos tradicionales de homología. Sin embargo, su aplicación a genes del segmento V es aún limitada, siendo IgDetective y VgeneFinder los únicos trabajos identificados que abordan específicamente este tipo de genes.

En tercer lugar, el uso de pipelines estructurados para integrar los resultados de múltiples herramientas es una práctica recurrente en los trabajos analizados. Sistemas como TOGA, Mgcod y el trabajo de Sinha et al., (2020) demuestran que la combinación de múltiples herramientas de alineamiento con modelos de inteligencia artificial permite obtener resultados más robustos que el uso aislado de cualquiera de estas técnicas. Esta tendencia valida el enfoque del método RIVAM, que integra los resultados de cinco herramientas de alineamiento en un pipeline unificado antes de aplicar las técnicas de reducción.

Finalmente, una constante identificada en la literatura es la ausencia de una metodología explícita, una arquitectura de software documentada o un modelado conceptual del proceso de anotación que explique el tratamiento de los datos y el

razonamiento utilizado para tomar decisiones durante la identificación de genes. La mayoría de los trabajos se centran en reportar los resultados sin describir el proceso que los genera, lo que dificulta su replicación y transferencia a otros contextos. Esta brecha metodológica es precisamente la que el método RIVAM busca cubrir, proponiendo un proceso estructurado, documentado y transferible que sistematiza el conocimiento experto en reglas y criterios formalizados.

4.3 Brechas identificadas en la literatura

El análisis de los trabajos presentados en las secciones anteriores permite identificar un conjunto de brechas metodológicas y científicas que justifican la necesidad del método RIVAM. Estas brechas no son limitaciones aisladas de trabajos individuales, sino patrones recurrentes que se observan de manera consistente a lo largo de la literatura revisada. Su identificación es fundamental para posicionar esta investigación dentro del estado del arte y demostrar que el método RIVAM representa una contribución original y relevante al campo de la bioinformática computacional.

4.3.1 Ausencia de métodos específicos para la reducción del espacio de búsqueda en genes V

Si bien existen numerosos sistemas para la identificación y anotación de genes, ninguno de los trabajos analizados aborda de manera explícita y sistemática la reducción del espacio de búsqueda generado por herramientas de alineamiento como etapa previa a la revisión experta.

Los sistemas existentes, como GeMoMa, TOGA o IgDetective, producen sus propios conjuntos de candidatos a partir del genoma, pero no ofrecen mecanismos para reducir o priorizar los resultados provenientes de múltiples herramientas de alineamiento dentro de un pipeline previamente establecido. Esta distinción es clave: RIVAM no se centra en identificar genes desde cero, sino en reducir el conjunto de candidatos previamente generado, una etapa que no ha sido tratada de forma sistemática en la literatura.

Esta ausencia es crítica en contextos donde el volumen de registros generado supera los cientos de miles, como ocurre en los genomas analizados en esta investigación, donde la proporción entre registros totales y genes válidos puede superar los 1,900 registros por cada gen real. En estos escenarios, la falta de un método de reducción obliga a los expertos a realizar una revisión manual exhaustiva de miles de candidatos, con el consecuente costo en tiempo, recursos y calidad de los resultados.

La reducción del espacio de búsqueda como estrategia metodológica ha sido abordada en otros dominios de la ciencia de datos, como la optimización multiobjetivo,

la minería de datos y el diseño de redes, donde existen métodos bien establecidos para acotar el espacio de soluciones antes de aplicar algoritmos de búsqueda o clasificación. Sin embargo, la transferencia de estas estrategias al dominio específico de la identificación de genes V en genomas de vertebrados no ha sido explorada de manera sistemática en la literatura, lo que representa una oportunidad de contribución original que esta investigación busca aprovechar.

4.3.2 Falta de métodos adaptables a especies poco estudiadas

La mayoría de los sistemas identificados en la revisión fueron diseñados y validados principalmente con datos de especies ya estudiadas, como el ratón o el humano, para las cuales existe una gran cantidad de secuencias de referencia de alta calidad disponibles en bases de datos públicas como IMGT o NCBI. Esta dependencia de referencias de alta calidad y abundancia es una limitación estructural de los métodos basados en homología, debido a que su desempeño está directamente condicionado por la disponibilidad y representatividad de las secuencias de referencia utilizadas durante el proceso de alineamiento.

Su aplicación a especies con menor representación en las bases de datos de referencia, como los quirópteros u otros vertebrados no convencionales, revela limitaciones importantes en términos de cobertura y precisión. En estos casos, la variabilidad genética específica de la especie analizada puede diferir significativamente de los patrones capturados en las referencias disponibles, lo que resulta en un mayor número de falsos positivos y falsos negativos durante el proceso de identificación. Esta situación es problemática en estudios de evolución inmunológica y diversidad del genoma, donde precisamente las especies menos estudiadas son las que ofrecen mayor interés científico.

Adicionalmente, la ausencia de métodos adaptables limita la escalabilidad del proceso de caracterización a medida que se incrementa el número de especies analizadas. En un contexto donde las tecnologías de secuenciación han reducido drásticamente el costo de obtener genomas completos, la comunidad científica enfrenta un creciente acervo de genomas de especies poco estudiadas que requieren caracterización. Contar con métodos que no dependan de referencias específicas de especie es, por tanto, una necesidad metodológica de primer orden.

El método RIVAM aborda esta brecha al basar su etapa de reducción en las características estructurales y estadísticas de los propios registros de alineamiento, en lugar de depender de referencias específicas de especie. Esta característica permite que

RIVAM tenga una mayor adaptabilidad a distintos contextos de la caracterización del genoma y lo hace adecuado para el análisis de especies con variabilidad genética particular o con escasa representación en las bases de datos de referencia.

4.3.3 Ausencia de sistematización del conocimiento experto

Una constante identificada en la literatura es la falta de documentación explícita del proceso de toma de decisiones durante la identificación y anotación de genes. Los trabajos revisados reportan resultados cuantitativos, como métricas de precisión y cobertura, pero omiten describir con detalle el razonamiento, las reglas biológicas y los criterios de selección que guían el proceso de anotación, lo que dificulta significativamente su replicación y transferencia a otros contextos o instituciones. Esta falta de descripción metodológica es una limitación recurrente que afecta tanto a los métodos basados en homología como a los enfoques de aprendizaje automático.

En los métodos basados en homología, las decisiones de filtrado y selección de candidatos suelen depender de umbrales y parámetros fijados de manera empírica, sin una justificación explícita de su origen o de las condiciones bajo las cuales son válidos. En los métodos de aprendizaje automático, la naturaleza de caja negra de los modelos más complejos, como las redes neuronales profundas, hace aún más difícil comprender y documentar el proceso de toma de decisiones que subyace a las predicciones generadas.

Esta brecha es particularmente relevante en el contexto del CISEI, donde el conocimiento experto acumulado durante años de trabajo en la caracterización de genes V en vertebrados no ha sido formalizado en un sistema computacional reproducible. Las reglas de identificación descritas en el Capítulo 3, que incluyen longitudes específicas, patrones de nucleótidos y criterios de funcionalidad para cada sección del gen V, representan un conocimiento especializado que actualmente reside principalmente en la experiencia individual de los expertos y no está disponible en forma de criterios computacionalmente aplicables.

La ausencia de esta sistematización tiene consecuencias prácticas importantes. Obliga a que cada nuevo análisis dependa de la disponibilidad y experiencia de los mismos expertos, limitando la escalabilidad del proceso. Dificulta la incorporación de nuevos investigadores al proceso de anotación, debido a que el conocimiento necesario no está documentado de manera accesible. E impide la automatización de las etapas más rutinarias del proceso, que podrían ser ejecutadas de manera consistente y eficiente por un sistema computacional si contaran con criterios explícitos y formalizados.

El método RIVAM contribuye a cerrar esta brecha al proponer un pipeline documentado y transferible que traduce el conocimiento experto en criterios formalizados y aplicables de manera automática y consistente. Al sistematizar las decisiones de filtrado y selección en reglas computacionales explícitas, RIVAM no solo mejora la eficiencia del proceso, sino que también contribuye a preservar y transferir el conocimiento experto acumulado, facilitando su aplicación en otros laboratorios y contextos de investigación.

En conjunto, las tres brechas identificadas evidencian que existe un espacio de contribución original y relevante para el método RIVAM dentro del estado del arte actual. La combinación de reducción sistemática del espacio de búsqueda, adaptabilidad a distintas especies y sistematización del conocimiento experto representa una propuesta que no tiene precedente directo en la literatura revisada, posicionando a RIVAM como una contribución novedosa y necesaria tanto para el campo de la bioinformática computacional como para las ciencias computacionales, en su aplicación a la caracterización de genes V en vertebrados.

Capítulo 5. Fundamentos conceptuales del método RIVAM

Este capítulo presenta la estructura conceptual del método RIVAM, describiendo los principios de diseño que guiaron su desarrollo, el razonamiento metodológico que justifica las decisiones tomadas y los componentes abstractos que lo conforman. En el marco de la metodología de la Ciencia del Diseño, este capítulo corresponde al inicio del Ciclo de Diseño, donde la propuesta metodológica pasa de una concepción teórica a una estructura formal que sirve de base para su implementación. A diferencia del capítulo siguiente, que detalla la implementación técnica del método, este capítulo se enfoca en el qué y el porqué del diseño, estableciendo el marco conceptual que da coherencia a todas las decisiones de implementación presentadas posteriormente.

5.1 Principios de diseño

El diseño del método RIVAM no surgió de manera arbitraria, sino a partir de un conjunto de principios que orientaron cada decisión metodológica tomada durante su desarrollo. Definir estos principios antes de comenzar la implementación fue una decisión consciente y deliberada, alineada con el enfoque de la Ciencia del Diseño que fundamenta esta investigación: antes de construir un artefacto, es necesario comprender profundamente el problema, identificar las restricciones del dominio y establecer los criterios que determinarán si la solución es válida y útil.

Estos principios no son abstracciones teóricas desconectadas de la práctica, sino que emergen directamente del análisis del proceso real de anotación descrito en el Capítulo 3 y de la revisión de la literatura presentada en el Capítulo 4. Cada uno de ellos responde a una necesidad concreta identificada en el dominio: la redundancia de los datos de alineamiento, la importancia de preservar la coherencia biológica durante el filtrado, la necesidad de que el método pueda aplicarse a distintas especies y genomas, y la importancia de que sus resultados sean compatibles con las herramientas que son utilizadas por los expertos.

En conjunto, estos principios actúan como restricciones y guías que garantizan que el método no solo sea técnicamente funcional, sino también biológicamente coherente, reutilizable en distintos contextos y compatible con los flujos de trabajo existentes en el dominio de la caracterización del genoma. Son, en última instancia, la expresión formal de lo que significa una buena solución al problema planteado en esta investigación. A continuación, se describen los cinco principios fundamentales que sustentan el diseño del método.

5.1.1 Base teórica de otros métodos de ciencia de datos

El diseño del método RIVAM toma como referencia metodológica los principios del proceso estándar entre industrias para la minería de datos, conocido como CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM es un marco metodológico adoptado en proyectos de ciencia de datos que estructura el proceso en seis fases iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (Chapman et al., 2000). La Figura 5.1 presenta este marco y la relación entre sus fases.

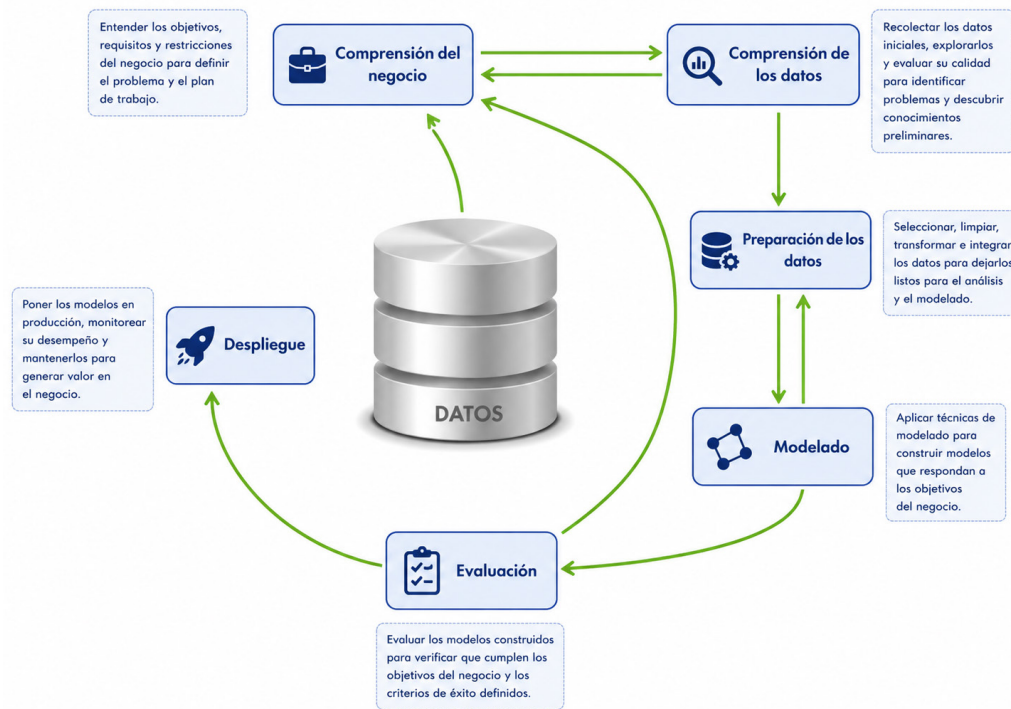


Figura 5.1 Diagrama del proceso CRISP-DM, adaptado de (Chapman et al., 2000)

Si bien CRISP-DM fue concebido para proyectos de minería de datos en contextos empresariales y no específicamente para la reducción del espacio de búsqueda, su relevancia para esta investigación radica en que proporciona un marco ordenado y flexible para abordar problemas con datos complejos, heterogéneos y provenientes de múltiples fuentes, características que definen precisamente el espacio de búsqueda de genes V en genomas de vertebrados.

No obstante, la aplicación directa de CRISP-DM al dominio de la caracterización del genoma presenta limitaciones importantes. El marco original fue diseñado para contextos donde el objetivo es la toma de decisiones basada en datos, sin contemplar las restricciones biológicas que deben respetarse durante el procesamiento. En el caso de los genes V, el filtrado y la reducción de datos no pueden realizarse únicamente con criterios estadísticos, sino que deben preservar la coherencia biológica de los candidatos seleccionados, lo que exige una adaptación del marco original al dominio específico de esta investigación.

De manera complementaria, el método RIVAM también toma como referencia la Metodología Fundamental para la Ciencia de Datos (MFCD), propuesta por Rollins (2015). La MFCD es un marco de diez fases orientado a la resolución de problemas basados en datos, que cubre desde la comprensión del problema hasta la retroalimentación y mejora continua del modelo. La Figura 5.2 muestra las fases que

componen este marco. Su aportación principal al diseño de RIVAM se concentra en las etapas de preparación de datos, modelado, evaluación y validación, proporcionando una guía estructurada para el tratamiento sistemático de los datos de alineamiento y la evaluación objetiva de los resultados.

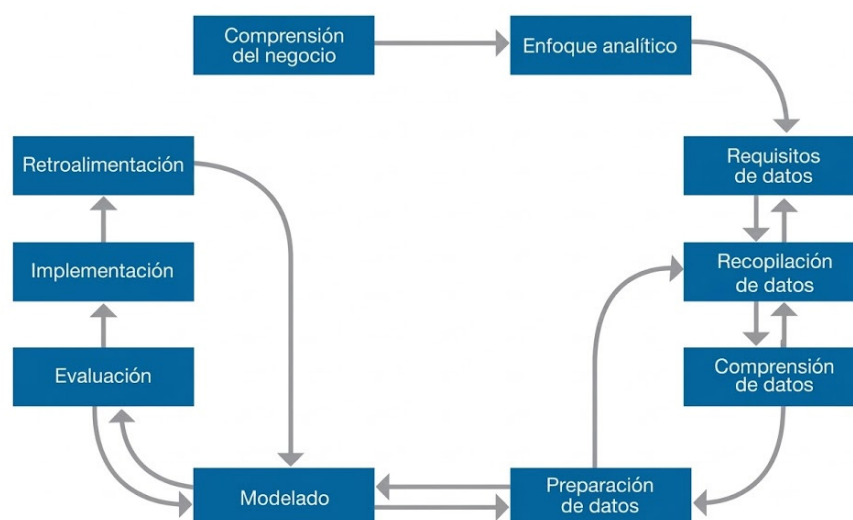


Figura 5.2 Diagrama de la Metodología Fundamental para la Ciencia de Datos, obtenida de (Rollins, 2015)

La combinación de CRISP-DM y MFCD como bases metodológicas permite que RIVAM cuente con un fundamento sólido en ciencia de datos, adaptado a las particularidades del dominio biológico. Mientras CRISP-DM aporta la estructura iterativa y el énfasis en la comprensión del dominio, MFCD complementa con una visión más detallada del tratamiento y evaluación de los datos. Por esta razón, el método RIVAM adopta los principios estructurales de ambos marcos, adaptándolos al contexto específico de la identificación de genes V. Esta adaptación incorpora restricciones biológicas en cada etapa del proceso, integra el conocimiento experto como criterio de validación y establece métricas de evaluación que consideran tanto la eficiencia computacional como la validez biológica de los resultados.

5.1.2 Reducción de redundancia

Uno de los problemas más significativos identificados en el análisis del proceso de anotación de genes V es la presencia de registros redundantes en los archivos GFF generados por las cinco herramientas de alineamiento. Dado que cada herramienta analiza el genoma desde una perspectiva distinta, es común que múltiples herramientas identifiquen la misma región del genoma como candidata, generando registros que representan la misma ubicación con pequeñas variaciones en sus coordenadas o puntuaciones. Esta redundancia incrementa artificialmente el volumen del espacio de búsqueda sin aportar información adicional relevante para la identificación de genes V.

La redundancia en datos de alineamiento no es un fenómeno trivial. En la práctica, un mismo gen V puede aparecer representado decenas de veces en el conjunto de datos, con variaciones mínimas en sus posiciones de inicio y fin que no corresponden a genes distintos sino a imprecisiones propias del proceso de alineamiento. Por ejemplo, la posición de inicio de un gen puede aparecer repetida múltiples veces dentro del mismo grupo con pequeñas diferencias en los valores asociados, como se muestra en la Tabla 5.1, donde registros con idéntico punto de inicio presentan variaciones menores en sus posiciones de fin. Si no se consolidan estos registros, el mismo gen puede ser contabilizado varias veces durante la evaluación, sesgando los resultados y dificultando la comparación con las anotaciones de referencia.

Tabla 5.1 Ejemplo de registros redundantes de dos genes V

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
247913721	247913760	247914148	247913759	247914065	247914189
247913721	247913762	247914142	247913759	247914070	247914183
247913721	247913762	247914143	247913759	247914067	247914190

El principio de reducción de redundancia establece que el método debe ser capaz de identificar y consolidar estos registros similares en representaciones únicas y más compactas, sin perder la información estructuralmente relevante que contienen. Esto implica que la reducción no debe realizarse de manera arbitraria, descartando registros al azar, sino mediante criterios biológicos y estadísticos que permitan determinar cuáles registros aportan información única y cuáles son duplicados o variaciones menores de otros ya presentes en el conjunto. Para ello, el método propone calcular el segundo cuartil de cada columna dentro del grupo de registros redundantes y utilizar ese valor como punto de referencia para seleccionar el registro más representativo, garantizando que el valor elegido mantenga coherencia con los datos reales y con la interpretación biológica.

En términos prácticos, este principio se materializa en la etapa de unificación de datos del pipeline de RIVAM. Como se evidenció en los genomas analizados, la aplicación de este principio permite reducciones superiores al 90% del volumen original de registros. Un ejemplo de este principio es el genoma de *Eptesicus nilssonii*, donde los 321,139 registros originales se reducen a 16,477 registros tras la unificación, lo que representa una reducción del 94.87% sin comprometer la cobertura del proceso de identificación. Este resultado demuestra que la redundancia en los datos de alineamiento es un problema de gran magnitud que, al ser abordado de manera sistemática, genera beneficios significativos en eficiencia y calidad de los resultados.

La reducción de redundancia no solo mejora la eficiencia computacional del proceso, sino que también facilita la aplicación de las etapas posteriores de agrupamiento y reducción de dimensionalidad, que operan de manera más estable y precisa sobre conjuntos de datos menos ruidosos y más representativos. El resultado empírico presentado, una reducción del 94.87% en el genoma de *Eptesicus nilssonii* sin comprometer la cobertura, demuestra que este principio no es un criterio teórico abstracto sino una característica verificable del método. Adicionalmente, dado que el proceso de consolidación opera sobre los archivos GFF estándar sin depender de secuencias de referencia específicas de especie, el principio garantiza que el método sea aplicable de manera consistente a cualquier genoma de vertebrado que utilice las mismas herramientas de alineamiento como entrada.

5.1.3 Preservación de coherencia biológica

La reducción del espacio de búsqueda en el contexto de la identificación de genes V no puede abordarse exclusivamente desde una perspectiva estadística o computacional. A diferencia de otros problemas de reducción de datos donde cualquier criterio de filtrado es válido siempre que mejore la eficiencia, en el dominio de la caracterización del genoma existe un conjunto de restricciones biológicas que deben respetarse durante todo el proceso. Estas restricciones no son opcionales: representan el conocimiento acumulado sobre la estructura del gen V y su violación puede resultar en la eliminación de candidatos biológicamente relevantes o en la retención de registros que no corresponden a genes reales.

El principio de preservación de coherencia biológica establece que cada decisión de filtrado, consolidación o reducción debe estar fundamentada en criterios consistentes con la biología del gen V. En el caso del método RIVAM, esto no implica analizar la secuencia interna del gen, debido a que el método opera sobre las posiciones de inicio y fin registradas en los archivos GFF, no sobre la secuencia de nucleótidos o aminoácidos. Lo que el método sí valida son las relaciones estructurales entre las tres secciones del gen: la presencia del péptido señal (SP), el exón V y la señal de recombinación (RSS), así como las distancias entre sus posiciones de inicio y fin.

La Figura 5.3 muestra esta representación estructural del gen V tal como es procesada por RIVAM. Como se puede observar, cada gen V se representa como un conjunto de seis atributos posicionales: Inicio_SP, Fin_SP, Inicio_Exon, Fin_Exon, Inicio_RSS y Fin_RSS. La coherencia biológica se evalúa a partir de dos criterios principales. Primero, la distancia total entre Inicio_SP y Fin_RSS debe oscilar entre aproximadamente 385 y 460 posiciones o caracteres, rango que corresponde a la longitud esperada de un gen V completo con sus tres secciones. Segundo, las distancias entre secciones adyacentes (d1 entre SP y Exón, y d2 entre Exón y RSS) deben ser

biológicamente plausibles, es decir, no deben exceder rangos que resultarían imposibles dado el tamaño conocido de cada sección del gen.

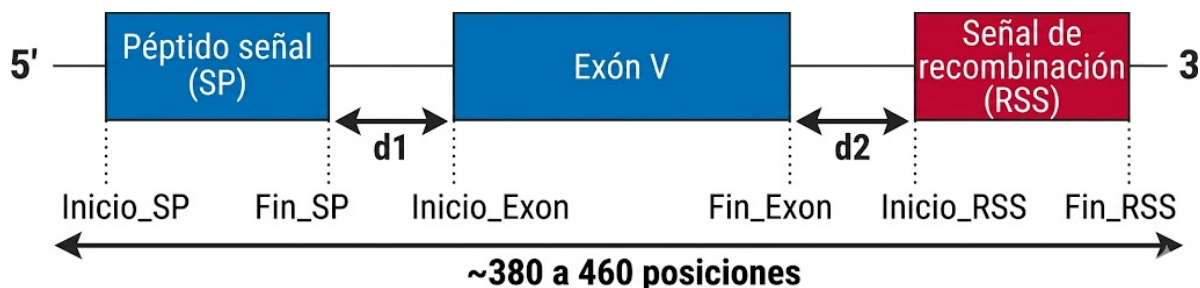


Figura 5.3 Representación del gen V y distancias validadas por RIVAM

Un registro que no cumpla con estas condiciones ya sea porque le falta alguna de las tres secciones o porque las distancias entre ellas son incoherentes, es clasificado como candidato no válido y excluido del conjunto reducido. Esta validación estructural garantiza que los candidatos conservados tras la reducción sean biológicamente plausibles, independientemente de si corresponden a genes funcionales o pseudogenes, debido a que ambos tipos pueden estar presentes con sus tres secciones estructurales completas, pero diferir en características internas de la secuencia que el método no evalúa.

Este principio establece una diferencia fundamental entre RIVAM y otros enfoques de reducción de datos basados exclusivamente en criterios estadísticos. RIVAM no busca únicamente disminuir el volumen de datos, sino hacerlo preservando candidatos que mantengan coherencia estructural con la biología del gen V. En este contexto, la preservación de la coherencia biológica actúa como un mecanismo de control de calidad, asegurando que la reducción de datos no comprometa la validez biológica de los resultados. De este modo, el conjunto reducido que se entrega al experto conserva candidatos verdaderamente relevantes y útiles para el proceso de anotación.

5.1.4 Reusabilidad

Uno de los objetivos centrales del diseño de RIVAM es que el método no sea una solución de uso único, desarrollada exclusivamente para un genoma o una especie específica, sino que pueda aplicarse de manera consistente a distintos contextos del genoma sin requerir modificaciones estructurales en su pipeline. Este principio de reusabilidad responde directamente a una de las brechas identificadas en la literatura: la mayoría de los métodos existentes fueron diseñados y validados para especies ampliamente estudiadas, lo que limita su aplicabilidad a nuevos organismos o genomas poco caracterizados.

En términos prácticos, la reusabilidad de RIVAM se sustenta en tres características de diseño. La primera es la independencia de referencias específicas de especie: dado que el método opera sobre las posiciones de inicio y fin registradas en los archivos GFF, no requiere secuencias de referencia particulares para cada especie analizada. Los mismos criterios estructurales y estadísticos que guían la reducción en un genoma de *Desmodus rotundus* son aplicables a cualquier otro genoma de vertebrado que utilice las mismas herramientas de alineamiento como entrada.

La segunda característica es la estandarización del formato de entrada y salida. Al tomar como entrada archivos GFF obtenidos por herramientas estándar y entregar como salida un archivo GFF reducido compatible con los mismos visores y herramientas utilizados en el proceso actual, RIVAM puede integrarse en cualquier pipeline para la caracterización del genoma que utilice estos formatos, independientemente de la especie o institución que lo aplique.

La tercera característica es la parametrización del método. Los hiperparámetros del modelo, optimizados mediante PSO durante el desarrollo, pueden ser ajustados para adaptarse a las características específicas de nuevos genomas sin necesidad de rediseñar el pipeline completo. Esta flexibilidad garantiza que el método pueda generalizarse a nuevos contextos manteniendo su estructura conceptual intacta.

En conjunto, estas tres características posicionan a RIVAM como un método reutilizable y transferible, cuya aplicación no está limitada al contexto específico del CISEI, sino que puede extenderse a cualquier laboratorio o institución que enfrente el mismo problema de reducción del espacio de búsqueda en la identificación de genes V.

5.1.5 Interoperabilidad

El principio de interoperabilidad establece que el método RIVAM debe ser capaz de integrarse de manera natural con las herramientas y flujos de trabajo existentes en el proceso de caracterización del genoma, sin requerir que los expertos abandonen sus prácticas actuales ni adopten nuevas plataformas o formatos. Este principio surge directamente del análisis del proceso real descrito en el Capítulo 3: cualquier intervención que interrumpa el flujo de trabajo establecido por los expertos del CISEI reduciría su utilidad práctica, independientemente de su eficiencia computacional.

En el diseño de RIVAM, la interoperabilidad se materializa en dos niveles. El primero es la interoperabilidad de datos, que se refiere a la capacidad del método de consumir y producir datos en formatos estándar reconocidos por las herramientas bioinformáticas más utilizadas. El método toma como entrada los archivos GFF obtenidos por BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y

Exonerate:est2genome, que son exactamente los mismos archivos que los expertos utilizan en el proceso actual. Su salida es igualmente un archivo GFF reducido que puede cargarse directamente en IGV sin ninguna conversión adicional, respetando así el flujo de trabajo del experto de la Sección 2 del proceso BPMN.

El segundo nivel es la interoperabilidad de proceso, que se refiere a la capacidad del método de insertarse en un punto específico del pipeline existente sin alterar las etapas anteriores ni posteriores. Como se describió en la sección 3.4.1, RIVAM se integra entre la Sección 1 y la Sección 2 del proceso representado en el BPMN, es decir, después de que los archivos GFF han sido entregados y antes de que el experto inicie la revisión manual. Esta ubicación garantiza que el método no interfiere con el proceso de alineamiento ni con el proceso de anotación final, actuando exclusivamente como una etapa de filtrado y reducción que mejora la eficiencia sin modificar la naturaleza del proceso.

La interoperabilidad no solo facilita la adopción del método por parte de los expertos, sino que también garantiza su compatibilidad con futuros desarrollos en el área. Al basarse en formatos estándar ampliamente adoptados en la comunidad bioinformática, RIVAM puede integrarse con nuevas herramientas de alineamiento o visualizadores que emerjan en el futuro, sin necesidad de modificar su diseño fundamental.

5.2 Justificación de las decisiones de diseño

Los principios descritos en la sección anterior establecen qué debe cumplir el método RIVAM, pero no explican por qué se eligieron las técnicas específicas que lo conforman. Esta sección responde precisamente a esa pregunta: ¿por qué reducción de dimensionalidad?, ¿por qué agrupamiento no supervisado con GMM?, ¿por qué explorar múltiples variantes?, ¿por qué PSO para la optimización de parámetros?

Cada una de estas decisiones no fue tomada de manera arbitraria, sino como resultado de un proceso iterativo de análisis, pruebas preliminares y retroalimentación con expertos del dominio. Documentar este razonamiento es fundamental para garantizar la transparencia metodológica del método y para facilitar su comprensión, replicación y eventual mejora por parte de otros investigadores.

5.2.1 ¿Por qué reducción de dimensionalidad?

El espacio de búsqueda de genes V , una vez consolidado y unificado a partir de las cinco herramientas de alineamiento, se representa como una matriz de datos. Cada registro se describe mediante un conjunto de atributos posicionales: *Inicio_SP*, *Fin_SP*, *Inicio_Exon*, *Fin_Exon*, *Inicio_RSS* y *Fin_RSS*, además de atributos derivados como distancias entre secciones y puntajes de alineamiento. Aunque el número de atributos

no es particularmente alto, la naturaleza de los datos introduce dificultades importantes para aplicar algoritmos de agrupamiento directamente sobre el espacio original.

En primer lugar, las posiciones presentan rangos de valores muy amplios debido al tamaño y organización de cada genoma. Por ejemplo, un gen V ubicado al inicio del genoma puede encontrarse en posiciones del orden de miles, mientras que otro localizado al final puede situarse en posiciones de cientos de millones. Como consecuencia, las distancias euclidianas calculadas en el espacio original no representan adecuadamente la similitud estructural entre registros. Dos genes V con estructuras prácticamente idénticas, pero localizados en regiones diferentes del genoma, aparecerían artificialmente alejados entre sí.

Este problema se relaciona con la denominada maldición de la dimensionalidad, descrita en la literatura de aprendizaje automático (Bellman, 1957). La situación se intensifica cuando los atributos manejan escalas muy distintas, como ocurre entre las posiciones del genoma que son absolutas y las distancias relativas entre secciones del gen.

En segundo lugar, la distribución de los registros en el espacio de atributos es altamente heterogénea. Los genes V reales constituyen únicamente una pequeña fracción del total de registros y suelen concentrarse en regiones específicas del espacio de datos. Estas regiones no son fácilmente separables mediante fronteras lineales simples. Debido a esta heterogeneidad, algoritmos de agrupamiento basados en distancias euclidianas, como K-Means, tienden a generar clusters con baja cohesión interna y alta superposición entre grupos. Esto dificulta la identificación confiable del grupo óptimo y produce variaciones significativas en la cobertura de genes reales entre diferentes ejecuciones (Jain, 2010).

En tercer lugar, las pruebas preliminares realizadas durante el desarrollo del método confirmaron estas limitaciones de manera empírica. Cuando el agrupamiento se aplicó directamente sobre el espacio original, sin realizar transformaciones previas, los resultados demostraron una alta variabilidad en la cobertura de genes reales entre distintos genomas. En algunos casos, el desempeño cambiaba considerablemente dependiendo de las características particulares del genoma analizado.

Esta inestabilidad evidenció la necesidad de transformar previamente el espacio de datos antes de aplicar el agrupamiento. La reducción de dimensionalidad permite abordar simultáneamente los problemas descritos. Al convertir el espacio original en una representación de menor dimensión, se preservan las relaciones estructurales entre registros independientemente de sus posiciones absolutas dentro del genoma.

Como resultado, se obtiene una representación más compacta, homogénea y adecuada para los algoritmos de agrupamiento. Esta transformación no elimina información relevante; por el contrario, reorganiza los datos en un espacio donde las similitudes estructurales pueden detectarse con mayor facilidad. Esto mejora tanto la cohesión interna de los clusters como la separación entre ellos (Jolliffe & Cadima, 2016).

Es importante señalar que la reducción de dimensionalidad cumple un propósito distinto al de la reducción del espacio de búsqueda, que constituye el objetivo central del método. La reducción del espacio de búsqueda limita el conjunto de regiones candidatas que deben ser revisadas por el experto, mientras que la reducción de dimensionalidad modifica la representación interna de esos datos para mejorar el desempeño del agrupamiento.

Ambos procesos son complementarios y necesarios. El primero disminuye el volumen del problema, mientras que el segundo mejora la calidad de la solución obtenida. Aunque esta diferencia puede parecer sutil, resulta fundamental para comprender correctamente el diseño y funcionamiento del método propuesto.

5.2.2 ¿Por qué agrupamiento no supervisado?

Una vez justificada la necesidad de transformar el espacio de datos mediante técnicas de reducción de dimensionalidad, la siguiente decisión de diseño consistió en seleccionar el mecanismo más adecuado para identificar el subconjunto de registros con mayor concentración de genes V reales dentro del espacio transformado. En este contexto, se optó por un enfoque de agrupamiento no supervisado debido a tres razones fundamentales relacionadas con la naturaleza del problema, la estructura de los datos y la interpretabilidad de los resultados.

La primera razón está asociada con la disponibilidad limitada de información etiquetada. En el problema de identificación de genes V no existen etiquetas previas para la totalidad de los registros que conforman el espacio de búsqueda. Aunque se dispone de anotaciones manuales de referencia para algunos genomas utilizados durante el desarrollo del método, dichas anotaciones corresponden únicamente a genes V previamente validados por expertos y representan solo una pequeña fracción del total de candidatos generados.

Bajo estas condiciones, utilizar un enfoque supervisado implicaría entrenar modelos a partir de conjuntos de datos parciales y potencialmente sesgados hacia ciertos genomas específicos. Esto reduciría la capacidad de generalización del método al aplicarlo sobre nuevos genomas sin anotaciones previas o con características estructurales distintas. En contraste, el agrupamiento no supervisado permite

identificar patrones inherentes en los datos sin depender de etiquetas conocidas, favoreciendo así la reutilización y adaptabilidad del método en diferentes contextos.

La segunda razón se relaciona con la propia estructura del espacio de datos. Los registros presentan una distribución altamente heterogénea, caracterizada por grupos de formas irregulares, densidades variables y regiones de superposición entre candidatos relevantes e irrelevantes. Estas propiedades dificultan la aplicación de algoritmos que asuman estructuras simples o grupos aproximadamente esféricos con varianzas homogéneas. La Tabla 5.2 presenta una comparativa teórica de los principales algoritmos de agrupamiento considerados durante el desarrollo del método, incluyendo sus parámetros, escalabilidad, caso de uso y métrica de distancia utilizada.

Tabla 5.2 Comparativa teórica de algoritmos de agrupamiento

Algoritmo	Parámetros principales	Escalabilidad	Caso de uso	Geometría / Métrica
K-Means	Número de grupos (k)	Muy alta respecto a $n_{samples}$	Adecuado para grupos bien definidos, de tamaño similar y geometría aproximadamente esférica. Método inductivo.	Distancia euclidiana
OPTICS	Tamaño mínimo de clúster y parámetros de densidad	Muy alta respecto a $n_{samples}$ y $n_{clusters}$	Útil para geometrías no planas, densidades variables y detección de <i>outliers</i> . Método transductivo.	Distancia euclidiana
Spectral Clustering	Número de grupos (k) y matriz de afinidad	Limitada (≈ 10 mil muestras); alto costo computacional	Adecuado para clústeres no convexos, estructuras complejas y análisis basado en similitud global.	Matrices de similitud
Gaussian Mixture Models (GMM)	Número de componentes y tipo de covarianza	Limitada; poco escalable para conjuntos muy grandes	Permite modelado probabilístico y representación de datos con distribuciones gaussianas. Método inductivo.	Distancia de Mahalanobis
Fuzzy C-Means	Número de grupos (c) y grado de difusidad (m)	Moderada; depende del número de iteraciones	Permite pertenencia parcial a múltiples grupos. Útil en problemas de clustering suave y bioinformática.	Distancia euclidiana, Minkowski o coseno

A partir de las características descritas en la tabla, GMM se eligió como el algoritmo más adecuado para este problema por su capacidad de modelar grupos de geometría no esférica mediante matrices de covarianza flexibles y su naturaleza inductiva que permite generar un modelo reutilizable. A diferencia de métodos como K-Means, GMM modela cada grupo como una distribución gaussiana independiente con su propia media y matriz de covarianza, lo que le permite adaptarse de manera más flexible a distribuciones complejas y estructuras no uniformes (Bishop & Nasrabadi, 2006). La

evaluación experimental detallada de estos algoritmos y la justificación cuantitativa de la elección final se presenta en el Capítulo 6.

La tercera razón corresponde a la interpretabilidad de los resultados del método. El propósito de RIVAM no consiste únicamente en agrupar registros similares, sino en identificar específicamente el grupo con mayor concentración de genes V reales para presentarlo al experto como un conjunto reducido de candidatos prioritarios.

En este sentido, GMM ofrece una ventaja metodológica importante: cada componente del modelo representa una región estadísticamente definida dentro del espacio transformado. Esto permite evaluar objetivamente qué componente concentra la mayor densidad de genes V mediante métricas cuantitativas reproducibles. Como consecuencia, la selección del subconjunto final no depende de criterios heurísticos sino de propiedades estadísticas claramente definidas y justificables.

Es importante destacar que la elección de GMM no fue arbitraria ni inmediata, sino el resultado de un proceso sistemático de evaluación comparativa entre distintos algoritmos presentados en la Tabla 5.2. Los resultados de esta comparación, descritos en detalle en el Capítulo 6, mostraron consistentemente que GMM proporcionaba el mejor equilibrio entre precisión, estabilidad y capacidad de adaptación a la heterogeneidad del espacio de búsqueda, confirmando la pertinencia de su elección como algoritmo central del método RIVAM.

5.2.3 ¿Por qué explorar múltiples variantes?

Una vez establecido el agrupamiento no supervisado con GMM como mecanismo central del método, surgió una pregunta de diseño adicional: ¿es suficiente aplicar GMM directamente sobre los datos transformados, o conviene explorar distintas combinaciones de técnicas de reducción de dimensionalidad previas al agrupamiento? La respuesta a esta pregunta motivó el diseño de múltiples variantes del método, cada una representando una configuración distinta del pipeline de reducción.

La exploración de múltiples variantes responde a una razón fundamental: la incertidumbre sobre cuál técnica de reducción de dimensionalidad produce la representación más favorable para el agrupamiento en el contexto específico de los datos de alineamiento de genes V. Como se describió en el Marco Conceptual, PCA realiza una reducción lineal que preserva la varianza global de los datos, mientras que UMAP realiza una reducción no lineal que preserva tanto la estructura local como la global.

Dado que los datos de alineamiento presentan una estructura compleja y potencialmente no lineal, no es posible determinar a priori cuál de estas técnicas producirá una mejor separabilidad entre el grupo de genes V reales y el resto de los candidatos. La única manera de responder esta pregunta con rigor científico es comparar el desempeño de cada combinación sobre los mismos datos.

Adicionalmente, la exploración de variantes permite evaluar si la optimización de hiperparámetros con PSO aporta mejoras significativas sobre la configuración base del método. Esto da origen a tres variantes concretas: GMM optimizado con PSO, PCA combinado con GMM y UMAP combinado con GMM. Cada variante representa una hipótesis de diseño específica sobre qué combinación de técnicas produce el mejor equilibrio entre reducción del espacio de búsqueda y cobertura de genes reales.

Esta decisión de explorar múltiples variantes está alineada con los principios metodológicos de CRISP-DM y MFCD descritos en la sección 5.1.1, que enfatizan la importancia de la evaluación comparativa como parte del proceso de desarrollo de métodos basados en datos. No se trata de una exploración arbitraria sino de un diseño experimental sistemático orientado a identificar la configuración óptima del método con evidencia empírica. Los resultados de esta comparación son presentados en detalle en los Capítulos 7 y 8, proporcionan la base cuantitativa para seleccionar la variante que mejor cumple con los principios de diseño establecidos en la sección 5.1.

5.2.4 ¿Por qué PSO para optimización?

El desempeño de los algoritmos de agrupamiento y reducción de dimensionalidad depende en gran medida de la configuración de sus hiperparámetros. En el caso de GMM, el número de componentes y el tipo de matriz de covarianza determinan directamente la forma y el tamaño de los grupos identificados. En el caso de UMAP, parámetros como `n_neighbors` y `min_dist` controlan el balance entre la preservación de la estructura local y global de los datos. Una configuración inadecuada de estos parámetros puede resultar en agrupamientos de baja calidad, con grupos que no reflejan la estructura real de los datos o que no corresponden a las regiones del espacio de búsqueda con mayor concentración de genes V reales.

La optimización de hiperparámetros es, por tanto, una etapa necesaria para garantizar que el método opere en su configuración más efectiva para cada genoma analizado. Sin embargo, la búsqueda exhaustiva de la configuración óptima en un espacio de múltiples parámetros es computacionalmente costosa y puede resultar inviable cuando el número de combinaciones posibles es grande. Esta limitación motivó la elección de un método de optimización metaheurística que permita explorar el

espacio de parámetros de manera eficiente sin necesidad de evaluar todas las combinaciones posibles.

La elección de PSO como algoritmo de optimización responde a sus características particulares. A diferencia de métodos de búsqueda exhaustiva como grid search o de métodos basados en gradiente que requieren funciones diferenciables, PSO es un algoritmo bioinspirado que explora el espacio de parámetros de manera colaborativa con un conjunto de partículas que se desplazan hacia las regiones de mayor desempeño. Esta característica lo hace adecuado para problemas de optimización donde la función objetivo no es diferenciable, como ocurre con las métricas de evaluación del método RIVAM, y donde el espacio de búsqueda es multimodal con múltiples óptimos locales (Imron & Prasetyo, 2020; Selvaraj & Choi, 2022).

Es importante destacar que la optimización de parámetros en el contexto de RIVAM no es un objetivo en sí mismo sino un medio para garantizar que el agrupamiento produzca los mejores resultados posibles en términos de cobertura de genes reales y reducción del espacio de búsqueda. La función objetivo que PSO optimiza integra ambos criterios, buscando el balance óptimo entre reducción máxima del espacio de candidatos y preservación de la mayor cantidad posible de genes V reales. Esta función objetivo, definida a partir del conocimiento experto del dominio, es precisamente la que conecta la optimización computacional con el principio de coherencia biológica descrito en la sección 5.1.3. Los detalles de su definición y los resultados de la optimización se presentan en el Capítulo 6.

5.3 Componentes metodológicos

Los principios de diseño descritos en la sección 5.1 y el razonamiento que fundamenta las decisiones técnicas presentadas en la sección 5.2 convergen en una estructura metodológica concreta que organiza el funcionamiento del método RIVAM. Esta estructura se compone de seis componentes secuenciales, cada uno con un rol específico dentro del proceso de reducción del espacio de búsqueda, que en conjunto transforman un conjunto amplio y heterogéneo de registros de alineamiento en un archivo GFF reducido y biológicamente coherente listo para la revisión experta.

La Figura 5.4 presenta el diagrama conceptual del método, donde se muestra la secuencia de componentes, sus entradas y salidas, y el flujo de información a lo largo del proceso. Como se puede observar, la progresión refleja una lógica clara: primero se comprende el dominio y se caracteriza la entrada, luego se integran y estandarizan los datos, posteriormente se reduce la dimensionalidad del espacio de datos para facilitar el modelado y la selección del conjunto óptimo, y finalmente se genera la representación de salida en un formato interoperable. Esta progresión está alineada con los marcos

metodológicos de CRISP-DM y MFCD descritos en la sección 5.1.1, garantizando que cada decisión tomada en un componente esté fundamentada en la información obtenida por las etapas anteriores.

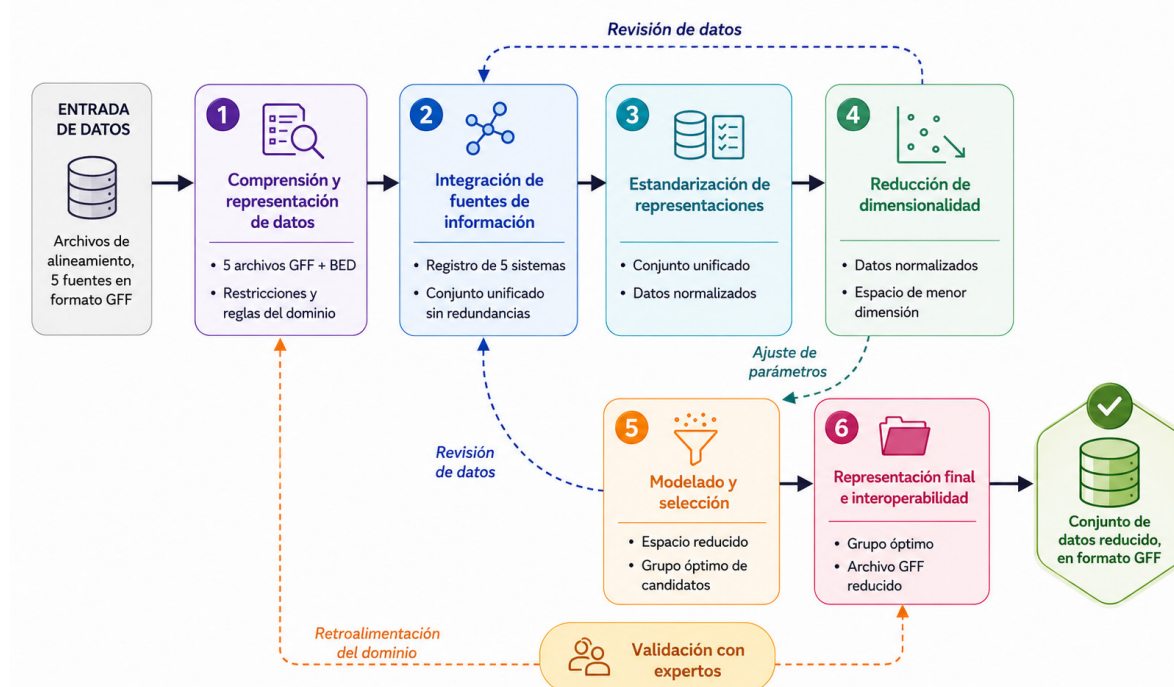


Figura 5.4 Diagrama conceptual de los componentes metodológicos del método RIVAM

A diferencia del Capítulo 6, que detalla la implementación técnica de cada componente con sus algoritmos, parámetros y especificaciones concretas, esta sección presenta una descripción abstracta y conceptual de qué hace cada componente, qué información recibe como entrada y qué entrega como salida. El objetivo es proporcionar una visión integral del método que permita al lector comprender su lógica de funcionamiento antes de entrar en los detalles de su implementación.

5.3.1 Comprensión y representación de datos

El primer componente del método establece el punto de partida del proceso de reducción, caracterizando la naturaleza y estructura de los datos de entrada antes de aplicar cualquier transformación. En cualquier problema de reducción del espacio de búsqueda, esta comprensión inicial es indispensable: sin conocer las características del espacio de entrada, sus dimensiones, su volumen y su estructura, no es posible definir criterios de reducción que sean efectivos y biológicamente válidos. Su función es precisamente consolidar ese conocimiento del dominio y traducirlo en criterios concretos que guíen las decisiones de los componentes posteriores.

La entrada de este componente, en términos generales, corresponde al conjunto completo de datos generados por las fuentes de información disponibles en el dominio del problema, junto con los datos de referencia que permiten evaluar la calidad de los resultados. En el caso particular de los genes V, esta entrada son los archivos obtenidos por los cinco sistemas de alineamiento: BLASTn, BLASTx, hmmscan, Exonerate:protein2genome:local y Exonerate:est2genome, así como los archivos de referencia en formato BED que contienen las anotaciones manuales de genes V validados por expertos. Estos archivos representan el espacio de búsqueda en su estado original, con todo su volumen y redundancia.

En términos generales, durante este componente se caracteriza el espacio de entrada analizando su volumen total, la distribución de los datos entre las distintas fuentes, la proporción entre soluciones relevantes e irrelevantes y las restricciones propias del dominio que deben respetarse durante la reducción. Esta caracterización no modifica los datos, sino que extrae el conocimiento necesario para tomar decisiones informadas en los siguientes componentes. En el caso de los genes V, esto implica analizar la distribución de registros entre los cinco sistemas de alineamiento, identificar la proporción entre candidatos que corresponden a genes reales y registros irrelevantes, y establecer las restricciones biológicas que definen qué constituye un candidato válido.

El resultado de este componente es un conjunto de restricciones y reglas del dominio que actúan como marco de referencia para todo el proceso de reducción. En cualquier problema de reducción del espacio de búsqueda, este tipo de conocimiento es fundamental: sin una comprensión clara de qué constituye una solución válida, cualquier criterio de filtrado puede eliminar información relevante o conservar registros irrelevantes. Las restricciones del dominio son precisamente las que permiten distinguir entre ambos casos y tomar decisiones fundamentadas en el conocimiento del problema.

En el caso de los genes V, estas restricciones se traducen en criterios concretos y verificables: las distancias esperadas entre las secciones del gen, los rangos válidos de longitud total, los criterios de coherencia biológica descritos en el principio 5.1.3 y las métricas de evaluación que determinarán la calidad del conjunto reducido final. Este conocimiento fluye de manera implícita hacia todos los componentes posteriores, garantizando que las decisiones de filtrado y selección sean consistentes con la biología del dominio y con las necesidades reales de los expertos que utilizarán el resultado.

5.3.2 Integración de fuentes de información

El segundo componente aborda uno de los desafíos más comunes en problemas de reducción del espacio de búsqueda: la diversidad de los datos provenientes de múltiples fuentes. En cualquier dominio donde se utilizan diversos sistemas o métodos para explorar un espacio de soluciones, los resultados suelen presentar diferencias en formato, escala y criterios de evaluación que dificultan su análisis en conjunto. La coexistencia de múltiples fuentes de información es, por un lado, una ventaja porque cada fuente aporta una perspectiva distinta sobre el espacio de soluciones, pero, por otro representa un reto metodológico porque los resultados no pueden compararse ni procesarse directamente sin una etapa previa de consolidación.

Antes de aplicar cualquier técnica de reducción, es necesario consolidar estas fuentes en una representación unificada que permita tratar todos los registros bajo los mismos criterios y en el mismo espacio de atributos. Este proceso de integración no consiste simplemente en concatenar los resultados de las distintas fuentes, sino en identificar relaciones entre ellos, resolver conflictos y construir una representación que preserve la información relevante de cada fuente sin duplicarla. Sin esta integración previa, los siguientes componentes operarían sobre datos inconsistentes y redundantes, comprometiendo tanto la calidad del agrupamiento como la validez biológica de los resultados finales.

En el caso particular de los genes V, la integración de fuentes implica consolidar los registros obtenidos por las cinco o más sistemas de alineamiento en un único conjunto estructurado. Cada sistema aporta información sobre distintas secciones del gen V desde perspectivas complementarias: hmmscan identifica señales de recombinación, BLASTn y BLASTx detectan regiones de similitud a nivel de nucleótidos y aminoácidos respectivamente, y las dos modalidades de Exonerate identifican la estructura exónica del gen. La integración de estas cinco perspectivas requiere identificar qué registros de distintas fuentes corresponden a la misma región del genoma y consolidarlos en una representación única.

Esta unificación implica varias operaciones específicas. Primero, se identifican los registros redundantes, es decir, aquellos que representan la misma región del genoma con pequeñas variaciones en sus coordenadas. Segundo, se construye una representación vectorial unificada que capture los atributos posicionales de cada candidato: las posiciones de inicio y fin de cada sección del gen (SP, Exón y RSS), las distancias entre secciones y atributos derivados adicionales. Tercero, se valida que los registros consolidados cumplan con las restricciones biológicas establecidas en el componente anterior, descartando aquellos que presenten incoherencias estructurales que los hagan biológicamente inviables.

La salida de este componente es un conjunto unificado de registros sin redundancias, donde cada entrada representa un candidato único con sus atributos posicionales completos y biológicamente coherentes. Este conjunto constituye la base sobre la cual operan todos los componentes posteriores del método, y su calidad determina directamente la efectividad de las etapas de estandarización, reducción y modelado que le siguen.

5.3.3 Estandarización de representaciones

El tercer componente aborda un problema fundamental en el procesamiento de datos provenientes de múltiples fuentes: la heterogeneidad en las escalas y rangos de los atributos. En cualquier problema donde los datos se describen mediante atributos con unidades o magnitudes distintas, aplicar directamente algoritmos de reducción de dimensionalidad o modelado de aprendizaje automático puede producir resultados sesgados hacia los atributos de mayor magnitud, que dominan las métricas de distancia independientemente de su relevancia real para el problema. La estandarización garantiza que todos los atributos se contribuyan de manera equitativa al proceso de reducción, independientemente de sus valores absolutos.

La entrada de este componente es el conjunto unificado de registros sin redundancias generado por el componente anterior. Cada registro está descrito por un conjunto de atributos posicionales cuyos valores pueden variar en rangos muy distintos: las posiciones absolutas en el genoma pueden alcanzar cientos de millones, mientras que las distancias entre secciones del gen se miden en decenas o cientos de posiciones. Esta disparidad de escalas hace que los atributos de mayor magnitud dominen las métricas de distancia utilizadas por los algoritmos de reducción y modelado, distorsionando la representación del espacio de datos.

En este componente se aplica una transformación de normalización que reescala todos los atributos a un rango común, preservando las relaciones relativas entre los valores de cada atributo sin modificar la estructura del conjunto de datos. Esta transformación es importante en el contexto de los genes V porque las posiciones absolutas en el genoma varían enormemente entre distintos genomas y entre distintas regiones del mismo genoma, mientras que las distancias relativas entre secciones del gen son mucho más consistentes entre especies. Al normalizar los atributos, el componente garantiza que el algoritmo de reducción de dimensionalidad pueda identificar patrones estructurales basados en las relaciones entre secciones del gen, sin verse influenciado por la ubicación absoluta de los candidatos en el genoma.

El resultado de este componente es un conjunto de datos normalizados donde todos los atributos se encuentran en una escala comparable, listos para ser procesados por el

componente de la reducción de dimensionalidad. En cualquier problema de reducción del espacio de búsqueda, la estandarización previa al modelado es una práctica metodológica establecida que mejora la estabilidad y la calidad de los resultados. En el caso particular de los genes V, esta estandarización es relevante porque permite que el método sea aplicable de manera consistente a genomas de distintos tamaños y especies, sin que las diferencias en la escala de las posiciones absolutas afecten la calidad del agrupamiento posterior.

5.3.4 Reducción de dimensionalidad

El cuarto componente transforma la representación de los datos normalizados en un espacio de menor dimensión que preserve las relaciones estructurales entre registros y mejore la separabilidad entre grupos. En cualquier problema de análisis de datos de alta dimensionalidad, la reducción de dimensionalidad es una estrategia ampliamente utilizada para superar las limitaciones que impone la maldición de la dimensionalidad: a medida que el número de atributos crece, las distancias entre puntos tienden a homogeneizarse, dificultando la identificación de patrones y la separación entre grupos mediante algoritmos de modelado convencionales (Bellman, 1957). Al proyectar los datos en un espacio de menor dimensión, permite que el componente de modelado opere sobre una representación más compacta y discriminativa.

La entrada del componente para la reducción de dimensiones es el conjunto de datos normalizados obtenido por el componente anterior. Aunque el número de atributos posicionales que describen cada candidato no es excesivamente grande en términos absolutos, la variabilidad entre genomas y la naturaleza heterogénea de los datos de alineamiento obtiene un espacio de datos cuya estructura no es directamente aprovechable por los algoritmos de agrupamiento sin una transformación previa. En particular, las posiciones absolutas de los candidatos en el genoma introducen una variabilidad de gran escala que dificulta la identificación de patrones estructurales comunes entre genes V de distintas regiones del genoma.

Durante este componente se aplica una técnica de reducción que transforma el espacio original en una representación de menor dimensión, preservando las relaciones de similitud entre registros que son relevantes para el problema. En términos generales, existen dos grandes enfoques para esta transformación: los métodos lineales, que proyectan los datos en un subespacio definido por combinaciones lineales de los atributos originales, y los métodos no lineales, que son capaces de capturar estructuras complejas que no pueden representarse con proyecciones lineales. La elección entre estos enfoques, explicada en la sección 5.2.1, influye en la calidad de la representación obtenida y en la efectividad del modelado posterior.

El resultado de este componente es una representación de los datos en un espacio de menor dimensión donde los registros con características estructurales similares se encuentran más próximos entre sí, independientemente de su ubicación absoluta en el genoma. En el caso particular de los genes V, esto significa que candidatos con distancias similares entre sus secciones estructurales y proporciones similares entre péptido señal, exón y señal de recombinación tenderán a agruparse en regiones contiguas del espacio reducido, facilitando que el componente de modelado y selección identifique con mayor precisión el subconjunto de candidatos con mayor concentración de genes V reales.

5.3.5 Modelado y selección

El quinto componente constituye el núcleo del proceso de reducción del espacio de búsqueda. Su función es identificar, dentro del espacio de menor dimensión obtenido en el componente anterior, el subconjunto de registros con mayor concentración de candidatos biológicamente relevantes, para entregárselo al experto como conjunto reducido y priorizado. En cualquier problema de reducción del espacio de soluciones, este componente responde a la pregunta central del método: ¿cuáles de todos los candidatos disponibles tienen mayor probabilidad de ser soluciones válidas al problema?

La entrada de este componente es la representación de datos en el espacio de menor dimensión obtenida en el componente anterior. Esta representación conserva las relaciones estructurales entre registros de manera que candidatos con características similares se encuentran próximos entre sí en el espacio transformado, lo que facilita la identificación de grupos naturales mediante técnicas de modelado. A diferencia de los componentes anteriores, que operan sobre todos los registros de manera uniforme, este componente introduce una distinción entre grupos de registros basada en sus propiedades estadísticas y estructurales.

Durante este componente se aplica un modelo que identifica los grupos naturales presentes en el espacio de datos transformado, asignando a cada registro una medida de pertenencia a cada grupo. Esta asignación no es binaria sino probabilística, lo que permite representar de manera más realista la incertidumbre inherente a los datos biológicos, donde la frontera entre candidatos relevantes e irrelevantes no siempre es nítida. Una vez identificados los grupos, se evalúa cuál de ellos concentra la mayor proporción de candidatos biológicamente válidos, aplicando los criterios de validación establecidos en el componente de comprensión del dominio. Esta evaluación combina criterios estadísticos, como la cohesión interna del grupo, con criterios biológicos, como la cobertura de genes de referencia validados por expertos.

El resultado de este componente es el grupo óptimo de candidatos, es decir, el subconjunto del espacio de búsqueda con mayor probabilidad de contener los genes V reales del genoma analizado. En cualquier problema de reducción del espacio de soluciones, la identificación de este subconjunto es el objetivo central del método: no se trata de encontrar todas las soluciones posibles sino de acotar el espacio de manera que el esfuerzo de revisión posterior se concentre en las regiones más prometedoras. En el caso particular de los genes V, este grupo óptimo representa una fracción pequeña del espacio de búsqueda original que conserva la gran mayoría de los genes V reales, permitiendo que el experto concentre su revisión en un conjunto manejable de candidatos sin comprometer la cobertura del proceso de anotación.

5.3.6 Representación final e interoperabilidad

El sexto y último componente cierra el proceso de reducción del espacio de búsqueda, traduciendo el grupo óptimo de candidatos identificado en el componente anterior en una representación de salida que pueda ser utilizada directamente por los expertos en su flujo de trabajo habitual. En cualquier método computacional orientado a apoyar la toma de decisiones de expertos, la utilidad práctica del resultado depende no solo de su calidad estadística sino también de su accesibilidad: un resultado técnicamente correcto pero presentado en un formato incompatible con las herramientas del dominio tiene un impacto real limitado. La interoperabilidad de la salida es, por tanto, tan importante como la calidad del proceso que la produce.

La entrada de este componente es el grupo óptimo de candidatos identificado en el componente anterior, representado en el espacio transformado del método. Este grupo contiene los registros seleccionados como los más probables de corresponder a genes V reales, pero en una representación interna del método que no es directamente utilizable por las herramientas bioinformáticas del dominio. La función de este componente es recuperar la información original de estos registros y traducirla a un formato estándar reconocido por el ecosistema de herramientas del proceso de anotación.

Durante este componente se recuperan los registros originales de alineamiento correspondientes a los candidatos del grupo óptimo, preservando toda la información posicional y los atributos originales de cada registro. A continuación, estos registros se organizan y estructuran en el formato de salida estándar del método, asegurando que la información sea completa, coherente y compatible con las herramientas de análisis utilizadas por los expertos. Este proceso de transformación no modifica el contenido biológico de los registros sino únicamente su representación, garantizando que la información entregada al experto sea fiel a los datos originales de alineamiento.

El resultado de este componente es un archivo en formato GFF reducido que contiene únicamente los registros del grupo óptimo, listo para ser cargado directamente en el IGV u otras herramientas utilizadas por los expertos sin ninguna conversión adicional. En cualquier método computacional que busque integrarse en un flujo de trabajo existente, esta compatibilidad con los formatos estándar del dominio es lo que determina su adopción práctica: los expertos pueden incorporar el resultado del método en su proceso habitual sin necesidad de aprender nuevas herramientas ni realizar conversiones adicionales. En el caso particular de los genes V, esto significa que el experto recibe un archivo GFF reducido que puede utilizar directamente en el IGV, exactamente como lo haría con los archivos originales de alineamiento, pero con un volumen de registros drásticamente menor que facilita y acelera el proceso de revisión y anotación.

5.4 Iteración y refinamiento

El método RIVAM no surgió en su forma final a partir de un único ejercicio de diseño, sino como resultado de un proceso iterativo de construcción, evaluación y ajuste que caracteriza al Ciclo de Diseño de la metodología DSR. Esta sección describe cómo el feedback continuo de los expertos en biología molecular del CISEI y los resultados de las pruebas preliminares sobre datos reales influyeron en las decisiones de diseño del método, modificando componentes, criterios y representaciones a lo largo del proceso.

La naturaleza iterativa del diseño no es una limitación sino una fortaleza metodológica. En problemas donde el dominio es complejo y el conocimiento experto es difícil de formalizar de manera completa desde el inicio, la iteración permite incorporar aprendizajes progresivos que enriquecen el método y garantizan que su diseño final sea el resultado de una comprensión profunda y validada del problema. Como lo establece el marco DSR, el artefacto se construye y evalúa en ciclos sucesivos hasta alcanzar un nivel de desempeño que satisfaga los criterios de utilidad definidos en el dominio.

5.4.1 Decisiones iniciales de diseño

En el marco de la Ciencia del Diseño, toda propuesta de artefacto comienza con una comprensión profunda del problema en su contexto real, antes de tomar cualquier decisión técnica. En el caso del método RIVAM, las decisiones iniciales de diseño surgieron directamente del Ciclo de Relevancia: de las interacciones con los expertos en biología molecular del CISEI, del análisis del proceso de anotación descrito en el Capítulo 3 y de la identificación de los cuellos de botella que limitan su eficiencia. Este punto de partida garantizó que las decisiones de diseño no fueran especulativas sino fundamentadas en necesidades reales y verificables del dominio.

Las primeras decisiones conceptuales del método giraron en torno a tres preguntas fundamentales: ¿dónde intervenir en el proceso de anotación?, ¿qué tipo de transformación aplicar a los datos? y ¿cómo evaluar que la intervención es útil para el experto? La respuesta a la primera pregunta, establecida en la sección 3.4.1, determinó que la intervención debía ubicarse entre la generación de archivos GFF y la revisión manual, sin alterar ninguna de las etapas anteriores ni posteriores del proceso. Esta decisión inicial definió la naturaleza del problema: no se trataba de identificar genes V desde cero, sino de reducir el conjunto de candidatos ya generados por los sistemas de alineamiento.

La respuesta a la segunda pregunta orientó el diseño hacia un enfoque de reducción en dos niveles: primero una reducción de redundancia mediante la unificación de datos, y luego una reducción estructural mediante agrupamiento. Esta decisión conceptual, aunque sencilla en apariencia, tuvo implicaciones profundas en el diseño de los componentes posteriores, ya que estableció una separación clara entre la etapa de preparación de datos y la etapa de modelado que se mantiene en la versión final del método. La respuesta a la tercera pregunta definió que el criterio de éxito del método no sería puramente estadístico sino biológicamente informado, lo que derivó directamente en el principio de coherencia biológica descrito en la sección 5.1.3.

5.4.2 Refinamiento a partir del conocimiento del dominio

Una vez establecidas las decisiones iniciales, el diseño del método entró en una fase de refinamiento guiada por el Ciclo de Rigor de la metodología DSR. Este ciclo establece que las decisiones de diseño deben fundamentarse no solo en la comprensión del problema real, sino también en el conocimiento acumulado en la literatura científica y en el dominio específico de aplicación. En el caso de RIVAM, este refinamiento se produjo en dos direcciones complementarias: por un lado, la revisión sistemática de la literatura presentada en el Capítulo 4 aportó fundamentos teóricos que validaron y enriquecieron las decisiones iniciales; por otro, las interacciones continuas con los expertos del CISEI permitieron incorporar restricciones y criterios del dominio que no eran evidentes en la primera conceptualización del método.

Desde la perspectiva de la literatura, el refinamiento más significativo fue la incorporación explícita de técnicas de reducción de dimensionalidad como etapa previa al agrupamiento. Las decisiones iniciales contemplaban únicamente la unificación de datos y el agrupamiento, sin considerar la transformación del espacio de representación. La revisión de trabajos sobre reducción de espacios de búsqueda en bioinformática y el análisis de las limitaciones de los métodos basados en agrupamiento directo evidenciaron la necesidad de esta etapa adicional, lo que llevó a reformular el

diseño conceptual para incluir la estandarización y la reducción de dimensionalidad como componentes independientes con roles bien definidos.

Desde la perspectiva del dominio, el refinamiento más importante fue la formalización de las restricciones biológicas como criterios de validación del método. En las primeras conceptualizaciones, los criterios de evaluación eran principalmente estadísticos: cohesión de los grupos, separabilidad y reducción porcentual del espacio de búsqueda. Sin embargo, las sesiones de trabajo con los expertos del CISEI evidenciaron que un método que optimiza únicamente criterios estadísticos puede generar grupos que, aunque compactos y bien definidos desde una perspectiva matemática, no conservan los genes *V* reales del genoma analizado. Esta observación llevó a incorporar la cobertura de genes de referencia validados por expertos como criterio central de evaluación, complementando los criterios estadísticos con una medida directa de la utilidad biológica del resultado.

5.4.3 Diseño conceptual consolidado

El proceso iterativo descrito en las secciones anteriores culminó en un diseño conceptual consolidado que integra de manera coherente las decisiones iniciales, los refinamientos derivados de la literatura y las restricciones incorporadas a partir del conocimiento experto del dominio. En el marco del Ciclo de Diseño de la metodología DSR, este diseño consolidado representa el artefacto conceptual que sirve de base para la implementación técnica descrita en el Capítulo 6, garantizando que cada decisión de implementación tenga un fundamento conceptual explícito y verificable.

El diseño consolidado se caracteriza por tres propiedades que surgieron del proceso iterativo y que no estaban completamente articuladas en las decisiones iniciales. La primera de estas propiedades es la modularidad del método: cada componente tiene una función específica, entradas y salidas claramente definidas, y puede ser evaluado de manera independiente. Esta propiedad, que se refleja directamente en el diagrama conceptual de la Figura 5.4, facilita la identificación de qué componente contribuye a qué aspecto del desempeño del método y permite ajustes focalizados sin necesidad de rediseñar el método completo.

La segunda propiedad es la complementariedad entre criterios estadísticos y biológicos. El diseño consolidado reconoce que la reducción del espacio de búsqueda en genes *V* no puede evaluarse únicamente con métricas computacionales, sino que requiere un balance entre la eficiencia estadística del agrupamiento y la validez biológica del conjunto resultante. Este balance, incorporado durante el proceso de refinamiento descrito en la sección anterior, está presente en cada componente del método y se materializa en los criterios de validación descritos en la sección 5.3.5.

La tercera propiedad es la transferibilidad. El diseño consolidado no está acoplado a un genoma específico ni a una especie particular, sino que define una estructura metodológica general que puede instanciarse en distintos contextos ajustando los parámetros del modelo sin modificar la arquitectura conceptual. Esta propiedad, que deriva directamente del principio de reusabilidad descrito en la sección 5.1.4, es la que permite que el método sea validado en múltiples genomas en el Capítulo 7 y evaluado en términos de su capacidad de generalización en el Capítulo 8.

En conjunto, estas tres propiedades definen un método conceptualmente robusto cuya implementación técnica, validación experimental y evaluación de resultados se presentan en los capítulos siguientes, completando así el Ciclo de Diseño de la metodología DSR adoptada en esta investigación.

Capítulo 6. Desarrollo iterativo del método RIVAM

Este capítulo describe el proceso con el cual los fundamentos conceptuales presentados en el Capítulo 5 se tradujeron en una implementación concreta y funcional del método RIVAM. A diferencia del capítulo anterior, que estableció el qué y el porqué del diseño, este capítulo se enfoca en el cómo: las decisiones técnicas tomadas, las iteraciones de desarrollo realizadas y los ajustes incorporados hasta llegar a las cuatro variantes del método que serán evaluadas en el Capítulo 7.

El desarrollo siguió los principios del Ciclo de Diseño de la metodología DSR, construyendo y evaluando el artefacto en iteraciones sucesivas, incorporando en cada ciclo evidencia empírica real y retroalimentación de los expertos del dominio. El capítulo se organiza en cuatro secciones: el proceso de desarrollo iterativo, el pipeline completo del método, las especificaciones técnicas de implementación y la validación con expertos en biología molecular.

6.1 Desarrollo iterativo del método RIVAM

El método RIVAM logró su configuración final después de un proceso iterativo de diseño, evaluación y ajuste, sustentado en evidencia empírica y retroalimentación continua. En esta sección se describe dicha evolución, abordando las decisiones adoptadas, los hallazgos, las limitaciones identificadas y las modificaciones incorporadas en cada etapa, hasta consolidar las cuatro variantes finales que posteriormente se analizan de manera sistemática en el Capítulo 7.

El proceso de desarrollo siguió una lógica acumulativa: cada fase partió de los aprendizajes de la anterior, incorporando nuevos componentes o modificando los existentes para superar las limitaciones identificadas. La Figura 6.1 muestra esta progresión, mostrando cómo el pipeline fue evolucionando desde una primera versión basada únicamente en unificación de datos y agrupamiento, hasta una versión consolidada que integra normalización, reducción de dimensionalidad y optimización de hiperparámetros.

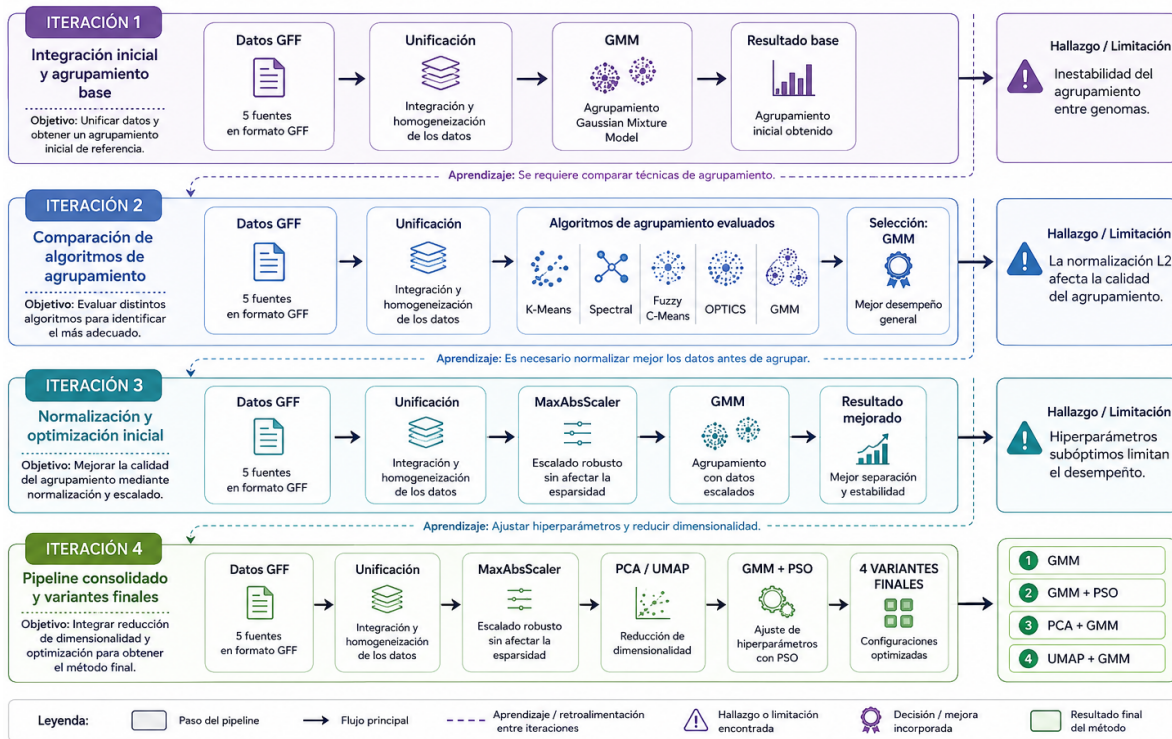


Figura 6.1 Etapas desarrolladas en el refinamiento del método

6.1.1 Primera iteración: unificación de datos y primera propuesta de agrupamiento

La primera iteración del desarrollo del método RIVAM abordó dos problemas fundamentales que debían resolverse antes de aplicar cualquier técnica de

agrupamiento: la dispersión de los datos provenientes de los cinco sistemas de alineamiento y la ausencia de una representación vectorial que permitiera aplicar algoritmos de aprendizaje automático sobre los registros de genes V. Esta iteración estableció las bases del método al desarrollar el proceso de unificación de datos y realizar las primeras pruebas de agrupamiento sobre un genoma de referencia.

6.1.1.1 Datos utilizados en el ámbito genómico

Uno de los primeros aspectos abordados en el desarrollo del método consistió en comprender el significado y la estructura de los datos contenidos en los archivos en formato GFF proporcionados por los cinco sistemas de alineamiento. Esta comprensión es fundamental para entender por qué surge la necesidad de unificar los datos antes de aplicar cualquier técnica de reducción.

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 particular de los genes V. 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 una caracterización más completa del gen V, pero genera un espacio de búsqueda amplio y heterogéneo que dificulta su análisis directo.

Estos sistemas proporcionan datos en formato GFF que representan las ubicaciones donde se produjeron alineamientos de secuencias. Los archivos GFF contienen nueve atributos que permiten la identificación de información clave, descritos en la Tabla 6.1. En esta investigación, las columnas Seqname, Start, End y Strand son relevantes, debido a que describen la ubicación y dirección de cada alineamiento en el genoma.

Tabla 6.1 Fragmento de datos de un archivo GFF

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	exonerate:est2genome	gene	1173811	1174122	NA	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174114	1239	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174122	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174113	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174119	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174125	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174125	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174119	2011	-	NA	NA
CM040297.1	exonerate:est2genome	gene	1173811	1174122	2011	-	NA	NA

Los atributos del formato GFF son los siguientes:

- **SeqID:** identificador único de la secuencia, como el cromosoma o scaffold donde se encuentra la característica.
- **Source:** sistema o método que generó el alineamiento, por ejemplo, Hmmer, Exonerate o BLAST.
- **Type:** tipo de característica alineada, como Gen, Exon, CDS o mRNA.
- **Start:** posición inicial de la característica en la secuencia.
- **End:** posición final de la característica en la secuencia.
- **Score:** valor de confianza asociado a la anotación.
- **Strand:** dirección de la característica en el ADN, indicada por + o -.
- **Phase:** especifica cómo deben leerse los codones. Los valores posibles son 0, 1 o 2.
- **Attributes:** información adicional sobre la característica, como identificadores únicos o nombres de genes.

En el contexto de esta investigación, los quirópteros han sido ampliamente estudiados debido a su diversidad genética y adaptaciones únicas. Específicamente, en el CISEI se ha trabajado con genomas que pueden contener más de 800 genes V, mientras que uno solo de sus archivos GFF puede superar los 167,165 registros. Para verificar la magnitud y distribución de los datos, se realizó un análisis de normalidad y densidad utilizando el atributo Start del archivo Tblastn de *Tadarida brasiliensis*, que contiene 167,165 registros. La Figura 6.2 presenta los gráficos QQ-normal y de densidad de las posiciones de inicio, que confirman la amplitud de la distribución y justifican la necesidad de un tratamiento especial de los datos.

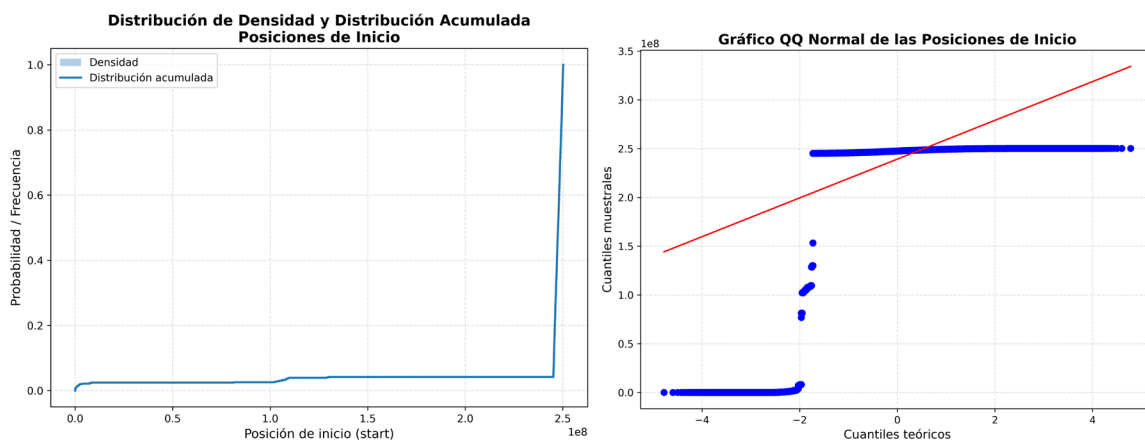


Figura 6.2 Gráfico QQnormal y de densidad de las posiciones de inicio

La Tabla 6.2 presenta las estadísticas básicas del conjunto de datos de Exonerate:est2genome, utilizado como ejemplo para mostrar la variabilidad y dispersión de los valores de inicio y fin de los alineamientos. Es importante señalar que

el archivo utilizado para el análisis de normalidad y densidad fue Tblastn, mientras que Exonerate:est2genome se presenta únicamente con fines descriptivos para mostrar las características generales de los datos de entrada y la variabilidad entre los archivos utilizados.

Tabla 6.2 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	-

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. El análisis de los datos originales en formato GFF muestra que su estructura dispersa dificulta la representación precisa del gen V, limitando la aplicación de algoritmos de aprendizaje automático. Por esta razón, fue necesario desarrollar un preprocesamiento que integre y unifique la información de todas las fuentes en registros completos que representen la estructura total del gen V, lo que da origen a la etapa de unificación de datos descrita en la siguiente sección.

6.1.1.2 Unificación de datos

El análisis de los datos originales en formato GFF reveló que su estructura dispersa dificulta la representación precisa del gen V, limitando la aplicación de algoritmos de aprendizaje automático. Cada sistema de alineamiento generaba su propio archivo GFF con registros parciales que cubrían únicamente una o dos secciones del gen, sin integrar la información de las demás fuentes. Esta fragmentación hacía imposible aplicar directamente técnicas de agrupamiento, ya que cada registro por sí solo no representaba un gen V completo sino únicamente una parte de él.

Con el fin de resolver este problema, se desarrolló un método de unificación de datos compuesto por cinco etapas interrelacionadas, orientadas a consolidar la información dispersa de los cinco sistemas de alineamiento en registros completos y estructurados que representan la estructura del gen V. La Figura 6.3 presenta el diagrama general del proceso de unificación y sus cinco etapas.

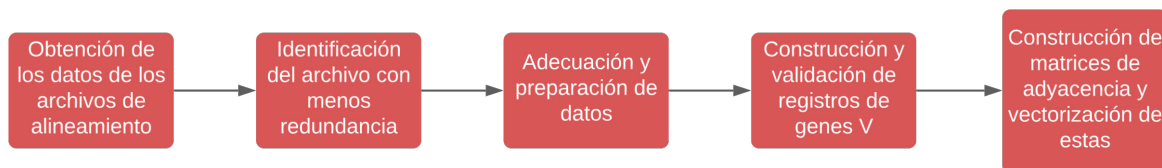


Figura 6.3 Proceso de unificación de datos y sus cinco etapas

Las cinco etapas del proceso fueron las siguientes, tal como se resume en la Tabla 6.3:

Tabla 6.3 Cinco etapas del proceso de unificación de datos

Etapas	Título	Descripción
1	Importación de datos GFF	Recuperación y adaptación de los archivos GFF, transformándolos en un formato tabular que facilita su manipulación y análisis posterior
2	Identificación del archivo con menor redundancia	Comparación de todos los archivos disponibles para determinar cuál tiene el menor número de registros redundantes; este archivo se selecciona como punto de partida por ofrecer la información más específica y confiable para construir 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 demás archivos para mejorar la representación del gen V
4	Construcción y validación de registros de genes V	Ensamblaje de registros de genes V según restricciones biológicas, asegurando que los registros generados no excedan distancias biológicamente coherentes entre sus secciones
5	Construcción de matrices de adyacencia y vectorización	Generación de una matriz de adyacencia por cada gen para representar datos relacionales; se derivan nuevos atributos basados en distancias que son vectorizados para facilitar su uso por algoritmos de aprendizaje automático

Estas cinco etapas se articularon para garantizar que los datos procesados fueran de alta calidad y estuvieran listos para su análisis en etapas posteriores, mejorando la precisión de los registros y reduciendo la redundancia del espacio de búsqueda original.

6.1.1.3 Etapas de unificación de datos

Las etapas del proceso de unificación se describen a continuación en detalle, incluyendo los subprocesos, variables y criterios aplicados en cada una. Es importante señalar que este proceso fue diseñado específicamente para el dominio de los genes V, incorporando restricciones biológicas en cada etapa que garantizan la coherencia estructural de los registros generados. La primera etapa consiste en la importación de datos GFF, donde se recuperan y adaptan los archivos generados por los cinco sistemas de alineamiento, transformándolos en un formato tabular que facilite su manipulación y análisis posterior. A continuación, se describen en detalle las etapas restantes.

Subproceso 1: Importación de datos GFF la primera etapa consiste en la recuperación y adaptación de los archivos GFF obtenidos por los cinco sistemas de alineamiento: HmmerScan para RSS, Exonerate:protein2genome, Exonerate:est2genome, Tblastx y Tblastn. Cada archivo contiene registros parciales que describen únicamente una o dos secciones del gen V desde la perspectiva del sistema que los generó.

En esta etapa los archivos son cargados y transformados en un formato tabular estructurado, donde cada fila representa un registro de alineamiento con sus columnas correspondientes: Seqname, Source, Feature, Start, End, Score, Strand, Frame y Attribute. Esta transformación establece la estructura base sobre la cual operarán las etapas siguientes, garantizando que todos los archivos sean procesados bajo el mismo esquema de datos independientemente del sistema que los generó. El proceso general de las cinco etapas se muestra en la Figura 6.3.

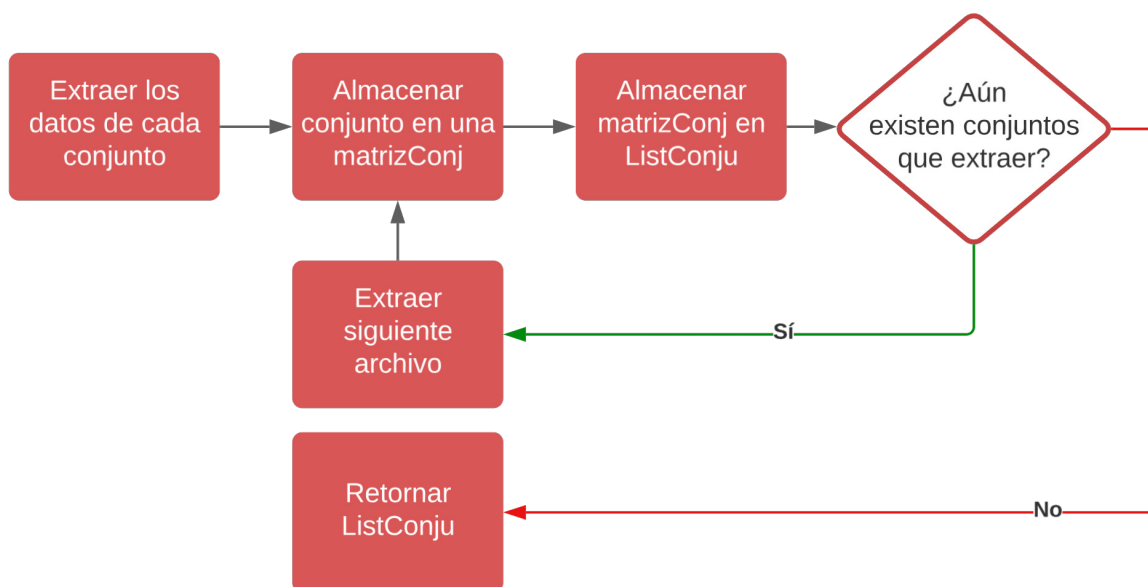


Figura 6.4 Subproceso de importación de datos para la unificación de datos

Subproceso 2: Identificación del archivo con menor redundancia el segundo subproceso tuvo como objetivo identificar el archivo con el menor nivel de redundancia, garantizando que el conjunto de datos seleccionado estuviera conformado únicamente por información no repetida. Para lograrlo, se evaluó la redundancia presente en cada conjunto GFF con el análisis de las columnas Inicio y Fin de los archivos GFF. En particular, se verificó si al menos una de estas columnas contiene valores únicos y si dichos valores correspondían a la totalidad de los registros del archivo analizado.

A partir de este análisis, se seleccionó el conjunto con menor redundancia, el cual fue almacenado en la variable DUMrev o DUMforw (*Data Unique Matrix*), según la orientación del alineamiento evaluado. De esta manera, los archivos asociados a alineamientos en dirección *Reverse* fueron asignados a DUMrev, mientras que aquellos correspondientes a alineamientos *Forward* se almacenaron en DUMforw. Finalmente, ambas matrices son retornadas para su uso en los subprocesos posteriores de la unificación, permitiendo trabajar con conjuntos de datos depurados y consistentes. La Figura 6.5 muestra el diagrama correspondiente a este subproceso.

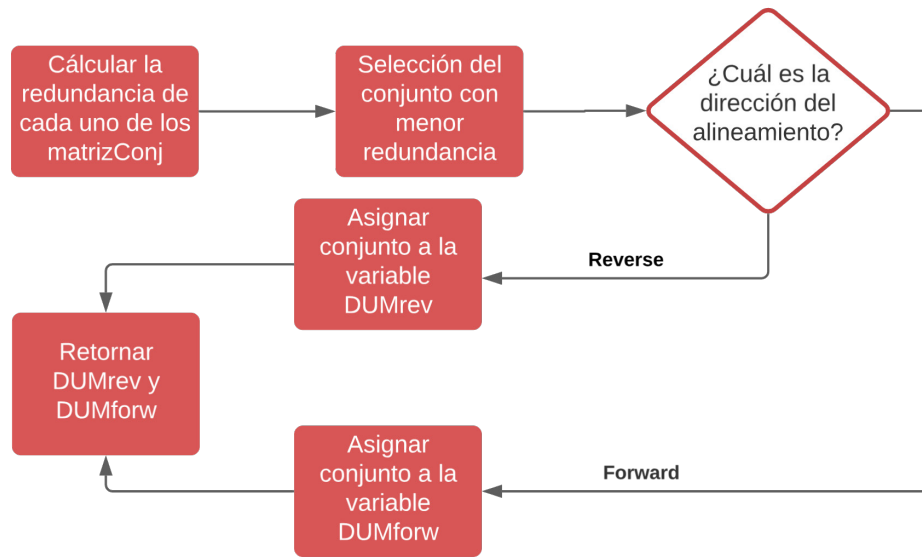


Figura 6.5 Identificación del archivo con menor redundancia

Subproceso 3: Extracción y preparación de datos el tercer subproceso se centró en la extracción y preparación de datos, asegurando que los registros fueran estructurados correctamente. Para ello, se extrajeron las posiciones de inicio de las matrices DUMrev y DUMforw en pares, utilizándose como rangos. Con estos rangos, se recuperaron los datos de los demás archivos GFF que coincidieran dentro de los intervalos establecidos.

Posteriormente, los registros extraídos fueron asignados a listas según su dirección: LISTMrev o LISTMforw. Estas listas almacenaron los registros de manera estructurada para su posterior análisis. El proceso continuó iterativamente hasta que no existieron datos pendientes en DUMrev o DUMforw. La Figura 6.6 muestra este subproceso.

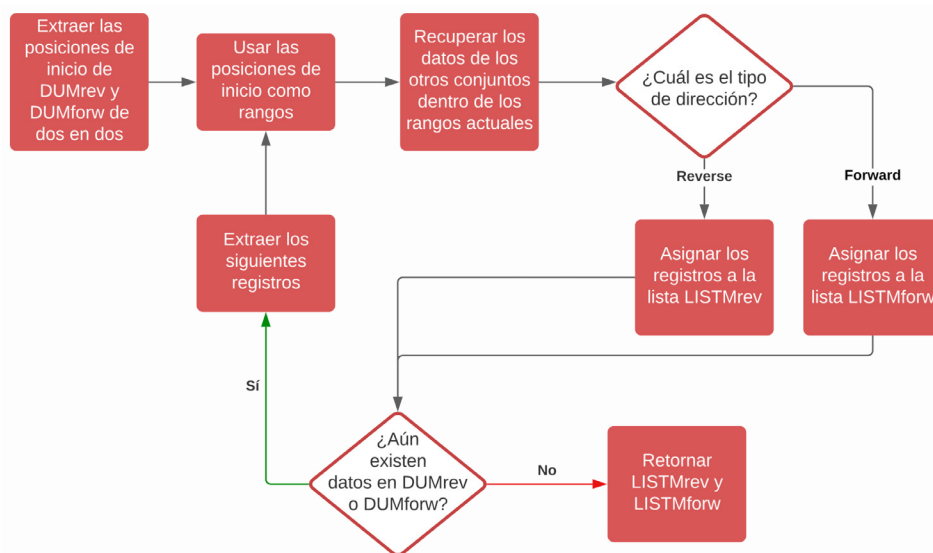


Figura 6.6 Extracción y preparación de datos

Subproceso 4: Construcción y validación de registros de genes V El cuarto subproceso, presentado en la Figura 6.7, estuvo centrado en la construcción y validación de registros de genes V. Su objetivo fue asegurar que los registros generados cumplieran con distancias biológicamente coherentes entre las secciones del gen, consolidando la información de los subprocesos previos en representaciones completas y estructuradas.

El subproceso comenzó con la extracción de un registro de DUMrev o DUMforw y su asignación a la variable RSS, que representa la señal de recombinación del gen. El siguiente paso procesó los datos de LISTMrev o LISTMforw según la dirección correspondiente: el primer registro de LISTM fue asignado a Exón y el siguiente a SP. Esta asignación secuencial garantizó que cada registro construido contuviera las tres secciones estructurales del gen V, siguiendo el orden biológico establecido en el dominio.

Un aspecto crítico de este subproceso fue la incorporación de conocimiento experto del dominio en forma de reglas explícitas que delimitan las distancias esperadas entre las secciones del gen. La construcción de registros se basó en una estructura genérica del gen V, asegurando que las tres secciones estuvieran correctamente representadas y que cada registro fuera biológicamente válido. Para mantener esta coherencia, se establecieron cuatro reglas fundamentales que se presentan en la Tabla 6.4.

Tabla 6.4 Reglas de validación biológica para la construcción de registros de genes V

Regla	Título	Descripción
1	Distancia entre RSS y Exón	La diferencia absoluta entre la posición de inicio de la RSS y el inicio del Exón debe ser ≥ 0 y ≤ 15 caracteres. Garantiza que las secciones estén próximas sin solapamientos ni separaciones excesivas; de lo contrario, se extrae el siguiente registro como Exón y se reevalúa
2	Distancia dentro del Exón	La diferencia absoluta entre el inicio y el fin del Exón debe estar entre 280 y 350 caracteres. Se usan valores absolutos porque los alineamientos Reverse aumentan y los Forward disminuyen; si no se cumple, se analiza el siguiente registro y se reinicia todo el proceso
3	Distancia dentro del SP	La diferencia absoluta entre el inicio y el fin del SP debe ser > 45 y ≤ 50 caracteres. Si esta condición no se cumple, el registro se extrae como SP y se reinicia todo el proceso de evaluación
4	Distancia entre SP y Exón	La distancia absoluta entre el inicio del SP y el fin del Exón debe estar entre 60 y 150 caracteres. Esto asegura una separación biológicamente coherente; si el criterio no se cumple, otro registro se trata como SP potencial y se reinicia todo el proceso

Si todos los criterios de validación se cumplen, los registros de RSS, SP y Exón se almacenan en la matriz Mat_Gen según la dirección del alineamiento y se almacenan 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 con los registros finales listos para su análisis en la siguiente fase. La Figura 6.7 presenta el diagrama completo de este subproceso.

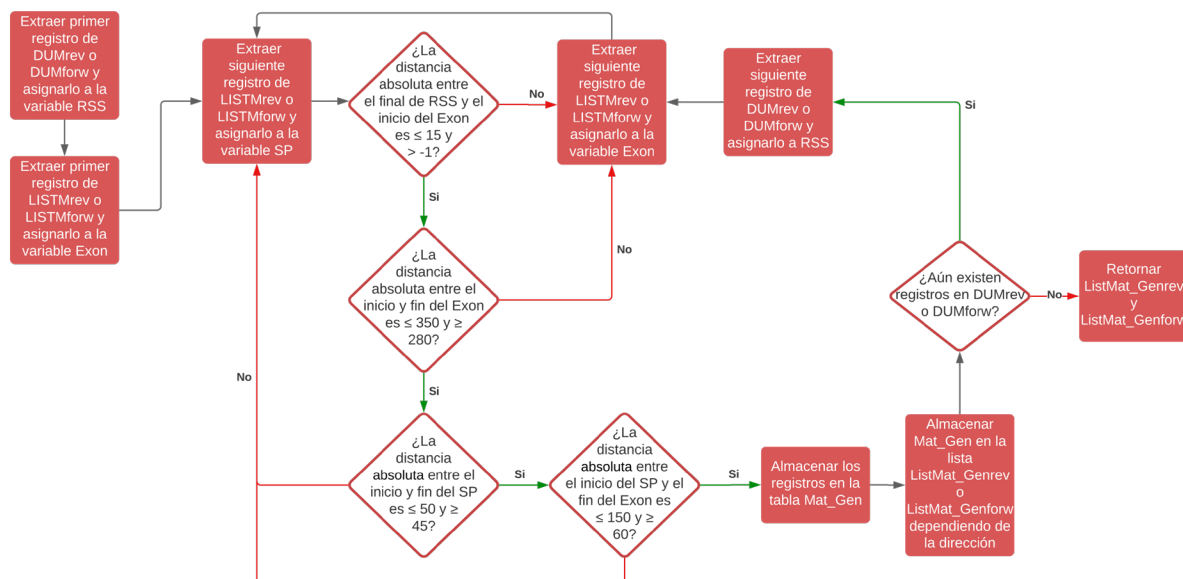


Figura 6.7 Construcción y validación de registros de genes V

Subproceso 5: Construcción de matrices de adyacencia y vectorización la Figura 6.8 representa el quinto 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 que optimice el análisis en etapas posteriores. El procedimiento inicia con la combinación de las listas ListMat_Genrev y ListMat_Genforw, obteniendo 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, específicamente entre sus campos de 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.

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 obtenidos para Totalgen se almacenan en MatFinal, que constituye el nuevo conjunto de datos conteniendo la estructura del gen, su dirección y sus distancias. El subproceso se repite hasta que no existan más registros en ListMat. La Figura 6.8 presenta el diagrama completo de este subproceso.

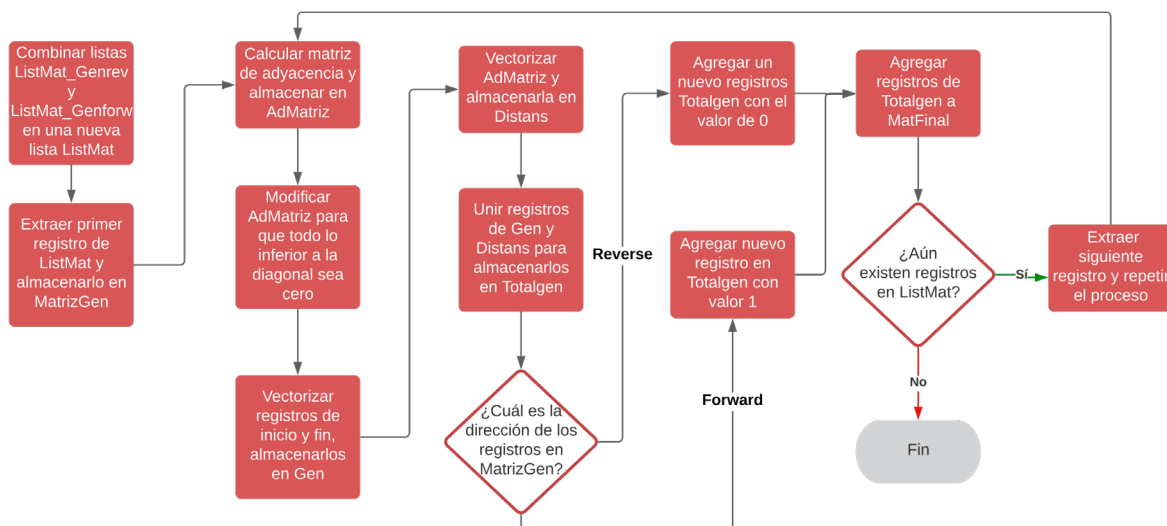


Figura 6.8 Construcción y validación de registros de genes V

6.1.1.4 Representación vectorial del gen v

Uno de los pasos fundamentales del método de unificación de datos es la formación de registros funcionales a partir de matrices estructuralmente válidas. Esta sección se centra en la construcción de estos registros con una matriz de adyacencia, con el objetivo de garantizar que cada registro dentro del conjunto represente distancias validadas entre las secciones del gen V y pueda ser procesado por algoritmos de aprendizaje automático.

La Tabla 6.5 presenta ejemplos de registros que pueden encontrarse en ListMat_Genforw o ListMat_Genrev, almacenados en MatrizGen. Cada grupo de tres registros representa un gen V con sus respectivas secciones y direcciones asociadas. Es importante señalar que, aunque la columna Feature especifica la sección del gen identificada, no siempre puede tomarse como criterio absoluto en la formación de registros.

Tabla 6.5 Ejemplo de registros para un gen V con todas sus secciones

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	hmmerscan	RSS	3712219	3712180	NA	-	NA	NA
CM040297.1	exonerate:es	Exon	3712181	3711873	NA	-	NA	NA
CM040297.1	exonerate:es	gene	3711792	3711746	NA	+	NA	NA
CM040297.1	hmmerscan	RSS	2168368	2168403	NA	-	NA	NA

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	exonerate:es t2genome	Exon	2168406	2168717	NA	-	NA	NA
CM040297.1	exonerate:es t2genome	gene	2168796	2168837	NA	+	NA	NA

El primer paso en la construcción de registros es la asignación de datos a la variable Gen, extrayendo únicamente las posiciones absolutas en el genoma y vectorizándolas. La variable Gen define las posiciones que se utilizarán para la matriz de adyacencia. Cada registro representa una sección del gen V: los primeros tres registros corresponden a las posiciones de inicio de la RSS, el Exón y el SP, mientras que los últimos tres corresponden a las posiciones de fin de estas secciones. La Tabla 6.6 presenta ejemplos de genes V organizados en sus seis secciones.

Tabla 6.6 Ejemplos de genes V organizados en sus seis secciones

StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
3712219	3712180	3712181	3711873	3711792	3711746
2168368	2168406	2168796	2168403	2168717	2168837

El cálculo de la distancia se basa en la sustracción progresiva de cada posición de un gen con respecto a todas las demás posiciones, repitiendo este procedimiento para cada posición. La ecuación 6.1 representa este cálculo:

$$d_{ij} = |x_i - x_j|, \quad \forall i, j \in \{1,2,3,4,5,6\} \quad (\text{Ec. 6.1})$$

Donde x_i y x_j representan las posiciones absolutas en el genoma de las secciones del gen V. El resultado de aplicar esta ecuación a cada par de posiciones genera la matriz de adyacencia completa mostrada en La ecuación 6.2:

$$\begin{bmatrix} 3712219 - 3712219 & 3711873 - 3712219 \\ 3712180 - 3712219 & 3711792 - 3712219 \\ 3712181 - 3712219 & 3711746 - 3712219 \end{bmatrix} = \{0 \quad 39 \quad 38 \quad 346 \quad 427 \quad 473\}$$

(Ec. 6.2)

El resultado es una matriz de adyacencia que representa las distancias absolutas entre las secciones del gen V. Cada posición se compara tanto con las otras cómo consigo misma, asegurando que todas las relaciones espaciales sean correctamente evaluadas. Para evitar valores redundantes, todos los valores por debajo de la diagonal principal se convierten en cero. La Tabla 6.7 presenta la matriz de adyacencia resultante.

Tabla 6.7 Ejemplo de matriz de adyacencia

Posiciones	3712219	3712180	3712181	3711873	3711792
3712219	0	39	38	346	427
3712180	0	0	1	307	388
3712181	0	0	0	308	389
3711873	0	0	0	0	81
3711792	0	0	0	0	0
3711746	0	0	0	0	0

Con la matriz de adyacencia generada, los valores de las distancias se vectorizan y se agregan 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, Totalgen se integra en la tabla MatFinal, donde se almacenan todos los registros de genes V.

Las posiciones resultantes de la vectorización pueden representarse como un grafo completamente conectado, donde cada nodo corresponde a una posición de inicio o fin de las secciones del gen V, y cada arista representa la distancia absoluta entre dos nodos. Esta representación gráfica, mostrada en la Figura 6.9, permite ver de manera intuitiva las relaciones espaciales entre las secciones del gen que son capturadas por la matriz de adyacencia. Los valores iguales o menores a cero son eliminados del grafo, resultando en una estructura completamente conectada que refleja la organización estructural del gen V.

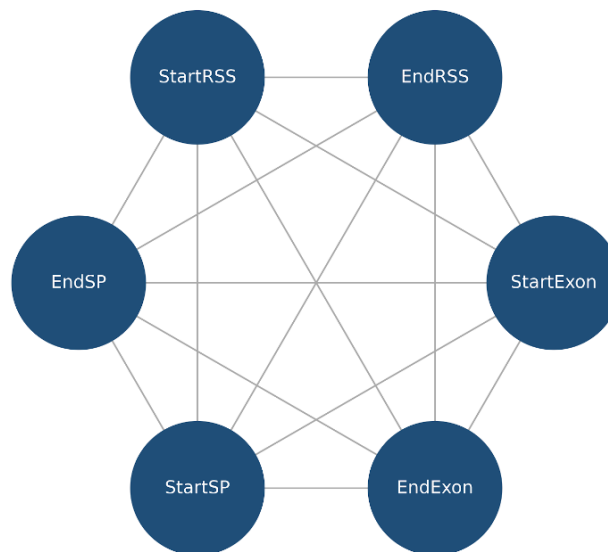


Figura 6.9 Grafo del gen V con sus posiciones como nodos y distancias como aristas

El nuevo conjunto de datos se compone de 43 atributos, de los cuales los primeros siete corresponden a las posiciones de los alineamientos en el genoma: 1) Strand, 2) StartRSS, 3) StartExon, 4) StartSP, 5) EndRSS, 6) EndExon y 7) EndSP. Estas primeras

características se mantienen para preservar la relación entre los registros y su ubicación en el genoma.

La Tabla 6.8 presenta los 36 atributos restantes, 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 por debajo de la diagonal de la matriz de adyacencia se convierten en cero. Con la integración de estos datos se obtiene el conjunto final listo para su uso en algoritmos de aprendizaje automático.

Tabla 6.8 Atributos resultantes de la unificación de datos

Atributos 8-13	Atributos 14-19	Atributos 20-25	Atributos 26-31	Atributos 32-37	Atributos 38-43
StartRSS-StartRSS	EndRSS-StartRSS	StartExon-StartRSS	EndExon-StartRSS	StartSP-StartRSS	EndSP-StartRSS
StartRSS-EndRSS	EndRSS-EndRSS	StartExon-EndRSS	EndExon-EndRSS	StartSP-EndRSS	EndSP-EndRSS
StartRSS-StartExon	EndRSS-StartExon	StartExon-StartExon	EndExon-StartExon	StartSP-StartExon	EndSP-StartExon
StartRSS-EndExon	EndRSS-EndExon	StartExon-EndExon	EndExon-EndExon	StartSP-EndExon	EndSP-EndExon
StartRSS-StartSP	EndRSS-StartSP	StartExon-StartSP	EndExon-StartSP	StartSP-StartSP	EndSP-StartSP
StartRSS-EndSP	EndRSS-EndSP	StartExon-EndSP	EndExon-EndSP	StartSP-EndSP	EndSP-EndSP

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. 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.

6.1.1.5 Normalización y preprocesamiento

Con la obtención del conjunto de datos unificado, fue necesario llevar a cabo un análisis exploratorio para detectar posibles inconsistencias, evaluar la normalidad, identificar valores atípicos y comprender la densidad de los datos. El primer paso consistió en obtener las estadísticas básicas del conjunto, verificando que no existieran datos faltantes o nulos.

La Tabla 6.9 presenta los estadísticos principales del conjunto antes de la normalización. Se observa que los atributos que representan distancias relativas entre secciones del gen, como InicioRSS-Inicio_Exon con una media de 42.26 y una desviación estándar de 3.25, presentan valores más estables y biológicamente coherentes. En contraste, atributos como InicioRSS-Fin_SP, que representa la distancia total del gen, presentan mayor variabilidad con valores que oscilan entre 432 y 577.

Tabla 6.9 Estadísticas básicas del conjunto de datos unificado

Estadístico	StartRSS- EndRSS	StartRSS- StartExon	StartRSS- EndExon	StartRSS- StartSP	StartRSS- EndSP	EndRSS- StartExon
count	47717	47717	47717	47717	47717	47717
mean	42.26	349.99	436.84	480.44	4.84	307.72
std	3.25	5.19	9.06	9.37	2.89	4.2
min	21	308	386	432	0	269
25%	40	346	429	473	3	305
50%	42	349	435	479	4	308
75%	44	354	446	490	7	311
max	53	409	531	577	14	370

Una vez verificada la integridad de los datos, se procedió a eliminar las columnas compuestas enteramente por valores cero, debido a que no aportan información relevante y podían interferir en el proceso de agrupamiento. Posteriormente, se aplicó un proceso de normalización utilizando la función Normalizer de scikit-learn, basada en la norma L2 o euclidiana, que ajusta los valores a un vector de norma unitaria.

Tras la normalización, la Tabla 6.10 muestra cómo los valores se reducen a un rango entre 0 y 0.42, con desviaciones estándar inferiores a 0.006. Esta homogeneización de la escala garantiza que todos los atributos contribuyan de manera equitativa al proceso de agrupamiento, sin que las distancias absolutas de mayor magnitud dominen el análisis. La Figura 6.10 presenta las gráficas de densidad en tres estados: datos en escala original, datos escalados y datos normalizados, mientras que la Figura 6.11 muestra los gráficos de dispersión antes y después de la normalización.

Tabla 6.10 Estadísticas básicas del conjunto de datos unificado

Estadístico	StartRSS- FinRSS	StartRSS- StartExon	StartRSS- EndExon	StartRSS- StartSP	StartRSS- EndSP	EndRSS- StartExon
count	47717	47717	47717	47717	47717	47717
mean	0.0365	0.3019	0.3768	0.4144	0.0042	0.2655
std	0.0028	0.0059	0.002	0.0018	0.0025	0.0052
min	0.018	0.2495	0.3655	0.404	0	0.2212
25%	0.0346	0.298	0.3756	0.4131	0.0025	0.2617
50%	0.0362	0.3024	0.377	0.4144	0.0035	0.2657
75%	0.0384	0.3061	0.3782	0.4155	0.0059	0.2696
max	0.0455	0.3328	0.3863	0.4229	0.0124	0.2995

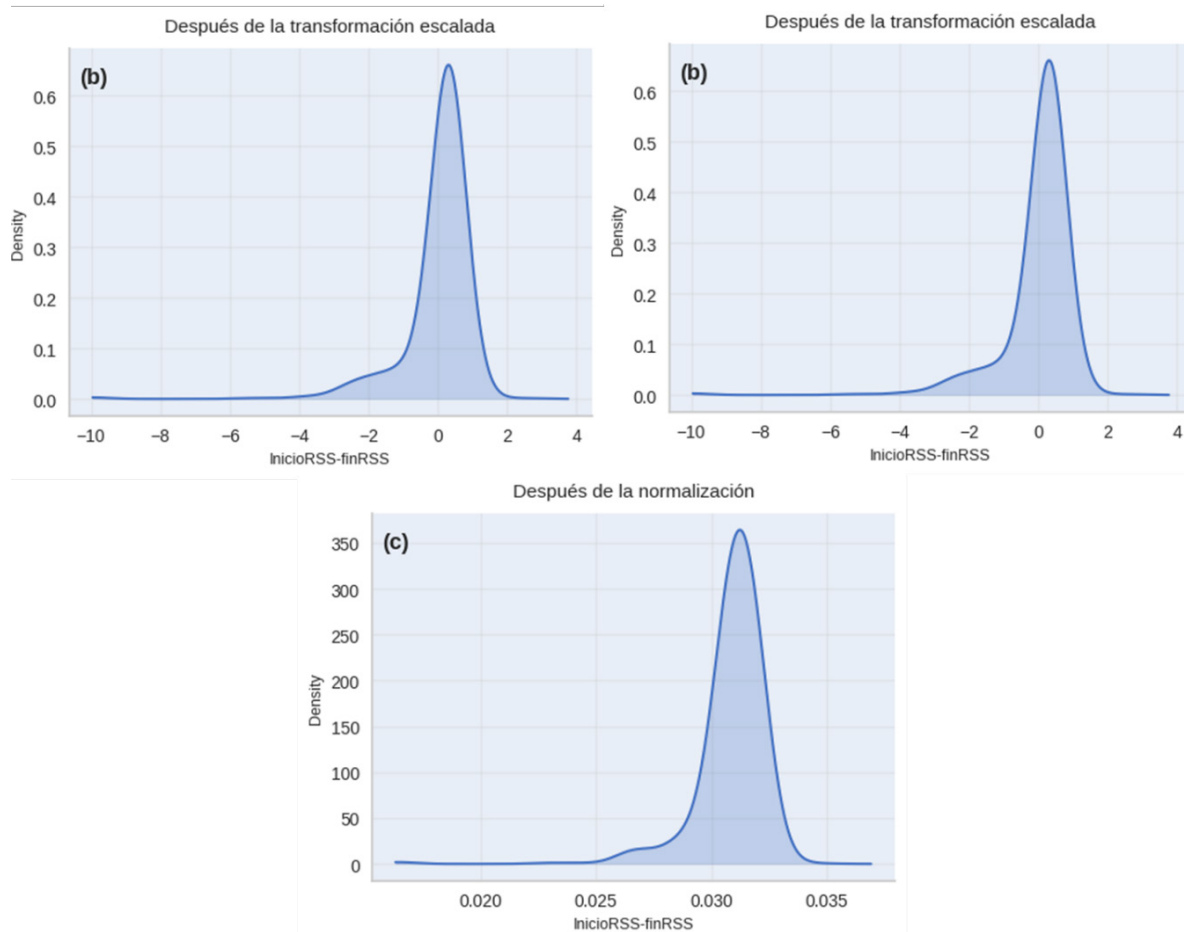


Figura 6.10 Gráficas de densidad del conjunto de datos: (A) escala original, (B) datos escalados, (C) datos escalados y normalizados

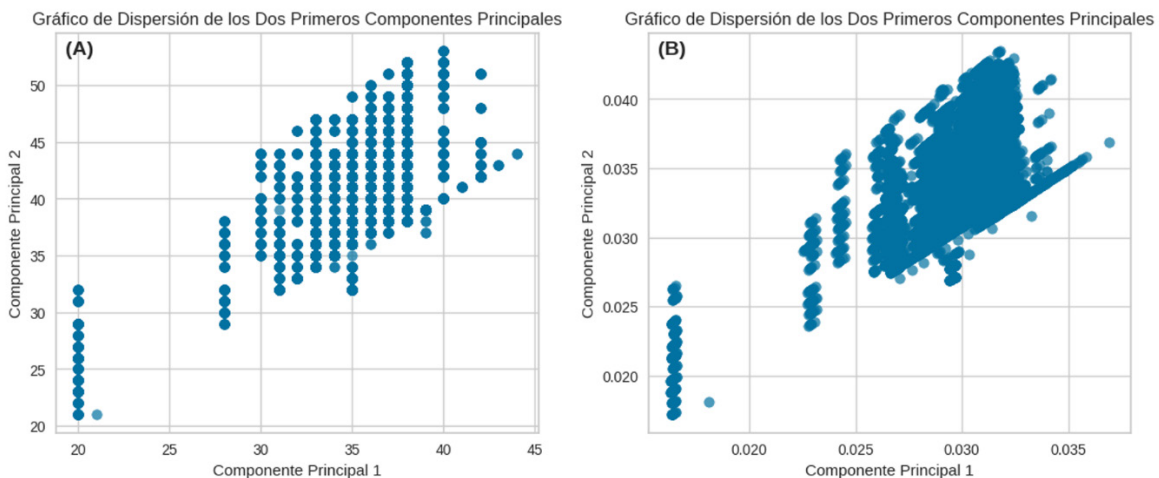


Figura 6.11 Gráficos de dispersión: (A) datos sin normalizar, (B) datos normalizados

La comparación entre las estadísticas básicas del conjunto antes y después de la normalización evidencia el impacto del proceso. La Tabla 6.9 presenta los estadísticos principales del conjunto original, donde se observa que los atributos que representan

distancias relativas entre secciones del gen presentan valores estables y biológicamente coherentes, como InicioRSS-Inicio_Exon con una media de 42.26 y una desviación estándar de 3.25. Sin embargo, la diferencia de escala entre atributos, con valores que oscilan entre 0 y 577, hace necesaria una normalización previa al agrupamiento para garantizar que todos los atributos contribuyan de manera equitativa al análisis.

6.1.1.6 Primer agrupamiento con K-Means

Con los datos normalizados, se procedió a realizar el primer agrupamiento utilizando el algoritmo K-Means. Para esta primera prueba se utilizó el conjunto de datos de *Tadarida brasiliensis*, una especie de murciélago cuyo genoma se encuentra publicado en la base de datos del NCBI y para el cual se disponía de anotaciones manuales realizadas por los expertos del CISEI. El conjunto de datos utilizado se compuso de dos subconjuntos: el conjunto principal con 47,717 registros obtenidos tras el proceso de unificación, y los datos de referencia validados por expertos con aproximadamente 850 registros que representan la ubicación real de cada gen V en el genoma. Este segundo subconjunto fue incorporado al proceso de agrupamiento como grupo adicional, con el objetivo de utilizarlo como punto de referencia para evaluar la calidad de los grupos generados.

En esta primera prueba se configuró K-Means con 20 grupos, donde el grupo 20 corresponde a los datos de referencia de los expertos. La selección de este número de grupos no se basó en un criterio métrico formal, sino en la necesidad de explorar una partición suficientemente granular del espacio de datos que permitiera identificar subconjuntos con alta concentración de genes V reales. La Figura 6.12 presenta el resultado del agrupamiento en un gráfico de dispersión, donde se aprecia la distribución de los 20 grupos generados.

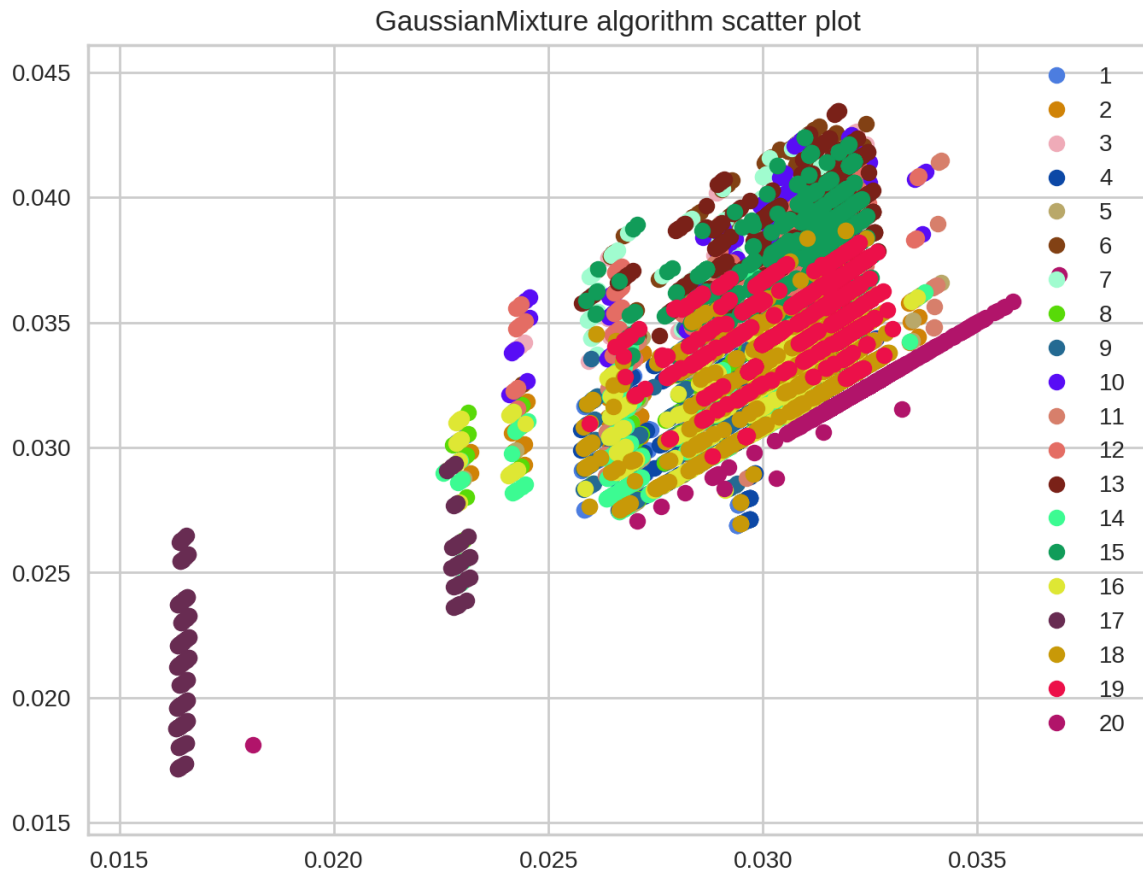


Figura 6.12 Gráfico de dispersión del primer agrupamiento con K-Means (20 grupos)

A partir del análisis visual del espacio tridimensional, se identificó que los grupos 1 y 16 eran los más cercanos a la referencia de los expertos. La Tabla 6.11 presenta el detalle de estos dos grupos, incluyendo el número de registros y la cantidad de genes identificados en cada uno.

Tabla 6.11 Grupos de interés identificados en el primer agrupamiento con K-Means

Algoritmo	Grupos de interés	Número de registros	Cantidad de genes
K-Means	Grupo 1	3595	575
	Grupo 16	3761	563

Como se observa en la tabla, ambos grupos presentan una cantidad similar de genes identificados y un volumen de registros comparable. La selección del grupo final se basó en dos criterios complementarios: la proximidad espacial del grupo a la referencia en el espacio tridimensional y el equilibrio entre la cantidad de genes identificados y el volumen de datos agrupados. Este equilibrio fue fundamental porque un grupo demasiado pequeño podría omitir genes válidos, mientras que un grupo demasiado grande incluiría un alto número de registros irrelevantes, reduciendo la efectividad de la reducción del espacio de búsqueda.

6.1.1.7 Resultados en IGV

Una vez identificado el grupo más relevante, se procedió a transformar los registros del grupo seleccionado al formato GFF para su análisis en IGV. Este proceso se basó en cinco pasos:

1. Extracción de las posiciones de inicio y fin del conjunto de datos
2. Identificación de las posiciones exactas en los cinco archivos de alineamiento
3. Unificación de los resultados en un único conjunto de datos en formato GFF
4. Eliminación de duplicados
5. Exportación del archivo GFF resultante

El resultado del proceso de transformación y unificación se presenta en la Figura 6.13, donde se muestra el sistema IGV con diferentes rieles. La sección resaltada en color azul corresponde a los genes identificados por el proceso de agrupamiento, mientras que la sección en color rojo corresponde a la referencia obtenida por los expertos en biología molecular.

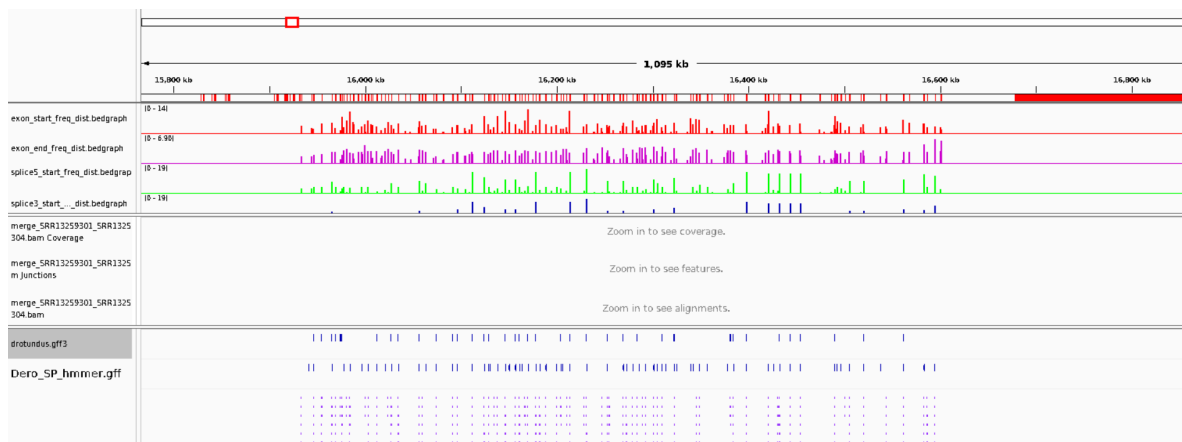


Figura 6.13 Resultados en IGV: vista general de genes identificados

La Figura 6.14 presenta uno de los resultados individuales comparado con la referencia de los expertos. Como se puede observar, la variación existente entre el resultado y la referencia es baja, limitándose a unas pocas posiciones. Esta variación, aunque presente, era esperada en esta primera iteración del método y sería corregida en iteraciones posteriores mediante la incorporación de métricas formales de error.

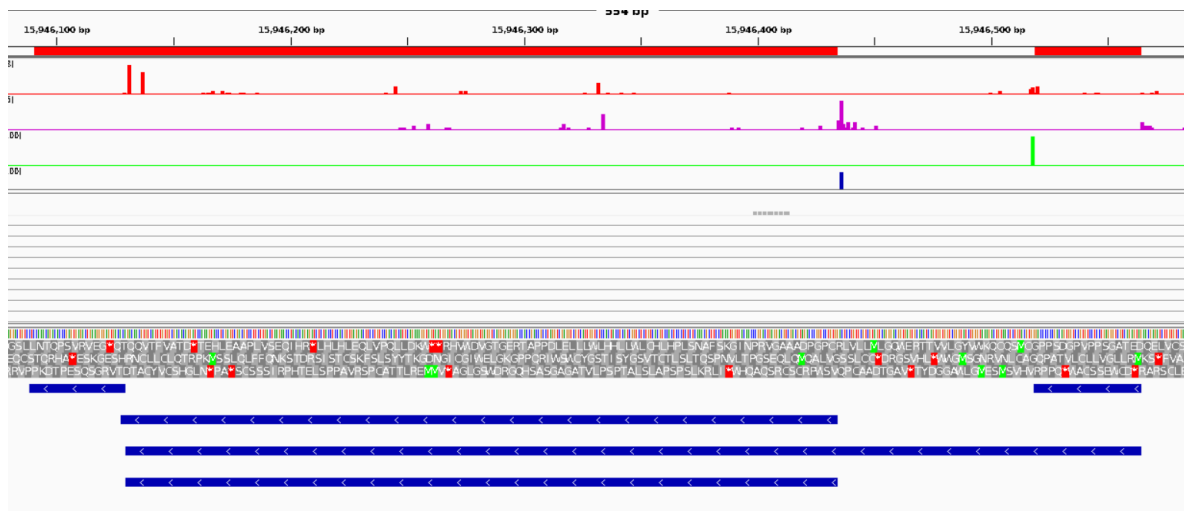


Figura 6.14 Resultados en IGV: comparación con referencia (gen 1)

Cada uno de los resultados presenta un comportamiento similar, con variaciones de un máximo de 20 posiciones respecto a la referencia. La Figura 6.15 presenta otro ejemplo de un gen identificado por el proceso de agrupamiento en IGV, donde se confirma la consistencia de los resultados. Es importante destacar que el total de registros presentes para cada gen fue de 9, lo que representa una reducción considerable respecto al volumen original de registros en el espacio de búsqueda.

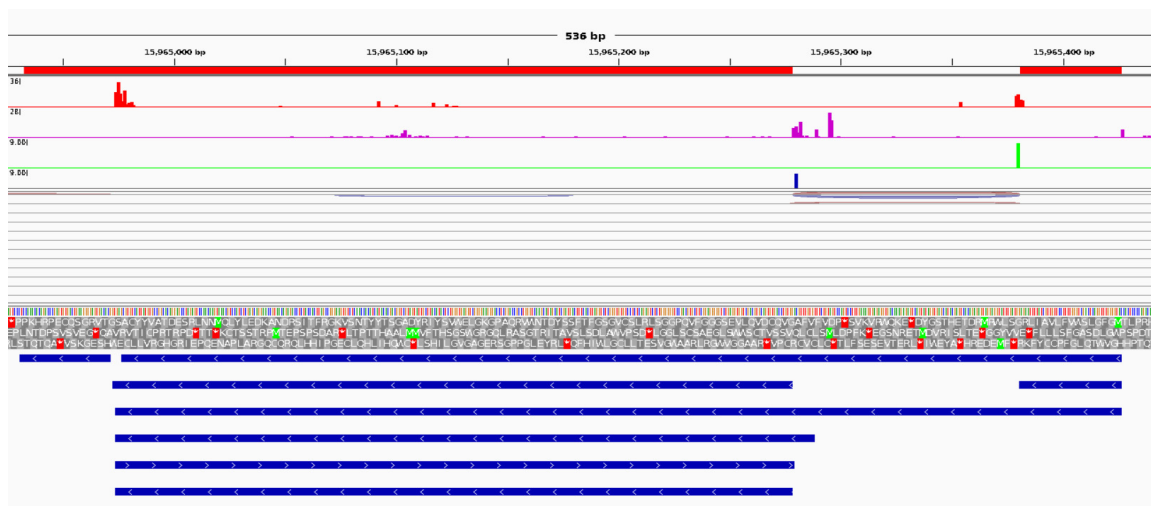


Figura 6.15 Resultados en IGV: comparación con referencia (gen 2)

6.1.1.8 Limitación identificada

Si bien los resultados en esta primera iteración demostraron que el método era capaz de identificar genes V reales a partir del espacio de búsqueda reducido, se identificó una limitación fundamental: la ausencia de una métrica formal de error que permitiera evaluar de manera objetiva y reproducible la calidad del grupo seleccionado.

La selección del grupo más relevante se basó únicamente en criterios visuales, específicamente en la proximidad del grupo a la referencia de los expertos en el espacio tridimensional. Si bien este criterio resultó útil como primera aproximación, presenta varias desventajas desde el punto de vista metodológico. En primer lugar, la selección visual es inherentemente subjetiva y puede variar entre distintos analistas o entre distintas ejecuciones del algoritmo. En segundo lugar, no permite cuantificar con precisión qué tan cercanas son las posiciones identificadas por el método a las posiciones reales de los genes V validados por los expertos. En tercer lugar, dificulta la comparación sistemática entre distintas configuraciones del método, debido a que no existe una métrica común que permita determinar objetivamente cuál configuración produce mejores resultados.

Esta limitación se volvió crítica al considerar que el método debía ser aplicado a diversos genomas con características distintas. En cada genoma, la distribución de los grupos en el espacio de datos puede variar significativamente, lo que hace que la selección visual basada en un genoma específico no sea generalizable a otros contextos. Para que el método fuera verdaderamente reutilizable y transferible, era necesario contar con criterios de selección formalizados que operaran de manera consistente independientemente de las características del genoma analizado.

Adicionalmente, la primera iteración evidenció que la normalización L2 utilizada, si bien mejoraba la distribución de los datos, no garantiza una separabilidad óptima entre el grupo de genes V reales y el resto de los candidatos. Esta observación, combinada con la ausencia de métricas formales, motivó el diseño de la segunda iteración, orientada a incorporar métricas cuantitativas de error que permitieran fundamentar las decisiones de selección en evidencia empírica objetiva y la exploración de distintos algoritmos de agrupamiento de manera sistemática y comparable.

6.1.2 Segunda iteración: incorporación de métricas y comparación de algoritmos

La segunda iteración del desarrollo del método RIVAM surgió directamente de la limitación identificada en la iteración anterior: la ausencia de una métrica formal que permitiera evaluar de manera objetiva la calidad del grupo seleccionado. Esta iteración

tuvo dos objetivos principales. El primero fue definir e incorporar métricas cuantitativas de error que fundamentaran las decisiones de selección en evidencia empírica objetiva. El segundo fue llevar a cabo una comparación sistemática de cinco algoritmos de agrupamiento bajo esas métricas, con el fin de identificar cuál de ellos ofrecía el mejor desempeño para el problema de reducción del espacio de búsqueda de genes V.

En esta iteración se utilizó el conjunto de datos de *Desmodus rotundus*, conformado por 2,904 registros correspondientes a los datos del genoma y 88 registros de referencia validados por los expertos del CISEI. El proceso seguido en esta iteración constó de seis etapas: unificación de datos de los alineamientos, unificación de datos de referencia, preprocesamiento de los datos, aplicación de los algoritmos de agrupamiento, unificación de datos por grupos y evaluación de resultados mediante métricas de error.

6.1.2.1 Motivación para la medición del error

La primera iteración demostró que el método era capaz de identificar genes V reales a partir del espacio de búsqueda reducido, pero la ausencia de una métrica formal de error limitaba la capacidad de evaluar y comparar de manera objetiva el desempeño de distintas configuraciones. La selección visual del grupo más relevante, si bien útil como primera aproximación, no permitía cuantificar con precisión qué tan cercanas eran las posiciones identificadas a las posiciones reales de los genes V validadas por los expertos, ni facilitaba la comparación sistemática entre distintos algoritmos o parámetros del método.

Buscando superar esta limitación, la segunda iteración se orientó hacia dos objetivos complementarios: la definición de métricas cuantitativas de error y la comparación sistemática de múltiples algoritmos de agrupamiento bajo esas métricas. Sin embargo, antes de poder calcular cualquier métrica de error, fue necesario resolver un problema previo: dentro de cada grupo obtenido por el algoritmo de agrupamiento, un mismo gen V podría estar representado por múltiples registros con ligeras variaciones en sus posiciones, producto del proceso de unificación de fuentes descrito en la sección 6.1.1. Esta redundancia residual dentro de los grupos hacía imposible establecer una correspondencia directa entre los resultados del método y la referencia de los expertos.

La solución adoptada retomó el principio de unificación aplicado previamente durante la integración de las fuentes de datos, cuyo objetivo consistió en consolidar múltiples registros que representan una misma entidad en un único registro representativo. No obstante, en esta ocasión la unificación no se realizó entre registros

provenientes de distintos sistemas de alineamiento, sino entre registros pertenecientes a un mismo grupo que hacían referencia al mismo gen V.

A este procedimiento se le denominó *unificación de datos por grupos*, y su incorporación representó un paso fundamental dentro del proceso metodológico. Gracias a esta estrategia fue posible reducir redundancias y homogenizar la representación de los datos dentro de cada grupo, lo que permitió establecer una base consistente para el análisis posterior. Esta etapa resultó indispensable para obtener los datos necesarios para el cálculo de métricas de error objetivas, reproducibles y comparables entre distintos algoritmos, facilitando así una evaluación más rigurosa del desempeño de cada aproximación implementada.

6.1.2.2 Unificación de datos por grupos

La unificación de datos por grupos es un paso crucial en el proceso de evaluación del método, debido a que permite minimizar la cantidad de resultados presentes en cada grupo y facilita la comparación de la calidad de cada agrupamiento con la referencia de los expertos.

El proceso parte de una observación fundamental: cada grupo generado por el algoritmo de agrupamiento identifica una cantidad N de genes, pero estos genes pueden aparecer representados por múltiples registros dentro del mismo grupo. Por ejemplo, la posición 15946089 puede ser la primera posición de un gen y repetirse múltiples veces con pequeñas variaciones en cada registro. La Tabla 6.X ilustra esta situación, mostrando múltiples registros con el mismo punto de inicio, pero con ligeras diferencias en sus demás posiciones.

Tabla 6.12 Ejemplo de registros redundantes de un gen dentro de un grupo

StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
15946089	15946130	15946516	15946129	15946438	15946566
15946089	15946130	15946516	15946129	15946441	15946566
15946089	15946130	15946516	15946129	15946444	15946566
247913721	247913760	247914148	247913759	247914065	247914189
247913721	247913762	247914142	247913759	247914070	247914183
247913721	247913762	247914143	247913759	247914067	247914190

Esta redundancia residual, si no se resuelve, dificulta la evaluación del error porque no existe una correspondencia directa entre los registros del grupo y los genes de referencia. Intentar realizar agrupamientos adicionales para aumentar la reducción del espacio sin resolver previamente este problema resultaría en pérdida de información relevante, debido a que los registros redundantes podrían distribuirse en distintos grupos durante las iteraciones subsecuentes.

Con el fin de resolver esta situación, se propuso una operación de unificación que calcula el promedio de cada columna considerando todos los registros que representan el mismo gen dentro de un grupo, obtenido así un único registro representativo. Sin embargo, este registro promedio puede contener valores decimales en algunas columnas, lo que impide su uso directo como posición exacta, dado que las posiciones en el genoma se expresan mediante valores enteros.

Por esta razón, una vez obtenido el promedio, se identifica dentro del grupo el registro real cuyas posiciones presentan la menor distancia respecto a dicho promedio, utilizándolo como representante definitivo del gen. Este enfoque garantiza que el registro seleccionado corresponda a una posición del genoma y no a una interpolación artificial derivada del proceso de promediado. La Tabla 6.13 muestra los dos registros finales seleccionados en este proceso.

Tabla 6.13 Resultado de la unificación de datos por grupos

StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
15946089	15946130	15946516	15946129	15946441	15946566
247913721	247913760	247914148	247913759	247914065	247914189

El resultado de esta operación es significativo en términos de reducción: si antes existían 500 registros para representar 50 genes, después de la unificación por grupos existen únicamente 50 registros, uno por gen. Esta reducción garantiza que la evaluación posterior sea precisa y comparable entre distintos algoritmos y configuraciones. La unificación de datos por grupos se repite para cada algoritmo y cada configuración de grupos evaluada, garantizando que todas las comparaciones se realicen bajo los mismos criterios y que los resultados sean comparables entre sí.

Es importante destacar que este proceso de unificación por grupos no modifica el espacio de búsqueda en sí mismo, sino que prepara los resultados del agrupamiento para su evaluación objetiva. La calidad del grupo seleccionado no depende únicamente de la cantidad de registros que contiene, sino también del grado de cercanía entre sus registros representativos y las posiciones reales de los genes V validadas por los expertos. Este criterio de calidad, que anteriormente solo podría evaluarse de manera visual, es el que las métricas de error descritas en la siguiente sección permiten cuantificar de forma objetiva, formal y reproducible.

6.1.2.3 Métricas de evaluación y selección del mejor grupo

Una vez establecido el proceso de unificación de datos por grupos, fue posible definir las métricas formales de error que permitirían evaluar y comparar de manera objetiva el desempeño de cada algoritmo y configuración. Las métricas seleccionadas fueron el

error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), ampliamente utilizadas en la literatura de aprendizaje automático para medir la discrepancia entre valores predichos y valores de referencia.

El error absoluto medio (MAE) calcula el promedio de las diferencias absolutas entre las posiciones identificadas por el método y las posiciones reales de la referencia, proporcionando una medida directa e interpretable del error promedio en términos de posiciones. Por ejemplo, aplicando el MAE a la Tabla 6.14, el error promedio de los cinco genes mostrados sería de 2.2 posiciones, lo que indica que en promedio el método identifica las posiciones con una desviación de 2.2 nucleótidos respecto a la referencia. La fórmula de MAE se presenta en la ecuación (Ec. 6.3).

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (\text{Ec. 6.3})$$

El error cuadrático medio (MSE) eleva al cuadrado las diferencias antes de promediarlas, penalizando con mayor fuerza los errores grandes y siendo más sensible a valores atípicos. Esta característica lo hace útil para detectar casos donde el método identifica posiciones muy alejadas de la referencia, debido a que estos errores grandes tienen un impacto desproporcionado en el valor final del MSE. En el contexto de los genes V, un error grande puede indicar que el método identificó una región incorrecta del genoma, lo que es más grave que una pequeña desviación en la posición exacta. La fórmula de MSE se presenta en la ecuación (Ec. 6.4).

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (\text{Ec. 6.4})$$

La raíz del error cuadrático medio (RMSE) es la raíz cuadrada del MSE, lo que permite interpretar el error en las mismas unidades que las posiciones originales, facilitando su comparación directa con los valores de referencia. El RMSE fue la métrica principal utilizada para la selección del mejor grupo en esta iteración, debido a que combina la sensibilidad del MSE a errores grandes con la interpretabilidad del MAE al expresar el error en posiciones o nucleótidos. Retomando el ejemplo de la Tabla 6.14, el RMSE de los cinco genes mostrados sería de 2.28 posiciones, ligeramente superior al MAE debido a la penalización de los errores mayores. La fórmula de RMSE se presenta en la ecuación (Ec. 6.5).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (\text{Ec. 6.5})$$

Donde y_i representa la posición real del gen V en la referencia de los expertos, $\hat{y}_i$ representa la posición identificada por el método tras la unificación por grupos, y n es el número de genes evaluados en el grupo.

En la comparación con la referencia, no siempre existe una posición idéntica a la identificada por el agrupamiento. Por ejemplo, la posición 15946089 en la referencia puede aparecer como 15946090 o 15946088. Para garantizar una evaluación justa, se buscó la posición más similar en la referencia para cada registro unificado. La Tabla 6.14 muestra cómo se realiza esta comparación, mostrando el vector de posiciones identificadas por el agrupamiento frente al vector de posiciones de referencia más similares, y el error absoluto resultante para cada gen.

Tabla 6.14 Ejemplo de comparación entre posiciones identificadas y referencia

Gen	Inicio_RSS (Agrupamiento)	Inicio_RSS (Referencia)	MAE	MSE	RMSE
Gen 1	15946089	15946090	1	1	1
Gen 2	247913721	247913719	2	4	2
Gen 3	312450100	312450098	2	4	2
Gen 4	89234567	89234570	3	9	3
Gen 5	456789012	456789015	3	9	3

Este proceso de comparación se repite para cada columna del vector de posiciones (Inicio_RSS, Fin_RSS, Inicio_Exon, Fin_Exon, Inicio_SP, Fin_SP), calculando el error absoluto entre cada posición identificada y su correspondiente valor en la referencia. El RMSE final del grupo se obtiene promediando estos errores sobre todas las posiciones y todos los genes evaluados. Adicionalmente, solo se consideraron los genes presentes en cada grupo para la comparación: si el grupo 6 del algoritmo K-Means identificó 55 genes, únicamente se compararon con los 55 genes de mayor similitud en la referencia.

El criterio de selección del mejor grupo combinó dos aspectos complementarios: el menor error RMSE posible y la mayor cantidad de genes identificables. Este criterio dual garantizó que el grupo seleccionado no solo fuera preciso en términos de posiciones, sino que también mantuviera una cobertura suficiente de genes reales. Un grupo con error muy bajo pero pocos genes identificados no serían útiles en la práctica, debido a que omitiría una fracción significativa de los genes V presentes en el genoma.

6.1.2.4 Selección del número de grupos

Cuando se trabaja con algoritmos de agrupamiento como K-Means, Gaussian Mixtures o Fuzzy C-Means, una de las decisiones críticas consiste en determinar el número de grupos que estructurarán los datos. Aunque en muchos estudios este número se define de manera arbitraria, una investigación rigurosa debe basarse en métodos cuantitativos y objetivos que permitan justificar la elección. La importancia de esta decisión radica en que un número inadecuado de grupos puede conducir a resultados poco interpretables, sobreajuste o pérdida de información relevante.

En esta iteración se evaluaron cuatro métodos ampliamente utilizados en la literatura para estimar el número óptimo de grupos:

- **Método del codo (Elbow Method):** consiste en graficar la suma de errores cuadráticos dentro de cada grupo (inercia) frente al número de grupos. El punto de inflexión o "codo" de la curva indica el número adecuado, ya que a partir de ese punto la ganancia en homogeneidad se reduce de manera marginal.
- **Coeficiente de Silueta:** evalúa simultáneamente la cohesión, es decir, qué tan cerca están los puntos dentro del mismo grupo, y la separación, qué tan distintos son respecto a otros grupos. Sus valores varían entre -1 y 1: valores cercanos a 1 indican agrupamientos bien definidos, mientras que valores negativos sugieren asignaciones erróneas.
- **Índice de Dunn:** busca maximizar la separación entre grupos y minimizar la dispersión interna de cada uno. Un valor alto del índice refleja una mejor partición de los datos, aunque puede ser sensible a ruido y valores atípicos.
- **Índice de Davies-Bouldin (DBI):** mide la relación entre la dispersión de cada grupo y la distancia entre los grupos más cercanos. Un valor bajo de DBI indica grupos compactos y bien separados, lo que lo convierte en una métrica confiable para la comparación de configuraciones.

La Figura 6.16 presenta los resultados con estos cuatro métodos aplicados al conjunto de datos de *Desmodus rotundus*. Los resultados fueron los siguientes:

- **Método del codo:** la curva de inercia presenta un punto de inflexión en tres grupos, lo que indica que a partir de este número las mejoras se estabilizan.
- **Coeficiente de Silueta:** el valor máximo se alcanza con dos grupos, sugiriendo que esta es la partición más cohesionada y separada.
- **Índice de Dunn:** el valor más alto se observa en nueve grupos, indicando una mayor separación entre grupos y compacidad interna.
- **Índice de Davies-Bouldin:** el valor mínimo se encuentra en diez grupos, lo que implica la mejor calidad de agrupamiento según esta métrica.

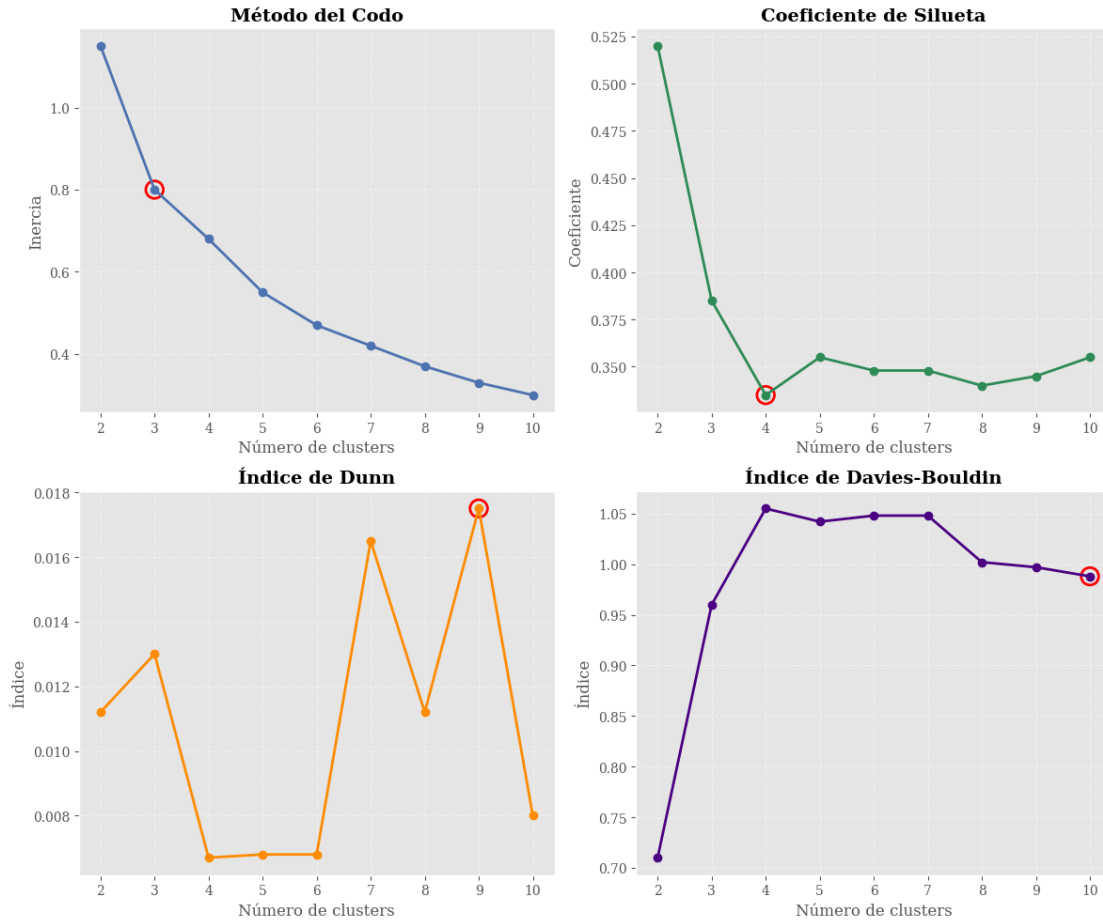


Figura 6.16 Métodos para definir el número de grupos

Esta diversidad de resultados refleja un aspecto clave en los estudios de agrupamiento: no existe un único criterio universalmente válido, sino que la decisión depende del equilibrio entre métricas estadísticas, características de los datos y conocimiento experto del dominio. En consecuencia, con el fin de asegurar una evaluación completa, se realizaron pruebas con configuraciones de 3 a 7 grupos, comparando sus resultados tanto en términos de las métricas de error definidas en la sección anterior como de la cobertura de genes V reales identificados. De este modo, la selección final no se limitó a un criterio matemático, sino que integró el juicio experto con los indicadores de calidad, garantizando una solución práctica y útil para el problema abordado.

6.1.2.5 Comparación de algoritmos

Con el proceso de unificación por grupos y las métricas de error definidas, se procedió a la comparación sistemática de cinco algoritmos de agrupamiento: K-Means, Spectral Clustering, Fuzzy C-Means, Gaussian Mixtures (GMM) y OPTICS. Cada algoritmo fue evaluado con configuraciones de 3, 4, 5, 6 y 7 grupos, aplicando el proceso de unificación por grupos y calculando las métricas MAE, MSE y RMSE para cada configuración. Los

resultados detallados de cada configuración se presentan en las Tablas de anexos de la 1 a la 5 del Anexo B.

La Figura 6.17 presenta los gráficos de dispersión de Spectral Clustering y K-Means para la configuración de 5 grupos, donde la referencia aparece representada como el grupo 10. Para mejorar los gráficos de los datos en dos dimensiones, se aplicó el algoritmo UMAP como técnica de reducción de dimensionalidad previa a la graficación, lo que permite apreciar con mayor claridad la distribución y separación de los grupos en el espacio de representación.

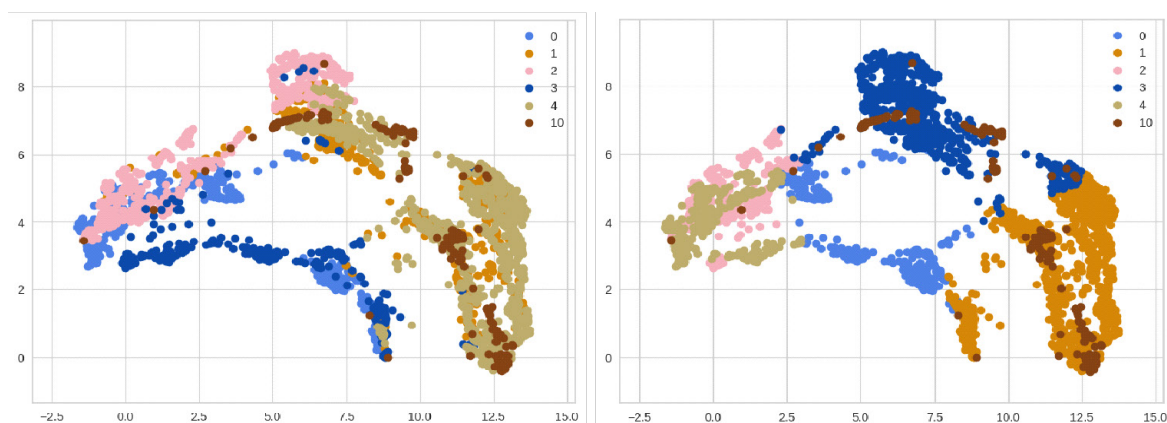


Figura 6.17 Gráficos de dispersión de Spectral clustering y K-Means

La Figura 6.18 presenta los gráficos de dispersión de Fuzzy C-Means y Gaussian Mixtures bajo la misma configuración y con la misma reducción de dimensionalidad aplicada. Como se puede observar, Gaussian Mixtures muestra una distribución más flexible de los grupos que se adapta mejor a la estructura heterogénea de los datos de alineamiento, mientras que Fuzzy C-Means presenta zonas de solapamiento entre grupos que reflejan su naturaleza de asignación probabilística suave.

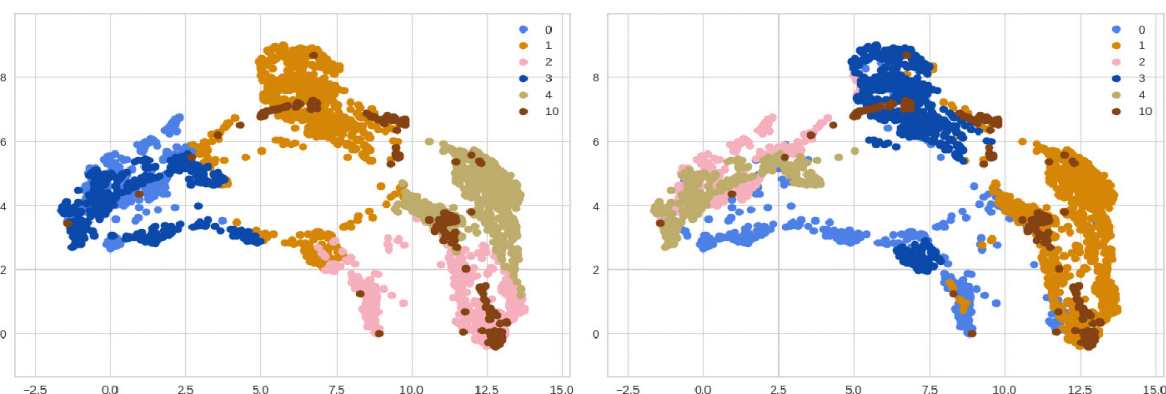


Figura 6.18 Gráficos de dispersión de Fuzzy C-Means y Gaussian Mixtures

El análisis de los resultados reveló patrones consistentes a lo largo de las distintas configuraciones de grupos. Para configuraciones de 3 grupos, Gaussian Mixtures logró identificar 58 genes con un RMSE de 5.15, siendo el algoritmo con mayor cobertura en esta configuración. Para 4 grupos, Spectral Clustering mostró resultados competitivos identificando 53 genes con un RMSE de 4.91. Para 5 grupos, Spectral Clustering alcanzó su mejor resultado con 55 genes identificados y el menor RMSE de todas las pruebas con 3.91. Para 6 y 7 grupos, Gaussian Mixtures identificó consistentemente 52 genes con errores RMSE de 5.35 y 5.33 respectivamente.

Es importante destacar que el criterio de selección no se basó únicamente en el menor error, sino en el equilibrio entre precisión y cobertura. Un grupo con error muy bajo pero pocos genes identificados no resulta útil en la práctica, debido a que omitiría una fracción significativa de los genes V presentes en el genoma. Por ello, se estableció como criterio mínimo la identificación de al menos 50 genes, combinado con el menor RMSE posible.

La Tabla 6.15 presenta los mejores resultados por cada configuración a lo largo de todas las pruebas realizadas, ordenados de mejor a peor resultado en términos de RMSE. Esta tabla constituye el punto de referencia principal para la evaluación del progreso del método a lo largo de las iteraciones, debido a que los resultados aquí presentados serán comparados con los obtenidos en la siguiente iteración para verificar si la incorporación de técnicas de reducción de dimensionalidad mejora la identificación de genes V. Es importante señalar que la Tabla 6.15 incluye no solo las métricas de error sino también las métricas de calidad del agrupamiento: el Silhouette Score, el índice de Davies-Bouldin y el índice de Calinski-Harabasz.

Tabla 6.15 Mejores resultados de los algoritmos para el conjunto de datos con todas sus dimensiones

Algs.	grps	N. datos	N. genes	N. genes ref.	MAE	RMSE	Aciertos	SS	DBI	CHI
SPC 3 grupos	2	1536	58	71	2.833	5.150	81.690	0.121	2.369	775.29
SPC 4 grupos	2	683	53	71	3.160	4.912	74.647	0.202	2.233	1591.17
SPC 5 grupos	5	979	55	71	2.266	3.915	77.464	0.365	1.145	2899.30
GMM 6 grupos	2	1096	52	71	3.125	5.347	73.239	0.340	1.137	2117.14
GMM 7 grupos	2	1096	52	71	3.112	5.330	73.239	0.271	1.541	1348.65

Estas métricas permiten evaluar la calidad intrínseca de los grupos obtenidos, independientemente de la referencia de los expertos, proporcionando una visión más completa del comportamiento de cada algoritmo. Un Silhouette Score (SS) alto indica grupos bien definidos y separados, un Davies-Bouldin Index (DBI) bajo indica grupos

compactos y distantes entre sí, y un Calinski-Harabasz Index (CHI) alto indica una buena relación entre la dispersión interna y externa de los grupos.

La combinación de métricas de error respecto a la referencia y métricas de calidad del agrupamiento permite una evaluación más robusta y completa que cualquiera de estas perspectivas por sí sola. Un algoritmo que minimiza el error respecto a la referencia, pero genera grupos de baja calidad estadística puede ser menos confiable que uno que equilibra ambos criterios de manera consistente, cuando se aplica a genomas con características distintas a las del conjunto de entrenamiento.

6.1.2.6 Selección de GMM

Los resultados presentados en la Tabla 6.15 muestran que Spectral Clustering obtuvo el menor RMSE de todas las pruebas con un 3.91 con 5 grupos, posicionándose como el algoritmo más preciso en términos estadísticos para el conjunto de datos de Desmodus rotundus. Sin embargo, la selección del algoritmo para el método RIVAM no podía basarse únicamente en el desempeño sobre un conjunto de datos específico, sino que debía considerar su viabilidad práctica en el contexto real de aplicación del método.

Durante las pruebas de escalabilidad realizadas con conjuntos de datos de mayor volumen, superiores a 10,000 registros, *Spectral Clustering* presentó un desbordamiento de pila que impedía su ejecución. Este problema es inherente a la naturaleza del algoritmo: *Spectral Clustering* requiere la construcción y descomposición de una matriz de similitud de tamaño $n \times n$, donde n es el número de registros, lo que hace que su costo computacional y su consumo de memoria crezcan de manera cuadrática con el tamaño del conjunto de datos. En el contexto del método RIVAM, donde los conjuntos de datos de genomas reales pueden superar fácilmente los 10,000 registros tras la unificación, esta limitación lo hace inviable como algoritmo central del método.

Gaussian Mixtures, en cambio, no presentó este problema durante las pruebas de escalabilidad. Al ser un modelo inductivo basado en la estimación de parámetros mediante el algoritmo *Expectation-Maximization* (EM), su costo computacional crece de manera lineal con el número de registros, lo que lo hace aplicable a conjuntos de datos de cualquier tamaño dentro del rango esperado para los genomas analizados en esta investigación. Adicionalmente, su naturaleza probabilística permite asignar a cada registro una probabilidad de pertenencia a cada grupo, lo que facilita la identificación objetiva del grupo con mayor concentración de genes V reales mediante criterios cuantitativos reproducibles.

Por estas razones, y a pesar de que Spectral Clustering mostró un RMSE ligeramente menor en el conjunto de prueba, se seleccionó Gaussian Mixtures como el algoritmo de agrupamiento central del método RIVAM. Esta decisión garantizó que el método fuera aplicable de manera consistente a genomas de distintos tamaños y especies, cumpliendo con el principio de reusabilidad establecido en el Capítulo 5.

6.1.2.7 Limitación identificada

La segunda iteración representó un avance significativo respecto a la primera, al incorporar métricas formales de error que permitieron evaluar y comparar de manera objetiva el desempeño de los algoritmos de agrupamiento. Sin embargo, el análisis de los resultados reveló una nueva limitación que motivó el diseño de la tercera iteración.

Si bien la normalización L2 aplicada en esta iteración mejoró la distribución de los datos respecto a los datos en escala original, los resultados sugirieron que esta técnica de normalización no era la más adecuada para el problema. La normalización L2 ajusta cada registro a un vector de norma unitaria, lo que garantiza que todos los registros tengan la misma magnitud global, pero no necesariamente preserva de manera óptima las relaciones relativas entre los atributos de cada registro. En el contexto de los genes V, donde las distancias relativas entre las secciones del gen son más informativas que sus posiciones absolutas, esta limitación se traduce en una separabilidad subóptima entre el grupo de genes V reales y el resto de los candidatos.

Adicionalmente, el proceso de unificación por grupos basado en el promedio, si bien resolvió el problema de la redundancia residual dentro de los grupos, introdujo una fuente adicional de imprecisión: el registro promedio puede no ser el más representativo del gen cuando la distribución de los registros dentro del grupo es asimétrica o presenta valores atípicos. Esta observación motivó la exploración de estrategias alternativas de unificación por grupos en la siguiente iteración.

En conjunto, estas dos limitaciones, la normalización L2 insuficiente y la estrategia de unificación por promedio, motivaron el diseño de la tercera iteración, orientada a explorar técnicas de normalización alternativas y estrategias de unificación más robustas que mejoraran la calidad del agrupamiento y la precisión de las posiciones identificadas.

6.1.3 Tercera iteración: normalizaciones y unificación por grupos

Los avances obtenidos en la segunda iteración demostraron que la incorporación de métricas formales de error y la comparación sistemática de algoritmos permitían tomar decisiones más fundamentadas sobre el diseño del método. Sin embargo, dos limitaciones identificadas al final de esa iteración motivaron la exploración de nuevas estrategias: la normalización L2 no garantizaba una separabilidad óptima entre el grupo de genes V reales y el resto de los candidatos, y la unificación por grupos basada en el promedio introducía imprecisiones cuando la distribución de registros dentro de un grupo era asimétrica.

Con el objetivo de abordar estas limitaciones, la tercera iteración se orientó hacia dos objetivos complementarios: la exploración de técnicas de normalización alternativas a L2 y la evaluación de estrategias de unificación por grupos más robustas. Adicionalmente, en esta iteración se incorporaron dos nuevos conjuntos de datos para ampliar la evaluación del método más allá del genoma de *Desmodus rotundus* utilizado en las iteraciones anteriores.

6.1.3.1 Motivación

La normalización L2, si bien es una técnica ampliamente utilizada en aprendizaje automático, presenta una limitación específica en el contexto de los datos de alineamiento de genes V: al ajustar cada registro a un vector de norma unitaria, normaliza la magnitud global del vector, pero no necesariamente preserva de manera óptima las relaciones relativas entre los atributos individuales. En el caso de los genes V, donde las distancias relativas entre las secciones del gen son más informativas que sus magnitudes absolutas, esta característica puede reducir la separabilidad entre grupos. Esto se evidenció en los resultados de la segunda iteración, donde a pesar de que la normalización L2 mejoró la distribución de los datos respecto a la escala original, los valores de RMSE obtenidos sugirieron que existía margen de mejora en la calidad del agrupamiento.

Con el objetivo de abordar esta limitación, se propuso explorar técnicas de normalización alternativas que pudieran preservar mejor las relaciones entre atributos relevantes para el problema. La biblioteca scikit-learn ofrece un conjunto de funciones de normalización con distintos comportamientos y supuestos, cada una adecuada para diferentes tipos de datos y problemas (Pedregosa et al., 2011). La comparación sistemática de estas técnicas sobre los datos de alineamiento de genes V permitiría identificar cuál de ellas produce la mejor separabilidad entre el grupo de genes V reales

y el resto de los candidatos, mejorando así la calidad del agrupamiento y reduciendo el error respecto a la referencia.

Por su parte, la unificación por grupos basada en el promedio presentó limitaciones cuando los registros dentro de un grupo no seguían una distribución simétrica. En estos casos, el promedio puede verse influenciado por valores atípicos o por la presencia de registros que representan genes distintos agrupados en la misma partición, produciendo un registro representativo que no corresponde fielmente al comportamiento central del grupo. Esta imprecisión se propaga directamente al cálculo del error respecto a la referencia, debido a que el registro utilizado para la comparación no es el más representativo del gen. Para resolver esta limitación, se propuso sustituir el promedio por el segundo cuartil, también conocido como mediana, que es una medida de tendencia central más robusta frente a valores atípicos y más representativa del comportamiento central de distribuciones asimétricas.

En esta iteración se incorporó un nuevo conjunto de datos con características distintas al de *Desmodus rotundus* utilizado en las iteraciones anteriores: el conjunto de *Eptesicus nilssonii* (EPNI), con 321,139 registros originales y 231 genes anotados manualmente como referencia. La incorporación de este nuevo genoma respondió a la necesidad de verificar que las mejoras introducidas en esta iteración fueran consistentes entre distintas especies y no específicas al conjunto de entrenamiento original, avanzando así hacia el principio de reusabilidad establecido en el Capítulo 5.

6.1.3.2 Comparación de técnicas de normalización

La normalización de los datos constituye una etapa crítica dentro del pipeline del método RIVAM, debido a que las técnicas de agrupamiento utilizadas son altamente sensibles a la escala, dispersión y distribución de los atributos. En el contexto de esta investigación, los registros derivados de alineamientos presentan características heterogéneas, incluyendo diferencias significativas en magnitud, alta dispersión y presencia de valores extremos. Estas propiedades pueden alterar las métricas de similitud utilizadas durante el agrupamiento, afectando directamente la estabilidad de los grupos y la coherencia biológica de los resultados.

Adicionalmente, los datos utilizados en el método RIVAM poseen una naturaleza parcialmente dispersa, debido a que determinados atributos representan ausencia de coincidencias, regiones no alineadas o distancias entre componentes del gen V. En consecuencia, la forma en que cada técnica transforma los atributos puede modificar la estructura geométrica del espacio de representación y alterar las relaciones relativas entre registros biológicos. Por esta razón, se realizó una comparación experimental de distintas estrategias de normalización con el objetivo de identificar aquella que

preservara de mejor manera las relaciones estructurales entre genes V y favoreciera la generación de agrupamientos consistentes mediante modelos GMM.

En esta evaluación se seleccionaron cinco técnicas de normalización disponibles en la biblioteca scikit-learn, considerando tanto métodos tradicionales de escalado como enfoques robustos orientados al tratamiento de datos dispersos y valores atípicos. La Tabla 6.16 resume las principales características, ventajas, limitaciones observadas y comportamiento experimental de cada técnica evaluada.

Tabla 6.16 Comparativa de técnicas de normalización evaluadas en el método RIVAM

Método	Ventaja	Desventaja	Resultado
StandardScaler	Homogeneiza escalas.	Sensible a outliers.	Agrupamientos inestables.
MinMaxScaler	Escala datos a [0,1].	Outliers reducen la separación.	Menor discriminación.
MaxAbsScaler	Preserva relaciones relativas.	No centra los datos.	Mejores agrupamientos.
Normalizer	Homogeneiza magnitudes.	Reduce diferencias relevantes.	Baja separación.
RobustScaler	Robusto ante outliers.	No preserva toda la estructura.	Resultados intermedios.

Los resultados experimentales mostraron que MaxAbsScaler produjo el comportamiento más favorable dentro del pipeline de RIVAM. A diferencia de los métodos que centran o redistribuyen los datos, esta técnica preservó la estructura dispersa de los registros y mantuvo los valores cero asociados a ausencia de coincidencias o relaciones biológicas relevantes. Experimentalmente, esta propiedad permitió conservar de mejor manera las relaciones geométricas entre atributos derivados de los alineamientos de secuencias, favoreciendo la estabilidad del agrupamiento mediante GMM y mejorando la separación entre grupos candidatos.

Asimismo, el uso de MaxAbsScaler contribuyó a obtener representaciones más adecuadas para las etapas posteriores de reducción de dimensionalidad y selección de grupos, reduciendo la distorsión introducida por otros métodos de normalización. A partir de estos resultados, esta técnica fue seleccionada como el método de normalización utilizado en las iteraciones posteriores de RIVAM y en las variantes finales evaluadas experimentalmente en los capítulos subsecuentes.

6.1.3.3 Segundo cuartil como estrategia de unificación

Paralelamente a la exploración de técnicas de normalización, se evaluó una estrategia alternativa para la unificación de datos por grupos. Como se describió en la Sección 6.1.2.2, el proceso de unificación por grupos consiste en obtener un único registro representativo por gen dentro de cada grupo, con el objetivo de facilitar la comparación con la referencia generada por los expertos y el cálculo de las métricas de error.

En la segunda iteración del método, este proceso se realizó con el cálculo del promedio de las posiciones de todos los registros que representaban el mismo gen dentro de cada grupo. Si bien este enfoque es simple y computacionalmente eficiente, presenta una limitación importante: el promedio es altamente sensible a la presencia de valores atípicos y puede dejar de ser representativo cuando la distribución de los registros es asimétrica. En el contexto de los datos de alineamiento de genes V, es común encontrar variaciones importantes en las posiciones de inicio y fin de los registros debido a diferencias entre herramientas de alineamiento, fragmentación de regiones o desplazamientos generados durante el proceso de anotación. Estas variaciones pueden provocar que algunos registros extremos desplacen el promedio y generen una representación poco precisa del gen dentro del grupo.

Con el objetivo de reducir este problema, se evaluaron diferentes estadísticos de tendencia central y dispersión como estrategias de unificación por grupos, incluyendo medidas como la media, moda, promedio, primer cuartil, tercer cuartil y segundo cuartil. Cada una de estas estrategias fue analizada experimentalmente con el propósito de determinar cuál generaba un registro representativo más preciso y estable para los genes V dentro de cada grupo.

Entre las estrategias evaluadas, el segundo cuartil, también conocido como mediana, mostró el comportamiento más favorable. El segundo cuartil corresponde al valor central de una distribución ordenada, dividiendo el conjunto de datos en dos partes iguales: el 50% de los registros queda por debajo de este valor y el otro 50% por encima. A diferencia del promedio, el segundo cuartil no se ve afectado significativamente por valores extremos, lo que lo convierte en una medida de tendencia central más robusta para distribuciones heterogéneas o asimétricas, características frecuentes en los datos derivados de alineamientos de secuencias.

Dentro del método RIVAM, el uso del segundo cuartil permitió seleccionar registros representativos más cercanos al comportamiento central real de cada grupo, evitando desplazamientos artificiales generados por alineamientos atípicos o posiciones extremas. Experimentalmente, esta estrategia produjo agrupamientos más estables y una reducción consistente del error en comparación con las estrategias basadas en el promedio y otros estadísticos evaluados, razón por la cual fue incorporada en las iteraciones posteriores del método.

Por ejemplo, si un grupo contiene registros con posiciones de inicio correspondientes a 100, 101, 102 y 300, el promedio produciría un valor desplazado debido a la influencia del registro extremo 300. En cambio, el segundo cuartil selecciona el valor central de la distribución ordenada, manteniendo una representación más cercana al comportamiento predominante del grupo. Este comportamiento resulta particularmente importante en los alineamientos, donde pequeños desplazamientos

pueden impactar directamente en el cálculo de error y en la evaluación biológica de los resultados.

La aplicación del segundo cuartil como estrategia de unificación por grupos demostró una reducción del error de algunas milésimas en comparación con el método basado en el promedio. Si bien esta mejora puede parecer pequeña en términos absolutos, representa un avance significativo desde el punto de vista metodológico, debido a que confirma que la selección de una medida de tendencia central robusta mejora la precisión del registro representativo y, por tanto, la calidad de la evaluación del agrupamiento. Esta mejora fue consistente entre los distintos conjuntos de datos evaluados, particularmente en los genomas *Pipistrellus* y EPNI, donde la combinación del segundo cuartil con la normalización *MaxAbsScaler* permitió obtener agrupamientos con menor RMSE y una mejor aproximación a las referencias generadas por los expertos.

En consecuencia, el segundo cuartil fue incorporado como parte de las configuraciones posteriores del método RIVAM, debido a su capacidad para generar registros representativos más estables, robustos y coherentes con la estructura biológica de los genes V presentes en los alineamientos.

6.1.3.4 Prueba con *eptesicus nilssonii*

Las pruebas de esta iteración se llevaron a cabo con el conjunto de datos de *Eptesicus nilssonii* (EPNI), un quiróptero cuyo genoma cuenta con una referencia de 231 genes anotados manualmente y 11,485 registros para el entrenamiento del algoritmo, de los cuales 200 genes pudieron ser utilizados para la evaluación. Este conjunto fue seleccionado como caso principal de validación de las mejoras introducidas en esta iteración debido a que contiene la mayor cantidad de genes en comparación con los demás conjuntos evaluados, lo que lo convierte en un escenario más representativo y exigente para demostrar la efectividad de las modificaciones realizadas. Adicionalmente, el volumen de registros de EPNI permitió evaluar el comportamiento del método en un espacio de búsqueda significativamente mayor al utilizado en las iteraciones anteriores con *Desmodus rotundus*.

Una de las observaciones más importantes de esta iteración fue que el algoritmo Spectral Clustering no pudo ser ejecutado sobre el conjunto de datos de EPNI debido a que el equipo utilizado para el entrenamiento no soportó el procesamiento requerido, presentando un desbordamiento de pila. Este resultado no fue inesperado: como se mencionó en la sección 6.1.2.6, Spectral Clustering presenta una complejidad computacional cuadrática respecto al número de registros, lo que lo hace inviable para conjuntos de datos que superen cierto umbral de tamaño. La imposibilidad de ejecutar Spectral Clustering sobre EPNI confirmó empíricamente esta limitación y validó

definitivamente la decisión de seleccionar GMM como algoritmo central del método RIVAM, ya que este último no presentó ningún problema de ejecución sobre el mismo conjunto de datos y produjo resultados consistentes y reproducibles.

La Tabla 6.17 presenta los resultados del algoritmo GMM aplicado al conjunto de EPNI comparando dos estrategias de unificación: el promedio y el segundo cuartil, ambas con normalización MaxAbsScaler. La Tabla permite observar de manera directa el impacto de cada estrategia sobre las métricas de error en cada uno de los cinco grupos generados. En esta prueba se configuró GMM con cinco grupos, número que corresponde a la configuración que produjo el mejor resultado en la iteración anterior con Spectral Clustering, lo que permite una comparación directa entre ambas iteraciones y facilita la evaluación del impacto de las modificaciones introducidas.

El hallazgo más relevante se observa en el grupo 1, que es el grupo de mayor interés por contener 87 genes identificados de 200 de referencia, representando la mayor cobertura de genes reales de todos los grupos. Con la estrategia de promedio, este grupo presentaba un RMSE de 4.62 y un MAE de 2.66. Al sustituir el promedio por el segundo cuartil, el RMSE se redujo a 4.36 y el MAE a 2.33, manteniendo exactamente la misma cantidad de genes identificados. Este resultado es significativo porque demuestra que el segundo cuartil mejora la precisión de las posiciones identificadas sin sacrificar la cobertura de genes reales, que es precisamente el equilibrio que el método RIVAM busca optimizar.

Tabla 6.17 Grupos del algoritmo GMM con los datos de EPNI con el segundo cuartil y MaxAbsScaler

Algoritmo	Grupo	N. datos	N. genes	N. genes ref	MAE	MSE	RMSE
Gaussian Mixture promedio	1	5134	87	200	2.66	21.38	4.62
	2	4154	62	200	2.47	13.52	3.68
	3	1430	48	200	3.94	47.28	6.88
	4	499	63	200	3.49	34.50	5.87
	5	268	12	200	5.42	64.75	8.05
Gaussian Mixture segundo cuartil	1	5134	87	200	2.33	19.02	4.36
	2	4154	62	200	2.15	10.75	3.28
	3	1430	48	200	3.91	47.55	6.90
	4	499	63	200	3.53	35.85	5.99
	5	268	12	200	5.25	64.22	8.01

El grupo 2, con 62 genes identificados, mostró la mejora porcentual más significativa: el RMSE se redujo de 3.68 a 3.28 y el MAE de 2.47 a 2.15, representando una reducción de aproximadamente un 10% en el error. Los grupos 3, 4 y 5 presentaron mejoras de menor magnitud, con algunos casos donde el MSE mostró ligeros incrementos compensados por reducciones en el MAE, lo que sugiere que en estos grupos la distribución de los registros es más homogénea y ambas estrategias producen resultados similares.

En conjunto, los resultados confirman que el segundo cuartil produce mejoras consistentes en los grupos con mayor cobertura de genes reales, validando su adopción como estrategia de unificación estándar para las iteraciones posteriores del método.

La Figura 6.19 presenta los gráficos de dispersión obtenidos a partir de los datos normalizados con MaxAbsScaler y procesados mediante UMAP para su representación en dos dimensiones. La aplicación de UMAP como técnica de reducción de dimensionalidad previa a la graficación permite proyectar el espacio de representación de alta dimensionalidad en un plano bidimensional, preservando las relaciones de similitud entre registros y facilitando la interpretación visual de la distribución de los grupos generados por GMM.

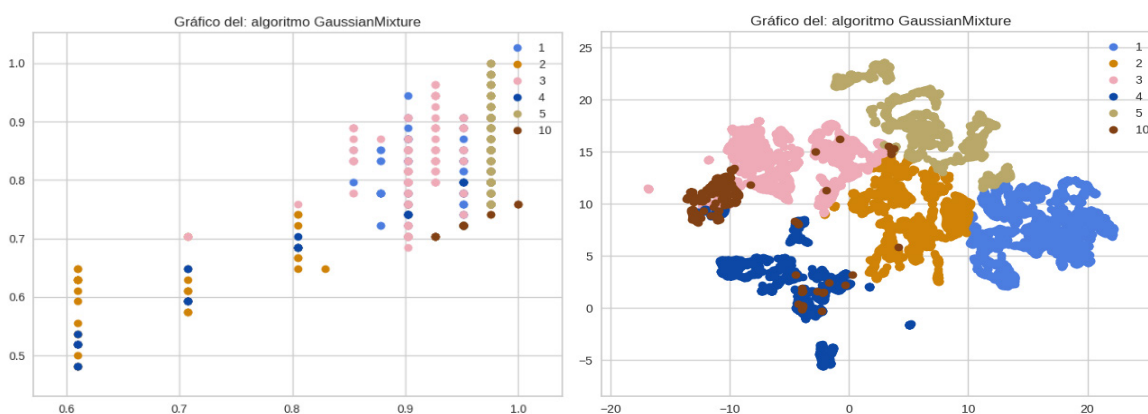


Figura 6.19 Gráficos de dispersión del conjunto EPNI con MaxAbsScaler y segundo cuartil

En estos gráficos se pueden identificar claramente cinco grupos con características diferenciadas en términos de tamaño, densidad y distribución espacial. Los grupos 3 y 4 muestran una mayor proximidad al grupo de referencia en el espacio de representación, lo que sugiere que estos grupos contienen la mayor concentración de genes V reales del genoma analizado. Esta observación es consistente con los resultados de la Tabla 6.17, donde los grupos 1 y 2 presentan los mayores números de genes identificados, confirmando que la proximidad visual al grupo de referencia en el espacio UMAP es un indicador confiable de la relevancia biológica del grupo.

La mejora observada en esta iteración puede atribuirse en gran medida a la combinación de tres factores: la normalización MaxAbsScaler, que preserva mejor las relaciones relativas entre los atributos de distancia del gen V; el segundo cuartil como criterio de unificación, que produce registros representativos más robustos frente a valores atípicos; y la mayor cantidad de datos disponibles en EPNI, que permite al algoritmo GMM identificar patrones más consistentes en el espacio de representación.

6.1.3.5 Limitación identificada

La tercera iteración representó un avance significativo respecto a las anteriores, al demostrar que la combinación de MaxAbsScaler y segundo cuartil mejoraba consistentemente la precisión del método sin comprometer la cobertura de genes reales. Sin embargo, el análisis de los resultados reveló una nueva limitación que motivó el diseño de la cuarta iteración.

Si bien la configuración de MaxAbsScaler y segundo cuartil mejoró las métricas de error respecto a las iteraciones anteriores, los valores de RMSE obtenidos aún presentaban un margen de mejora considerable. El grupo 1 de EPNI, que es el grupo de mayor interés por contener 87 genes identificados, presentaba un RMSE de 4.36, lo que indica que las posiciones identificadas se desviaban en promedio 4.36 posiciones de la referencia de los expertos. Este nivel de error, si bien aceptable como primera aproximación, podía reducirse con una optimización más sistemática de los hiperparámetros del algoritmo GMM.

La configuración de GMM utilizada hasta este punto se basaba en parámetros definidos de manera empírica, es decir, sin un proceso formal de búsqueda del valor óptimo para cada hiperparámetro. El número de componentes, el tipo de matriz de covarianza y otros parámetros del modelo se seleccionaban con base en la prueba y error, lo que no garantizaba que la configuración elegida fuera adecuada para cada conjunto de datos. Esta ausencia de un proceso formal de optimización de hiperparámetros representaba la principal limitación de la tercera iteración y motivó la incorporación de una técnica de optimización metaheurística en la iteración siguiente.

Adicionalmente, los resultados sugirieron que la distribución de los datos en el espacio de representación original, aunque mejorada por MaxAbsScaler, podía beneficiarse de una transformación adicional que mejorara la separabilidad entre el grupo de genes V reales y el resto de los candidatos. Esta observación motivó la exploración de técnicas de reducción de dimensionalidad como PCA y UMAP en la cuarta iteración, con el objetivo de identificar una representación del espacio de datos que facilitara aún más la tarea del algoritmo de agrupamiento.

6.1.4 Cuarta iteración: incorporación de reducción de dimensionalidad

La tercera iteración demostró que la combinación de MaxAbsScaler y segundo cuartil mejoraba consistentemente la precisión del método, pero evidenció dos limitaciones que debían abordarse de manera simultánea: los hiperparámetros del algoritmo GMM se definían de manera empírica sin un proceso formal de optimización, y la distribución de los datos en el espacio de representación original podía beneficiarse de una

transformación adicional que mejorara la separabilidad entre el grupo de genes V reales y el resto de candidatos. Estas dos limitaciones motivaron el diseño de la cuarta iteración, cuya estructura general se muestra en la Figura 6.20.

Como se puede observar en la Figura 6.20, la cuarta iteración se organiza en tres etapas principales. La primera corresponde al preprocesamiento común, que mantiene la unificación de datos y la normalización con MaxAbsScaler establecidas en la iteración anterior, obteniendo un conjunto de datos normalizado listo para el agrupamiento. La segunda etapa corresponde al agrupamiento experimental con tres modelos distintos, donde PSO actúa como mecanismo de optimización global de hiperparámetros para cada modelo de manera colaborativa. La tercera etapa corresponde a la evaluación y comparación de los modelos con métricas de desempeño como el RMSE, con el objetivo de seleccionar la configuración con mejor desempeño general.

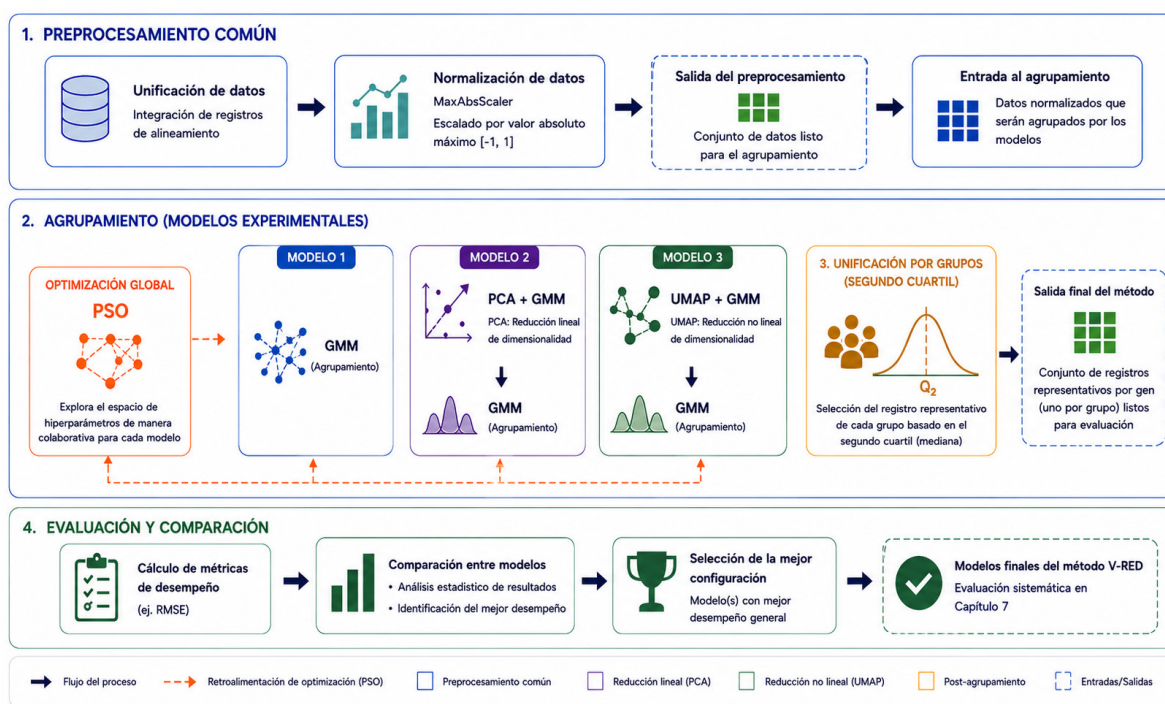


Figura 6.20 Cuarta iteración del método RIVAM

Con el objetivo de abordar la primera limitación, se incorporó el algoritmo de Optimización por Enjambre de Partículas (PSO, por sus siglas en inglés *Particle Swarm Optimization*), una técnica bioinspirada que explora el espacio de hiperparámetros de manera colaborativa mediante un conjunto de partículas que se desplazan hacia regiones de mayor desempeño (Selvaraj & Choi, 2022). La elección de PSO sobre otras técnicas de optimización, como la búsqueda en cuadrícula o la búsqueda aleatoria, se fundamentó en su capacidad para explorar eficientemente espacios multimodales sin necesidad de evaluar todas las combinaciones posibles, característica relevante cuando el costo computacional de cada evaluación es significativo.

La segunda limitación se abordó mediante la evaluación de dos técnicas de reducción de dimensionalidad con características complementarias: el *Principal Component Analysis* (PCA), que realiza una reducción lineal preservando la mayor proporción posible de la varianza de los datos, y *Uniform Manifold Approximation and Projection* (UMAP), que lleva a cabo una reducción no lineal preservando tanto la estructura local como parte de la estructura global del espacio de datos. La combinación de estas técnicas con PSO y GMM permitió analizar si la transformación del espacio de representación mejoraba adicionalmente la calidad del agrupamiento.

Esta iteración exploró tres modelos utilizando el genoma de *Tadarida brasiliensis* como conjunto de datos base para la optimización de hiperparámetros. Este genoma fue seleccionado por contar con una referencia de 1,006 genes validados por expertos, siendo el conjunto con mayor cantidad de genes de referencia disponible, lo que lo convierte en el escenario más representativo y exigente para calibrar los parámetros del método.

El primer modelo, denominado Modelo GMM, corresponde a un modelo de mezclas gaussianas (GMM) con optimización de hiperparámetros mediante PSO sobre el espacio original de datos, sin incorporar técnicas previas de reducción de dimensionalidad. Esta configuración representa la variante base con optimización formal de hiperparámetros y sirve como punto de referencia para evaluar el impacto de la reducción de dimensionalidad en los modelos posteriores.

El segundo modelo, denominado Modelo PCA-GMM, incorpora PCA como etapa previa al agrupamiento, realizando una reducción lineal de dimensionalidad antes de aplicar GMM. En esta configuración, PSO optimizó conjuntamente los hiperparámetros de PCA y GMM, buscando la combinación óptima entre el número de componentes principales y la configuración del modelo de mezclas gaussianas.

El tercer modelo, denominado Modelo UMAP-GMM, sustituye PCA por UMAP como técnica de reducción de dimensionalidad, realizando una reducción no lineal que preserva la estructura local y, en cierta medida, la estructura global de los datos. Al igual que en el modelo anterior, PSO optimizó conjuntamente los hiperparámetros de ambos algoritmos, evaluando si la reducción no lineal proporcionada por UMAP producía una mejor separabilidad entre grupos en comparación con la reducción lineal obtenida con PCA.

Los resultados de estos tres modelos, presentados en las secciones siguientes, junto con el modelo GMM base desarrollado en la iteración anterior, permitieron consolidar las cuatro variantes finales del método, las cuales serán evaluadas de manera sistemática en el Capítulo 7.

6.1.4.1 Motivación para la mejora de la separabilidad con la reducción de la dimensionalidad

Las iteraciones anteriores demostraron que el método era capaz de identificar genes V reales con una precisión aceptable, pero el análisis de los resultados reveló que la distribución de los datos en el espacio de representación original presentaba características que limitaban la separabilidad entre el grupo de genes V reales y el resto de los candidatos. En particular, los 43 atributos que conforman la representación vectorial del gen V, derivados de la vectorización de las distancias entre las posiciones de inicio y fin de las tres secciones del gen, obtienen un espacio de representación donde las diferencias entre registros relevantes e irrelevantes no siempre son fácilmente detectables por el algoritmo de agrupamiento.

Este problema es conocido en la literatura de aprendizaje automático como la maldición de la dimensionalidad (Bellman, 1957): a medida que el número de atributos crece, las distancias entre puntos en el espacio tienden a homogeneizarse, dificultando la identificación de patrones y la separación entre grupos. En el caso de los datos de alineamiento de genes V, donde los 43 atributos incluyen tanto posiciones absolutas como distancias relativas entre secciones, esta homogeneización puede reducir la capacidad del algoritmo GMM para distinguir entre el grupo de genes V reales y los grupos de registros irrelevantes.

La reducción de dimensionalidad ofrece una solución a este problema al transformar el espacio de representación original en un espacio de menor dimensión que preserve las relaciones estructurales más relevantes entre registros, mejorando la separabilidad entre grupos. Para esta iteración se evaluaron dos técnicas con enfoques complementarios, ambas descritas en el Marco Conceptual del Capítulo 2.

PCA fue seleccionada como primera alternativa por su capacidad de capturar las direcciones de mayor varianza en los datos con una transformación lineal. En el contexto de los datos de alineamiento de genes V, donde las distancias relativas entre secciones del gen presentan patrones de varianza consistentes entre distintos genomas, PCA puede identificar las combinaciones lineales de atributos que mejor discriminan entre grupos, reduciendo el ruido introducido por atributos menos informativos.

UMAP fue seleccionada como segunda alternativa por su capacidad de capturar estructuras no lineales en los datos, preservando tanto las relaciones locales entre registros similares como la estructura global del espacio de datos. A diferencia de PCA, UMAP no asume que las relaciones entre atributos son lineales, lo que le permite adaptarse mejor a la estructura compleja y heterogénea de los datos de alineamiento. Esta característica es relevante en el contexto de los genes V, donde las relaciones entre

las secciones del gen pueden presentar patrones no lineales que PCA no puede capturar completamente.

La evaluación comparativa de PCA y UMAP en combinación con GMM y PSO permitió determinar cuál de estas técnicas produce la representación más favorable para el agrupamiento en el contexto específico de los datos de alineamiento de genes V, proporcionando la evidencia empírica necesaria para justificar la selección de las variantes finales del método.

6.1.4.2 Función objetivo

La optimización con PSO requirió definir dos elementos fundamentales: la configuración del algoritmo y la función objetivo que guía la búsqueda de hiperparámetros óptimos.

Configuración de PSO, este es un algoritmo bioinspirado en el comportamiento social de bandadas de aves o cardúmenes de peces, donde cada partícula del enjambre representa una posible configuración de hiperparámetros y se desplaza en el espacio de búsqueda influenciada por su mejor posición histórica y la mejor posición encontrada por el enjambre (Selvaraj & Choi, 2022). La configuración del algoritmo PSO utilizada en esta investigación se presenta en la Tabla 6.1.

Tabla 6.18 Configuración del algoritmo PSO

Hiperparámetro	Valor	Descripción
C1 (coeficiente cognitivo)	1.5	Controla la influencia de la mejor experiencia individual de cada partícula sobre su nueva posición. Un valor más alto hace que cada partícula dependa más de su propio historial de éxito.
C2 (coeficiente social)	1.5	Controla la influencia del mejor valor global encontrado por el enjambre sobre cada partícula. Un valor más alto hace que las partículas sigan más fuertemente al grupo.
w (factor de inercia)	0.5	Controla la contribución de la velocidad previa en el movimiento de la partícula, permitiendo un balance entre exploración y explotación.
Dimensions	N	Define el número de variables o hiperparámetros que cada partícula debe optimizar. El valor cambia dinámicamente según el modelo evaluado: GMM, PCA-GMM o UMAP-GMM.
Iterations	100	Define el número máximo de actualizaciones que realizará cada partícula. Un número mayor incrementa la posibilidad de convergencia hacia un óptimo, pero también aumenta el tiempo de procesamiento.

Función objetivo, la función objetivo establece el criterio para evaluar la calidad de cada configuración de hiperparámetros explorada por PSO. En esta investigación, la función objetivo debía equilibrar dos criterios complementarios: minimizar el error de las posiciones identificadas respecto a la referencia y maximizar la cantidad de genes identificados. La función objetivo-propuesta se expresa de la siguiente manera (Ec. 6.6):

$$f = \sum_{i=1}^N \frac{RMSE_i}{G_i} - \alpha \cdot C \quad (\text{Ec. 6.6})$$

Donde:

- $RMSE_i$: raíz del error cuadrático medio del grupo i
- G_i : cantidad de genes en el grupo i
- C : cantidad de grupos que cumplen simultáneamente $RMSE_i < \theta$ y $G_i > \mu$
- α : parámetro de recompensa que incentiva grupos de alta calidad
- θ : umbral máximo de error aceptable
- μ : mínimo de genes requerido por grupo

El primer término $\sum_{i=1}^N \frac{RMSE_i}{G_i}$ penaliza grupos con alto error y pocos genes, mientras que el segundo término $\alpha \cdot C$ premia aquellos que cumplen simultáneamente con los criterios de calidad definidos por θ y μ . Esta formulación garantiza que PSO no favorezca configuraciones que maximicen la cantidad de genes a costa de la precisión, ni configuraciones que minimicen el error agrupando pocos genes. Los valores de θ , μ y α pueden ajustarse según las características del genoma analizado.

6.1.4.3 Resultados del modelo GMM optimizado con PSO

La primera prueba de optimización se realizó con el conjunto de datos de EPNI por ser el conjunto con mayor cantidad de genes disponibles para la evaluación. La configuración de PSO utilizó 6 dimensiones, correspondientes directamente a los seis parámetros que definen el comportamiento de GMM, y se ejecutó durante 100 iteraciones, resultando en un tiempo de ejecución total de 2 horas y 55 minutos.

La Tabla 6.19 presenta los hiperparámetros óptimos encontrados por PSO para GMM. La configuración óptima indicó que el mejor número de componentes es 2, con una matriz de covarianza de tipo tied, un máximo de 4,049 iteraciones, una tolerancia de convergencia de 0.005492, inicialización mediante K-Means y 5 inicializaciones aleatorias.

Tabla 6.19 Hiperparámetros óptimos encontrados por PSO para el Modelo GMM

Hiperparámetro GMM	Valor óptimo	Descripción
n_components	3	Número de componentes gaussianas del modelo
covariance_type	tied	Tipo de matriz de covarianza compartida entre todos los componentes
max_iter	16551	Número máximo de iteraciones del algoritmo EM
tol	0.001587	Tolerancia de convergencia del algoritmo
init_params	K-Means	Método de inicialización de los parámetros
n_init	7	Número de inicializaciones aleatorias

La Tabla 6.20 presenta los resultados comparativos entre la prueba anterior sin optimización (segunda prueba) y la prueba con PSO (tercera prueba). La segunda prueba corresponde a los resultados en la iteración anterior con MaxAbsScaler y segundo cuartil, pero sin optimización formal de hiperparámetros, mientras que la

tercera prueba corresponde a los resultados con la configuración óptima encontrada por PSO.

Tabla 6.20 Resultados comparativos de EPNI antes y después de la optimización con PSO

Algoritmo	Grupo	N. datos	N. genes identificados	N. genes de referencia	MAE	MSE	RMSE
GMM sin optimizar	1	5134	87	248	2.33	19.02	4.36
GMM optimizado	1	8270	191	248	0.770161	1.7329	2.907

Como se puede observar, la incorporación de PSO produjo una mejora significativa en todos los aspectos evaluados. El número de genes identificados aumentó de 87 a 191, el RMSE se redujo de 4.36 a 2.907 y el MAE disminuyó de 2.33 a 0.770. Esta mejora integral demuestra que la optimización de hiperparámetros con PSO no solo incrementa la cobertura de genes reales identificados, sino que también mejora considerablemente la precisión de las posiciones, logrando el mejor equilibrio entre ambos criterios obtenido hasta este punto del desarrollo del método.

La Figura 6.21 presenta el gráfico de dispersión obtenidos con UMAP, donde se muestra que el grupo 1 es el más próximo al grupo de referencia, confirmando visualmente los resultados cuantitativos. Los gráficos de densidad por sección del gen V se presentan en el Anexo X para referencia del lector.

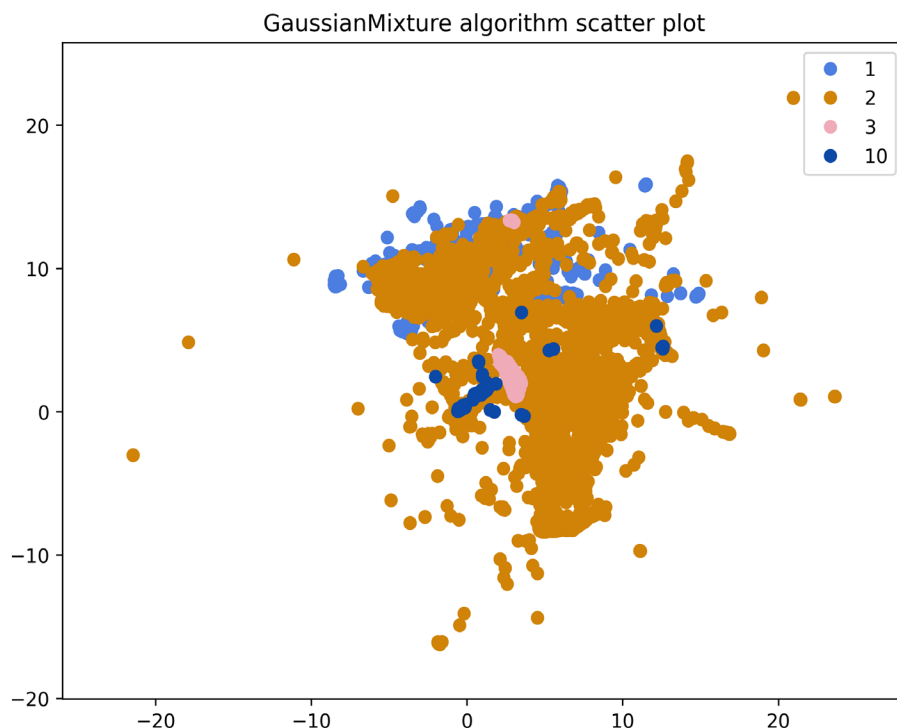


Figura 6.21 Gráficos de dispersión de la prueba de optimización del conjunto de datos EPNI

6.1.4.4 Resultados del modelo PCA-GMM optimizado con PSO

El segundo modelo evaluado en esta iteración incorporó PCA como técnica de reducción de dimensionalidad previa al agrupamiento con GMM. En este caso, PSO optimizó los hiperparámetros de ambos algoritmos de manera conjunta, utilizando 6 dimensiones que corresponden a los parámetros combinados de GMM y PCA. La reducción de dimensionalidad mediante PCA transforma el espacio original de 43 atributos en un espacio de menor dimensión que preserva la mayor varianza posible de los datos, con el objetivo de mejorar la separabilidad entre grupos antes de aplicar el agrupamiento.

Las Tablas 6.21 y 6.22 presentan los hiperparámetros óptimos encontrados por PSO para este modelo. La configuración óptima de GMM indicó 4 componentes gaussianas con matriz de covarianza compartida de tipo tied, mientras que PCA redujo el espacio a 2 componentes principales con normalización de las componentes mediante whitening.

Tabla 6.21 Hiperparámetros óptimos encontrados por PSO para el Modelo GMM

Hiperparámetro GMM	Valor óptimo	Descripción
n_components	4	Número de componentes gaussianas
covariance_type	tied	Matriz de covarianza compartida entre componentes
max_iter	11799	Máximo de iteraciones del algoritmo EM
tol	0.008127	Tolerancia de convergencia
init_params	K-Means	Método de inicialización
n_init	6	Número de inicializaciones aleatorias

Tabla 6.22 Hiperparámetros óptimos encontrados por PSO para el Modelo PCA

Hiperparámetro GMM	Valor óptimo	Descripción
n_components	2	Número de componentes principales
svd_solver	auto	Método de descomposición en valores singulares
whiten	True	Normalización de las componentes principales

La Tabla 6.23 presenta los resultados del Modelo PCA-GMM sobre el conjunto de datos de *Eptesicus nilssonii* (EPNI), que cuenta con 248 genes de referencia validados por expertos. Este conjunto fue seleccionado como caso de evaluación por ser el más representativo de los conjuntos disponibles, permitiendo una evaluación robusta del impacto de la reducción de dimensionalidad con PCA sobre la calidad del agrupamiento. La Figura 6.23 presenta el gráfico de dispersión obtenido con UMAP para la representación de los grupos.

Tabla 6.23 Resultados del Modelo PCA-GMM sobre EPNI

Algoritmo	Grupo	N. datos	N. genes identificados	N. genes de referencia	MAE	MSE	RMSE
GMM sin optimizar	1	5134	87	248	2.33	19.02	4.36
PCA-GMM optimizado	4	4801	112	248	0.451	2.107	3.129

Los resultados muestran que el Modelo PCA-GMM identificó 112 genes de 248 de referencia con un RMSE de 3.129 y un MAE de 0.451. Las métricas de calidad del agrupamiento presentan valores favorables: un Silhouette Score de 0.482 indica grupos bien definidos, un Davies-Bouldin Index de 0.722 refleja grupos compactos y bien separados, y un Calinski-Harabasz Index de 18542 indica una buena relación entre la dispersión interna y externa de los grupos.

Si bien la reducción lineal con PCA mejoró la estructura del agrupamiento en términos de métricas de calidad, la cobertura de genes identificados resultó inferior a la esperada. Esta observación motivó la evaluación del Modelo UMAP-GMM, que utiliza una reducción no lineal capaz de preservar estructuras más complejas en los datos, con el objetivo de mejorar tanto la cobertura como la precisión del agrupamiento. La Figura 6.22 presenta los gráficos de dispersión generados mediante UMAP para la visualización de los grupos obtenidos.

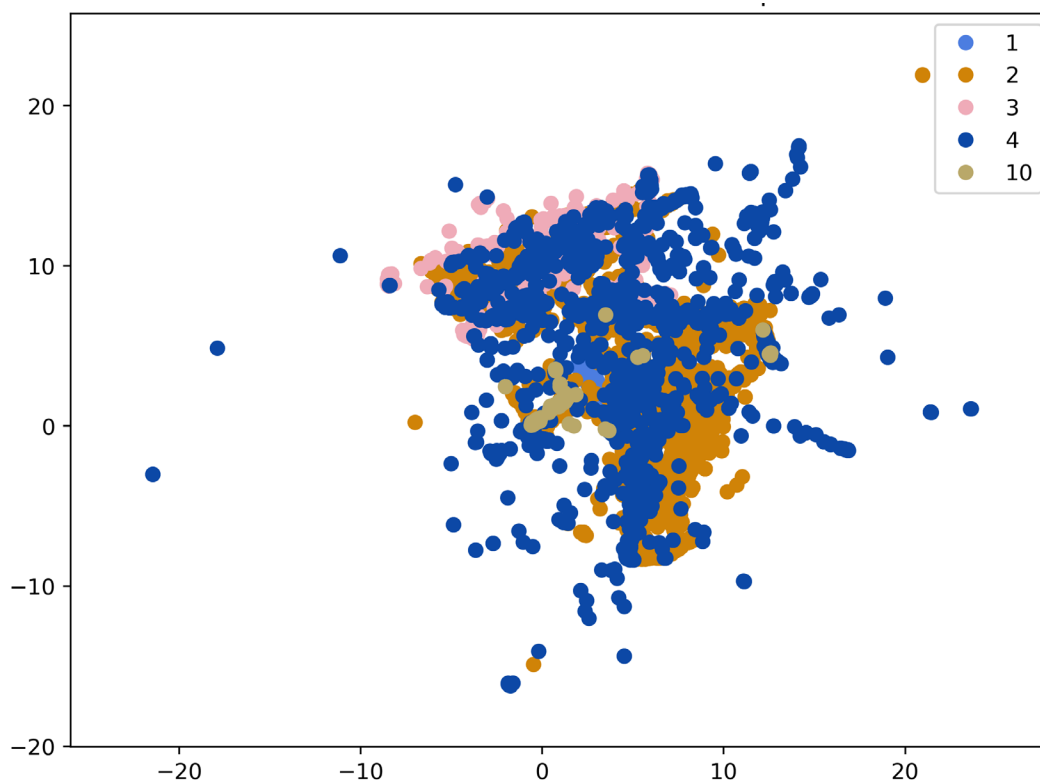


Figura 6.22 Gráficos de dispersión del Modelo PCA-GMM sobre EPNI

6.1.4.5 Resultados del modelo UMAP-GMM con PSO

El tercer modelo evaluado en esta iteración incorporó UMAP como técnica de reducción de dimensionalidad previa al agrupamiento con GMM, sustituyendo la reducción lineal de PCA por una reducción no lineal capaz de preservar tanto la estructura local como global del espacio de datos. A diferencia de PCA, que proyecta los datos en direcciones

de máxima varianza con transformaciones lineales, UMAP construye una representación de menor dimensión que mantiene las relaciones de similitud entre registros cercanos y la estructura topológica global del espacio, lo que produce una mejor separabilidad entre grupos de datos con estructuras complejas como en los alineamientos de genes V.

Al igual que en el modelo anterior, PSO optimizó los hiperparámetros de UMAP y GMM de manera conjunta. Las Tablas 6.24 y 6.25 presentan los hiperparámetros óptimos encontrados por PSO para cada componente del modelo.

Tabla 6.24 Hiperparámetros óptimos de GMM para el Modelo UMAP-GMM

Hiperparámetro GMM	Valor óptimo	Descripción
n_components	3	Número de componentes gaussianas
covariance_type	tied	Matriz de covarianza compartida entre componentes
max_iter	8283	Máximo de iteraciones del algoritmo EM
tol	0.001457	Tolerancia de convergencia
init_params	K-Means	Método de inicialización
n_init	7	Número de inicializaciones aleatorias
random_state	42	Semilla para reproducibilidad

Tabla 6.25 Hiperparámetros óptimos de UMAP para el Modelo UMAP-GMM

Hiperparámetro GMM	Valor óptimo	Descripción
n_components	2	Número de dimensiones del espacio reducido
n_neighbors	63	Número de vecinos para estructura local
min_dist	0.159598	Distancia mínima entre puntos en espacio reducido
metric	cosine	Métrica de distancia utilizada
angular_rp_forest	True	Uso de bosque de proyección aleatoria angular
random_state	42	Semilla para reproducibilidad

La Tabla 6.26 presenta los resultados con el Modelo UMAP-GMM sobre el conjunto de datos de *Eptesicus nilssonii* (EPNI), que cuenta con 248 genes de referencia validados por expertos. La Figura 6.23 presenta los gráficos de dispersión obtenidos para los diferentes grupos de la configuración.

Tabla 6.26 Resultados del Modelo UMAP-GMM sobre EPNI

Algoritmo	Grupo	N. datos	N. genes identificados	N. genes de referencia	MAE	MSE	RMSE
GMM sin optimizar	1	5134	87	248	2.33	19.02	4.36
UMAP-GMM optimizado	2	8169	194	248	1.708763	9.53866	3.088472

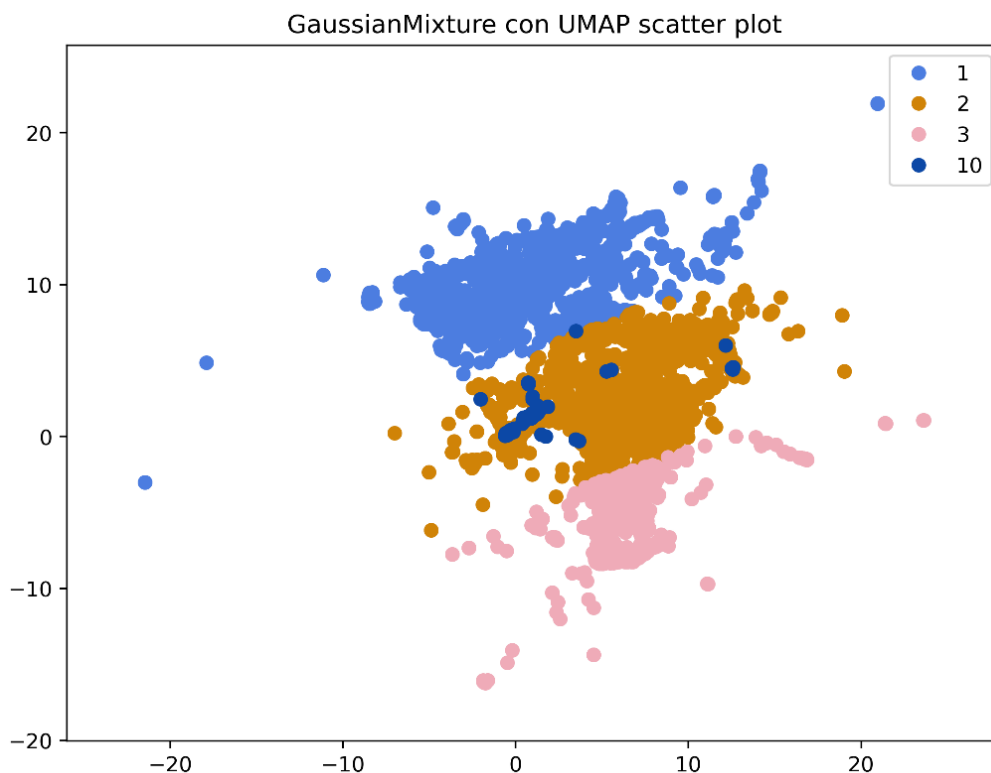


Figura 6.23 Gráficos de dispersión del Modelo UMAP-GMM sobre EPNI

Los resultados muestran que el Modelo UMAP-GMM identificó 194 genes de 248 de referencia con un RMSE de 3.088 y un MAE de 1.708, representando una mejora significativa respecto al modelo sin optimizar que identificaba únicamente 87 genes con un RMSE de 4.36. Comparado con el Modelo PCA-GMM, UMAP-GMM logró una cobertura significativamente mayor, identificando 194 genes frente a los 112 del modelo anterior, con un RMSE similar de 3.088 frente a 3.129. Este resultado confirma que la reducción no lineal de UMAP produce una representación más favorable para el agrupamiento en el contexto de los datos de alineamiento de genes V, al preservar mejor las relaciones estructurales entre registros biológicamente similares que la reducción lineal de PCA.

Esta cercanía en el desempeño entre ambos modelos justifica la inclusión de ambas configuraciones como variantes finales del método. La variante PCA-GMM, si bien identificó menos genes en este conjunto, se incluye igualmente como variante de comparación debido a que su menor MAE sugiere un comportamiento potencialmente más preciso en genomas con características distintas a las de EPNI, y porque la comparación entre reducción lineal y no lineal representa una pregunta metodológica relevante que solo puede responderse con una evaluación sistemática sobre múltiples genomas. Los resultados de estas tres variantes se comparan de manera sistemática en el Capítulo 7.

6.1.5 Consolidación de las tres variantes finales

El proceso de desarrollo iterativo descrito en las secciones anteriores permitió consolidar tres variantes finales del método RIVAM, cada una con características particulares en términos de precisión, cobertura y complejidad computacional. La Figura 6.24 presenta el diagrama consolidado del método, mostrando el flujo de procesamiento común y las tres configuraciones de agrupamiento resultantes.

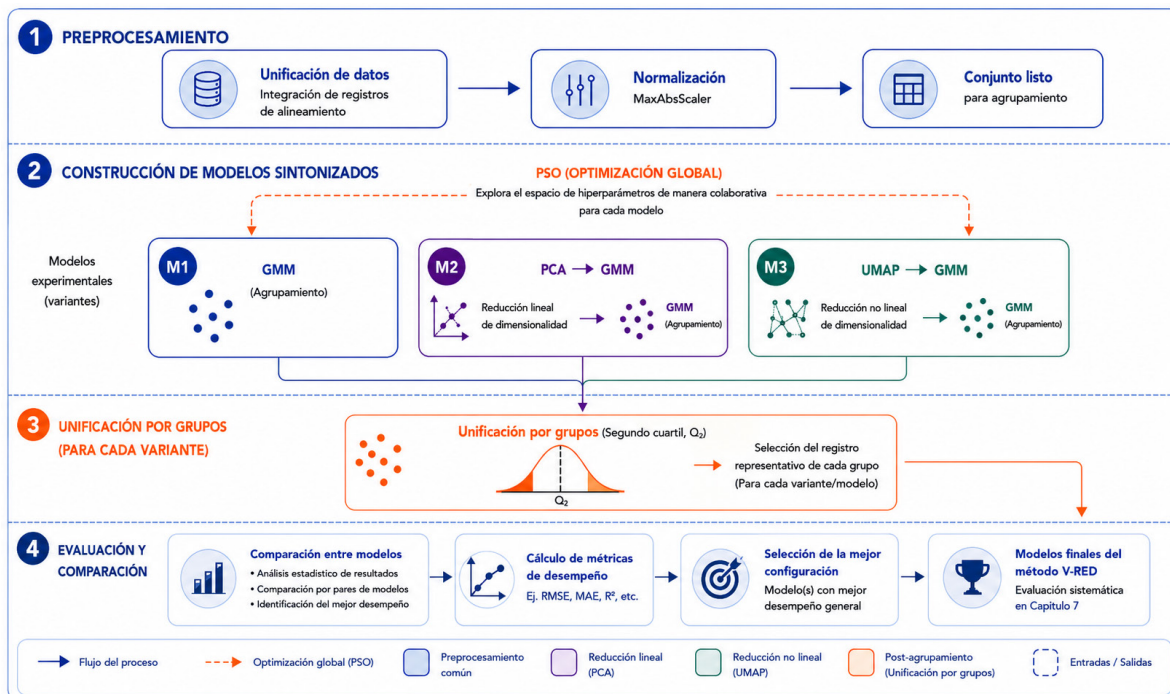


Figura 6.24 Diagrama consolidado de las tres variantes finales del método RIVAM

Como se puede observar en la figura, el método se organiza en cuatro etapas principales. La primera corresponde al preprocesamiento común, donde los registros de alineamiento son unificados y normalizados con MaxAbsScaler, produciendo un conjunto de datos listo para el agrupamiento. La segunda etapa corresponde a la obtención de los modelos sintonizados, donde PSO optimiza los hiperparámetros de cada una de las tres variantes de manera independiente. La tercera etapa aplica la unificación por grupos con el segundo cuartil para cada variante, seleccionando el registro representativo de cada grupo. La cuarta etapa corresponde a la evaluación y comparación de los modelos con métricas de desempeño como RMSE, MAE y R^2 , con el objetivo de identificar la configuración con mejor desempeño general.

La Tabla 6.27 presenta una comparativa de los resultados con cada variante sobre el conjunto de datos de EPNI, que sirvió como conjunto de referencia para la consolidación del método. Es importante destacar que estos resultados corresponden a la evaluación sobre un único genoma y tienen como propósito ilustrar las diferencias de

comportamiento entre las variantes, no determinar cuál es la mejor de manera definitiva.

Tabla 6.27 Comparativa de las tres variantes finales del método RIVAM sobre EPNI

Algoritmo	Grupo	N. datos	N. genes identificados	N. genes de referencia	MAE	MSE	RMSE
GMM + PSO	2	8270	191	248	0.770	1.732	2.907
PCA-GMM + PSO	4	4801	112	248	0.451	2.107	3.129
UMAP-GMM + PSO	3	8169	194	248	0.762	1.688	2.941

El análisis comparativo de las tres variantes revela patrones complementarios que justifican su evaluación sistemática en el Capítulo 7. La variante GMM con PSO y la variante UMAP-GMM con PSO presentan resultados muy similares en términos de cobertura, identificando 191 y 194 genes respectivamente, con RMSE de 2.907 y 2.941. Esta similitud en el desempeño sugiere que ambas variantes capturan de manera efectiva la estructura del espacio de búsqueda de genes V , aunque mediante mecanismos distintos: GMM opera directamente sobre el espacio normalizado, mientras que UMAP-GMM transforma previamente ese espacio mediante una reducción no lineal que preserva las relaciones topológicas entre registros.

La variante PCA-GMM, en cambio, identificó únicamente 112 genes, pero con el menor MAE de las tres (0.451), lo que sugiere que cuando logra identificar un gen, lo hace con mayor precisión en las posiciones. Este comportamiento puede explicarse por la naturaleza de la reducción lineal de PCA: al proyectar los datos en las direcciones de mayor varianza, PCA puede eliminar ruido que afecta la precisión del agrupamiento, pero también puede descartar información relevante para la identificación de genes con características menos frecuentes en el conjunto de datos, reduciendo así la cobertura.

Desde el punto de vista de la complejidad computacional, las tres variantes presentan diferencias importantes. GMM con PSO es la variante más simple y eficiente computacionalmente, debido a que no requiere ninguna transformación previa del espacio de datos. PCA-GMM añade un costo computacional moderado por la descomposición en componentes principales. UMAP-GMM es la variante más costosa computacionalmente debido a la construcción del grafo de vecindad que requiere UMAP, aunque este costo es aceptable para los volúmenes de datos manejados en esta investigación.

Estas diferencias en el comportamiento y la complejidad de cada variante refuerzan la decisión de evaluarlas de manera independiente en el Capítulo 7, donde serán aplicadas a un conjunto de 16 genomas distintos con el objetivo de determinar cuál de ellas ofrece el mejor desempeño general y en qué condiciones cada variante es más adecuada para el problema de reducción del espacio de búsqueda de genes V .

6.2 Especificaciones técnicas

Buscando garantizar la reproducibilidad y claridad de los experimentos realizados, esta sección describe los entornos de ejecución utilizados durante la implementación, optimización y evaluación del método RIVAM, así como las tecnologías y librerías empleadas en su desarrollo.

6.2.1 Entornos de ejecución

Se utilizaron dos entornos de ejecución de manera complementaria, cada uno con características distintas que permitieron abordar diferentes etapas del proceso de desarrollo y evaluación. La selección de estos entornos respondió a criterios prácticos relacionados con la disponibilidad de recursos, la duración de los experimentos y la necesidad de garantizar la continuidad en procesos de optimización de larga duración.

Google Colab fue utilizado como entorno de desarrollo en la nube, ofreciendo acceso flexible a recursos de cómputo sin requerir instalación local ni configuración de hardware especializado. Su principal ventaja radica en la disponibilidad inmediata de recursos de procesamiento y en la posibilidad de ejecutar experimentos desde cualquier dispositivo con acceso a internet, lo que facilitó el desarrollo y las pruebas preliminares de las distintas variantes del método. Sus características principales se presentan en la Tabla 6.28.

Tabla 6.28 Especificaciones del entorno Google Colab

Componente	Especificación
Procesador	Intel Xeon virtualizado o equivalente
RAM disponible	12 según sesión
GPU	NVIDIA Tesla T4 o P100 (cuando se activó aceleración)
Entorno de desarrollo	Python 3.11.13, scikit-learn, pyswarms, UMAP-learn, NumPy, Pandas
Tiempo de procesamiento	De 2 a 8 horas por optimización

El equipo personal ASUS TUF Gaming F15 (FX517ZR) fue utilizado como entorno de ejecución local, para las optimizaciones de mayor duración que requerían continuidad en la ejecución sin las restricciones de tiempo de sesión propias de los entornos en la nube. Este equipo permitió ejecutar procesos de optimización de manera ininterrumpida durante periodos prolongados, garantizando la convergencia de PSO en los modelos más complejos. Sus especificaciones se presentan en la Tabla 6.29.

Tabla 6.29 Especificaciones del equipo personal ASUS TUF Gaming F15

Componente	Especificación
Procesador	Intel Core i7-12700H (14 núcleos, 20 hilos)
Memoria RAM	32 GB DDR5
Almacenamiento	SSD NVMe de alta velocidad
Sistema operativo	Windows 11 Home
Entorno de desarrollo	JupyterLab, Python 3.11.13
Tiempo de procesamiento	Entre 2 horas y hasta 65 horas continuas

La combinación de ambos entornos permitió balancear la accesibilidad a través de la nube y el poder de cómputo local, garantizando la eficiencia en el procesamiento, la continuidad de los experimentos de largo plazo y la reproducibilidad de los resultados. Google Colab fue utilizado principalmente para el desarrollo y pruebas preliminares, mientras que el equipo personal se reservó para las optimizaciones de mayor complejidad que requerían sesiones de ejecución continua.

La optimización de hiperparámetros con PSO implicó tiempos de ejecución altamente variables dependiendo de la complejidad del modelo y el tamaño del conjunto de datos: desde procesos relativamente rápidos de aproximadamente 2 horas para configuraciones simples hasta optimizaciones que requirieron hasta 65 horas continuas de ejecución para los modelos más complejos como UMAP-GMM sobre genomas de mayor volumen. Esta variabilidad en los tiempos de procesamiento justifica la disponibilidad de ambos entornos como infraestructura complementaria para el desarrollo del método.

6.2.2 Tecnologías y bibliotecas

El desarrollo del método RIVAM se realizó íntegramente en Python 3.11.13, haciendo uso de un conjunto de bibliotecas especializadas para el procesamiento de datos, la reducción de dimensionalidad, el agrupamiento y la optimización de hiperparámetros. La selección de estas bibliotecas respondió a criterios de madurez, documentación, compatibilidad entre versiones y adopción amplia en la comunidad científica de aprendizaje automático y bioinformática computacional. La Tabla 6.30 presenta las principales bibliotecas utilizadas junto con sus versiones exactas y su función dentro del pipeline del método.

Tabla 6.30 Bibliotecas y versiones utilizadas en el desarrollo del método RIVAM

biblioteca	Versión	Función principal
Python	3.11.13	Lenguaje de programación base
scikit-learn	1.6.1	GMM, PCA, MaxAbsScaler, métricas de evaluación
UMAP-learn	0.5.9	Reducción no lineal de dimensionalidad
pyswarms	1.3.0	Optimización por enjambre de partículas (PSO)
NumPy	2.1.3	Operaciones matriciales y numéricas
Pandas	2.2.3	Manipulación y procesamiento de datos tabulares

La reproducibilidad de los experimentos se garantizó con el uso de una semilla aleatoria fija (random_state=42) en todos los modelos y algoritmos que incorporan procesos estocásticos, así como mediante el guardado de los modelos optimizados en formato .joblib para su reutilización en la etapa de evaluación sistemática descrita en el Capítulo 7.

6.2.3 Tiempos de ejecución

Es importante distinguir entre dos tipos de tiempos de ejecución en el contexto del método RIVAM. El primero corresponde al tiempo de optimización de hiperparámetros con PSO, que se realiza una única vez por variante y puede extenderse desde 2 horas hasta 65 horas continuas dependiendo de la complejidad del modelo y el tamaño del conjunto de datos utilizado para la optimización. El segundo corresponde al tiempo de aplicación del modelo ya optimizado sobre un nuevo genoma, que es significativamente menor y representa el tiempo real de uso del método en producción.

La Tabla 6.31 presenta los tiempos de ejecución medidos para cada variante del método sobre el conjunto de datos de EPNI, considerando únicamente la aplicación del modelo ya optimizado sobre el genoma.

Tabla 6.31 Tiempos de ejecución por variante del método RIVAM

Variante	Tiempo de ejecución
GMM	3.6 segundos (0.06 minutos)
PCA-GMM	2.8 segundos (0.05 minutos)
UMAP-GMM	21.7 segundos (0.36 minutos)

Los tiempos de ejecución confirman que las tres variantes son computacionalmente eficientes para su aplicación en nuevos genomas. GMM y PCA-GMM presentan tiempos prácticamente equivalentes de menos de 4 segundos, mientras que UMAP-GMM requiere aproximadamente 22 segundos debido al proceso de construcción del grafo de vecindad de UMAP. En todos los casos, los tiempos son significativamente menores al tiempo de revisión manual que el método busca reducir, lo que confirma su viabilidad práctica en el contexto del proceso de anotación de genes V.

Capítulo 7. Diseño del experimento

Este capítulo describe el diseño experimental utilizado para evaluar de manera sistemática las tres variantes finales del método RIVAM desarrolladas en el Capítulo 6. A diferencia del capítulo anterior, que se centró en el proceso de construcción y refinamiento iterativo del método, este capítulo se enfoca en la validación formal de su desempeño con la aplicación de las tres variantes a un conjunto amplio y diverso de genomas, con el objetivo de determinar cuál de ellas ofrece el mejor equilibrio entre precisión, cobertura y generalización a distintas especies.

El diseño experimental sigue los principios del Ciclo de Evaluación de la metodología DSR, que establece que el artefacto desarrollado debe ser evaluado de manera rigurosa y sistemática para demostrar su utilidad en el contexto del problema que busca resolver. Para garantizar la validez de los resultados, se definieron criterios explícitos de selección de genomas, métricas de evaluación formales y análisis estadísticos que permiten comparar el desempeño de las variantes de manera objetiva y reproducible.

7.1 Datos experimentales

El conjunto de datos utilizado en este estudio comprende genomas de quirópteros y del ajolote obtenidos de la plataforma National Center for Biotechnology Information (NCBI), clasificados en dos grupos según su función dentro del estudio: genomas de evaluación y genomas de validación. Los primeros se emplearon durante las fases de comparación estadística de las configuraciones, mientras que los segundos se reservaron exclusivamente para la fase de validación externa, garantizando una evaluación independiente de la capacidad de generalización del método. Los genomas se encuentran ensamblados en distintos niveles, incluyendo ensamblajes a nivel de cromosoma y a nivel de contig, lo que introduce una fuente adicional de heterogeneidad estructural que el método debe ser capaz de manejar de manera consistente.

7.1.1 Genomas de evaluación

El conjunto de genomas de evaluación está conformado por 12 especies de quirópteros anotadas por expertos del CISEI. Estas especies abarcan una amplia diversidad taxonómica dentro del orden Chiroptera, así como una variabilidad significativa en el número de genes V reportados por especie y en el volumen de registros obtenidos con los cinco sistemas de alineamiento, lo que permite evaluar el desempeño de las variantes del método en distintos escenarios de complejidad real.

La selección de estas 12 especies buscó representar distintas familias dentro del orden Chiroptera, incluyendo especies de los géneros *Eptesicus*, *Pipistrellus*, *Nyctalus*, *Myotis*, *Hipposideros* y *Rhinolophus*, entre otros. Esta diversidad taxonómica garantiza que los resultados no sean específicos de un linaje particular sino representativos del orden en su conjunto, fortaleciendo la validez externa de las conclusiones derivadas del experimento. La Tabla 7.1 presenta los 12 genomas utilizados, incluyendo su nombre científico, identificador, número de genes de referencia validados por expertos, número de registros originales y su acceso en el NCBI.

Tabla 7.1 Genomas de evaluación utilizados en el estudio

N.	Especie	ID	Genes ref.	Registros originales	Acceso NCBI
1	<i>Antrozous pallidus</i>	Anpa	90	163,001	GCA_027563665.1
2	<i>Aselliscus stoliczkanus</i>	ASST	48	137,955	GCA_033961575.1
3	<i>Eptesicus fuscus</i>	Epfu	168	240,377	GCA_027574615.1
4	<i>Eptesicus nilssonii</i>	Epni	248	321,140	GCA_951640355.1
5	<i>Hipposideros larvatus</i>	Hila	121	210,888	GCA_031876335.1
6	<i>Miniopterus schreibersii</i>	Misc	135	203,100	GCA_964146895.1
7	<i>Nyctalus aviator</i>	Nyav	300	342,239	GCA_036971965.2
8	<i>Pipistrellus nathusii</i>	Pina	228	308,848	GCA_963693515.1
9	<i>Pipistrellus pygmaeus</i>	Pipy	268	327,576	GCA_949987585.1

N.	Especie	ID	Genes ref.	Registros originales	Acceso NCBI
10	<i>Rhynchonycteris naso</i>	Rhna	117	199,075	GCA_037038545.1
11	<i>Saccopteryx leptura</i>	Sale	159	292501	GCA_036850995.1
12	<i>Vespertilio murinus</i>	Vemu	325	421862	GCA_963924515.1

Como se observa en la Tabla 7.1, el número de registros originales varía desde 137,955 en *Aselliscus stoliczkanus* hasta 421,862 en *Vespertilio murinus*, lo que refleja la heterogeneidad en el tamaño y la complejidad de los genomas analizados. De manera similar, el número de genes V de referencia oscila entre 48 y 325, representando una variabilidad considerable que pone a prueba la capacidad del método para mantener un desempeño consistente en distintos escenarios.

Dentro del total de genes V anotados por especie, los expertos distinguen entre genes funcionales, que presentan una estructura completa y capacidad de expresión, y genes no funcionales, que incluyen pseudogenes y secuencias truncadas. Esta distinción cobra especial relevancia en la evaluación del método, debido a que la recuperación de genes funcionales constituye un criterio clave de calidad en el proceso de anotación. El desglose detallado por tipo de gen se presenta en el Capítulo 8, junto con los resultados del proceso de reducción aplicado a cada genoma.

7.1.2 Genomas de validación externa

Con el propósito de evaluar la capacidad de generalización del método RIVAM más allá de los genomas utilizados durante la fase de evaluación, se incorporaron tres genomas adicionales con características distintas entre sí. Estos genomas no fueron considerados durante la fase de optimización de hiperparámetros ni durante la comparación estadística de las variantes, lo que garantiza una validación externa independiente y representativa de condiciones reales de la caracterización del genoma donde el método opera sobre datos completamente nuevos.

La selección de estos tres genomas responde a escenarios distintos que permiten evaluar la generalización del método desde diferentes perspectivas. El primero es un quiróptero con mayor repertorio de genes V que los del conjunto de evaluación. El segundo es una especie filogenéticamente distante a los quirópteros. El tercero es un genoma sin anotación experta disponible. La Tabla 7.2 presenta los genomas de validación utilizados.

Tabla 7.2 Genomas de validación externa

Especie	ID	Genes referencia	Registros originales	Acceso NCBI
<i>Plecotus auritus</i>	Plau	850	909,584	GCF_963455305.1
<i>Ambystoma mexicanum</i>	Amme	87	13,410	GCF_040938575.1
<i>Molossus nigricans</i>	Moni	No disponible	570,842	GCA_039880945.1

El primer genoma corresponde a *Plecotus auritus* (Plau), un quiróptero de la familia *Vespertilionidae* con una referencia anotada por expertos de 850 genes V y un volumen de 909,584 registros originales, siendo el genoma con mayor cantidad de registros de todo el estudio. Su inclusión permite verificar que el método mantiene su desempeño en un individuo perteneciente al mismo orden taxonómico que los genomas de evaluación, pero con un repertorio de genes V considerablemente mayor, lo que representa una prueba crítica de la escalabilidad del método.

El segundo genoma corresponde a *Ambystoma mexicanum* (Amme), especie nativa de México conocida comúnmente como ajolote. A diferencia de los quirópteros, este organismo pertenece al orden *Caudata* de la clase *Amphibia*, lo que lo hace filogenéticamente distante de las especies utilizadas durante la optimización del método. Su genoma es considerablemente más grande y estructuralmente más complejo, con características arquitectónicas del gen V significativamente distintas a las observadas en mamíferos.

La región intrónica (la zona que corresponde al fin del SP y el inicio del exón) del gen V en esta especie presenta una longitud aproximada de 200 nucleótidos, dando lugar a una estructura completa del gen superior a los 600 nucleótidos, excediendo la estructura típica observada en genes V de mamíferos. La inclusión de este genoma permite evaluar si el flujo metodológico propuesto mantiene su coherencia y capacidad discriminativa al enfrentarse a un dominio filogenético completamente distinto al utilizado durante el diseño del método.

El tercer genoma es el de *Molossus nigricans* (Moni), un quiróptero de la familia Molossidae sin anotación experta disponible, cuyo número total de genes V es desconocido. Este genoma representa un escenario completamente exploratorio, donde no existe información previa que guíe el proceso de identificación ni una referencia con la cual comparar directamente los resultados.

Con 570,842 registros originales, constituye el segundo conjunto de datos más grande del estudio después de *Plecotus auritus*, lo que lo convierte en el caso más exigente desde el punto de vista computacional entre los genomas de validación. Su inclusión permite evaluar el comportamiento del método en condiciones reales de

caracterización de nuevas especies, donde la ausencia de anotaciones previas es precisamente el escenario que el método busca facilitar.

7.1.3 Criterios de selección

La selección de los genomas de evaluación respondió a cuatro criterios que en conjunto garantizan la validez, representatividad y comparabilidad de los resultados. El primero fue la disponibilidad de anotaciones manuales validadas por expertos del CISEI, lo que garantiza la existencia de una referencia confiable para la evaluación cuantitativa del método. Sin esta referencia no sería posible calcular las métricas de error ni el porcentaje de identificación necesarios para comparar objetivamente el desempeño de las tres variantes propuestas. La calidad de estas anotaciones es fundamental porque determina directamente la confiabilidad de los resultados: una referencia imprecisa introduciría sesgos en la evaluación que dificultarían la interpretación de los resultados.

El segundo criterio fue la diversidad taxonómica dentro del orden *Chiroptera*. Esta diversidad permite evaluar la robustez del método frente a variaciones en el tamaño del genoma, el número de genes V y la complejidad estructural de cada especie. La inclusión de especies pertenecientes a distintas familias, como *Vespertilionidae*, *Molossidae*, *Rhinolophidae* y *Emballonuridae*, asegura que los resultados no dependan de un linaje particular, sino que reflejen un comportamiento representativo del dominio de estudio en su conjunto.

El tercer criterio fue la variabilidad en el número de genes V reportados por especie, que oscila entre 48 genes en *Aselliscus stoliczkanus* y 325 genes en *Vespertilio murinus*. Esta diferencia de casi siete veces en el tamaño del repertorio inmunológico permite analizar el desempeño del método bajo distintos niveles de complejidad, evaluando si las variantes mantienen un comportamiento consistente independientemente del número de genes que deben identificarse.

El cuarto criterio, de carácter práctico, fue la disponibilidad de los archivos GFF generados por los cinco sistemas de alineamiento para cada genoma, requisito indispensable para la aplicación del método. Todos los genomas seleccionados contaban con estos archivos previamente procesados por el equipo del CISEI, lo que garantizó condiciones homogéneas de entrada y permitió mantener la comparabilidad de los resultados entre especies, eliminando la variabilidad introducida por diferencias en el proceso de alineamiento.

En el caso de los genomas de validación, los criterios de selección estuvieron orientados a evaluar la capacidad de generalización del método en escenarios complementarios y progresivamente más exigentes, cada uno diseñado para estresar

un aspecto distinto de la robustez del método. El primer escenario corresponde a un quiróptero con un repertorio de genes V considerablemente mayor al observado en los genomas de evaluación, con 850 genes de referencia frente a un máximo de 325 en el conjunto de evaluación, lo que permite analizar la escalabilidad del método ante conjuntos de datos significativamente más grandes dentro del mismo dominio taxonómico.

El segundo escenario corresponde a *Ambystoma mexicanum*, una especie filogenéticamente distante perteneciente al orden *Caudata*, utilizada para evaluar la transferibilidad del método hacia dominios con arquitecturas del gen V diferentes a las empleadas durante la etapa de optimización. Este escenario es exigente porque el método no fue optimizado para este dominio y las diferencias estructurales entre anfibios y mamíferos implican que los patrones aprendidos durante la optimización no se transfieren de manera directa.

Finalmente, el tercer escenario corresponde a *Molossus nigricans*, un quiróptero sin anotación experta disponible, que simula condiciones reales de caracterización de nuevas especies donde el método debe operar de manera exploratoria sin información de referencia que guíe el proceso de identificación. Este es el escenario más cercano al uso real del método en investigación del área biológica de frontera, donde la ausencia de anotaciones previas es precisamente la condición que el método busca facilitar.

La combinación de estos criterios garantiza que la evaluación del método RIVAM cubra tanto su desempeño en condiciones controladas como su capacidad de generalización en escenarios no considerados durante el diseño, proporcionando una base sólida y rigurosa para las conclusiones presentadas en el Capítulo 8.

7.2 Proceso de evaluación

El proceso de evaluación del método RIVAM se diseñó con el objetivo de comparar el desempeño de las tres variantes de manera sistemática, objetiva y reproducible sobre el conjunto de 12 genomas de evaluación descrito en la sección anterior. Para garantizar la validez de los resultados, se definió un protocolo experimental explícito que estandariza las condiciones de aplicación de cada variante y un esquema de configuración que asegura la comparabilidad entre ellas.

7.2.1 Protocolo experimental

El protocolo experimental define la secuencia de pasos seguidos para la aplicación de cada variante del método sobre cada genoma de evaluación. Este protocolo fue diseñado para garantizar que todas las variantes sean evaluadas bajo las mismas condiciones de entrada y con los mismos criterios de evaluación, eliminando fuentes de variabilidad

que pudieran introducir sesgos en los resultados. La Figura 7.1 presenta el diagrama completo del protocolo, donde se muestran las seis etapas del proceso y su relación con las tres variantes evaluadas.

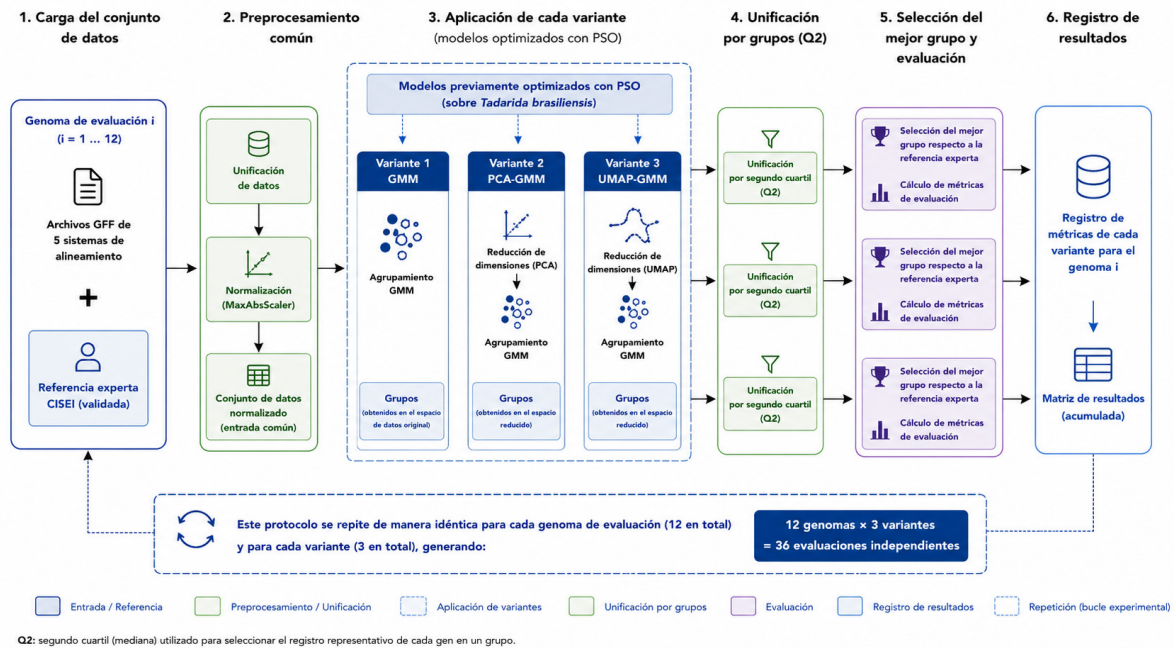


Figura 7.1 Diagrama del protocolo experimental del método RIVAM

El protocolo consta de las siguientes etapas:

1. **Carga del conjunto de datos:** para cada genoma i (donde $i = 1, \dots, 12$) se cargan los archivos GFF obtenidos por los cinco sistemas de alineamiento junto con los datos de referencia validados por los expertos del CISEI. Esta referencia experta constituye el estándar de comparación para la evaluación cuantitativa de cada variante.
2. **Preprocesamiento común:** se aplica el proceso de unificación de datos y normalización con MaxAbsScaler descrito en el Capítulo 6, obteniendo un conjunto de datos normalizado que sirve como entrada común para las tres variantes. Al ser idéntico para todas las variantes, este preprocesamiento garantiza que las diferencias en los resultados sean atribuibles exclusivamente a las distintas configuraciones de agrupamiento y no a diferencias en el procesamiento de los datos de entrada.
3. **Aplicación de cada variante:** se aplican de manera independiente las tres variantes del método utilizando los modelos previamente optimizados con PSO sobre *Tadarida brasiliensis*. La Variante 1 (GMM) aplica el agrupamiento directamente sobre el espacio de datos original normalizado. La Variante 2 (PCA-GMM) aplica primero una reducción lineal de dimensionalidad con PCA y luego el agrupamiento con GMM. La Variante 3 (UMAP-GMM) aplica una reducción no lineal con UMAP antes del agrupamiento. Cada variante obtiene un conjunto de grupos sobre el espacio de datos del genoma analizado.

4. **Unificación por grupos:** se aplica el proceso de unificación por el segundo cuartil (Q2) sobre cada grupo obtenido por cada variante, para consolidar los resultados en único registro por gen para cada grupo. Este paso garantiza la correspondencia directa entre los resultados del método y la referencia experta, habilitando el cálculo de las métricas de error.
5. **Selección del mejor grupo y evaluación:** se identifica el grupo con mayor número de genes identificados respecto a la referencia experta y se calculan las métricas de evaluación correspondientes: porcentaje de identificación, MAE, MSE y RMSE.
6. **Registro de resultados:** se registran las métricas de evaluación de cada variante para el genoma analizado, acumulándolas en la matriz de resultados que será analizada estadísticamente en la sección 8.

Este protocolo se repite de manera idéntica para cada uno de los 12 genomas de evaluación y para cada una de las tres variantes, produciendo un total de 36 evaluaciones independientes que constituyen la base del análisis estadístico posterior.

7.2.2 Configuración del experimento

El experimento se diseñó siguiendo un esquema de evaluación cruzada entre variantes y genomas, donde cada una de las tres variantes del método es aplicada sobre cada uno de los 12 genomas de evaluación de manera independiente. Este diseño produce una matriz de resultados de $3 \text{ variantes} \times 12 \text{ genomas} = 36 \text{ observaciones}$, que constituye la base del análisis estadístico descrito en la sección 7.4.

El diseño experimental adoptado corresponde a un diseño de bloques completos, donde los genomas actúan como bloques y las variantes como tratamientos. Este tipo de diseño es adecuado cuando se desea comparar múltiples tratamientos bajo condiciones heterogéneas, debido a que permite controlar la variabilidad entre bloques y aislar el efecto de cada tratamiento de manera independiente (Fleiss, 2011). En este contexto, cada genoma representa un escenario distinto con características propias en términos de tamaño, complejidad y número de genes V , lo que introduce una fuente natural de variabilidad entre bloques que el diseño permite controlar estadísticamente sin eliminarla del análisis.

La elección de este diseño responde a una característica fundamental del problema: no es posible asumir que todos los genomas son equivalentes entre sí, debido a que difieren significativamente en su tamaño, en el número de genes V que contienen y en la complejidad de su arquitectura. Si se ignorara esta variabilidad y se trataran todas las observaciones como independientes, el análisis estadístico sobreestimaría la varianza del error y reduciría la potencia de las pruebas de comparación entre variantes. El diseño de bloques completos permite separar la variabilidad debida a las diferencias entre genomas de la variabilidad debida a las diferencias entre variantes, produciendo estimaciones más precisas del efecto de cada configuración del método.

La variable de respuesta principal es el porcentaje de identificación de genes V respecto a la referencia experta, que mide la cobertura del método. Esta métrica fue seleccionada como variable principal porque refleja directamente la utilidad práctica del método para el experto: un método que identifica un mayor porcentaje de genes reales reduce proporcionalmente el esfuerzo de revisión manual. Como variable de respuesta secundaria se utilizó el RMSE, que mide la precisión de las posiciones identificadas respecto a la referencia. La combinación de ambas métricas permite una evaluación integral del desempeño del método que considera tanto su capacidad de recuperar genes reales como la exactitud con la que delimita sus posiciones en el genoma.

Adicionalmente, se registraron las métricas de calidad del agrupamiento para cada variante y genoma, incluyendo el Silhouette Score, el índice de Davies-Bouldin y el índice de Calinski-Harabasz. Estas métricas complementan la evaluación basada en la referencia experta con una perspectiva estadística interna que no depende de las anotaciones manuales, permitiendo analizar si la calidad del agrupamiento está correlacionada con el desempeño en la identificación de genes.

7.3 Análisis estadístico

La evaluación comparativa de las tres variantes del método RIVAM requiere un análisis estadístico riguroso que permita determinar si las diferencias observadas en el desempeño son estadísticamente significativas o atribuibles a la variabilidad natural entre genomas. El análisis se estructuró en cuatro etapas complementarias: estadística descriptiva, evaluación de normalidad, prueba de *Friedman* y análisis post hoc de *Nemenyi*.

7.3.1 Estadística descriptiva

Como primera etapa del análisis, se realizó un análisis descriptivo de los resultados con diagramas de caja y bigotes, con el objetivo de explorar la distribución de las métricas de porcentaje de identificación y RMSE para cada variante y cada genoma. Esta representación permite identificar tendencias generales, valores atípicos y diferencias visuales entre variantes antes de proceder al análisis estadístico formal, proporcionando una primera aproximación al comportamiento del método bajo distintas condiciones.

Los diagramas de caja y bigotes son útiles en este contexto porque permiten analizar simultáneamente la mediana, el rango intercuartílico y los valores extremos de cada distribución, facilitando la comparación visual entre las tres variantes sin asumir ningún supuesto sobre la distribución subyacente de los datos. A diferencia de los gráficos de barras que representan únicamente promedios, los diagramas de caja y bigotes capturan la variabilidad interna de cada distribución, lo que es relevante en este estudio donde la heterogeneidad entre genomas puede ser considerable.

El análisis descriptivo se realizó de manera independiente para cada una de las dos métricas. Para el porcentaje de identificación, el diagrama permitió observar qué variante tiende a recuperar un mayor número de genes V reales de manera consistente entre genomas, identificando si alguna variante presenta una mediana significativamente superior a las demás o si existe una alta variabilidad en su desempeño dependiendo del genoma analizado. Para el RMSE, el diagrama permitió comparar la precisión de las posiciones identificadas entre variantes, observando si alguna configuración produce sistemáticamente errores menores o si las diferencias son marginales.

Adicionalmente, el análisis descriptivo incluyó el cálculo de las estadísticas básicas para cada variante y métrica: media, mediana, desviación estándar, valores mínimo y máximo, y rango intercuartílico. Estas estadísticas complementan la información visual de los diagramas de caja y bigotes y proporcionan una base cuantitativa para la interpretación de los resultados antes de proceder al análisis inferencial. La comparación de estas estadísticas entre variantes permitió identificar patrones generales en el desempeño del método y formular hipótesis sobre las diferencias observadas que serán contrastadas formalmente en las secciones siguientes.

7.3.2 Evaluación de normalidad

Con el fin de determinar el tipo de pruebas estadísticas más adecuadas para la comparación de las variantes, se evaluará el supuesto de normalidad de las métricas analizadas con el uso de dos herramientas complementarias: gráficos de normalidad (QQ-plots) y la prueba de Shapiro-Wilk, aplicada de manera independiente a cada variante y métrica.

Los QQ-plots permiten evaluar visualmente el ajuste de los datos a una distribución normal, comparando los cuantiles observados con los cuantiles teóricos esperados bajo normalidad. La prueba de Shapiro-Wilk complementa esta evaluación con una prueba estadística formal, donde valores de p inferiores a 0.05 indican el rechazo de la hipótesis nula de normalidad. La combinación de ambas herramientas proporciona una evaluación robusta del supuesto de normalidad que no depende exclusivamente de ninguno de los dos enfoques por separado.

Esta evaluación es necesaria porque determina el tipo de prueba estadística más adecuada para la comparación de las variantes: si los datos siguen una distribución normal, se podría emplear un ANOVA de medidas repetidas; en caso contrario, se requiere una prueba no paramétrica que no asuma normalidad en las distribuciones.

7.3.3 Prueba de friedman

Dado que los datos de evaluación corresponden a un diseño de medidas repetidas donde los mismos 12 genomas son evaluados bajo las tres variantes, y considerando que la evaluación de normalidad puede indicar que las métricas no siguen una

distribución normal, se empleó la prueba no paramétrica de *Friedman* como alternativa al ANOVA de medidas repetidas (Friedman, 1937).

La prueba de *Friedman* evalúa la hipótesis nula de que las tres variantes tienen el mismo desempeño promedio sobre los 12 genomas, frente a la hipótesis alternativa de que al menos una variante difiere significativamente de las demás. El estadístico de la prueba se calcula a partir de los rangos asignados a las variantes dentro de cada bloque (genoma), lo que elimina el efecto de las diferencias entre genomas y permite aislar el efecto de cada variante de manera independiente. Este enfoque basado en rangos es robusto cuando las distribuciones de las métricas son asimétricas o presentan valores atípicos, condiciones frecuentes en datos de alineamiento de secuencias donde algunos genomas pueden producir resultados significativamente distintos al resto.

El procedimiento de la prueba consiste en asignar rangos a las tres variantes dentro de cada genoma de manera independiente, de modo que la variante con menor valor recibe el rango 1 y la de mayor valor el rango 3 para las métricas donde un valor mayor es mejor, como el porcentaje de identificación, o a la inversa para las métricas donde un valor menor es mejor, como el RMSE. Una vez asignados los rangos, se calcula el estadístico de *Friedman* χ_F^2 con la siguiente expresión (Ec. 7.1):

$$\chi_F^2 = \left(\frac{12}{Nk(k+1)} \right) \sum_{j=1}^k R_j^2 - 3N(k+1) \quad (\text{Ec. 7.1})$$

Donde N es el número de bloques (genomas), k es el número de variantes y R_j es la suma de rangos de la variante j sobre todos los bloques. Este estadístico sigue aproximadamente una distribución chi-cuadrado con $k - 1$ grados de libertad, lo que permite determinar la significancia estadística de las diferencias observadas.

La prueba se aplicó de manera independiente para cada una de las dos métricas principales: porcentaje de identificación y RMSE. Un valor de p inferior a 0.05 indicará la existencia de diferencias globales estadísticamente significativas entre las variantes para esa métrica, motivando la aplicación del análisis post hoc descrito en la sección siguiente. En caso de que el valor de p sea superior a 0.05, se concluirá que no existen evidencias estadísticas suficientes para afirmar que alguna variante supera a las demás en esa métrica, lo que también constituye un resultado relevante desde el punto de vista metodológico.

La elección de la prueba de *Friedman* sobre otras alternativas no paramétricas se fundamenta en tres características del diseño experimental: el número de variantes comparadas es superior a dos, los datos no garantizan normalidad y se requiere comparar resultados sobre las mismas muestras (genomas). Estas condiciones hacen de *Friedman* la prueba más adecuada para este escenario, siendo ampliamente utilizada

en estudios de comparación de algoritmos de aprendizaje automático sobre múltiples conjuntos de datos (Demšar, 2006).

7.3.4 Análisis post hoc de Nemenyi

Dado que la prueba de *Friedman* solo indica si existen diferencias globales entre las variantes, pero no identifica específicamente entre qué pares se concentran dichas diferencias, se aplicó el análisis post hoc de *Nemenyi* una vez detectadas diferencias globales significativas (Demšar, 2006).

El análisis de *Nemenyi* realiza comparaciones pareadas entre todas las combinaciones posibles de variantes a partir de los rangos promedio obtenidos en la prueba de *Friedman*. Para cada par de variantes, se calcula la diferencia absoluta entre sus rangos promedio y se compara con una diferencia crítica (CD) calculada para un nivel de significancia de $\alpha = 0.05$. Si la diferencia de rangos supera la CD, se concluye que las dos variantes difieren estadísticamente de manera significativa para esa métrica. La diferencia crítica se calcula con la siguiente expresión (Ec. 7.2):

$$CD = q_{\alpha} \sqrt{\frac{k(k+1)}{6N}} \quad (\text{Ec. 7.2})$$

Donde q_{α} es el valor crítico de la distribución *Studentized Range* para el nivel de significancia α y el número de variantes k , y N es el número de bloques (genomas). Para este estudio, con $k = 3$ variantes, $N = 12$ genomas y $\alpha = 0.05$, la diferencia crítica resulta en $CD = 0.8002$.

La elección del análisis de *Nemenyi* sobre otras pruebas post hoc se debe a que no requiere asumir normalidad ni homogeneidad de varianzas, siendo adecuada para comparaciones múltiples derivadas de la prueba de *Friedman* en contextos no paramétricos. Los resultados de este análisis permitirán identificar cuál o cuáles variantes presentan un desempeño estadísticamente superior, orientando la selección de la configuración óptima del método RIVAM para su aplicación en nuevos genomas.

7.4 Comparación con método de referencia

Con el propósito de demostrar el desempeño del método RIVAM dentro del estado del arte, se realizó una comparación directa con IgDetective, el método de referencia más relevante identificado durante la revisión sistemática de la literatura presentada en el Capítulo 4. Esta comparación es fundamental desde la perspectiva de la metodología DSR, que establece que el artefacto desarrollado debe demostrar su utilidad no solo en términos absolutos sino también en relación con las soluciones existentes en el dominio del problema. Sin esta comparación, los resultados del método RIVAM carecerían del

contexto necesario para evaluar su contribución real al campo de la bioinformática computacional aplicada a la caracterización de genes V en genomas de quirópteros.

7.4.1 Descripción de IgDetective

IgDetective, introducido en la Sección 3.3.1 como la herramienta de mayor relevancia dentro del estado del arte para el problema abordado, es el método de referencia seleccionado para esta comparación, el método fue desarrollado por Sirupurapu et al. (2022). Su arquitectura se basa en la combinación de modelos de Markov ocultos construidos a partir de perfiles de dominios conservados de inmunoglobulinas y alineamientos de secuencias contra bases de datos de referencia, lo que le permite identificar regiones candidatas directamente sobre la secuencia sin requerir la ejecución previa de sistemas de alineamiento especializados.

El método opera en tres etapas. En la primera, utiliza perfiles HMM construidos a partir de secuencias de referencia de genes V, D y J de distintas especies de vertebrados para identificar regiones del genoma con similitud a los dominios conservados de las inmunoglobulinas. En la segunda etapa, refina las regiones candidatas con alineamientos de secuencias que permiten delimitar con mayor precisión las fronteras de cada gen candidato. En la tercera etapa, aplica un conjunto de reglas biológicas para filtrar y clasificar los candidatos identificados, eliminando aquellos que no cumplen con las restricciones estructurales esperadas para genes V funcionales o pseudogenes.

Una característica relevante de IgDetective es su capacidad de identificar no solo genes V sino también genes D y J, así como las regiones constantes de las inmunoglobulinas, lo que lo convierte en un método de anotación más amplio que el método RIVAM, cuyo alcance está específicamente delimitado a los genes V. Esta diferencia en el alcance es importante para contextualizar la comparación: mientras IgDetective aborda el problema de la anotación completa del locus de inmunoglobulinas, RIVAM se centra específicamente en la reducción del espacio de búsqueda de genes V como etapa previa a la anotación manual por expertos.

A diferencia del método RIVAM, que opera como una etapa de reducción y selección sobre los resultados de múltiples sistemas de alineamiento previamente ejecutados, IgDetective realiza la identificación de genes V en un único paso integrado que va directamente de la secuencia al conjunto de genes candidatos. Esta diferencia fundamental en el enfoque implica que ambos métodos no son directamente comparables en términos de pipeline completo, sino que representan filosofías distintas de abordar el problema: RIVAM busca reducir y organizar un espacio de búsqueda ya existente, mientras que IgDetective construye su propio espacio de candidatos desde cero. Sin embargo, ambos métodos pueden ser comparados de manera significativa en términos del conjunto de genes V que cada uno es capaz de

recuperar respecto a la referencia de los expertos, debido a que este es el criterio final de utilidad en el contexto de la caracterización del genoma.

Desde el punto de vista de los requisitos de entrada, IgDetective requiere únicamente la secuencia en formato FASTA y las bases de datos de referencia de inmunoglobulinas, mientras que el método RIVAM requiere los archivos GFF obtenidos por cinco sistemas de alineamiento especializados. Esta diferencia tiene implicaciones prácticas en dos sentidos: por un lado, IgDetective es más accesible para laboratorios que no disponen de la infraestructura para ejecutar los cinco sistemas de alineamiento; por otro, su estrategia basada en señales conservadas presenta, como se documentó en la Sección 3.3.1, limitaciones de generalización en especies con variabilidad genética distinta a la de los organismos de entrenamiento, condición que caracteriza a la mayoría de los quirópteros analizados en esta investigación.

7.4.2 Criterios de comparación

La comparación entre el método RIVAM y IgDetective se realizó sobre la base de criterios que permitan una evaluación justa y representativa de las fortalezas y limitaciones de cada enfoque, considerando tanto su desempeño cuantitativo como su aplicabilidad práctica en el contexto de la caracterización del genoma en quirópteros. La selección de estos criterios responde a la naturaleza dual del problema: por un lado, la necesidad de recuperar el mayor número posible de genes V reales, y por otro, la necesidad de que las posiciones identificadas sean precisas y estructuralmente coherentes con la arquitectura del gen V.

El primer criterio es el porcentaje de identificación de genes funcionales, definido como la proporción de genes V funcionales anotados por expertos que son correctamente recuperados por cada método respecto al total de genes funcionales de referencia. Este criterio es el más relevante desde el punto de vista biológico, debido a que los genes funcionales son los que presentan una estructura completa y capacidad de expresión, constituyendo el objeto de mayor interés para el experto en inmunogenética.

A diferencia del porcentaje de identificación general utilizado en la evaluación comparativa entre variantes, este criterio se restringe exclusivamente a genes funcionales, lo que representa una evaluación más exigente y biológicamente significativa. Un método que recupera un alto porcentaje de genes funcionales reduce directamente la carga de revisión manual del experto, que debe verificar cada candidato para determinar su funcionalidad. Adicionalmente, se reportó el número absoluto de genes funcionales recuperados por cada método, que permite contextualizar las diferencias porcentuales en términos del volumen real de genes que cada método es capaz de identificar correctamente.

El segundo criterio es la precisión estructural en la delimitación del gen V, que evalúa la capacidad de cada método para identificar correctamente las tres secciones que componen el gen V: el péptido señal (SP), el exón V y la señal de recombinación (RSS). Este criterio va más allá de la simple detección del gen y evalúa la calidad de la delimitación estructural, que tiene implicaciones directas en la calidad de la anotación funcional y en estudios posteriores de diversidad inmunogenética.

Un método que detecta el gen, pero delimita incorrectamente sus secciones puede producir anotaciones que, si bien identifican la región general del gen, no son suficientemente precisas para su uso en análisis funcionales o comparativos. La comparación de este criterio entre ambos métodos se realizó con un análisis visual en IGV sobre ejemplos representativos de cada genoma evaluado, donde se podrá observar directamente la concordancia entre las predicciones de cada método y la referencia experta en términos de la delimitación de las tres secciones del gen V.

Como criterio complementario, aplicable únicamente en el escenario del genoma sin anotación experta (*Molossus nigricans*), se consideró el tiempo de ejecución de cada método. En este escenario específico, donde no es posible calcular métricas de identificación directas al no contar con una referencia validada, la eficiencia computacional constituye el criterio principal de comparación.

Este criterio es relevante en el contexto de la caracterización del genoma exploratoria, donde los tiempos de procesamiento pueden ser determinantes para la factibilidad del análisis en laboratorios con recursos computacionales limitados. La comparación del tiempo de ejecución entre ambos métodos permitirá evaluar la viabilidad práctica de cada enfoque en condiciones de alta complejidad y ausencia de información de referencia previa.

7.4.3 Genomas seleccionados para la comparación

La comparación con IgDetective se realizó sobre un subconjunto de seis genomas seleccionados del conjunto de evaluación y validación, cubriendo distintos niveles de complejidad estructural y tamaños de repertorio de genes V. La selección de este subconjunto respondió a tres criterios principales: la diversidad taxonómica dentro del orden *Chiroptera*, la variabilidad en el número de genes funcionales de referencia disponibles y la representatividad de los escenarios evaluados respecto al dominio de aplicación del método.

Los seis genomas seleccionados abarcan un rango amplio en términos de tamaño de repertorio, desde *Aselliscus stoliczkanus* con 38 genes funcionales de referencia hasta *Plecotus auritus* con 653 genes funcionales, lo que permite evaluar el comportamiento comparativo de ambos métodos bajo distintos niveles de complejidad. La inclusión de *Plecotus auritus*, perteneciente al conjunto de validación externa, responde a la

necesidad de evaluar la comparación en un escenario con un repertorio de genes V considerablemente mayor al observado en los genomas de evaluación, proporcionando una prueba crítica de la escalabilidad de ambos métodos.

Adicionalmente, se incluyó el genoma de *Molossus nigricans* como sexto caso de comparación exploratoria. A diferencia de los cinco genomas anteriores, este genoma no cuenta con anotación experta disponible, por lo que la comparación se limita al número de regiones candidatas identificadas y al tiempo de ejecución requerido por cada método, tal como se describió en la sección 7.4.2.

La Tabla 7.3 presenta los genomas seleccionados para la comparación con IgDetective, incluyendo el número total de genes de referencia, el número de genes funcionales de referencia y el conjunto al que pertenece cada genoma dentro del estudio.

Tabla 7.3 Genomas seleccionados para la comparación con IgDetective

Genoma	ID	Genes ref. totales	Genes func. ref.	Conjunto
<i>Antrozous pallidus</i>	Anpa	90	56	Evaluación
<i>Aselliscus stoliczkanus</i>	ASST	48	38	Evaluación
<i>Eptesicus fuscus</i>	Epfu	168	77	Evaluación
<i>Hipposideros larvatus</i>	Hila	121	86	Evaluación
<i>Plecotus auritus</i>	Plau	850	653	Validación externa
<i>Molossus nigricans</i>	Moni	No disponible	No disponible	Validación externa

Los resultados de esta comparación, incluyendo el análisis detallado del desempeño de ambos métodos sobre cada genoma y la discusión de las diferencias observadas, se presentan en el Capítulo 8.

Capítulo 8. Resultados y discusión

Este capítulo presenta los resultados de la evaluación sistemática del método RIVAM organizados en cuatro bloques: la evaluación comparativa de las tres variantes sobre los 12 genomas de evaluación, la validación externa sobre tres genomas adicionales, la comparación con IgDetective como referencia del estado del arte, y la discusión de los hallazgos en el contexto del problema planteado. A lo largo del capítulo se emplean tres métricas principales: el porcentaje de identificación, que mide la proporción de genes V de referencia correctamente recuperados; el número de genes identificados, que reporta el valor absoluto de genes V recuperados; y el RMSE, que evalúa la precisión en la localización de los genes identificados. Estas métricas permiten analizar tanto la cobertura como la precisión del método en cada escenario evaluado.

8.1 Resultados de la evaluación comparativa

La evaluación comparativa se llevó a cabo sobre los 12 genomas de evaluación descritos en la Sección 7.1.1, aplicando de manera consistente las tres variantes del método RIVAM con sus hiperparámetros optimizados: GMM, PCA+GMM y UMAP+GMM. El proceso se estructuró en cuatro etapas complementarias: un análisis descriptivo inicial, una evaluación de los supuestos estadísticos, un análisis inferencial con la prueba de *Friedman* y el análisis post hoc de Nemenyi, y finalmente la selección de la variante óptima para la fase de validación externa.

8.1.1 Estadística descriptiva y análisis visual

Como primera etapa del análisis, se calcularon las estadísticas descriptivas de las tres métricas principales porcentaje de identificación, RMSE y número de genes identificados para cada una de las tres variantes evaluadas. La Tabla 8.1 resume los valores de tendencia central y dispersión obtenidos sobre los 12 genomas de evaluación.

Tabla 8.1 Estadísticas descriptivas de las métricas de evaluación por variante

Métrica	Variante	Media	Mediana	Desv. Est.	Mín.	Máx.
Porcentaje de identificación	GMM	0.7054	0.7357	0.0953	0.5185	0.8125
	PCA+GMM	0.6220	0.6426	0.1212	0.4395	0.7736
	UMAP+GMM	0.7498	0.7347	0.0709	0.6708	0.8994
Error cuadrático medio de la raíz (RMSE)	GMM	2.6931	2.6655	0.6033	1.6159	4.0151
	PCA+GMM	2.7673	2.8398	0.4790	1.8670	3.6597
	UMAP+GMM	2.7470	2.7519	0.5770	1.5021	3.7579
Genes identificados	GMM	129.00	105.50	64.71	39	220
	PCA+GMM	115.67	103.00	60.55	23	214
	UMAP+GMM	135.08	133.50	59.98	39	218

En términos de porcentaje de identificación, UMAP+GMM produce la media más alta (0.7498) y la menor dispersión ($\sigma = 0.0709$), lo que indica no solo una mayor cobertura promedio sino también un comportamiento más estable entre genomas. GMM presenta una media intermedia (0.7054) con una dispersión moderada, mientras que PCA+GMM muestra la media más baja (0.6220) y la mayor variabilidad ($\sigma = 0.1212$), sugiriendo un desempeño menos consistente en distintos escenarios de la identificación de los genes V.

En cuanto al RMSE, los tres modelos presentan valores medios similares, comprendidos entre 2.69 y 2.77, lo que indica que la precisión en la localización de los genes identificados es comparable entre las tres variantes. La incorporación de técnicas

de reducción de dimensionalidad, tanto lineal como no lineal, no parece afectar de forma relevante el error de localización respecto al modelo GMM base.

Respecto al número de genes identificados, UMAP+GMM presenta la media más alta (135.08) con una mediana de 133.50, mientras que GMM (129.00) y PCA+GMM (115.67) recuperan en promedio un número menor de genes. Esta diferencia absoluta es consistente con las tendencias observadas en el porcentaje de identificación y refuerza la ventaja de UMAP+GMM en términos de cobertura.

La Tabla 8.2 presenta los resultados de identificación obtenidos por las tres variantes del método RIVAM sobre los 12 genomas de evaluación. Cada fila corresponde a un genoma identificado por su abreviatura. La columna Referencia indica el número total de genes V anotados manualmente por expertos del CISEI, que constituye el estándar de comparación. Las columnas Genes identificados reportan cuántos de esos genes fueron correctamente recuperados por cada variante, mientras que las columnas Cobertura % representa ese valor como proporción del total de la referencia. Las tres variantes evaluadas son: GMM como configuración base sin reducción de dimensionalidad, PCA+GMM con reducción lineal, y UMAP+GMM con reducción no lineal.

Tabla 8.2 Resultados detallados por genoma para las tres variantes del método RIVAM

ID	Referencia	Genes identificados (GMM)	Cobertura % (GMM)	Genes identificados (PCA+GMM)	Cobertura % (PCA+GMM)	Genes identificados (UMAP+GMM)	Cobertura % (UMAP+GMM)
Anpa	90	73	81.11	69	76.67	76	84.44
Asst	48	39	81.25	23	47.92	39	81.25
Epfu	168	124	73.81	97	57.74	124	73.81
Epni	248	191	77.02	109	43.95	192	77.42
Hila	121	84	69.42	57	47.11	84	69.42
Misc	135	70	51.85	71	52.59	93	68.89
Nyav	300	220	73.33	214	71.33	211	70.33
Pina	228	162	71.05	161	70.61	157	68.86
Pipy	268	199	74.25	182	67.91	196	73.13
Rhna	117	87	74.36	85	72.65	88	75.21
Sale	159	84	52.83	123	77.36	143	89.94
Vemu	325	215	66.15	197	60.62	218	67.08

Un análisis por genoma revela comportamientos diferenciados entre variantes. En la mayoría de los genomas, UMAP+GMM mantiene o supera el porcentaje de identificación de GMM; los casos más destacados son *Saccolpteryx leptura* (Sale), donde UMAP+GMM alcanza un 89.94% frente al 52.83% de GMM, y *Miniopterus schreibersii* (Misc), donde

UMAP+GMM obtiene un 68.89% frente al 51.85% de GMM. PCA+GMM, por su parte, muestra caídas importantes en algunos genomas, con porcentajes tan bajos como 43.95% en *Eptesicus nilssonii* y 47.11% en *Hipposideros larvatus*, ambos notablemente inferiores a las otras dos variantes.

Existen, no obstante, genomas donde GMM o PCA+GMM superan a UMAP+GMM. En *Nyctalus aviator* (Nyav), GMM identifica 220 genes frente a los 211 de UMAP+GMM. En *Pipistrellus nathusii* (Pina), las tres variantes se comportan de forma similar, con diferencias menores al 3%. Estos casos aislados no revierten la tendencia general pero sí evidencian que el comportamiento del método es sensible a la estructura particular de cada genoma.

Las Figuras 8.1, 8.2 y 8.3 presentan los diagramas de caja del porcentaje de identificación, el RMSE y el número de genes identificados respectivamente, permitiendo ver la distribución y dispersión de cada variante sobre el conjunto de 12 genomas.

La Figura 8.1 confirma la ventaja de UMAP+GMM en términos de cobertura: su caja se sitúa en la parte alta del rango con una mediana cercana a 0.74 y una dispersión reducida, lo que indica un comportamiento consistente entre genomas. GMM presenta una distribución similar en mediana, pero con tres valores atípicos por debajo de 0.55, correspondientes a los genomas donde la configuración base tiene mayor dificultad. PCA+GMM es la variante con mayor variabilidad, su caja es la más extendida y su mediana cae alrededor de 0.64, con un mínimo cercano al 44% correspondiente al genoma de *Eptesicus nilssonii*, lo que evidencia su menor consistencia entre genomas.

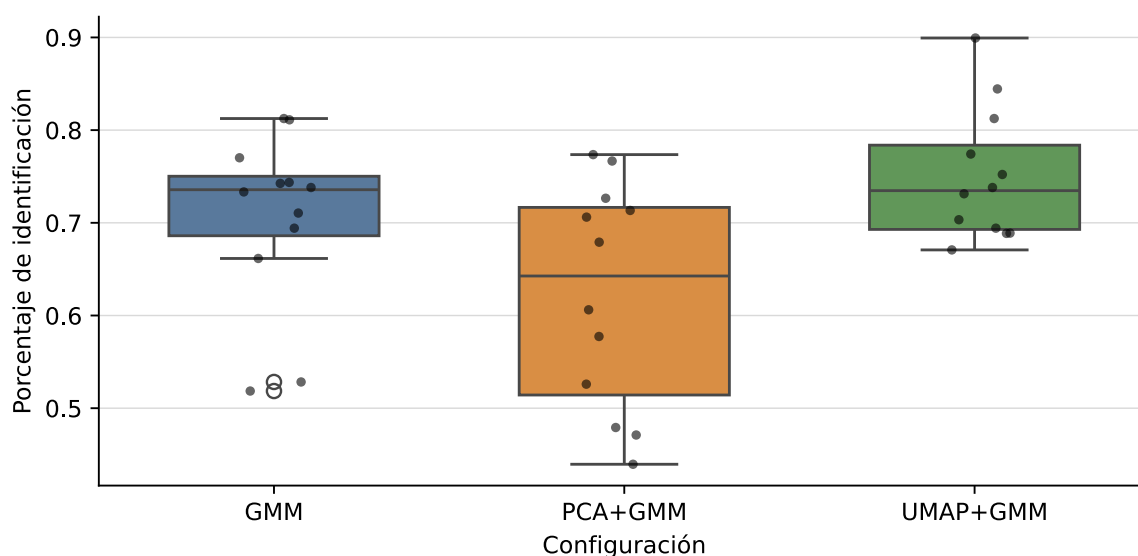


Figura 8.1 Boxplot del porcentaje de identificación por variante

La Figura 8.2 muestra que las tres variantes presentan distribuciones de RMSE muy similares, con medianas comprendidas entre 2.7 y 2.85 y rangos intercuartílicos que se solapan ampliamente. Destaca que GMM concentra sus valores en una banda más estrecha, mientras que PCA+GMM y UMAP+GMM presentan una dispersión ligeramente mayor. Los valores atípicos aparecen en las tres variantes sin un patrón consistente, lo que anticipa la ausencia de diferencias estadísticamente significativas en esta métrica.

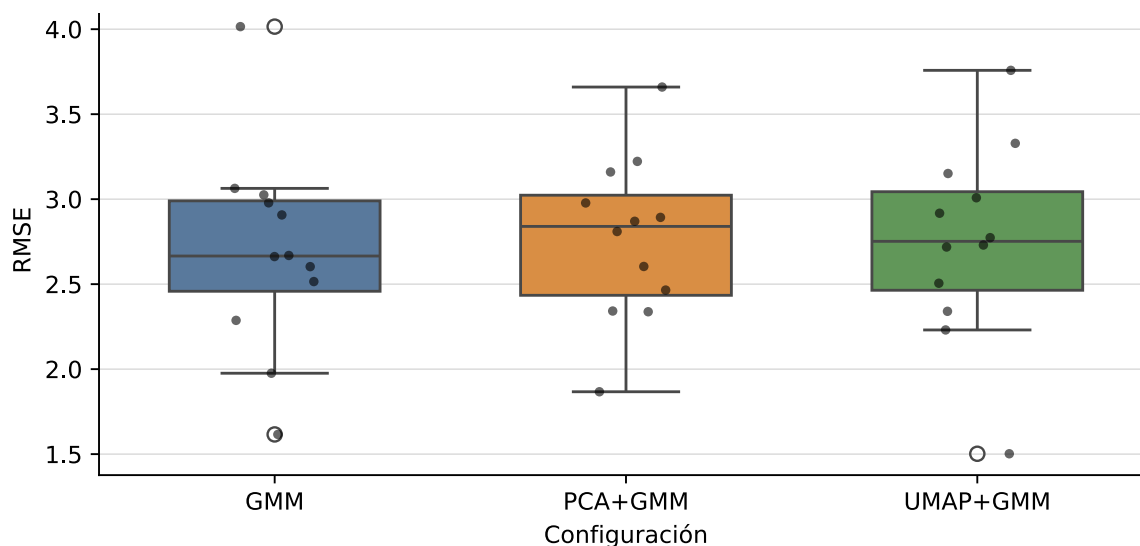


Figura 8.2 Boxplot de RMSE por variante

La Figura 8.3 refuerza las tendencias observadas en cobertura. UMAP+GMM presenta la mediana más alta en número de genes identificados, aproximadamente 133, con una distribución más homogénea que GMM. Este último muestra una caja considerablemente más asimétrica, con la mediana desplazada hacia la parte baja de la caja, indicando que en varios genomas la identificación cae por debajo de la mitad del rango. PCA+GMM mantiene la mediana más baja y la mayor dispersión de las tres variantes, reforzando la conclusión de que la reducción lineal de dimensionalidad no captura de manera efectiva la estructura del espacio de búsqueda de genes V en todos los genomas evaluados.

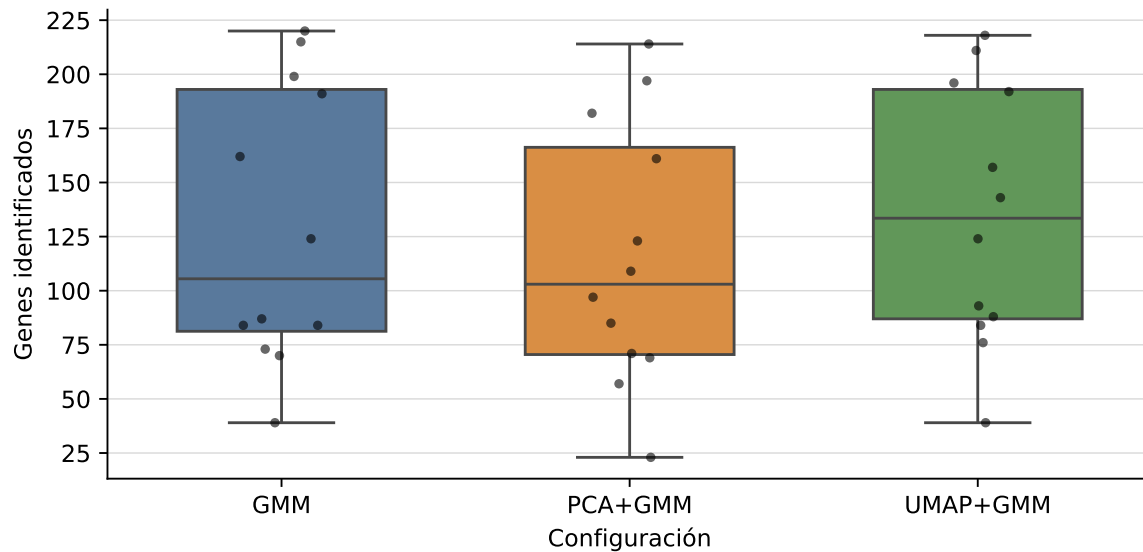


Figura 8.3 Boxplot del número de genes identificados por variante

Los diagramas de caja confirman las tendencias descritas en la Tabla 8.1. UMAP+GMM presenta la distribución más compacta y centrada en valores altos para el porcentaje de identificación, mientras que PCA+GMM muestra la caja más extendida y desplazada hacia valores inferiores. Para el RMSE, las tres distribuciones se solapan considerablemente, lo que anticipa la ausencia de diferencias estadísticamente significativas en esta métrica, como se verifica en la Sección 8.1.3.

8.1.2 Evaluación de normalidad

Con el propósito de determinar el tipo de prueba estadística más adecuada para la comparación de las variantes, se evaluó el supuesto de normalidad de las métricas analizadas con dos herramientas complementarias: gráficos de normalidad (QQ-plots) y la prueba de Shapiro-Wilk, aplicada de manera independiente a cada variante y métrica.

La Figura 8.4 muestra los QQ-plots del RMSE para las tres variantes. En los tres casos los puntos siguen de cerca la línea teórica a lo largo de todo el rango, con desviaciones mínimas en los extremos, lo que es consistente con una distribución aproximadamente normal. GMM presenta el punto más alejado en el extremo superior, correspondiente al valor atípico de *Pipistrellus nathusii* (RMSE = 4.015) observado en la Tabla 8.2, aunque esta desviación no es suficiente para rechazar la normalidad.

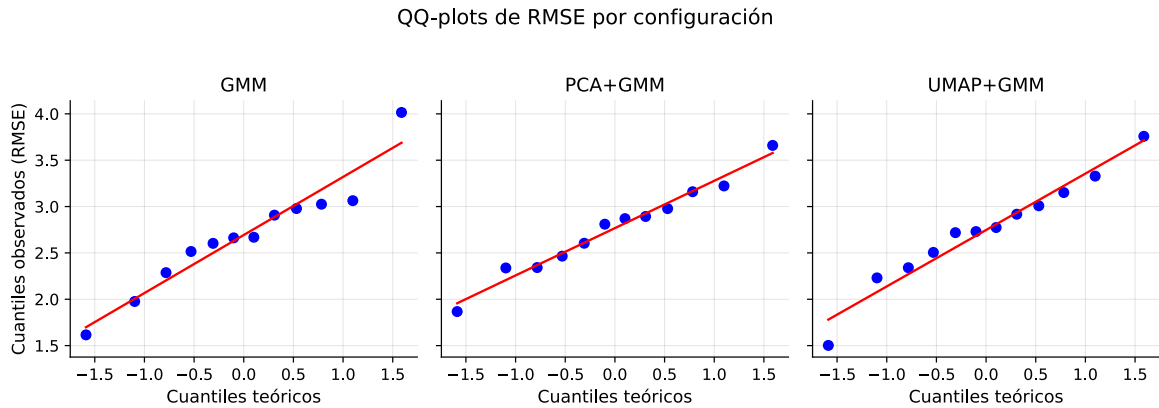


Figura 8.4 QQ-plots de RMSE por variante

La Figura 8.5 muestra los QQ-plots del porcentaje de identificación. Aquí el comportamiento es notablemente distinto entre variantes. En GMM se observa una curvatura en el extremo inferior, donde dos puntos se alejan marcadamente de la línea teórica hacia valores más bajos de lo esperado, correspondientes a los genomas Sale y Misc donde la cobertura cae por debajo del 55%. Esta desviación es la que determina la no normalidad de GMM en esta métrica. PCA+GMM y UMAP+GMM presentan una distribución más cercana a la línea teórica, aunque PCA+GMM muestra cierta dispersión en el extremo inferior que refleja su mayor variabilidad general.

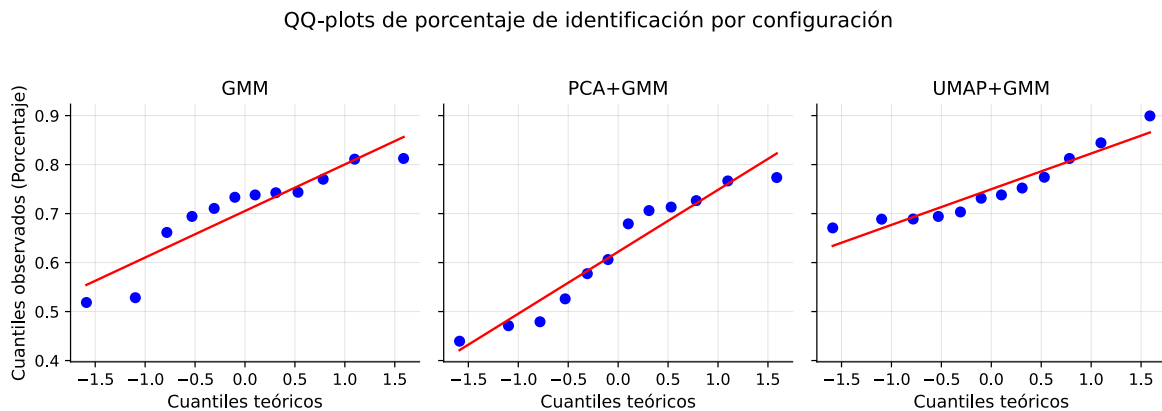


Figura 8.5 QQ-plots de porcentaje de identificación por variante

Los resultados formales de la prueba de Shapiro-Wilk se presentan en las Tablas 8.3, 8.4 y 8.5, organizadas por métrica. Para cada variante se reporta el estadístico W , cuyo valor máximo es 1 e indica perfecta normalidad, y el p-valor asociado. Un p-valor menor a 0.05 lleva a rechazar la hipótesis de normalidad. La columna Normalidad resume el resultado de esta decisión para cada caso.

Tabla 8.3 Prueba de Shapiro-Wilk para el porcentaje de identificación

Variante	Estadístico W	p-valor	Normalidad
GMM	0.8502	0.0370	No normal
PCA+GMM	0.9090	0.2070	Normal
UMAP+GMM	0.9024	0.1706	Normal

Tabla 8.4 Prueba de Shapiro-Wilk para el RMSE

Variante	Estadístico W	p-valor	Normalidad
GMM	0.9500	0.6364	Normal
PCA+GMM	0.9858	0.9975	Normal
UMAP+GMM	0.9751	0.9560	Normal

Tabla 8.5 Prueba de Shapiro-Wilk para el número de genes identificados

Variante	Estadístico W	p-valor	Normalidad
GMM	0.8875	0.1093	Normal
PCA+GMM	0.9506	0.6458	Normal
UMAP+GMM	0.9331	0.4147	Normal

En el RMSE y el número de genes identificados las tres variantes cumplen el supuesto de normalidad. Sin embargo, el porcentaje de identificación de GMM no sigue una distribución normal ($W = 0.8502$, $p = 0.037$), resultado consistente con la curvatura observada en la Figura 8.5.

Con base en estos resultados, se optó por la prueba no paramétrica de *Friedman* para la comparación entre variantes, decisión que se sustenta en tres razones complementarias. Primero, la métrica principal de evaluación viola el supuesto de normalidad en una de las variantes, lo que invalida el uso consistente de pruebas paramétricas para todas las métricas.

Segundo, el diseño experimental corresponde a un esquema de medidas repetidas sobre los mismos 12 genomas, para el cual *Friedman* es la alternativa no paramétrica más adecuada. Tercero, con un tamaño de muestra de 12 bloques, los *Tests* de normalidad tienen potencia limitada y los resultados de pruebas paramétricas son menos confiables. Por estas razones, *Friedman* se aplica de forma consistente a las tres métricas, garantizando la comparabilidad de los resultados (Friedman, 1937).

8.1.3 Prueba de friedman

La prueba de *Friedman* se aplicó de manera independiente a cada métrica de evaluación con el objetivo de determinar si existen diferencias estadísticamente significativas entre

las tres variantes del método RIVAM sobre los 12 genomas de evaluación. La Tabla 8.6 resume los resultados.

Tabla 8.6 Resultados de la prueba de *Friedman* por métrica

Métrica	χ^2	p-valor	Diferencias significativas
Porcentaje de identificación	8.9333	0.0115	Sí ($p < 0.05$)
RMSE	0.6667	0.7165	No ($p \geq 0.05$)
Genes identificados	8.9333	0.0115	Sí ($p < 0.05$)

En primer lugar, para el porcentaje de identificación la prueba arrojó un estadístico $\chi^2 = 8.9333$ con un p-valor de 0.0115, por debajo del nivel de significancia $\alpha = 0.05$. Este resultado indica que al menos un par de variantes difiere significativamente en términos de cobertura, lo que justifica la aplicación del análisis post hoc de *Nemenyi* descrito en la Sección 8.1.4.

En segundo lugar, para el RMSE el estadístico fue $\chi^2 = 0.6667$ con un p-valor de 0.7165, muy por encima del umbral de significancia. Este resultado confirma que las tres variantes son estadísticamente equivalentes en cuanto a la precisión de localización de los genes V identificados. La incorporación de reducción de dimensionalidad ya sea lineal con PCA o no lineal con UMAP, no compromete la capacidad del método para delimitar correctamente las posiciones de los genes V.

Finalmente, para el número de genes identificados el estadístico fue $\chi^2 = 8.9333$ con un p-valor de 0.0115, coincidiendo exactamente con el resultado del porcentaje de identificación. Esta concordancia es esperable dado que ambas métricas cuantifican dimensiones complementarias de la cobertura del método y están directamente relacionadas entre sí.

8.1.4 Análisis post hoc de *Nemenyi*

Dado que la prueba de *Friedman* detectó diferencias significativas en el porcentaje de identificación y en el número de genes identificados, se aplicó el análisis post hoc de *Nemenyi* para determinar específicamente entre qué pares de variantes se concentran dichas diferencias (Demšar, 2006). El análisis se realizó con una diferencia crítica (CD) de 0.8002, calculada para un nivel de significancia de $\alpha = 0.05$ con 12 genomas evaluados. Las comparaciones se realizaron de manera independiente para cada métrica. En el RMSE, tal como anticipaba el resultado de *Friedman*, ninguna comparación pareada superó la diferencia crítica, como se muestra en la Tabla 8.7.

Tabla 8.7 Análisis post hoc de *Nemenyi* para RMSE

Variante 1	Variante 2	Diferencia de rangos	CD	Diferencia significativa
GMM	PCA+GMM	0.50	0.8002	No
GMM	UMAP+GMM	0.75	0.8002	No
PCA+GMM	UMAP+GMM	0.25	0.8002	No

La Figura 8.6 muestra el diagrama de diferencia crítica para el RMSE. Las tres variantes quedan conectadas por una única barra gruesa, lo que confirma visualmente que ningún par supera la diferencia crítica de 0.800. Los puntos de GMM y UMAP+GMM se sitúan en los extremos del eje con PCA+GMM en el centro, pero las diferencias entre ellos son menores a la mitad de la CD, evidenciando una equivalencia estadística sólida en esta métrica.

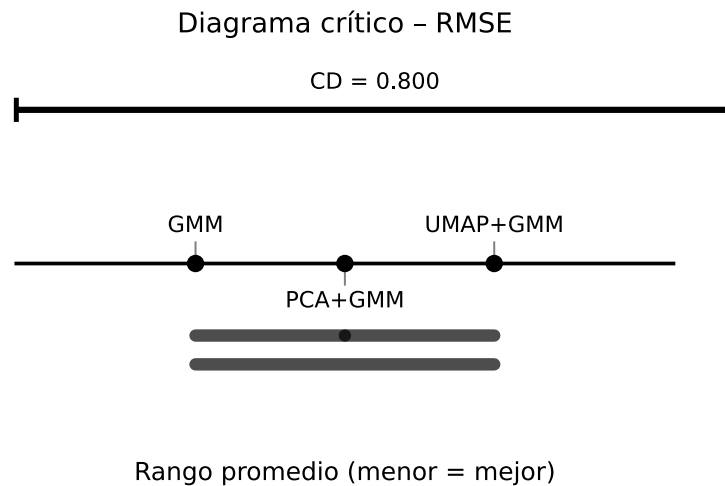


Figura 8.6 Diagrama de diferencia crítica para RMSE

En el porcentaje de identificación, el análisis reveló diferencias estadísticamente significativas en dos de las tres comparaciones pareadas, como se muestra en la Tabla 8.8.

Tabla 8.8 Análisis post hoc de *Nemenyi* para porcentaje de identificación

Variante 1	Variante 2	Diferencia de rangos	CD	Diferencia significativa
UMAP+GMM	GMM	0.25	0.8002	No
UMAP+GMM	PCA+GMM	1.13	0.8002	Sí
GMM	PCA+GMM	0.88	0.8002	Sí

La Figura 8.7 muestra el diagrama de diferencia crítica para el porcentaje de identificación. UMAP+GMM ocupa el extremo izquierdo del eje, correspondiente al mejor rango promedio, seguido de GMM al centro y PCA+GMM en el extremo derecho con el peor rango. Una barra conecta UMAP+GMM con GMM, indicando que su diferencia no supera la CD y por tanto no son estadísticamente distintos. PCA+GMM en

cambio queda desconectado de UMAP+GMM, confirmando la diferencia significativa entre ambos. La barra que conecta GMM con PCA+GMM es más corta que la CD pero la supera ligeramente, lo que refleja la diferencia significativa entre estas dos variantes también.

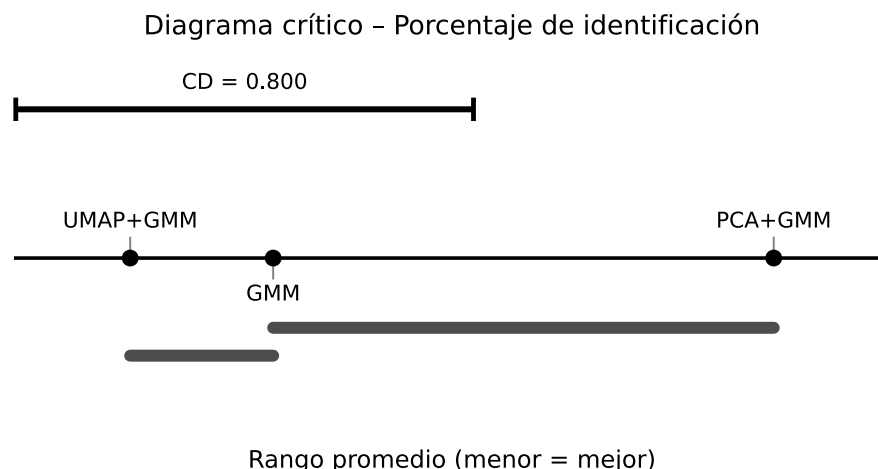


Figura 8.7 Diagrama de diferencia crítica para porcentaje de identificación

Las diferencias significativas se concentran en las comparaciones que involucran a PCA+GMM, lo que evidencia que esta variante presenta un desempeño estadísticamente inferior en términos de cobertura. La reducción lineal mediante PCA no preserva de manera adecuada la información estructural relevante para la identificación de genes V: al proyectar los datos en las direcciones de mayor varianza global, descarta variaciones locales que resultan determinantes para separar las regiones candidatas con genes V reales del resto del espacio de búsqueda.

Por el contrario, UMAP+GMM y GMM son estadísticamente equivalentes en cobertura, aunque UMAP+GMM obtiene el mejor rango promedio en esta métrica sin incrementar el error de localización, lo que le otorga una ventaja práctica en escenarios de mayor complejidad estructural.

8.1.5 Selección de la variante óptima

Los resultados del análisis estadístico permiten establecer una jerarquía clara entre las tres variantes evaluadas, con implicaciones tanto metodológicas como prácticas para el problema de reducción del espacio de búsqueda en la identificación de genes V. La variante PCA+GMM queda descartada como opción óptima a partir de evidencia estadística consistente. Su desempeño en porcentaje de identificación es significativamente inferior tanto frente a GMM (diferencia de rangos = 0.88, $p < 0.05$) como con UMAP+GMM (diferencia de rangos = 1.13, $p < 0.05$), y su dispersión es la más alta de las tres variantes en todas las métricas evaluadas.

Esta limitación no es trivial ni atribuible a casos aislados; se presenta de forma sistemática en nueve de los doce genomas evaluados, con caídas de cobertura que en algunos casos superan los 30 puntos porcentuales respecto a las otras variantes. El origen de este comportamiento reside en la naturaleza de la transformación lineal que PCA aplica sobre el espacio de representación: al proyectar los datos en las direcciones de máxima varianza global, descarta estructuras locales que, aunque de menor magnitud estadística, son biológicamente relevantes para discriminar las regiones candidatas con genes V reales del resto del espacio de búsqueda. En un dominio donde la variabilidad genética es precisamente el rasgo definitorio de los genes V, una transformación que prioriza la varianza global resulta inadecuada como paso previo al agrupamiento.

La comparación entre GMM y UMAP+GMM es más sutil pero igualmente informativa. Desde el punto de vista estadístico, ambas variantes son equivalentes en las tres métricas evaluadas: la prueba de *Friedman* no detecta diferencias globales en RMSE, y el análisis de *Nemenyi* confirma que la diferencia de rangos en cobertura entre ambas (0.25) se mantiene por debajo de la diferencia crítica. Sin embargo, la equivalencia estadística no implica equivalencia práctica. UMAP+GMM obtiene el mejor rango promedio en porcentaje de identificación y supera a GMM en casos concretos: en *Saccopteryx leptura* la cobertura de UMAP+GMM alcanza el 89.94% frente al 52.83% de GMM, una diferencia de 37 puntos porcentuales, y en *Miniopterus schreibersii* obtiene un 68.89% frente al 51.85%, con una diferencia de 17 puntos. En ninguno de los doce genomas evaluados GMM logra una ventaja comparable sobre UMAP+GMM.

Esta asimetría en los casos extremos revela una propiedad importante del método. Mientras GMM opera directamente sobre el espacio original de alta dimensionalidad, UMAP proyecta los datos en una representación latente que preserva estructuras topológicas no lineales, lo que le confiere mayor robustez frente a la heterogeneidad estructural entre genomas. Esta capacidad de adaptación resulta relevante en el contexto del método RIVAM, cuyo propósito es generalizar a genomas con distintos niveles de complejidad, incluyendo especies no caracterizadas previamente.

Por estas razones, UMAP+GMM se selecciona como la variante óptima para la fase de validación externa descrita en la Sección 8.2. Esta decisión está respaldada por la evidencia estadística, es consistente con los principios de diseño establecidos en el Capítulo 5, en particular con los principios de coherencia biológica y reusabilidad, y refleja una elección metodológicamente fundamentada que privilegia la estabilidad y la capacidad de generalización del método sobre el dominio biológico en el que opera.

8.2 Resultados de la validación externa

Con el propósito de evaluar la capacidad de generalización del método RIVAM más allá de los genomas utilizados durante la fase de evaluación comparativa, se llevó a cabo una validación externa empleando la configuración UMAP+GMM, seleccionada como variante óptima en la Sección 8.1.5. Esta validación se realizó sobre tres genomas con características deliberadamente distintas entre sí, siguiendo el mismo flujo metodológico descrito en el Capítulo 6, sin modificaciones en el proceso de agrupamiento ni en los hiperparámetros optimizados durante la fase de desarrollo.

Los tres escenarios de validación fueron seleccionados para cubrir condiciones representativas de la caracterización del genoma en un entorno real. El primer genoma corresponde a un quiróptero con anotación experta y un repertorio de genes V considerablemente mayor al observado durante la evaluación, lo que permite verificar si el método escala adecuadamente ante un volumen de referencia superior. El segundo corresponde a una especie filogenéticamente distinta a los quirópteros, con características estructurales del gen V marcadamente diferentes, lo que permite evaluar la transferibilidad del método a dominios taxonómicos no contemplados durante el diseño. El tercero corresponde a un quiróptero sin anotación experta disponible, escenario en el que el análisis se orienta hacia la eficiencia computacional y la coherencia estructural de las regiones candidatas identificadas.

8.2.1 Validación con *Plecotus Auritus*

La primera validación externa se llevó a cabo sobre el genoma de *Plecotus auritus* (Plau), un quiróptero con una referencia anotada por expertos de 850 genes V, que incluye tanto genes funcionales como no funcionales. Este genoma representa el escenario más cercano a los utilizados durante la fase de evaluación, aunque con un repertorio de genes V considerablemente mayor al observado durante la optimización, lo que permite verificar si el método mantiene su capacidad de identificación al enfrentarse a un volumen de referencia superior. Cabe señalar que en la Sección 8.3, dedicada a la comparación con IgDetective, el análisis se centra exclusivamente en los 653 genes funcionales del mismo genoma, por lo que los resultados de ambas secciones no son directamente comparables en términos absolutos.

Con el propósito de analizar si el comportamiento observado durante la evaluación comparativa se mantiene en este escenario de validación, se reportan los resultados de las tres variantes del método RIVAM, tal como se presenta en la Tabla 8.9.

Tabla 8.9 Resultados de la validación sobre el genoma de *Plecotus auritus*

Variante	Referencia	Genes identificados	Cobertura (%)	RMSE
GMM	850	686	80.71	3.0734
PCA+GMM	850	707	83.18	2.6066
UMAP+GMM	850	774	91.06	3.4594

La Figura 8.8 muestra la proyección UMAP del espacio de búsqueda sobre el genoma de *Plecotus auritus*, con los grupos identificados por el modelo GMM superpuestos. Los ejes representan las dos componentes del espacio latente generado por UMAP, denominadas Componente UMAP 1 y Componente UMAP 2, y no corresponden a variables biológicas directas sino a dimensiones abstractas que preservan las relaciones de proximidad estructural entre las regiones candidatas del espacio de búsqueda original. La distancia entre puntos en la proyección refleja similitud en los atributos de alineamiento utilizados para representar cada región candidata: puntos cercanos comparten características estructurales similares, mientras que puntos distantes presentan perfiles marcadamente distintos.

En la proyección se distinguen tres agrupamientos identificados por el modelo. El Grupo 1 (azul oscuro) agrupa 11,062 regiones candidatas en la zona central superior del espacio proyectado, con una distribución densa y relativamente compacta alrededor de su centroide. El Grupo 2 (azul claro) es el de mayor extensión, con 20,616 regiones candidatas distribuidas de forma más dispersa hacia la derecha del espacio, con ramificaciones que se extienden en dirección diagonal. El Grupo 3 (naranja) concentra 3,712 regiones en la zona inferior izquierda, con una distribución elongada verticalmente y un núcleo denso bien definido alrededor de su centroide.

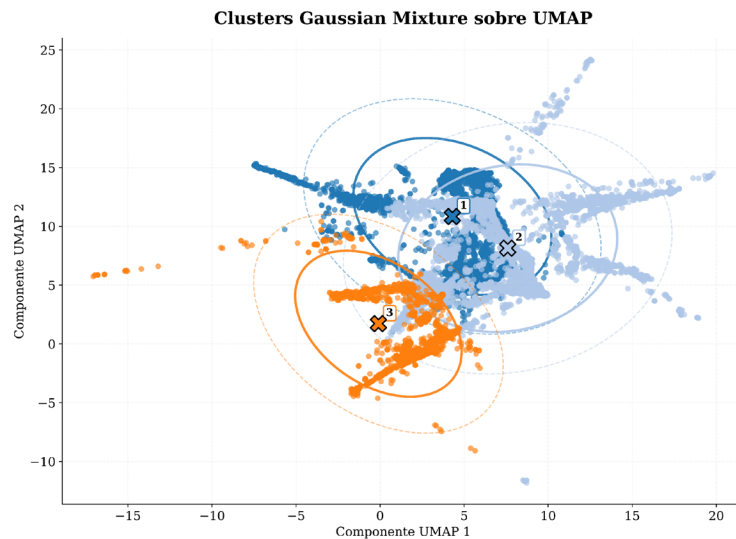


Figura 8.8 Proyección UMAP con grupos GMM sobre el genoma de *Plecotus auritus*

Las elipses continuas representan las componentes gaussianas ajustadas por el modelo GMM sobre el espacio UMAP, mientras que las elipses discontinuas corresponden a niveles de contorno adicionales que ilustran la extensión de cada distribución. Las cruces indican los centroides de cada grupo. El solapamiento parcial entre las elipses de los grupos 1 y 2 refleja que la frontera de decisión entre ambos no es completamente nítida en el espacio proyectado, lo que es consistente con la naturaleza probabilística del modelo GMM y con la presencia de regiones candidatas con características estructurales intermedias.

La separación visual entre agrupamientos confirma que UMAP logra organizar el espacio de búsqueda en estructuras discriminables incluso ante un repertorio de 850 genes de referencia, superior al observado durante la optimización, sustentando la cobertura del 91.06% alcanzada por UMAP+GMM en este genoma.

En términos de RMSE, PCA+GMM presenta el valor más bajo (2.6066), seguida de GMM (3.0734) y UMAP+GMM (3.4594). Este comportamiento es consistente con los resultados de la evaluación comparativa, donde las diferencias en RMSE entre variantes no alcanzaron significancia estadística. El incremento en el RMSE de UMAP+GMM se mantiene dentro de un rango controlado, lo que permite afirmar que la mejora en cobertura no compromete de forma relevante la precisión espacial del método.

En conjunto, los resultados sobre *Plecotus auritus* constituyen una primera evidencia sólida de la capacidad de generalización del método RIVAM: la configuración óptima seleccionada estadísticamente en la fase de evaluación mantiene y acentúa su ventaja en un escenario de mayor escala, sin haber sido reoptimizada para este genoma.

8.2.2 Ambystoma Mexicanum

La segunda validación externa se llevó a cabo sobre el genoma de *Ambystoma mexicanum* (Amme), especie nativa de México conocida comúnmente como ajolote. A diferencia de los quirópteros, este organismo posee un genoma considerablemente más grande y estructuralmente complejo, con características arquitectónicas del gen V significativamente distintas a las observadas durante la fase de optimización. Esta prueba permite evaluar si el flujo metodológico propuesto mantiene su coherencia y capacidad discriminativa cuando se enfrenta a un dominio filogenético distinto al utilizado durante el diseño del método.

Una diferencia estructural relevante en este escenario es la longitud de la región intrónica del gen V en *Ambystoma mexicanum*, que presenta una extensión aproximada de 200 nucleótidos, lo que da lugar a una estructura completa del gen V superior a los 600 nucleótidos. Esta longitud excede la estructura típica observada en genes V de mamíferos, que generalmente no supera los 500 nucleótidos. Por ello, se ajustó el

umbral de distancia intrónica para adaptarse a estas características estructurales particulares, sin realizar modificaciones en el flujo metodológico general ni en los parámetros de agrupamiento.

La configuración UMAP+GMM identificó 48 de los 87 genes V reportados en la referencia experta, alcanzando una tasa de identificación del 55.17%. Si bien este resultado es inferior al observado en los genomas de quirópteros, confirma que el método mantiene capacidad operativa en un genoma filogenéticamente distinto, logrando recuperar más de la mitad de los genes de referencia sin haber sido optimizado para este dominio.

La disminución en el porcentaje de identificación respecto a los quirópteros es esperable. En primer lugar, las diferencias estructurales en la arquitectura del gen V entre anfibios y mamíferos implican que los patrones aprendidos durante la optimización sobre *Tadarida brasiliensis* no se transfieren de manera completa a este dominio, lo que limita la capacidad discriminativa del agrupamiento en regiones con características estructurales no observadas previamente. En segundo lugar, el mayor tamaño y complejidad del genoma del ajolote obtiene un espacio de búsqueda más heterogéneo, lo que incrementa la dificultad para detectar estructuras subyacentes mediante agrupamiento no supervisado.

A pesar de estas limitaciones, esta validación demuestra que el método mantiene su aplicabilidad en condiciones estructurales marcadamente distintas a las empleadas durante la optimización, logrando reducir el espacio de búsqueda y recuperar un subconjunto relevante de genes V en un organismo con características arquitectónicas diferentes. Este comportamiento constituye evidencia de la flexibilidad del marco metodológico propuesto más allá del dominio taxonómico de los quirópteros.

8.2.3 *Molossus Nigricans*

La tercera validación externa se llevó a cabo sobre el genoma de *Molossus nigricans* (Moni), un quiróptero sin anotación experta disponible cuyo número total de genes V es desconocido. A diferencia de las validaciones anteriores, este escenario no permite calcular métricas de identificación directas al carecer de una referencia validada por expertos. Por ello, el análisis se orienta hacia dos aspectos principales: la coherencia estructural de las regiones candidatas identificadas y la capacidad del método para operar en condiciones exploratorias sin referencia conocida.

Este escenario es particularmente relevante desde el punto de vista metodológico, debido a que representa una condición frecuente en estudios reales donde la caracterización del locus de inmunoglobulinas aún no ha sido realizada. En este contexto, el objetivo no es medir precisión frente a una referencia conocida, sino determinar si el método es capaz de identificar regiones candidatas estructuralmente

consistentes y biológicamente plausibles a partir de un volumen inicial de alineamientos considerablemente alto.

La Figura 8.9 muestra la proyección UMAP del espacio de búsqueda sobre el genoma de *Molossus nigricans*, con los grupos identificados por el modelo GMM superpuestos. A diferencia de lo observado en *Plecotus auritus*, donde los grupos presentaban una separación visual clara, en este genoma los tres agrupamientos muestran un solapamiento considerable en el espacio proyectado. El grupo 1 (azul oscuro) se concentra en la región superior central con una distribución relativamente compacta. El grupo 2 (azul claro) ocupa la zona central e inferior con mayor dispersión. El grupo 3 (naranja) es el de mayor extensión y domina gran parte del espacio proyectado, con ramificaciones que se extienden en múltiples direcciones.

El solapamiento entre grupos refleja la mayor heterogeneidad estructural del espacio de búsqueda en este genoma, lo que es consistente con la ausencia de anotación experta y con la complejidad característica de genomas no caracterizados previamente. A pesar de esta heterogeneidad, el modelo GMM logra establecer fronteras de decisión probabilísticas que permiten seleccionar un subconjunto de regiones candidatas estructuralmente coherentes.

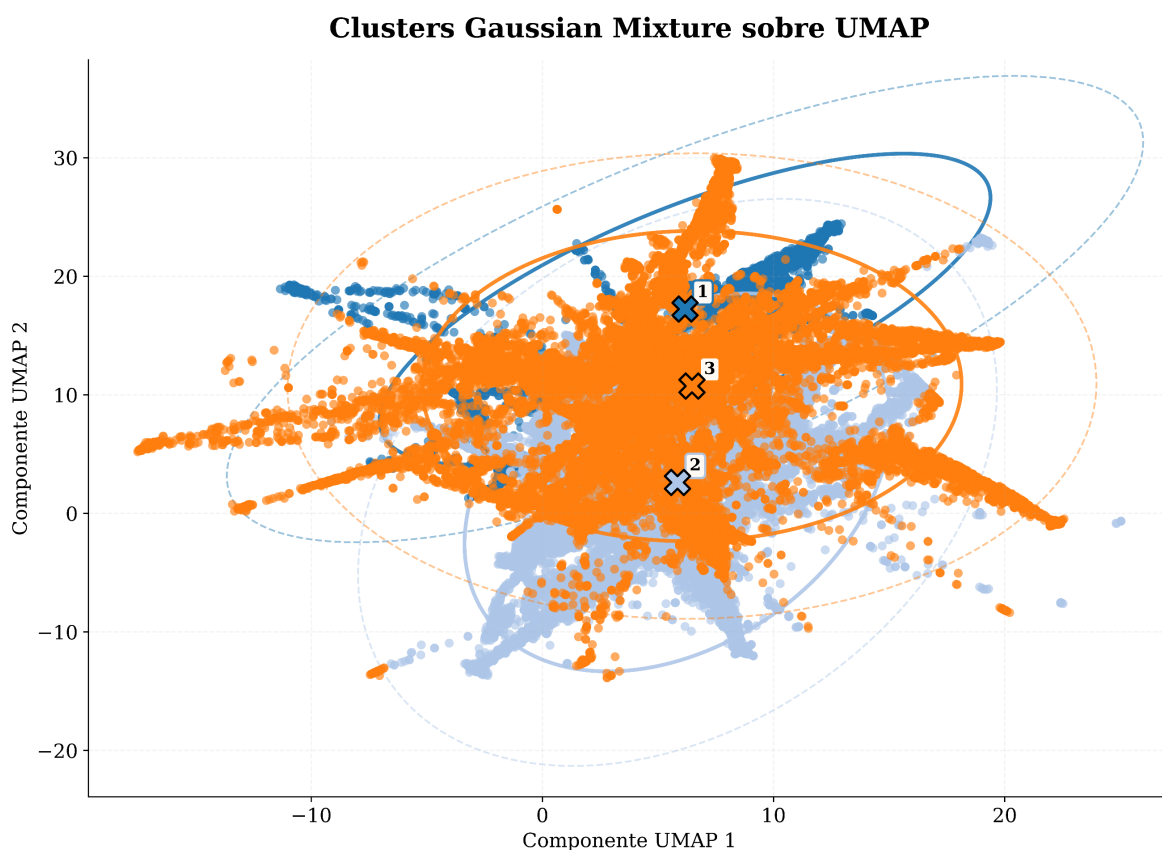


Figura 8.9 Proyección UMAP con grupos GMM sobre el genoma de *Molossus nigricans*

La Tabla 8.10 presenta los resultados por grupo, reportando el número de regiones candidatas agrupadas y el número de genes identificados en cada caso.

Tabla 8.10 Resultados por cluster sobre el genoma de *Molossus nigricans*

Grupo	Regiones candidatas	Genes identificados
1	22,942	1,206
2	77,495	1,414
3	151,751	1,470

Los tres grupos presentan diferencias tanto en el volumen de regiones agrupadas como en el número de genes identificados. El grupo 3 concentra el mayor número de regiones candidatas (151,751) y es el que mayor número de genes identifica (1,470), seguido del grupo 2 con 77,495 regiones y 1,414 genes, y del grupo 1 con 22,942 regiones y 1,206 genes.

Con base en el principio de optimización del método, el grupo seleccionado como de interés biológico corresponde a aquel que maximiza el número de genes identificados dentro del espacio de búsqueda reducido. Bajo este criterio, el grupo 3 representa el grupo óptimo en este escenario, debido a que concentra la mayor cantidad de regiones con características estructurales consistentes con genes V reales.

El análisis de las regiones identificadas mostró que el método mantiene una delimitación estable de las tres secciones estructurales del gen V tanto en orientación *forward* como *reverse*. Este resultado es particularmente relevante en genomas no anotados, donde la variabilidad en la orientación de los alineamientos puede dificultar la detección precisa de regiones funcionales. El análisis visual detallado de estas regiones se presenta en la Sección 8.4.

8.3 Comparación con IgDetective

La comparación con IgDetective, cuyas características y criterios de selección como método de referencia se describen en la Sección 7.4.1, se realizó en dos etapas complementarias. La primera etapa consistió en comparar la cantidad de genes funcionales identificados en cinco genomas con anotación experta disponible. La segunda se centró en evaluar, con IGV, la precisión con la que cada método identifica las tres secciones que conforman el gen V, utilizando la anotación de referencia como punto de comparación.

8.3.1 Resultados comparativos por genoma

Con el objetivo de contrastar el desempeño del método con una herramienta utilizada en la literatura, se realizó una comparación formal con IgDetective sobre un conjunto de cinco genomas con anotación experta disponible. A diferencia del análisis exploratorio presentado en la Sección 8.2.3, esta comparación se centra exclusivamente

en genes funcionales, lo que permite evaluar de manera directa la capacidad de ambos métodos para recuperar elementos biológicamente validados dentro del genoma.

Los seis genomas seleccionados cubren distintos niveles de complejidad estructural y tamaños de repertorio, incluyendo tanto genomas del conjunto de evaluación como el genoma de validación *Plecotus auritus*. Es importante señalar que la comparación se realizó a nivel de cromosoma, es decir, sobre los contigs o scaffolds donde se localizan los genes V anotados por expertos, sin procesar el genoma completo. Este enfoque permite una comparación directa y controlada entre ambos métodos sobre las mismas regiones de interés.

Cabe señalar que *Ambystoma mexicanum* no se incluye en esta comparación debido a que IgDetective no logró completarse exitosamente sobre ese genoma, lo que evidencia una limitación operativa de las herramientas diseñadas bajo supuestos biológicos específicos de mamíferos. En el caso de *Molossus nigricans*, la ausencia de anotación experta impide calcular métricas de identificación de genes funcionales, por lo que su comparación con IgDetective se limitó al análisis de eficiencia computacional presentado en la Sección 8.2.3.

La Tabla 8.11 presenta los resultados por ambos métodos sobre los cinco genomas comparables, incluyendo el número de genes funcionales de referencia, el número de genes identificados y la tasa de identificación. El análisis de precisión estructural de las predicciones se realizó con IGV y se presenta en la Sección 8.3.2.

Tabla 8.11 Comparación del desempeño de RIVAM e IgDetective en la identificación de genes V funcionales

Genom a	Ref. total	Genes funcionale s	RIVA M	IgDetectiv e	RIVA M (%)	IgDetectiv e (%)	Diferenci a (%)
Anpa	95	56	55	47	98.21	83.93	+14.29
Asst	43	38	37	38	97.37	100.00	-2.63
Epfu	168	77	77	72	100.00	93.51	+6.49
Hila	121	86	63	68	73.26	79.07	-5.81
Plau	850	653	637	601	97.55	92.04	+5.51

El método supera a IgDetective en tres de los cinco genomas evaluados, con mejoras que oscilan entre +5.51% y +14.29%. El caso más destacado es *Antrozous pallidus* (Anpa), donde se observa la mejora más sustancial con una diferencia de 14.29 puntos porcentuales. En *Eptesicus fuscus* (Epfu) el método identifica el 100% de los genes funcionales de referencia, superando en 6.49 puntos a IgDetective. En *Plecotus auritus*

(Plau), el genoma de mayor repertorio evaluado, el método recupera el 97.55% de los genes funcionales anotados frente al 92.04% de IgDetective.

Sin embargo, el método no supera a IgDetective en todos los casos. En *Hipposideros larvatus* (Hila) IgDetective presenta una tasa de identificación mayor, con una diferencia de 5.81 puntos porcentuales a su favor, lo que sugiere que en genomas con ciertas características estructurales particulares los métodos basados en señales conservadas pueden ofrecer ventajas específicas. En *Aselliscus stoliczkanus* (Asst) IgDetective recupera la totalidad de los genes funcionales anotados, presentando una ventaja de 2.63 puntos porcentuales respecto al método propuesto.

En conjunto, el método supera a IgDetective en los genomas con repertorios más amplios y mayor complejidad estructural, mientras que IgDetective mantiene una ventaja en genomas de menor tamaño donde las señales estructurales conservadas están mejor preservadas. Este comportamiento es consistente con la naturaleza de ambos enfoques: mientras IgDetective se apoya en señales de referencia específicas de un linaje, el método propuesto opera sobre atributos estructurales generales que resultan más informativos en escenarios de mayor heterogeneidad.

8.3.2 Análisis de precisión estructural en IGV

Más allá de las diferencias cuantitativas en la tasa de identificación, resulta relevante analizar la precisión estructural en la delimitación del gen V. Este análisis permite evaluar tanto la cantidad de genes identificados como la calidad de su identificación, considerando la precisión con la que cada método localiza las tres secciones de la arquitectura del gen V: el exón señal (SP_exon), el exón variable (IGHV_exon) y la secuencia de señal de recombinación (RSS3). El análisis de las secciones se realizó en IGV, empleando tres *tracks* por genoma: la anotación de referencia experta, las predicciones del método propuesto y las predicciones de IgDetective.

Un hallazgo consistente a lo largo de todos los genomas analizados es que IgDetective no identifica el exón señal (SP_exon) en ninguno de los genes evaluados. Esta limitación no es un caso aislado sino una característica sistemática de la herramienta, que reporta únicamente el exón y la señal de recombinación, pero omite la región SP que precede al exón variable en la arquitectura completa del gen V. El método propuesto, en cambio, identifica las tres secciones de forma consistente en todos los genomas evaluados, lo que representa una ventaja estructural relevante independientemente de las diferencias en la tasa de identificación cuantitativa.

La Figura 8.10 muestra la vista general del cromosoma CM050515.1 de *Antrozous pallidus* (ANPA), abarcando una región de aproximadamente 138 kb. En el *track* de referencia se observan múltiples genes V anotados en orientación inversa, identificados por sus tres secciones estructurales. El *track* del método reproduce de forma

consistente las mismas posiciones con las tres secciones bien delimitadas. La pista de IgDetective muestra predicciones que cubren las mismas regiones, pero reporta únicamente RSS_nonamer y V_exon, sin incluir el SP_exon en ningún caso.

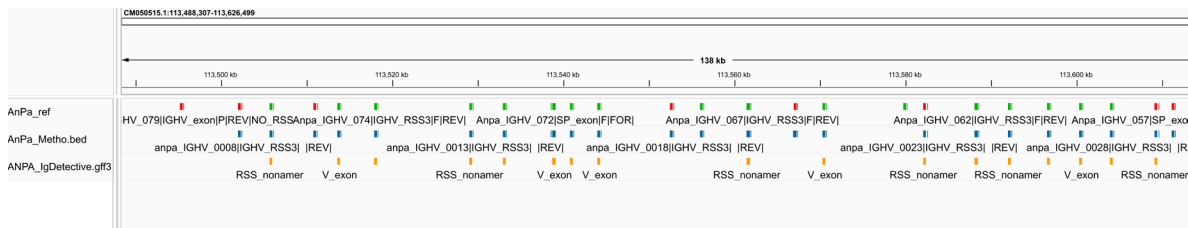


Figura 8.10 Vista general del cromosoma CM050515.1 de *Antrozous pallidus*

La Figura 8.11 muestra un acercamiento sobre el gen 73 de *Antrozous pallidus* en orientación inversa, en una región de 493 bp. En esta vista se muestra con claridad que el método identifica correctamente el RSS3, el IGHV_exon y el SP_exon en posiciones concordantes con la referencia. IgDetective identifica el RSS_nonamer y el V_exon pero no incluye el SP_exon, y presenta una ligera diferencia en la delimitación del extremo del exón variable respecto a la referencia.

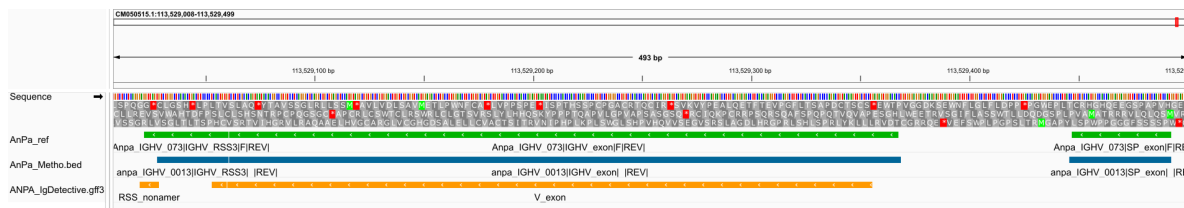


Figura 8.11 Gen 73 de *Antrozous pallidus* en orientación inversa

La Figura 8.12 muestra el gen 72 de *Antrozous pallidus* en orientación forward, en una región de 540 bp. El método identifica las tres secciones con alta concordancia respecto a la referencia. IgDetective identifica el V_exon y el RSS_heptamer pero nuevamente omite el SP_exon, y muestra una delimitación del extremo 5' del exón variable ligeramente desplazada respecto a la anotación de referencia.

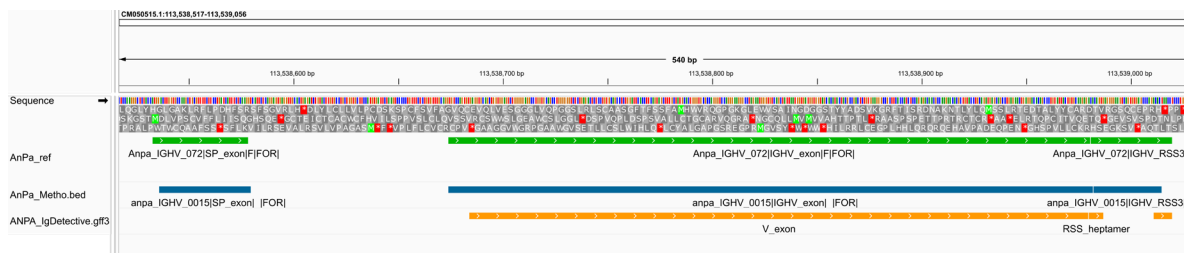


Figura 8.12 Gen 72 de *Antrozous pallidus* en orientación forward

La Figura 8.13 muestra la vista general del cromosoma CM063093.1 de *Hipposideros larvatus*, abarcando una región de aproximadamente 969 kb. En esta vista se observa que ambos métodos cubren de forma densa las regiones candidatas a lo largo del cromosoma con una distribución similar de predicciones. Sin embargo, al igual que

en *Antrozous pallidus*, IgDetective reporta únicamente V_exon sin incluir el SP_exon en ninguna de las predicciones visibles.

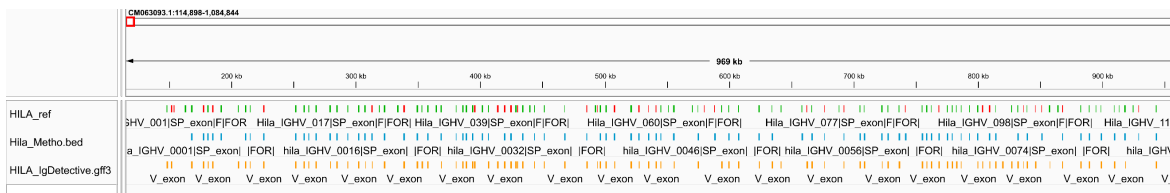


Figura 8.13 Vista general del cromosoma CM063093.1 de Hipposideros larvatus

La Figura 8.14 muestra al gen 004 de *Hipposideros larvatus* en orientación forward, en una región de 568 bp. El método identifica correctamente las tres secciones del gen V en posiciones concordantes con la referencia. IgDetective identifica el V_exon y el RSS_heptamer con una delimitación comparable en el exón variable, aunque omite nuevamente el SP_exon.

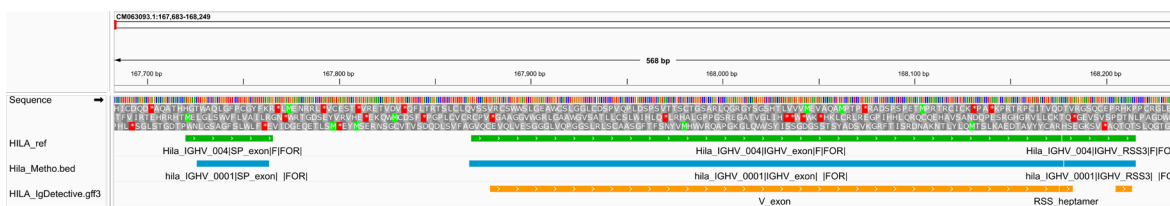


Figura 8.14 Gen 004 de Hipposideros larvatus en orientación forward

La Figura 8.15 muestra al gen 104 de *Hipposideros larvatus* en orientación forward, en una región de 595 bp. En este caso el método identifica el SP_exon y el IGHV_exon con buena concordancia respecto a la referencia, pero no logra delimitar el RSS3 en este gen particular. IgDetective en cambio sí identifica el V_exon y el RSS_heptamer, aunque sin el SP_exon. Este caso muestra que la ventaja estructural del método no es absoluta en todos los genes individuales, y que ambas herramientas presentan casos específicos donde una supera a la otra en términos de completitud de la predicción.

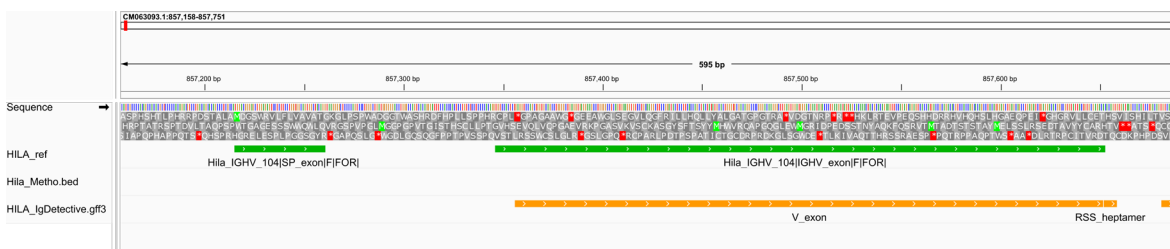


Figura 8.15 Gen 104 de Hipposideros larvatus en orientación forward

La Figura 8.16 muestra la vista expandida del cromosoma OY734022.1 de *Plecotus auritus*, abarcando una región de aproximadamente 3,172 kb. En esta vista panorámica se aprecia que el método genera predicciones con las tres secciones estructurales bien diferenciadas, mientras que IgDetective reporta únicamente spacer y V_gene sin incluir el SP_exon.

método con las de IgDetective. Ambos métodos obtienen una cobertura densa y distribuida a lo largo de todo el cromosoma con un número comparable de regiones candidatas identificadas, lo que es consistente con los datos cuantitativos reportados en la Sección 8.2.3.

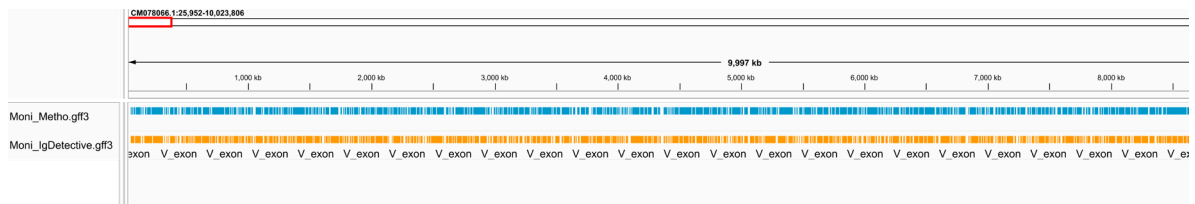


Figura 8.19 Vista general del cromosoma CM078066.1 de *Molossus nigricans*

La Figura 8.20 muestra al gen 1430 de *Molossus nigricans* en orientación forward, en una región de 505 bp. En este caso el método identifica correctamente las tres secciones estructurales del gen V, mientras que IgDetective no genera ninguna predicción para esta región. Este es un ejemplo representativo de genes identificados exclusivamente por el método propuesto en este genoma no anotado.

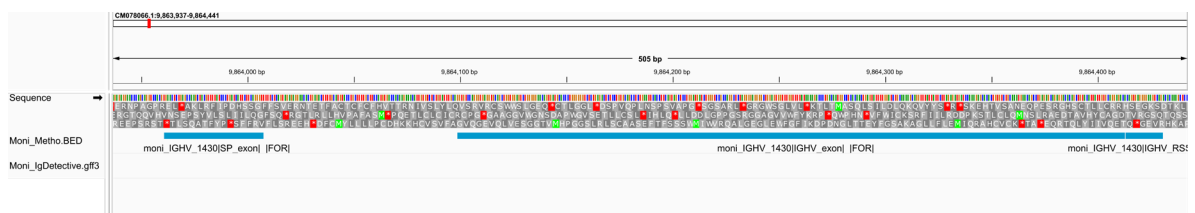


Figura 8.20 Gen 1430 de *Molossus nigricans*, identificado únicamente por el método propuesto

La Figura 8.21 muestra al gen 1429 de *Molossus nigricans* en orientación forward, en una región de 525 bp. En este caso ambos métodos identifican el gen, pero con diferencias estructurales notables. El método delimita el SP_exon, el IGHV_exon y el RSS3 de forma precisa y consistente con la arquitectura esperada del gen V. IgDetective identifica el V_exon y el RSS_heptamer pero omite el SP_exon y presenta una delimitación del exón variable con límites ligeramente distintos.

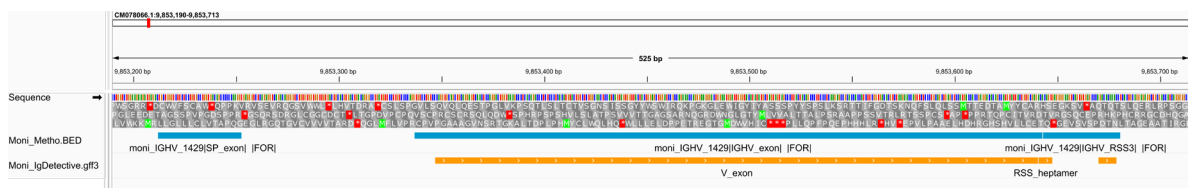


Figura 8.21 Gen 1429 de *Molossus nigricans* en orientación forward

La Figura 8.22 muestra al gen 0067 de *Molossus nigricans* en orientación forward, en una región de 502 bp. El comportamiento es similar al observado en el gen 1429: el método identifica las tres secciones estructurales con coherencia, mientras que IgDetective reporta únicamente el V_exon y el RSS_heptamer sin incluir el SP_exon. La delimitación del exón variable por parte de IgDetective presenta en este caso un

desplazamiento respecto al inicio del exón identificado por el método, lo que podría afectar la precisión de la anotación funcional en estudios posteriores.

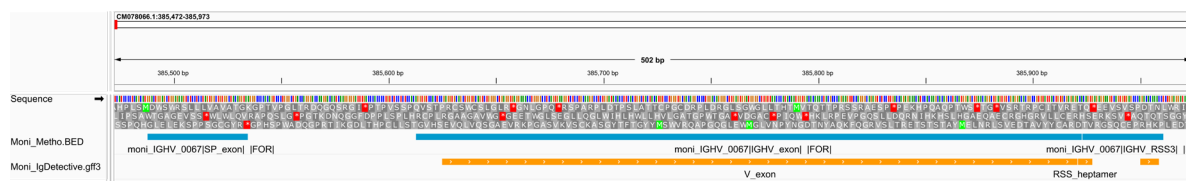


Figura 8.22 Gen 0067 de *Molossus nigricans* en orientación forward

En conjunto, el análisis visual en IGV confirma que la ventaja del método propuesto sobre IgDetective no se limita a las diferencias cuantitativas en la tasa de identificación. La identificación sistemática del SP_exon en todos los genomas evaluados, ausente en todas las predicciones de IgDetective, representa una contribución estructural relevante que mejora la completitud de la anotación del gen V. Adicionalmente, en varios genes donde IgDetective no obtiene ninguna predicción, el método logra identificar correctamente las tres secciones estructurales, tanto en genes funcionales como en elementos no funcionales.

Esta capacidad se deriva del enfoque de agrupamiento sobre atributos estructurales generales, en lugar de señales de referencia específicas de un linaje, lo que le confiere al método una mayor independencia respecto a los supuestos biológicos incorporados en herramientas especializadas. En este sentido, el método no debe entenderse como un sustituto de IgDetective sino como una herramienta complementaria que aporta información estructural adicional, en escenarios donde la precisión en la delimitación completa del gen V es crítica para estudios de diversidad inmunogenética.

8.3.3 Discusión de diferencias

Las diferencias observadas entre el método propuesto e IgDetective en el análisis estructural no son aleatorias, sino que responden a decisiones de diseño fundamentales de cada herramienta, con implicaciones directas sobre el tipo de información que cada una es capaz de recuperar. La ausencia sistemática del SP_exon en todas las predicciones de IgDetective tiene una explicación metodológica clara. IgDetective basa su estrategia de detección en la identificación de secuencias de señal de recombinación (RSS), elementos conservados ubicados en el extremo 3' del gen V. A partir de estas señales, la herramienta delimita el exón variable hacia el extremo 5' con un proceso iterativo, pero sin un modelo explícito para la detección del exón señal.

El SP_exon, al estar separado del exón variable por una región intrónica y presentar mayor variabilidad de secuencia entre especies, queda sistemáticamente fuera del alcance del modelo de detección de IgDetective. Esta limitación no es un defecto puntual sino una consecuencia directa de la estrategia de anclaje en señales conservadas: la

herramienta detecta con precisión lo que está cerca de la RSS, pero no tiene mecanismos para extender la predicción hacia regiones más distales y variables como el SP_exon.

El método propuesto supera esta limitación de forma natural y sin requerir modificaciones específicas para cada especie. Al operar sobre atributos estructurales generales derivados de los alineamientos de secuencias, el agrupamiento no supervisado captura el perfil completo del gen V tal como aparece en los datos, incluyendo el SP_exon, el IGHV_exon y el RSS3 como unidades integradas en una misma representación. Esta característica no depende de la conservación de señales específicas en una posición determinada del gen, lo que explica por qué el método logra delimitar las tres secciones de forma consistente en todos los genomas evaluados, independientemente del grado de conservación de las señales RSS en cada especie. La reducción del espacio de búsqueda basada en agrupamiento permite que el modelo identifique regiones con el perfil estructural completo del gen V, en lugar de buscar elementos individuales de forma secuencial.

Esta diferencia en el enfoque tiene consecuencias adicionales más allá de la detección del SP_exon. En genomas con menor grado de conservación de las señales RSS, como ocurre en especies filogenéticamente distantes o en linajes con mayor divergencia, la estrategia basada en señales conservadas de IgDetective puede perder sensibilidad de forma progresiva. El método desarrollado, al no depender de la conservación de señales específicas, mantiene su capacidad de identificación en estos escenarios con mayor robustez. Esto se evidencia en los resultados de *Ambystoma mexicanum*, donde IgDetective no logró completarse exitosamente mientras que el método propuesto identificó el 55.17% de los genes de referencia, demostrando una mayor tolerancia a la divergencia estructural entre linajes taxonómicos distantes.

La ventaja en eficiencia computacional es otro aspecto diferencial relevante. Como se reportó en la Sección 8.2.3, el método propuesto completa el análisis de *Molossus nigricans* en aproximadamente 40 minutos, frente a las 24 horas requeridas por IgDetective sobre el mismo genoma. Esta diferencia no es marginal sino de un orden de magnitud, y tiene implicaciones prácticas importantes en estudios de inmunogenómica comparativa a gran escala donde se analizan decenas o cientos de genomas simultáneamente.

La etapa de reducción del espacio de búsqueda no solo mejora la precisión estructural de las predicciones, sino que también reduce de forma sustancial el costo computacional del análisis, lo que lo hace más accesible en entornos con recursos limitados. Desde el punto de vista de la anotación del genoma, la detección del SP_exon tiene implicaciones relevantes para la calidad del resultado final. Una delimitación incompleta del gen V que omita esta región puede afectar la identificación precisa del marco de lectura y la correcta caracterización de la estructura completa del gen, con

consecuencias directas sobre estudios posteriores de expresión de genes y análisis de repertorio. En este sentido, la capacidad del método para recuperar sistemáticamente las tres secciones del gen V representa una mejora en la calidad de la anotación, independientemente del enfoque biológico con el que se utilicen los resultados.

No obstante, el análisis de los resultados también revela escenarios donde IgDetective presenta ventajas específicas. En genomas de menor tamaño con señales RSS bien conservadas, como *Aselliscus stoliczkanus*, IgDetective logra recuperar la totalidad de los genes funcionales anotados, superando al método propuesto en 2.63 puntos porcentuales. En *Hipposideros larvatus*, IgDetective muestra una tasa de identificación mayor en 5.81 puntos porcentuales, y en el análisis individual del gen 104 logra delimitar el V_exon y el RSS_heptamer en un gen donde el método propuesto no alcanza a identificar el RSS3. Estos casos evidencian que la estrategia basada en señales conservadas puede ofrecer ventajas de precisión en escenarios donde dichas señales están bien preservadas y el genoma presenta características estructurales homogéneas.

En este sentido, el método propuesto e IgDetective no deben entenderse como herramientas mutuamente excluyentes sino como enfoques complementarios con fortalezas diferenciadas. IgDetective ofrece alta precisión en la detección del exón variable y la señal de recombinación en organismos bien caracterizados, mientras que el método propuesto aporta mayor completitud estructural al incluir el SP_exon, mayor robustez en escenarios de divergencia taxonómica y mayor eficiencia computacional. La integración de ambos enfoques en un flujo de anotación complementario representa una línea de trabajo futura con potencial para mejorar la calidad y completitud de los catálogos de genes V en vertebrados.

8.4 Flujo de procesamiento del método propuesto

La Figura 8.23 presenta el flujo de procesamiento del método propuesto, organizado en cuatro etapas secuenciales que integran la comprensión de los datos de entrada, la preparación y transformación, el agrupamiento y selección del grupo óptimo, y la evaluación y obtención de la salida final. Este flujo representa la materialización computacional de los principios de diseño establecidos en el Capítulo 5, donde cada etapa corresponde a una decisión metodológica fundamentada en los requisitos del problema de identificación de genes V en genomas de vertebrados.

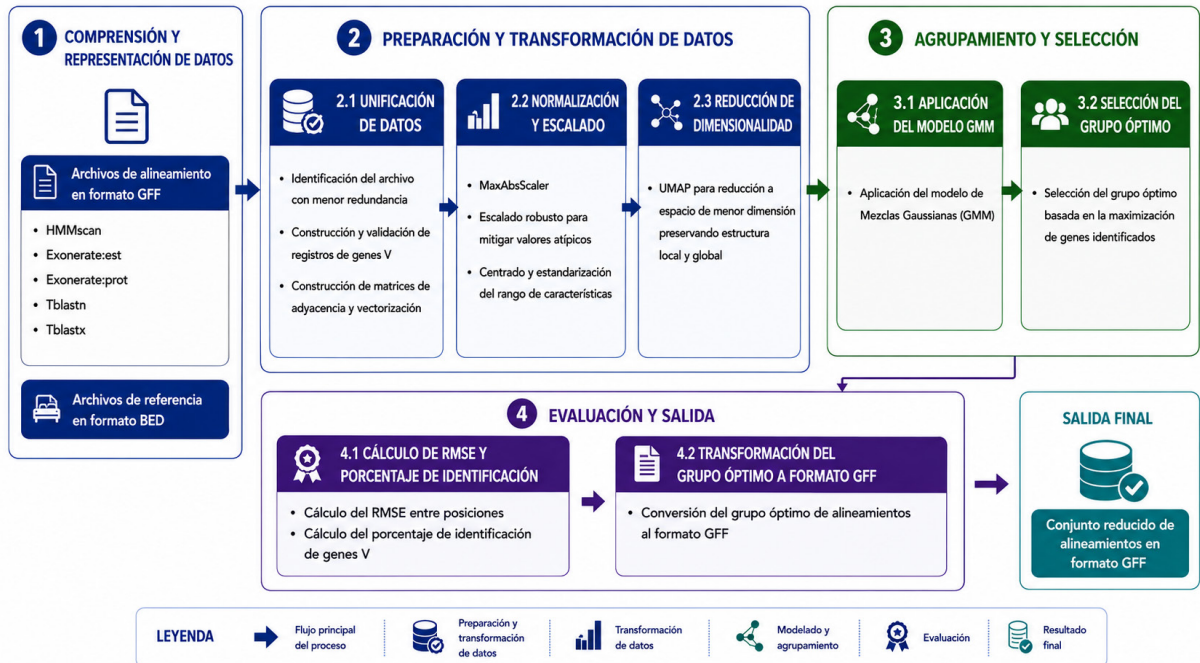


Figura 8.23 Flujo de procesamiento del método propuesto

La primera etapa corresponde a la comprensión y representación de datos, donde se reciben como entrada los archivos de alineamiento en formato GFF generados por HMMScan, Exonerate:est, Exonerate:prot, Tblastn y Tblastx, junto con los archivos de referencia en formato BED. Esta etapa establece el punto de partida del método a partir de los resultados de múltiples herramientas de alineamiento, lo que permite una representación más completa del espacio de búsqueda.

La segunda etapa integra tres subprocesos de preparación y transformación de datos. En primer lugar, la unificación de datos consolida los archivos de alineamiento identificando el de menor redundancia, construye y valida los registros de genes V y genera las matrices de adyacencia y su vectorización. En segundo lugar, la normalización y escalado aplica MaxAbsScaler junto con escalado robusto para mitigar el efecto de valores atípicos y estandarizar el rango de características. En tercer lugar, la reducción de dimensionalidad mediante UMAP proyecta los datos a un espacio de menor dimensión preservando la estructura local y global de las relaciones entre regiones candidatas.

La tercera etapa corresponde al agrupamiento y selección, donde el modelo de Mezclas Gaussianas (GMM) se aplica sobre el espacio reducido para identificar agrupamientos estructuralmente coherentes. A partir de los grupos generados, se selecciona el grupo óptimo mediante el criterio de maximización del número de genes identificados, tal como se describió en las secciones anteriores.

La cuarta etapa evalúa los resultados del grupo óptimo calculando el RMSE entre posiciones y el porcentaje de identificación de genes V, y transforma el grupo seleccionado al formato GFF para su uso en análisis posteriores. La salida final es un conjunto reducido de alineamientos en formato GFF que representa el espacio de búsqueda comprimido con las regiones candidatas más relevantes.

La implementación del método está disponible en el repositorio público: <https://github.com/antonio160672/RIVAM.git>

8.5 Discusión

Los resultados presentados en las secciones anteriores permiten construir una interpretación integral del desempeño del método propuesto, más allá de los valores numéricos individuales. Esta sección examina el significado de esos resultados en el contexto del problema de identificación de genes V, identifica las limitaciones que deben considerarse al extrapolar las conclusiones a nuevos escenarios, y verifica en qué medida los principios de diseño establecidos en el Capítulo 5 fueron efectivamente satisfechos por el método en su configuración final. Este análisis cierra el ciclo de diseño de la metodología DSR adoptada en la investigación, articulando la evidencia experimental con las decisiones metodológicas que guiaron el desarrollo del método desde sus etapas iniciales.

8.5.1 Interpretación de los resultados en el contexto del problema

El problema central que motivó el desarrollo del método propuesto es la ineficiencia computacional que surge al aplicar herramientas de alineamiento de secuencias sobre la totalidad del espacio de genoma sin una etapa previa de reducción. En genomas de vertebrados, el número de regiones candidatas obtenidas por herramientas como HMMscan, Exonerate o BLAST puede alcanzar decenas de miles de registros, de los cuales solo una fracción corresponde a genes V reales. Sin un mecanismo de filtrado previo, el experto o la herramienta de anotación posterior debe procesar un volumen de información desproporcionado respecto al resultado esperado, lo que incrementa tanto el costo computacional como el riesgo de error en la anotación.

Los resultados presentados a lo largo del Capítulo 8 demuestran que el método propuesto aborda este problema de manera efectiva y generalizable. En la evaluación comparativa sobre 12 genomas, la configuración UMAP+GMM alcanzó una cobertura media del 74.98% con una desviación estándar de 7.09%, lo que indica no solo una alta tasa de recuperación de genes V reales sino también un comportamiento consistente entre genomas con distintos niveles de complejidad. Esta consistencia es relevante desde el punto de vista metodológico, porque evidencia que el método no está sobre

ajustado a las características de un subconjunto particular de genomas, sino que captura patrones estructurales generalizables.

La selección de UMAP+GMM como variante óptima con el análisis estadístico de *Friedman* y *Nemenyi* no es un resultado trivial. El análisis reveló que la reducción no lineal de dimensionalidad aporta una ventaja real sobre la reducción lineal de PCA, la cual mostró un desempeño significativamente inferior en términos de cobertura. Este hallazgo es consistente con la naturaleza de los datos de alineamiento de genes V, cuya estructura no puede ser capturada adecuadamente con el uso de transformaciones lineales que preservan únicamente la varianza global. UMAP, al preservar tanto la estructura local como la global del espacio de datos, obtiene una representación más favorable para la separabilidad de los grupos, lo que se traduce directamente en una mayor capacidad de recuperación de genes V reales.

La validación externa sobre tres genomas con características deliberadamente distintas amplía la perspectiva de los resultados. En *Plecotus auritus*, el método alcanzó una cobertura del 91.06% sobre un repertorio de 850 genes, muy superior al utilizado durante la optimización, lo que demuestra que el método escala efectivamente sin requerir reoptimización. En *Ambystoma mexicanum*, la cobertura del 55.17% en un dominio filogenéticamente distinto evidencia que el método mantiene capacidad operativa más allá del grupo taxonómico para el que fue desarrollado, con las limitaciones esperables derivadas de las diferencias estructurales entre anfibios y mamíferos. En *Molossus nigricans*, la identificación de 1,434 regiones candidatas en 40 minutos frente a las 24 horas de IgDetective demuestra que la reducción del espacio de búsqueda no solo mejora la precisión de la anotación, sino que también transforma radicalmente la eficiencia del proceso.

En conjunto, estos resultados posicionan al método propuesto como una contribución metodológica sólida al problema de la identificación de genes V en genomas de vertebrados. No se trata de una herramienta que reemplaza a las existentes, sino de una etapa previa de reducción que hace más eficiente y preciso el trabajo de las herramientas de anotación posteriores, independientemente del genoma o la especie sobre la que se aplique.

8.5.2 Limitaciones identificadas

El análisis crítico de los resultados permite identificar un conjunto de limitaciones que deben considerarse al interpretar el alcance del método propuesto y al planificar su aplicación en nuevos contextos.

La primera limitación está relacionada con el tamaño del conjunto de evaluación. Los 12 genomas utilizados en la fase de evaluación comparativa corresponden exclusivamente a quirópteros, un grupo taxonómico con características

inmunogenéticas relativamente homogéneas entre sí. Si bien la validación externa demostró cierta capacidad de generalización fuera de este grupo, el conjunto de evaluación no es suficientemente diverso para garantizar que el comportamiento observado se mantendrá en todos los órdenes de vertebrados. Grupos como peces, reptiles o aves presentan características estructurales del gen V potencialmente distintas que podrían afectar el desempeño del agrupamiento de formas no anticipadas por el diseño actual.

La segunda limitación concierne al proceso de optimización de hiperparámetros. Los parámetros de UMAP y GMM fueron optimizados con PSO sobre el genoma de *Tadarida brasiliensis*, lo que introduce un sesgo hacia las características estructurales de ese genoma específico. Aunque los resultados de validación sugieren que este sesgo no compromete gravemente la capacidad de generalización, no puede descartarse que en genomas con características muy alejadas del genoma de optimización el desempeño sea inferior al observado en la evaluación comparativa.

La tercera limitación está asociada a la ausencia de un mecanismo de adaptación automática del umbral de distancia intrónica. Como se demostró en la validación con *Ambystoma mexicanum*, la longitud de la región intrónica varía entre grupos taxonómicos y requirió un ajuste manual para este genoma. En su configuración actual, el método no incorpora un mecanismo automático para detectar y adaptar este umbral a las características específicas de cada genoma, lo que limita su aplicabilidad directa en especies con arquitecturas del gen V muy divergentes sin intervención del usuario.

La cuarta limitación corresponde al criterio de selección del grupo óptimo. La maximización del número de genes identificados es un criterio efectivo en genomas con anotación experta disponible, donde puede verificarse empíricamente. Sin embargo, en genomas no anotados como *Molossus nigricans*, este criterio no puede validarse de forma directa y su aplicación depende de que el grupo con mayor número de genes identificados en los genomas de evaluación corresponde también al grupo más relevante en el nuevo genoma. Esta situación, aunque razonable, introduce una incertidumbre que no puede cuantificarse sin una referencia experta.

Finalmente, la escalabilidad del método en genomas de gran tamaño representa una limitación potencial. Si bien el tiempo de procesamiento de 40 minutos sobre *Molossus nigricans* es considerablemente menor al de IgDetective, genomas con tamaños varios órdenes de magnitud superiores, como el genoma del ajolote o algunos genomas de plantas y peces, podrían requerir estrategias adicionales de particionamiento o procesamiento paralelo que el método en su configuración actual no implementa explícitamente.

8.5.3 Verificación de los principios de diseño del método

El diseño del método propuesto estuvo guiado desde sus etapas iniciales por cinco principios establecidos en el Capítulo 5, derivados de los requerimientos del problema y de las restricciones impuestas por el dominio de aplicación. Los resultados experimentales presentados a lo largo de este capítulo permiten verificar en qué medida cada uno de esos principios fue efectivamente satisfecho por el método en su configuración final, cerrando así el ciclo de diseño de la metodología DSR adoptada en esta investigación.

El primer principio, la adopción del marco CRISP-DM adaptado con la Metodología de Flujo de Conocimiento del Dominio, se refleja en la estructura completa del flujo de procesamiento presentado en la Sección 8.4. Cada etapa del flujo corresponde a una fase del proceso de minería de datos, desde la comprensión y representación de los datos de entrada hasta la evaluación y obtención de la salida final. La integración del conocimiento experto del dominio en el diseño del método, concretada en la validación con especialistas en biología molecular descrita en el Capítulo 6, garantizó que las decisiones técnicas fueran consistentes con la realidad del problema de anotación de genes V y no derivaran en soluciones técnicamente correctas, pero biológicamente irrelevantes.

El segundo principio, la reducción de redundancia en el espacio de búsqueda, fue verificado cuantitativamente en todos los genomas evaluados. La etapa de unificación de datos reduce el número de registros candidatos al identificar el archivo de menor redundancia entre las fuentes disponibles, y el agrupamiento posterior concentra el análisis en el subconjunto con mayor densidad de genes V reales. Los resultados demuestran coberturas superiores al 90% en varios genomas con repertorios de referencia de hasta 850 genes, lo que confirma que la reducción del espacio de búsqueda no compromete la recuperación de los elementos de interés.

El tercer principio, la coherencia biológica de las predicciones, fue verificado tanto cuantitativamente con las métricas de cobertura y RMSE como cualitativamente con el análisis visual en IGV presentado en la Sección 8.3.2. La identificación sistemática de las tres secciones estructurales del gen V, incluyendo el SP_exon ausente en todas las predicciones de IgDetective, demuestra que el agrupamiento captura patrones estructuralmente coherentes con la arquitectura real del gen V y no simplemente regiones con similitud superficial a las secuencias de referencia.

El cuarto principio, la reusabilidad del método en distintos contextos del genoma, fue verificado a través de la validación externa sobre tres genomas con características deliberadamente distintas. El método mantuvo su capacidad operativa en *Plecotus auritus* con un repertorio significativamente mayor al utilizado durante la optimización,

en *Ambystoma mexicanum* fuera del dominio taxonómico de los quirópteros y en *Molossus nigricans* sin anotación experta disponible, en todos los casos sin reoptimización de los parámetros del modelo. Esta propiedad es la que distingue al método de enfoques más específicos y lo posiciona como una herramienta de uso general para la reducción del espacio de búsqueda en genes V de vertebrados.

El quinto principio, la interoperabilidad con los formatos y herramientas estándar del dominio, se materializa en la adopción del formato GFF como entrada y salida del método. Esta decisión garantiza que el método puede integrarse directamente en flujos de trabajo existentes de anotación del genoma sin requerir conversiones adicionales ni modificaciones en las herramientas adyacentes. Los resultados de la comparación con IgDetective confirman que esta interoperabilidad no solo es técnicamente funcional, sino que permite una comparación directa y justa entre ambos enfoques sobre las mismas regiones de interés.

En conjunto, la verificación empírica de los cinco principios de diseño a través de los resultados experimentales cierra el ciclo de diseño de la metodología DSR. El método propuesto no solo resuelve el problema planteado de forma efectiva y generalizable, sino que lo hace de manera metodológicamente fundamentada, con cada decisión de diseño respaldada por evidencia experimental y coherente con los principios establecidos desde el inicio del proceso de desarrollo.

Capítulo 9. Conclusiones

Esta investigación desarrolló y validó un método computacional para la reducción del espacio de búsqueda en la identificación de genes variables en genomas de vertebrados, integrando principios de diseño derivados del dominio biológico con técnicas de aprendizaje automático no supervisado. Los resultados a lo largo del proceso de evaluación y validación permiten formular las siguientes conclusiones.

El objetivo general de la investigación fue desarrollar un método para la identificación de genes V con la reducción del espacio de búsqueda de las posiciones de inicio y fin en alineamientos de secuencias, integrando principios de diseño y técnicas de aprendizaje automático para su implementación y validación. Este objetivo fue alcanzado en su totalidad. El método propuesto reduce el espacio de búsqueda generado por múltiples herramientas de alineamiento mediante agrupamiento no supervisado con UMAP+GMM, alcanzando una cobertura media del 74.98% sobre 12 genomas de evaluación con una desviación estándar del 7.09%, lo que evidencia tanto una alta tasa de recuperación de genes V reales como un comportamiento consistente entre genomas con distintos niveles de complejidad estructural.

El primer objetivo específico planteó definir los principios de diseño del método a partir del análisis del proceso de caracterización. Este objetivo fue cumplido en el Capítulo 5, donde se establecieron cinco principios de diseño que guiaron el desarrollo e implementación del método: la adopción del marco CRISP-DM con integración de conocimiento experto del dominio, la reducción de redundancia en el espacio de búsqueda, la coherencia biológica de las predicciones, la reusabilidad en distintos entornos del genoma y la interoperabilidad con los formatos estándar del dominio. La verificación empírica de estos principios a través de los resultados experimentales del Capítulo 8 confirmó que cada uno de ellos fue efectivamente satisfecho por el método en su configuración final.

El segundo objetivo específico planteó construir un conjunto de datos unificado a partir de diversas fuentes de alineamiento de secuencias con técnicas de integración y normalización que garanticen la representación completa de los genes V. Este objetivo fue cumplido en el Capítulo 6, donde se desarrolló el proceso de unificación de datos que integra los resultados de cinco herramientas de alineamiento como HMMscan, Exonerate:est, Exonerate:prot, Tblastn y Tblastx, en una representación vectorial unificada basada en matrices de adyacencia, normalización con MaxAbsScaler y escalado robusto. El conjunto de datos resultante fue aplicado de forma consistente sobre los 15 genomas analizados en la investigación.

El tercer objetivo específico planteó identificar la variante óptima del método con la comparación sistemática de múltiples variantes, evaluando su precisión en las ubicaciones de inicio y fin de genes V. Este objetivo fue cumplido en la Sección 8.1, donde la comparación de las tres variantes con el uso de las pruebas de Friedman y *Nemenyi*

demostró con evidencia estadística sólida que UMAP+GMM es la variante óptima. El análisis reveló diferencias estadísticamente significativas en el porcentaje de identificación ($\chi^2 = 8.9333$, $p = 0.0115$), con UMAP+GMM superando a PCA+GMM en todos los genomas evaluados y mostrando ventajas concretas de hasta 37 puntos porcentuales en casos específicos como *Saccopteryx leptura*.

El cuarto objetivo específico planteó comprobar la efectividad del método con su aplicación en genomas anotados manualmente, usando métricas cuantitativas y validación comparativa con expertos. Este objetivo fue cumplido en las Secciones 8.2 y 8.3. La validación externa sobre *Plecotus auritus* alcanzó una cobertura del 91.06% sobre 850 genes de referencia sin reoptimización del modelo. La validación en *Ambystoma mexicanum* demostró transferibilidad inter-especie con una cobertura del 55.17% en un anfibio filogenéticamente distante.

La comparación con IgDetective sobre cinco genomas con anotación experta mostró que el método supera a IgDetective en tres de los cinco genomas evaluados. Más allá de las métricas cuantitativas, el análisis estructural en IGV reveló que el método identifica de forma sistemática las tres secciones que componen la arquitectura completa del gen V ausente en todas las predicciones de IgDetective en todos los genomas analizados. Esta capacidad representa una contribución estructural con implicaciones directas sobre la calidad de la anotación en estudios de diversidad inmunogenética.

En síntesis, el método propuesto constituye una contribución metodológica sólida y verificada al problema de la identificación de genes V en genomas de vertebrados. La combinación de reducción sistemática del espacio de búsqueda, completitud estructural en la delimitación del gen V y capacidad de generalización a nuevos genomas y dominios taxonómicos posiciona al método como una herramienta con potencial de adopción real en el proceso de caracterización del genoma, tanto en el contexto institucional del CISEI como en otros laboratorios que enfrenten el mismo problema en distintas especies de vertebrados.

9.1 Trabajo futuro

Los resultados y las limitaciones identificadas a lo largo de esta investigación abren un conjunto de líneas de trabajo futuro que pueden ampliar el alcance y la capacidad del método propuesto. La primera línea de trabajo corresponde al desarrollo de un mecanismo de adaptación automática del umbral de distancia intrónica. En su configuración actual, el método requiere un ajuste manual de este parámetro cuando se aplica en organismos con arquitecturas del gen V significativamente distintas a las de los quirópteros, como se evidenció en la validación con *Ambystoma mexicanum*. El diseño de un componente que estime este umbral de forma automática a partir de las propiedades estructurales del genoma de entrada eliminaría esta limitación y

aumentaría la aplicabilidad directa del método en nuevos grupos taxonómicos sin intervención del usuario, acercando el método a un flujo de procesamiento completamente automatizado.

La segunda línea de trabajo es la expansión del conjunto de evaluación a otros órdenes de vertebrados. Los 12 genomas utilizados durante la evaluación comparativa corresponden exclusivamente a quirópteros, lo que introduce un sesgo taxonómico que debe considerarse al analizar las conclusiones. La inclusión de genomas de peces, reptiles, aves y otros mamíferos en futuras evaluaciones permitirá caracterizar con mayor precisión el comportamiento del método en distintos contextos evolutivos y estructurales, y podría motivar ajustes en el proceso de optimización de hiperparámetros orientados a mejorar la generalización entre grupos taxonómicos más distantes.

La tercera línea de trabajo es la integración formal del método propuesto en el flujo de la caracterización del genoma en el CISEI. Si bien el método fue desarrollado a partir del análisis del proceso real de anotación de genes V llevado a cabo en el CISEI y fue validado con la participación de sus expertos en biología molecular, su incorporación como etapa sistemática dentro del flujo institucional representa un paso adicional que requiere el desarrollo de interfaces de usuario, documentación operativa y protocolos de uso adaptados a las necesidades específicas del equipo. Esta integración permitiría que los beneficios demostrados en términos de reducción del espacio de búsqueda, completitud estructural y eficiencia computacional se trasladen directamente a la práctica cotidiana de la caracterización del genoma, reduciendo los tiempos de anotación y el esfuerzo manual requerido por los especialistas del dominio. Dado que el método ya fue validado con la participación del equipo del CISEI, esta línea de trabajo representa la continuación de la investigación.

La cuarta línea de trabajo corresponde a la mejora de la escalabilidad del método en genomas de gran tamaño. El tiempo de procesamiento observado en los genomas evaluados es razonable para los tamaños de repertorio analizados, pero genomas con decenas de miles de genes V o varias órdenes de magnitud superiores podrían requerir estrategias de particionamiento, procesamiento paralelo o reducción progresiva del espacio de búsqueda que el método en su configuración actual no implementa explícitamente. El desarrollo de estas estrategias ampliaría el alcance del método a organismos como algunos anfibios, peces cartilaginosos o mamíferos con genomas de alta complejidad.

La quinta línea de trabajo es la exploración de criterios alternativos para la selección del grupo óptimo en genomas no anotados. El criterio actual de maximización del número de genes identificados es efectivo cuando existe una referencia experta disponible para su validación, pero introduce incertidumbre en escenarios

exploratorios como el de *Molossus nigricans*. El desarrollo de criterios basados en propiedades intrínsecas del agrupamiento, como la cohesión interna, la separabilidad entre grupos o la estabilidad de las predicciones ante perturbaciones del conjunto de datos, proporcionaría una base más sólida para la selección del grupo óptimo en genomas sin anotación previa, consolidando la aplicabilidad del método en contextos de caracterización del genoma de novo.

Referencias

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433–459.
- Ahmed, F., Robinson, S., & Tako, A. A. (2014). Using the structured analysis and design technique (SADT) in simulation conceptual modeling. *Proceedings of the Winter Simulation Conference 2014*, 1038–1049.
- Alioto, T., Blanco, E., Parra, G., & Guigó, R. (2018). Using geneid to identify genes. *Current protocols in bioinformatics*, 64(1), e56.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410.
- Ania, I., & Kuri, A. (2007). Agrupamiento en Grandes Bases de Datos vía la Reducción Estadística del Espacio de Búsqueda. *AMCIS 2007 Proceedings*, 266.
- Arunkumar, M., & Zielinski, C. E. (2021). T-Cell Receptor Repertoire Analysis with Computational Tools—An Immunologist’s Perspective. *Cells*, 10(12), 3582.
- Bayas-Jiménez, L., Martínez-Solano, F. J., Iglesias-Rey, P. L., & Mora-Meliá, D. (2021). Search space reduction for genetic algorithms applied to drainage network optimization problems. *Water*, 13(15), 2008.
- Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: theory and algorithms*. John Wiley & sons.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1), 38–44.
- Bellman, R. (1957). *Dynamic programming*, Princeton Univ. Press Princeton, New Jersey, 39.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Número 4). Springer.
- Bode, F., Reed, P., Reuschen, S., & Nowak, W. (2019). Search space representation and reduction methods to enhance multiobjective water supply monitoring design. *Water Resources Research*, 55(3), 2257–2278.
- Brown, T. (2008). *Genomas/Genome*. Ed. Médica Panamericana.
- Chao, J., Tang, F., & Xu, L. (2022). Developments in Algorithms for Sequence Alignment: A Review. *Biomolecules*, 12(4), 546.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. C.-D. (2000). 1.0: Step-by-step data mining guide. *SPSS Inc.: Chicago, IL, USA*, 9(13), 1–73.

- Chavez Serret, N. R. (2025). *Modelo conceptual del proceso de anotación del gen V de anticuerpos basado en la Ciencia del Diseño*.
- Cho, M., & Hegde, C. (2019). Reducing the Search Space for Hyperparameter Optimization Using Group Sparsity. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3627–3631. <https://doi.org/10.1109/ICASSP.2019.8682434>
- Choudhary, S. (2017). *Predicting protein coding boundaries using RNNs*.
- Chowdhury, B., & Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5–6), 419–431.
- De Amorim, R. C., & Mirkin, B. (2016). A clustering-based approach to reduce feature redundancy. *Knowledge, Information and Creativity Support Systems: Recent Trends, Advances and Solutions: Selected Papers from KICSS'2013-8th International Conference on Knowledge, Information, and Creativity Support Systems, November 7-9, 2013, Kraków, Poland*, 465–475.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1), 1–22.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan), 1–30.
- Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, 7(10), e1002195.
- EMBL-EBI. (2022). *GFF/GTF File Format - Definition and supported options*. <https://www.ensembl.org/info/website/upload/gff.html>
- Feng, L., Gupta, A., Tan, K., & Ong, Y. (2022). Multi-Space Evolutionary Search for Large-Scale Single-Objective Optimization. En *Evolutionary Multi-Task Optimization: Foundations and Methodologies* (pp. 147–169). Springer.
- Fleiss, J. L. (2011). *Design and analysis of clinical experiments* (Vol. 73). John Wiley & Sons.
- Friedman, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200), 675–701. <https://doi.org/10.1080/01621459.1937.10503522>
- Ganesh, P., Gupta, G., Saini, S., & Paul, K. (2018). Nucl2Vec: Local alignment of DNA sequences using Distributed Vector Representation. *Biorxiv*, 401851.
- García Merino, A. (2011). Anticuerpos monoclonales. Aspectos básicos. *Neurología*, 26(5), 301–306. <https://doi.org/10.1016/J.NRL.2010.10.005>
- García Simón, A., Casamayor Ródenas, J. C., & Pastor López, O. (2018). *Gestión de datos genómicos basada en Modelos Conceptuales*. <https://riunet.upv.es/handle/10251/111666>
- Gervásio, J., Ferreira, A., & Felicori, L. F. (2022). Yclon: Ultrafast clustering of B cell clones from high-throughput immunoglobulin repertoire sequencing data. *bioRxiv*, 2022.
- gmod. (2017, abril). *GFF2*. <http://gmod.org/wiki/GFF2>
- Gouret, P., Vitiello, V., Balandraud, N., Gilles, A., Pontarotti, P., & Danchin, E. G. J. (2005). FIGENIX: intelligent automation of genomic annotation: expertise integration in a new software platform. *BMC bioinformatics*, 6(1), 198.
- Grady Booch, I. J., & Rumbaugh, J. (2011). Unified Modeling Language Reference Manual, The (2nd Edition). Available on: <http://www.omg.org/spec/UML/2.1>.

- https://books.google.com/books/about/The_Unified_Modeling_Language_Reference.html?hl=es&id=0a5QAAAAMAAJ
- Hevner, A., Chatterjee, S., Hevner, A., & Chatterjee, S. (2010). Design science research in information systems. *Design research in information systems: theory and practice*, 9–22.
- Imron, M. A., & Prasetyo, B. (2020). Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn. *Journal of Soft Computing Exploration*, 1(1), 56–62.
- Ishioka, T. (2000). Extended K-means with an efficient estimation of the number of clusters. *Ouyou toukeigaku*, 29(3), 141–149.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
- José Luis Rodríguez Pliego. (2009). *Recuperación y Procesamiento de Datos Genómicos para la Carga de Información Clasificada en una Base de Datos Basada en un Modelo Conceptual del Genoma Humano*. <https://www.cenidet.edu.mx/subplan/biblio/seleccion/Tesis/MC%20Jose%20Luis%20Rodriguez%20Pliego%202009.pdf>
- 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.
- Kirilenko, B. M., Munegowda, C., Osipova, E., Jebb, D., Sharma, V., Blumer, M., Morales, A., Ahmed, A.-W., Kontopoulos, D.-G., & Hilgers, L. (2022). Integrating gene annotation with orthology inference at scale. *bioRxiv*, 2022–2029.
- Lefranc, M.-P., & Lefranc, G. (2001). *The immunoglobulin factsbook*. Academic press.
- Lemoine, F., Blassel, L., Voznica, J., & Gascuel, O. (2021). COVID-Align: Accurate online alignment of hCoV-19 genomes using a profile HMM. *Bioinformatics*, 37(12), 1761–1762.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.
- Marca, D. A., & McGowan, C. L. (1987). *SADT: structured analysis and design technique*. McGraw-Hill, Inc.
- McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- 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*.

- Mikocziova, I., Greiff, V., & Sollid, L. M. (2021). Immunoglobulin germline gene variation and its impact on human disease. *Genes & Immunity*, 22(4), 205–217.
- Moncada Vázquez, C. A. (2019). *Selección de variantes genéticas asociadas a enfermedades crónico degenerativas usando técnicas de aprendizaje automático*. Tecnológico Nacional de México. <https://rinacional.tecnm.mx/jspui/handle/TecNM/1365>
- Mount, D. W. (2004). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press. <https://books.google.com.mx/books?id=bvY21DGa1OwC>
- Nescakova, Z., & Bystrycky, S. (2011). B cells–ontogenesis and immune memory development. *Gen. Physiol. Biophys*, 30, 1–10.
- Object Management Group. (s/f). *Business Process Model & Notation™ (BPMN™)*. Recuperado el 18 de agosto de 2024, de <https://www.omg.org/bpmn/>
- 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>
- Orozco-Arias, S., Isaza, G., & Guyot, R. (2019). Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning. *International journal of molecular sciences*, 20(15), 3837.
- Oscar Contreras Carrasco. (2024). *Gaussian Mixture Model Explained* . <https://builtin.com/articles/gaussian-mixture-model>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12(null), 2825–2830.
- Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2020). Design science research process: A model for producing and presenting information systems research. *arXiv preprint arXiv:2006.02763*.
- 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.
- Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics (Oxford, England)*, 39(1). <https://doi.org/10.1093/BIOINFORMATICS/BTAC830>
- Rollins, J. B. (2015). Metodología fundamental para la ciencia de datos. *Obtenido de IBM*.
- Sandoval, S. R., Carvajal, H. V., & Zeledón, L. N. (2019). Adaptación de la metodología de ciencia de diseño en el desarrollo de luminarias. *Tecnología Vital*, 2(6).
- Selvaraj, S., & Choi, E. (2022). Dynamic Sub-Swarm Approach of PSO Algorithms for Text Document Clustering. *Sensors*, 22(24), 9653.
- Simarro, F. S. (2022). Avances en genética clínica y sus retos actuales. *Anales de Pediatría*, 97(4), 281-e1.
- Sinha, S., Lynn, A. M., & Desai, D. K. (2020). Implementation of homology based and non-homology based computational methods for the identification and annotation of orphan

- enzymes: using *Mycobacterium tuberculosis* H37Rv as a case study. *BMC bioinformatics*, 21, 1–18.
- Sirupurapu, V., Safonova, Y., & Pevzner, P. A. (2022). Gene prediction in the immunoglobulin loci. *Genome Research*, 32(6), 1152–1169.
- Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC bioinformatics*, 6, 1–11.
- Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. En L. T. Wille (Ed.), *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (pp. 273–309). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-08968-2_16
- Stiehler, F., Steinborn, 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.
- Universidad de Gent. (2015). *Formato de archivo FASTA*. <http://bioinformatics.intec.ugent.be/MotifSuite/fastafomat.php>
- Villa, P. E. D. la C. V. de, Loayza, D. E. B., & Huamán, M. R. (2010). Documentación de Proyectos Con UML 2. *Revista de investigación de Sistemas e Informática*, 7(1), 27–36. <https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/3263>
- Virrueta Gordillo, A. O. (2012). *Esquema Conceptual del Genoma Humano: una herramienta para la integración y gestión de su información*.
- White, S. A. (2004). *Introduction to BPMN*. www.bptrends.com
- Zeledón, L. N. (2020). Big data y ciencia de diseño: oportunidades de investigación. *Technology Inside by CPIC*, 4, 1–10.
- Zhang, K.-Y., Gao, Y.-Z., Du, M.-Z., Liu, S., Dong, C., & Guo, F.-B. (2019). Vgas: a viral genome annotation system. *Frontiers in microbiology*, 10, 184.
- 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.

Anexo A resúmenes del estado del arte

Resúmenes del estado del arte para el alineamiento de secuencias

Esta subsección presenta los resúmenes de los cuatro trabajos seleccionados durante la revisión sistemática correspondientes al primer núcleo temático: el uso del alineamiento de secuencias para la identificación de genes. Los trabajos incluidos abordan el alineamiento desde distintos ángulos: revisiones de algoritmos de alineamiento por pares y múltiple (Chao et al., 2022; Chowdhury & Garai, 2017), una herramienta de alineamiento de alto rendimiento para secuencias de larga longitud (Li, 2018) y una aplicación de modelos ocultos de Markov para el alineamiento de genomas virales (Lemoine et al., 2021). En conjunto, estos trabajos proporcionan el contexto técnico que fundamenta el uso de herramientas como BLAST, Exonerate y HMMER dentro del proceso de identificación de genes V descrito en el Capítulo 3.

Desarrollos en algoritmos para la alineación de secuencias: una revisión

En la investigación realizada por Chao et al. (2022), se abordan los avances en algoritmos para la alineación de secuencias, los cuales son fundamentales para el procesamiento de secuencias biológicas. El estudio se centra en los algoritmos de alineación de secuencias por pares; algoritmos heurísticos para la alineación de secuencias múltiples; y los métodos de estimación de calidad utilizados para evaluar el software de alineación. Se discute la importancia del sistema de puntuación en la calidad de la alineación, considerando matrices de sustitución y penalizaciones por huecos.

El artículo también explora estrategias de alineamiento como el alineamiento en estrella y el alineamiento progresivo. Se analizan los posibles defectos de los algoritmos heurísticos y se presentan contramedidas, como la función objetivo de consistencia y el refinamiento iterativo. Se resalta la relevancia del alineamiento de secuencias en áreas como la filogenética, la genómica comparativa y la predicción de la función y estructura de las proteínas.

En el artículo se presenta una visión general de las diversas herramientas y métodos utilizados en la alineación de secuencias, destacando sus ventajas y desventajas, y su aplicación en diferentes áreas de la biología.

Minimap2: alineación por pares para secuencias de nucleótidos

En su investigación, Li (2018) se propuso desarrollar Minimap2, un sistema de alineación de secuencias de ADN y ARN de larga longitud frente a una extensa base de datos de referencia. Minimap2 es un alineador de propósito general que puede manejar secuencias cortas de 100 caracteres y secuencias de 1000 caracteres con una tasa de error del 15 %.

Este programa utiliza el costo de la brecha cóncava para manejar inserciones y deleciones largas, y también introduce nuevas heurísticas para mejorar la precisión de las alineaciones. En términos de rendimiento, Minimap2 es de 3 a 4 veces más rápido que los principales mapeadores de lecturas cortas, manteniendo una precisión comparable. Además, supera a los mapeadores genómicos o de ADNc de lecturas largas en términos de velocidad, siendo 30 veces más rápido, sin comprometer la precisión. Asimismo, Minimap2 supera a la mayoría de los alineadores especializados en un solo tipo de alineación. La utilización de alineación de lectura dividida y las nuevas heurísticas implementadas en el programa contribuyen a reducir las alineaciones espurias, lo que garantiza una herramienta altamente precisa y eficiente para el mapeo de secuencias de ADN y ARNm.

En conclusión, Minimap2 se destaca como una solución valiosa para la alineación de secuencias de ADN y ARN de larga longitud, ofreciendo un rendimiento rápido y preciso en comparación con otros alineadores existentes. Sus características innovadoras y su capacidad para reducir las alineaciones erróneas lo convierten en una herramienta esencial en la investigación del genoma.

Una revisión sobre el alineamiento de secuencias múltiples desde la perspectiva del algoritmo genético

En el artículo desarrollado por los autores Chowdhury & Garai en (2017) se realizó un estudio exhaustivo de los métodos y algoritmos utilizados en la alineación de secuencias múltiples (MSA). Se abarcan tanto los enfoques tradicionales, como el alineamiento progresivo, los árboles guía, los enfoques iterativos y alineaciones basadas en los modelos ocultos de Markov, como también los métodos más recientes, como los algoritmos genéticos (GA) y el recocido simulado.

El artículo examina diferentes técnicas de puntuación, como la suma de pares (SOP), las penalizaciones por brecha y la puntuación basada en la consistencia. También se exploran enfoques estocásticos, como los algoritmos evolutivos (EA) y los algoritmos genéticos (GA). Se destaca la importancia del MSA en la predicción de la función y estructura de las macromoléculas biológicas y se discuten las ventajas del MSA en comparación con la alineación de secuencias por pares (PSA).

El artículo concluye que los algoritmos basados en GA muestran un gran potencial para el MSA, cuando se aplican como un problema de optimización multiobjetivo. Se proporciona un análisis comparativo de diferentes algoritmos basados en GA, junto con Clustal W/X en BALiBASE 2.0. Además, se aborda la evaluación comparativa y el uso de alineaciones de referencia para evaluar el rendimiento de diversas herramientas de alineación. Por último, se destacan las posibles direcciones de investigación futura en este campo.

En resumen, este artículo ofrece una revisión completa de los métodos y algoritmos utilizados en el MSA en bioinformática. Proporciona una visión general de las técnicas utilizadas, las ventajas del MSA sobre el PSA, y destaca el potencial de los algoritmos basados en GA. Además, se abordan aspectos clave como la evaluación comparativa y las futuras direcciones de investigación.

Covid-align: alineación precisa en línea de los genomas de hcov-19 utilizando un perfil HMM

En el presente trabajo los autores Lemoine et al. en (2021) realizaron el desarrollo de un sistema basado en modelos ocultos de Markov (HMM) para el alineamiento de secuencias de hCoV-19 causante del COVID-19. Esta herramienta, diseñada para ser fácil de usar, proporciona estadísticas resumidas para evaluar la calidad de la secuenciación y la novedad evolutiva de los genomas de entrada. Su enfoque es adecuado para alinear nuevos genomas debido a la alta conservación del virus y la falta de recombinación.

COVID-Align ofrece una notable eficiencia, siendo capaz de procesar 1000 genomas en aproximadamente 50 minutos. Esto la convierte en una herramienta altamente eficiente para los investigadores que estudian la evolución y propagación del virus. Además, la herramienta se encuentra disponible en línea y también puede ser utilizada de manera independiente a través de Docker, lo que la hace fácilmente accesible para investigadores de todo el mundo.

En general, COVID-Align es una herramienta valiosa para la comunidad científica en la lucha contra el COVID-19. Su capacidad para alinear con precisión nuevos genomas y proporcionar estadísticas resumidas la convierte en una herramienta esencial para los investigadores que estudian el virus y su evolución.

Resúmenes del estado del arte del proceso de anotación del genoma e identificación de genes

Esta subsección presenta los resúmenes de los diecisiete trabajos seleccionados durante la revisión sistemática correspondientes al segundo núcleo temático: la sistematización del proceso de anotación de genes e identificación de regiones genómicas funcionales. Los trabajos cubren un espectro amplio de enfoques y

dominios: métodos basados en homología y predicción ab initio (GeMoMa, geneid, TOGA), sistemas especializados en genomas virales y microbianos (Mgcod, VGAS, GOODORFS), herramientas de aprendizaje profundo para la identificación de señales genómicas (DeepAnnotator, DeepGSR, Helixer), métodos de análisis de repertorios de inmunoglobulinas (IgDetective, VgeneFinder, RAPID, YClon) y enfoques basados en RNN para la predicción de límites de codificación. La revisión de estos trabajos permitió identificar las brechas metodológicas descritas en la Sección 4.3, que justifican el desarrollo del método propuesto en esta investigación.

GEMOMA: predicción de genes basada en homología utilizando datos de secuenciación de ARN y secuenciación de intrones poco conservados

En la investigación desarrollada por Keilwagen et al. en (2019), se tuvo el objetivo de desarrollar un proceso de anotación de genes que utilice una combinación de enfoques basados en la homología y la conservación de secciones del genoma para predecir genes. *Gene Model Mapper* (GeMoMa) es un pipeline, implementado en Java, que utiliza la herramienta TBLASTN para identificar regiones genómicas similares a exones de especies de referencia. Además, GeMoMa aprovecha datos de RNA-seq mapeados para una predicción con una mayor presión e identificar sitios de empalme.

GeMoMa filtra todos los resultados de TBLASTN que se producen, en especial para secciones parecidas a exones cortos con respecto a su posición. De este modo, GeMoMa identifica las regiones que son similares a secuencias de exones de referencia. GeMoMa fue comparado con otros programas como BRAKER1, MAKER2 y CodingQuarry, GeMoMa demostró un rendimiento igual o incluso mejor al incluir datos de RNA-seq.

En resumen, GeMoMa es una herramienta valiosa para la anotación de genes, debido a que combina diferentes enfoques para mejorar la precisión y el rendimiento en este proceso. Aunque en el artículo no se explica cómo se compararon las secuencias, es de gran interés este sistema para la identificación y anotación de genes presentes en el genoma de un individuo.

Uso de geneid para identificar genes

En el protocolo desarrollado por Alioto et al. (2018), presentó geneid, un programa de búsqueda de genes que facilita el análisis de secuencias genómicas extensas y la mejora de anotaciones existentes. El protocolo detalla de manera exhaustiva los pasos necesarios para utilizar el software geneid, tanto para la predicción de genes como para la visualización de los resultados con herramientas como Integrative Genomics Viewer (IGV). Además, se proporcionan instrucciones sobre cómo incorporar información externa con el fin de mejorar las predicciones realizadas por geneid.

El programa utiliza como primer método de identificación matrices de pesos de posición (PWM) para predecir posibles marcadores de inicio, corte o unión de secuencias del ADN. El segundo método consiste en los modelos ocultos de Markov para asignar puntuaciones a posibles exones. Geneid distingue cuatro tipos de exones: iniciales, internos, terminales y únicos. Si bien geneid ofrece una alta especificidad en la búsqueda de genes, puede pasar por alto algunos exones reales si su secuencia es muy corta. La puntuación de un gen se calcula mediante la suma de los cocientes de verosimilitud de sus exones, con un factor de corrección que considera el número de exones.

Aunque geneid es un programa de búsqueda de genes altamente efectivo. Se debe considerar que existen algunas limitaciones en la detección de exones cortos. Su enfoque basado en PWM y modelos ocultos de Markov ofrece una alta especificidad para la identificación y anotación de genes.

Anotación de genomas de fagos con múltiples códigos genéticos

En la investigación realizada por Pfennig et al. (2022) se desarrolló Mgcod, una herramienta para anotar genomas de fagos con múltiples códigos genéticos. Con ventanas deslizantes y datos de secuencias codificantes de proteínas (PES), Mgcod segmenta el genoma en bloques de códigos genéticos y anota las regiones codificadoras de proteínas. Se evaluó su precisión en genomas simulados y se aplicó a la detección de cambios de codones de stop en secciones de genomas. Los resultados revelaron varios contigs de fagos con múltiples códigos genéticos, aportando nuevos conocimientos sobre la diversidad y evolución de los fagos.

Mgcod es una herramienta que procesa archivos FASTA y proporciona anotaciones genéticas en formato GTF/GFF2. El sistema genera un archivo separado por tabulaciones que muestra el número de codones de stop reasignados por gen, junto con sus coordenadas correspondientes. La funcionalidad principal de Mgcod es analizar las secuencias de ADN contenidas en los archivos FASTA y realizar anotaciones genéticas para identificar y caracterizar los genes presentes en cada registro.

En resumen, Mgcod es una potente herramienta para la anotación precisa de genomas de fagos con múltiples códigos genéticos, brindando nuevos conocimientos sobre la diversidad y evolución de los fagos. Además, tiene posibles aplicaciones en la anotación de regiones codificadoras en secuencias metagenómicas, y podría ser un recurso valioso para investigadores que estudian fagos y otros organismos con códigos genéticos no estándar.

Uso de la composición de aminoácidos y la entropía de posibles marcos de lectura abiertos para identificar genes codificadores de proteínas

El artículo realizado por McNair et al. (2021), introduce el método GOODORFS para identificar genes codificadores de proteínas en genomas de procariontes. GOODORFS utiliza una secuencia en aminoácidos y la entropía de los marcos abiertos de lectura (ORF) codificantes para su agrupación e identificación. El algoritmo incluye la búsqueda de ORF, el cálculo del perfil de densidad de entropía (EDP), la agrupación y la selección del conjunto con ORFs codificantes.

GOODORFS utiliza una métrica EDP de 24 dimensiones, que considera los codones de stop y la longitud del ORF, para agrupar los ORF potenciales en categorías codificantes y no codificantes. El método se basa en aprendizaje no supervisado y no requiere referencias el algoritmo utilizado para la implementación de GOODORFS es K-Means.

En una prueba con 14,179 genomas de fagos anotados, GOODORFS superó a otros cuatro métodos similares, los cuales son Glimmer, MED2, PHANOTATE, Prodigal. Las métricas utilizadas y sus resultados consisten en precisión, recall, accuracy y puntuación F1, obteniendo una precisión de 0.92, un recall de 0.96, un accuracy de 0.94 y una puntuación F1 media de 0.85.

TOGA integración de la anotación de genes con la inferencia de ortología a escala

En la investigación realizada por Kirilenko et al. (2022), se desarrolló el método TOGA, el cual integra la anotación de genes y la inferencia de secuencias compartidas entre especies (ortología), superando los desafíos tradicionales de abordar estos dos aspectos por separado. El método utiliza una técnica basada en alineamiento de secuencias de ADN no codificante y regiones entre dos genes, para inferir genes ortólogos, mejorando la detección y la anotación de genes en comparación con otros métodos.

El método de TOGA consta de varios pasos, los cuales consisten en: la detección de secuencias compartidas entre especies, la anotación, clasificación, y determinación de la clase identificada. La calidad de las anotaciones generadas por TOGA, se compararon con las anotaciones de Ensembl y NCBI. Los resultados mostraron que las anotaciones de TOGA tenían una mayor integridad en las especies evaluadas, superando a los métodos existentes. Además, se utilizó la herramienta BUSCO para evaluar la integridad de las anotaciones de genes, y se encontró que las anotaciones de TOGA eran más completas en promedio. En resumen, TOGA es un método innovador y escalable que integra la anotación de genes y la inferencia de ortología.

Predicción de genes en los loci de inmunoglobulinas

En la investigación desarrollada por Sirupurapu et al. en (2021) se tuvo el objetivo de desarrollar el algoritmo IGDetective, diseñado para identificar genes del segmento V(D)J en genomas de vertebrados. El algoritmo utiliza una combinación de métodos computacionales, como algoritmos de predicción de genes, análisis de razón de verosimilitud y técnicas de agrupamiento. En la investigación se identificaron genes del segmento V(D)J basándose en la presencia de secuencias de señales de recombinación (RSS) en el genoma. El rendimiento del algoritmo se evaluó mediante métricas como tasas de verdaderos positivos y falsos positivos.

Los resultados mostraron que IGDetective pudo identificar los genes presentes en los segmentos V(D)J en diferentes especies de vertebrados, como humanos, ratones, vacas y un conjunto de 20 vertebrados diferentes. Además, el algoritmo descubrió nuevos genes de inmunoglobulinas (IG) no anotados previamente en el genoma. Mediante técnicas de agrupamiento, los autores también analizaron los patrones de uso de diferentes RSS en el genoma, proporcionando información sobre la historia evolutiva de las familias de genes IG.

Finalmente, la investigación destaca la utilidad de los métodos computacionales para la identificación y análisis de genes IG en genomas de vertebrados. El algoritmo IGDetective ofrece una herramienta para los investigadores que estén interesados en el sistema inmunológico y su evolución, con aplicaciones importantes en campos como la inmunología, la genética y la biotecnología.

Descubrimiento iterativo de genes variables a partir de la secuenciación del genoma completo con un algoritmo de resolución múltiple basado en pruebas de arranque

En su estudio, Olivieri y Gambón-Deza (2019) se propusieron desarrollar un algoritmo iterativo de aprendizaje automático supervisado llamado VgeneFinder para la identificación precisa de secuencias de genes V. Este enfoque utiliza la traducción de las secuencias en vectores numéricos de características y emplea un clasificador Random Forest para determinar la validez de los genes V y su ubicación en el genoma. Una mejora significativa de VgeneFinder con respecto a métodos anteriores radica en la inclusión de genes V de diversas familias de vertebrados en su base de conocimientos, lo que permite un descubrimiento de genes V con mayor precisión.

El artículo también aborda las limitaciones de otros programas informáticos de búsqueda de genes V y resalta la necesidad de un algoritmo con mayor precisión debido a la complejidad y variabilidad de los límites de los exones presentes en los genes V. Además, se presentan métricas de evaluación, como las puntuaciones de

multiresolución (MR) y las distribuciones de probabilidad de clase, que respaldan el éxito de VgeneFinder en la identificación de genes V con alta especificidad.

En resumen, el algoritmo VgeneFinder desarrollado en el estudio representa un avance significativo en la identificación precisa de secuencias de genes V. El uso de genes V de diferentes familias de vertebrados y su enfoque de aprendizaje automático supervisado muestran resultados prometedores y beneficios potenciales para la investigación genómica y la identificación de homologías de genes V.

DEEPGSR: una estructura optimizada de aprendizaje profundo para el reconocimiento de señales y regiones genómicas

El artículo realizado por Kalkatawi, Magana Mora, et al. (2019), introduce DeepGenGrep, un sistema basado en aprendizaje profundo para la predicción de múltiples señales de regiones genómicas (GSR) a partir de secuencias de ADN. Este sistema utiliza una red neuronal convolucional de tres capas (CNN) y una red de memoria a corto plazo de dos capas (LSTM) para extraer características relevantes de las secuencias en un enfoque híbrido.

DeepGenGrep logró un rendimiento sobresaliente en la predicción de sitios de iniciación de la transcripción (TIS), sitios de poliadenilación (PAS) y sitios de empalme en diferentes especies. Los investigadores llevaron a cabo experimentos para evaluar el rendimiento de DeepGenGrep en tres tipos de GSR (TIS, PAS y aceptor de empalme) en cuatro especies (*H.sapiens*, *M.musculus*, *B.taurus* y *D.melanogaster*), y lo compararon con otros métodos. Los resultados demuestran que DeepGenGrep supera a los otros métodos en términos de precisión, exactitud y puntuación F1.

El artículo proporciona una descripción detallada de la metodología utilizada, los resultados y el análisis de los experimentos realizados. Este sistema puede ser utilizado para identificar regiones de señal genómica potenciales en genomas eucarióticos. Las métricas de evaluación utilizadas en los experimentos incluyen precisión, exactitud y puntuación F1. En general, DeepGenGrep es una herramienta potente basada en aprendizaje profundo para la predicción precisa de diferentes tipos de regiones de señal genómica a partir de secuencias de ADN.

HELIXER: anotación de genes entre especies de grandes genomas eucariotas mediante aprendizaje profundo

En el artículo realizado por Stiehler et al. (2020), presentaron Helixer, una herramienta basada en aprendizaje profundo con redes neuronales convolucionales (CNN) para la predicción de genes en genomas. Esta herramienta fue entrenada utilizando un amplio conjunto de datos de genomas de animales y plantas, logrando una alta precisión en la predicción de estructuras de genes. Helixer supera la dependencia de modelos de genes

específicos y mejora la calidad predictiva mediante una arquitectura completamente nueva. Los investigadores desarrollaron y entrenaron dos modelos, uno para la anotación de genomas animales y otro para genomas vegetales, y evaluaron su rendimiento utilizando un conjunto de genomas de prueba. La herramienta también fue evaluada con datos independientes de RNAseq, demostrando un buen rendimiento.

En comparación con AUGUSTUS, una herramienta ampliamente utilizada para los genes caller, Helixer demostró superarlo en términos de precisión. Además, Helixer obtuvo mejores resultados en métricas de pares de bases y mostró mayor consistencia con datos de RNAseq independientes. Los autores sugieren que el desarrollo de un método de posprocesamiento para convertir predicciones basadas en pares de bases en predicciones de transcritos completos en cada loci podría mejorar aún más el rendimiento y ampliar las aplicaciones en el mundo real.

En resumen, Helixer es una herramienta prometedora para la predicción precisa de genes en genomas, basada en aprendizaje profundo, y ha demostrado su superioridad sobre métodos existentes en términos de precisión y consistencia con datos independientes.

VGAS: un sistema de anotación de genomas virales

El artículo desarrollado por K.-Y. Zhang et al. (2019), Vgas, un sistema de anotación de genes de virus que utiliza una combinación de algoritmos de predicción de genes basados en la homología y de novo. El sistema fue evaluado utilizando un amplio conjunto de datos de 5,705 genomas virales y se comparó con otros programas de búsqueda de genes, como GeneMark, Glimmer y Prodigal. Se utilizaron parámetros de evaluación como sensibilidad, especificidad y precisión.

Los resultados demostraron que Vgas obtuvo los mejores resultados, en virus pequeños, y fue capaz de identificar nuevos genes no anotados previamente en RefSeq. Además, se encontró que Vgas era más adecuado para la anotación de genes de virus en comparación con otros sistemas diseñados para bacterias. El sistema Vgas consta de varios módulos que se basan en homología, predicción de genes de novo, y un módulo de fusión y filtrado de predicciones de genes.

El módulo basado en la homología utiliza BLAST para buscar secuencias homólogas en una base de datos de referencia, mientras que el módulo de novo utiliza modelos ocultos de Markov y redes neuronales para predecir genes basándose en las características de la secuencia. La evaluación del sistema utilizando los genomas virales demostró que Vgas superó a otros programas de búsqueda de genes en términos de sensibilidad, especificidad y precisión. Además, Vgas fue capaz de identificar nuevos genes no anotados previamente en RefSeq, lo que demuestra su capacidad para mejorar la anotación de genomas virales.

En resumen, el sistema Vgas representa una mejora significativa en la anotación de genes virales y tiene el potencial de facilitar el descubrimiento de nuevos genes y funciones en virus.

Predicción de los límites de codificación de proteínas utilizando rnn

En la investigación realizada por Choudhary en (2017) se realizó una revisión de las redes neuronales recurrentes (RNNs) para predecir los dominios de codificación de proteínas en genes. El modelo propuesto emplea un enfoque de incrustación de palabras junto con modelos de memoria a largo plazo o a corto plazo (LSTM) bidireccionales para aprender los límites de los dominios de codificación de proteínas en los genes. Se entrenó el modelo con 5,000 genes humanos y se probó con secuencias de ratones y *C. elegans*. Según el estudio, el modelo logró un accuracy media global de 0.67 en el conjunto de datos humanos, y de 0.70 y 0.71 en los conjuntos de ratones y *C. elegans*, respectivamente.

El artículo sugiere la implementación de autocodificadores, convolucionales y campos aleatorios condicionales para mejorar la precisión del modelo. Sin embargo, se destacan algunas limitaciones, como la lentitud del entrenamiento de los LSTM y la necesidad de contar con más datos de entrenamiento. El estudio encontró que el modelo era capaz de aprender características globales que determinan los límites de codificación de proteínas entre diferentes especies.

En resumen, el artículo propone un enfoque prometedor para predecir los dominios de codificación de proteínas en genes utilizando RNNs, y sugiere posibles mejoras para aumentar la precisión del modelo. No obstante, se requiere más investigación para abordar las limitaciones del estudio y validar el rendimiento del modelo en conjuntos de datos más amplios.

DEEPANNOTATOR: anotación del genoma con aprendizaje profundo

En el estudio desarrollado por Amin et al. (2018), se presenta un sistema llamado DeepAnnotator que utiliza el aprendizaje profundo para la anotación del genoma. El sistema alcanza una puntuación F del 94%, eliminando la necesidad de características diseñadas manualmente. El aprendizaje profundo fue utilizado para escanear genomas completos de la base de datos NCBI. Se utilizaron incrustaciones de palabras, que son una representación eficiente de las relaciones semánticas y sintácticas, para capturar la información de las secuencias largas de ADN. La arquitectura del sistema incluye capas LSTM bidireccionales y capas densas con activación sigmoide.

Las características de la experimentación, donde se utilizaron conjuntos de datos positivos y negativos para entrenar y evaluar los modelos del sistema. Los conjuntos de datos positivos contienen secuencias de genes reales, mientras que los conjuntos de datos negativos incluyen secuencias de codificación y regiones intergénicas. Además, se

realizó una comparación del sistema DeepAnnotator con otros tres sistemas de anotación de genomas utilizando genomas de especies como *Bacillus subtilis*, *Escherichia coli* y *Synechocystis* sp.

Los resultados del sistema DeepAnnotator se presentan en términos de métricas de precisión para cada uno de los modelos del sistema. Además, se realiza una comparación con otros sistemas de anotación utilizando métricas como *both_index*, *only_start*, *only_stop*, *contained* y *overlap*, que evalúan diferentes aspectos de las anotaciones realizadas.

En conclusión, el sistema DeepAnnotator representa un enfoque prometedor para la anotación del genoma utilizando el aprendizaje profundo. Los resultados demuestran la efectividad del sistema y se sugiere el uso de métricas adicionales para evaluar los modelos de inteligencia artificial.

Implementación de métodos computacionales basados en la homología y no basados en la homología para la identificación y anotación de enzimas huérfanas: utilizando mycobacterium tuberculosis h37rv como caso de estudio

El estudio realizado por Sinha et al. (2020), propone un enfoque sistemático para identificar las enzimas que faltan en los genomas microbianos, utilizando métodos basados y no basados en la homología. Se realizó una prueba en el genoma de *Mycobacterium tuberculosis* H37Rv, donde se identificaron 56 predicciones de enzimas nuevas utilizando el método basado en homología y 87 enzimas/agujeros ausentes utilizando el método no basado en homología.

Los autores desarrollaron un "protocolo de relleno de agujeros" para identificar proteínas candidatas para las enzimas que faltaban, utilizando una puntuación combinada de interacciones de la base de datos STRING. Este protocolo se validó utilizando los genomas de *E. coli* y otros ocho organismos.

El estudio resalta la importancia de los métodos no basados en la homología, que utilizan el contexto de la red metabólica para predecir la función de las proteínas cuando no hay similitud de secuencia significativa. Los autores desarrollaron métodos para identificar las enzimas que faltan en una vía metabólica y asociar secuencias de proteínas a los huecos utilizando un conjunto de secuencias candidatas no metabólicas e hipotéticas.

En conclusión, los investigadores han desarrollado un protocolo denominado "Hole finding protocol" para identificar las enzimas que faltan en las rutas metabólicas de los organismos. Este enfoque combina métodos basados y no basados en la homología y utiliza el contexto de la red metabólica para predecir la función de las proteínas. El

protocolo desarrollado podría ser útil en el descubrimiento de fármacos al identificar posibles dianas farmacológicas en las rutas metabólicas de los microorganismos.

Rapid: una plataforma de análisis de conjuntos de datos rep-seq con una base de datos de anticuerpos integrada

El repertorio de anticuerpos desempeña un papel crucial en el sistema inmune adaptativo y su comprensión ha mejorado la investigación en inmunidad humoral, el descubrimiento de anticuerpos y el diagnóstico y tratamiento de enfermedades. A pesar de esto, hasta el momento no había existido una herramienta que integrará los grandes conjuntos de datos de Rep-Seq y el conocimiento previo de los anticuerpos funcionales para aclarar el diverso repertorio de anticuerpos.

En respuesta a esta necesidad los investigadores Y. Zhang et al. (2021), desarrollaron RAPID, una plataforma en línea y gratuita que permite a los investigadores procesar y analizar conjuntos de datos de Rep-Seq. RAPID incorpora una base de datos de anticuerpos que incluye 521 anticuerpos terapéuticos reconocidos por la OMS, 88,059 anticuerpos específicos de antígenos o enfermedades y 306 millones de clones derivados de 2449 conjuntos de datos IGH Rep-Seq humanos.

Además, RAPID ofrece un pipeline de análisis de conjuntos de datos Rep-Seq estandarizada que permite a los usuarios cargar, analizar y comparar sus propios conjuntos de datos con los repertorios existentes. La plataforma anota automáticamente los clones basándose en los anticuerpos terapéuticos y conocidos integrados, y permite a los usuarios buscar y consultar fácilmente los anticuerpos y repertorios utilizando secuencias o palabras clave.

En resumen, RAPID se presenta como una herramienta esencial para explorar y aplicar el repertorio inmunitario a través del análisis de conjuntos de datos Rep-Seq y el acceso a una base de datos integral de anticuerpos.

Análisis del repertorio del receptor de células t con herramientas computacionales: una perspectiva del inmunólogo

Este artículo desarrollado por Arunkumar & Zielinski (2021), se aborda el uso de herramientas computacionales en el análisis de los repertorios de receptores de células T. El análisis brinda información sobre la exposición previa a antígenos de un individuo y pueden servir como biomarcadores de perturbación inmunológica. Se presenta una visión general de las metodologías disponibles para el análisis de repertorios y se comparan las herramientas bioinformáticas comúnmente utilizadas en el análisis de repertorios de TCR, resaltando sus ventajas y limitaciones desde la perspectiva de un inmunólogo.

En el artículo se discute la importancia de los índices de diversidad en el análisis de los repertorios de TCR, así como los posibles sesgos y limitaciones asociadas a estos índices. Además, se mencionan varias herramientas y métodos para analizar los repertorios de células T y B a partir de datos de secuenciación de células individuales. Estas herramientas incluyen el coeficiente de Gini, DE50, el índice de Morisita-Horn, el índice de Jaccard, el índice de Yu-Clayton. Las herramientas utilizadas en el análisis de repertorios de TCR y BCR en inmunología son cinco y consisten en: Scirpy, Immunarch, ImmunoSEQ Analyzer 3.0, Immcantation portal y VDJtools.

Cada herramienta tiene sus propias fortalezas y debilidades, y se advierte a los investigadores sobre la importancia de ser cuidadosos al comparar resultados de diferentes herramientas que utilizan diferentes técnicas.

YCLON: agrupación ultrarrápida de clones de células b a partir de datos de secuenciación del repertorio de inmunoglobulinas de alto rendimiento

En la investigación realizada por Gervásio et al. (2022), se desarrolló YClon, la cual es una herramienta novedosa para el clonotipado de repertorios de anticuerpos que utiliza un enfoque de agrupación jerárquica. Implementada en Python3 y disponible de forma gratuita en GitHub, esta herramienta ha sido desarrollada por JG, AF y LFF, y ha sido respaldada por varias subvenciones.

En el estudio realizado, los autores determinaron el mejor tamaño de n-grama y umbral de corte para vectorizar y agrupar secuencias utilizando 74 repertorios de anticuerpos simulados. Compararon YClon con otros métodos de clonotipado y encontraron que YClon mostraba un rendimiento mejorado y una velocidad de clonotipado más rápida.

La funcionalidad de YClon se evaluó utilizando tres conjuntos de datos de repertorios de anticuerpos de diferentes tamaños. El sistema de YClon realiza un preprocesamiento de los datos de entrada para eliminar secuencias de baja calidad y filtrar aquellas que no corresponden a anticuerpos. Luego, las secuencias restantes se vectorizan utilizando un enfoque k-mer, donde el tamaño de n-grama es determinado por los propios autores. A continuación, las secuencias vectorizadas se agrupan mediante un algoritmo de agrupación jerárquica.

Las métricas de evaluación utilizadas en el estudio incluyen precisión, sensibilidad y puntuación F1. Además, se descubrió que YClon es más de 130 veces más rápido que los métodos existentes, pudiendo procesar hasta 2 millones de secuencias de Ig en un ordenador portátil. En conclusión, YClon es un método eficaz y novedoso para identificar clones en grandes conjuntos de datos de repertorios de anticuerpos. Ha sido

probado en conjuntos de datos con más de 2 millones de secuencias y ha demostrado un rendimiento superior en términos de velocidad y precisión en comparación con otros métodos de clonotipado.

Retrotransposones en genomas de plantas: estructura, identificación y clasificación mediante bioinformática y aprendizaje automático

En el artículo realizado por Orozco-Arias et al. (2019), se revisa la importancia de los elementos transponibles (TEs) en la arquitectura y evolución del genoma de las plantas, centrándose en los retrotransposones, que son el tipo más común de TEs en los genomas de las plantas. Se discuten la estructura, diversidad, dinámica y función de los retrotransposones, así como los métodos bioinformáticos y algoritmos utilizados actualmente para detectar y clasificar los TEs.

También se explora el potencial de los enfoques de aprendizaje automático y aprendizaje profundo para mejorar la detección y clasificación de los TEs. Se enfatiza la importancia de la clasificación de los TEs para comprender la dinámica y los mecanismos de evolución del genoma.

En resumen, el artículo proporciona una visión general de la importancia de los elementos transponibles en el genoma de las plantas, con un enfoque en los retrotransposones. Se abordan aspectos relacionados con su estructura, función, detección, clasificación y su papel en la evolución del genoma y la biotecnología vegetal.

Anexo B resultado de comparativa de algoritmos

Este anexo presenta los resultados detallados de la comparativa de algoritmos de agrupamiento realizada durante la segunda iteración del método, descrita en la Sección 6.1.2. Se evaluaron cinco algoritmos siendo estos K-Means, Fuzzy c-means, Mezclas Gaussianas, OPTICS y Spectral Clustering, bajo configuraciones de 3 a 7 grupos, aplicados sobre el conjunto de datos del genoma *Eptesicus nilssonii* (EPNI). Para cada configuración se reportan el número de registros por grupo, el número de genes identificados, el número de genes de referencia, el MAE, el RMSE y el porcentaje de identificación. Los resultados de esta comparativa fundamentaron la selección del algoritmo GMM como componente central del método RIVAM.

Tabla de anexos 1 Resultados de la comparación de algoritmos para 3 grupos

Algoritmo	Grupo	N. datos	N. genes	N. de genes de referencia	MAE	RMSE	Porcentaje de identificación
K-Means	1	897	59	71	5.60452	11.414878	83.09859155
	2	796	47	71	14.312057	23.90792	66.1971831
	3	1210	65	71	3.74359	7.474058	91.54929577
Fuzzy c means	1	801	47	71	14.234043	23.746444	66.1971831
	2	905	60	71	5.555556	11.339704	84.50704225
	3	1197	65	71	3.723077	7.434155	91.54929577
Gaussian mixtures	1	735	54	71	4.966049	10.108821	76.05633803
	2	882	50	71	13.773333	22.989853	70.42253521
	3	1286	65	71	3.774359	7.485028	91.54929577
OPTICS	1	2809	71	71	5.429577	11.253481	100
	2	42	17	71	0.72549	1.102582	23.94366197
	3	52	21	71	3.103175	6.12502	29.57746479
Spectral clustering	1	789	47	71	14.620567	24.123721	66.1971831
	2	1536	58	71	2.833333	5.150605	81.69014085
	3	578	61	71	4.008197	7.66637	85.91549296

Tabla de anexos 2 Resultados de la comparación de algoritmos para 4 grupos

Algoritmo	Grupo	N. datos	N. genes	N. de genes de referencia	MAE	RMSE	Porcentaje de identificación
K-Means	1	291	33	71	11.106061	18.262535	46.47887324
	2	1170	64	71	3.713542	7.48575	90.14084507
	3	679	45	71	14.537037	24.334475	63.38028169
	4	763	56	71	4.758929	10.090808	78.87323944
Fuzzy c means	1	696	50	71	3.043333	5.018964	70.42253521
	2	755	44	71	14.996212	24.7142	61.97183099
	3	836	57	71	5.672515	11.611902	80.28169014
	4	616	61	71	4.095628	8.142347	85.91549296
Gaussian mixtures	1	413	25	71	6.28	10.648318	35.21126761
	2	1109	53	71	3.160377	5.415735	74.64788732
	3	721	39	71	14.662393	24.571977	54.92957746
	4	660	42	71	4.198413	8.92384	59.15492958
OPTICS	1	2681	71	71	4.312207	9.042285	100
	2	67	12	71	12.930556	21.625281	16.90140845
	3	95	22	71	14.386364	21.775194	30.98591549
	4	60	6	71	3.722222	6.159906	8.450704225
Spectral clustering	1	1158	64	71	3.114583	6.368461	90.14084507
	2	683	53	71	3.160377	4.912761	74.64788732
	3	693	46	71	14.286232	23.832827	64.78873239
	4	369	34	71	9.779412	16.659067	47.88732394

Tabla de anexos 3 Resultados de la comparación de algoritmos para 5 grupos

Algoritmo	Grupo	N. datos	N. genes	N. de genes de referencia	MAE	RMSE	Porcentaje de identificación
K-Means	1	281	33	71	11.09596	18.257004	46.47887324
	2	1162	64	71	3.744792	7.506941	90.14084507
	3	365	33	71	13.919192	24.124676	46.47887324
	4	759	56	71	4.657738	9.805398	78.87323944
	5	336	29	71	14.752874	24.030034	40.84507042
Fuzzy c means	1	371	33	71	13.919192	24.124676	46.47887324
	2	809	58	71	5.698276	11.488	81.69014085
	3	589	61	71	4.051913	8.024382	85.91549296
	4	444	31	71	14.22043	23.210004	43.66197183
	5	690	49	71	3.153061	5.195825	69.01408451
Gaussian mixtures	1	418	26	71	6.788462	12.059415	36.61971831
	2	1107	53	71	3.160377	5.415735	74.64788732
	3	357	27	71	14.882716	25.506354	38.02816901
	4	659	42	71	4.18254	8.917612	59.15492958
	5	362	27	71	14.290123	23.555687	38.02816901
OPTICS	1	2489	71	71	3.948357	8.141357	100
	2	189	14	71	17.25	28.054157	19.71830986
	3	67	12	71	12.930556	21.625281	16.90140845
	4	95	22	71	14.386364	21.775194	30.98591549
	5	63	6	71	3.638889	6.125992	8.450704225
Spectral clustering	1	494	35	71	12.019048	20.960961	49.29577465
	2	600	52	71	3.455128	5.96517	73.23943662
	3	566	42	71	9.123016	18.625273	59.15492958
	4	264	12	71	7.583333	12.234968	16.90140845
	5	979	55	71	2.266667	3.915006	77.46478873

Tabla de anexos 4 Resultados de la comparación de algoritmos para 6 grupos

Algoritmo	Grupo	N. datos	N. genes	N. de genes de referencia	MAE	RMSE	Porcentaje de identificación
K-Means	1	275	32	71	11.3125	18.57025	45.07042254
	2	892	53	71	3.267296	5.411669	74.64788732
	3	365	33	71	13.919192	24.12467	46.47887324
	4	673	52	71	5.076923	10.509458	73.23943662
	5	335	29	71	14.752874	24.036012	40.84507042
	6	363	56	71	4.0625	8.190623	78.87323944
Fuzzy c means	1	555	60	71	4.102778	8.143675	84.50704225
	2	488	52	71	4.467949	9.667882	73.23943662
	3	369	33	71	13.90404	24.123106	46.47887324
	4	664	45	71	3.174074	5.231316	63.38028169
	5	406	39	71	5.833333	11.64264	54.92957746
	6	421	30	71	14.555556	23.664789	42.25352113
Gaussian mixtures	1	200	7	71	10.309524	17.255089	9.85915493
	2	1096	52	71	3.125	5.347837	73.23943662
	3	44	8	71	11.854167	20.764051	11.26760563
	4	657	42	71	4.18254	8.917612	59.15492958
	5	720	38	71	14.491228	24.312476	53.52112676
	6	186	21	71	5.206349	8.618916	29.57746479
OPTICS	1	2612	71	71	4.910798	10.38304	100
	2	28	9	71	1.111111	1.527525	12.67605634
	3	28	16	71	3.385417	6.512008	22.53521127
	4	31	17	71	0.882353	1.371989	23.94366197
	5	27	11	71	3.393939	7.904736	15.49295775
	6	69	11	71	12.848485	21.589139	15.49295775
Spectral clustering	1	532	53	71	3.455975	5.858123	74.64788732
	2	657	43	71	3.244186	7.415937	60.56338028
	3	503	53	71	3.742138	8.963939	74.64788732
	4	228	11	71	9.30303	14.395917	15.49295775
	5	506	37	71	15.977477	26.092093	52.11267606
	6	477	55	71	3.821212	7.379538	77.46478873

Tabla de anexos 5 Resultados de la comparación de algoritmos para 7 grupos

Algoritmo	Grupo	N. datos	N. genes	N. de genes de referencia	MAE	RMSE	Porcentaje de identificación
K-Means	1	170	20	71	6.258333	11.148617	28.16901408
	2	941	53	71	3.292453	5.54541	74.64788732
	3	357	32	71	13.833333	24.068911	45.07042254
	4	666	52	71	5.080128	10.543712	73.23943662
	5	284	28	71	14.797619	24.277905	39.43661972
	6	274	51	71	4.212418	8.612007	71.83098592
	7	211	28	71	14.982143	22.730172	39.43661972
Fuzzy c means	1	393	39	71	5.790598	11.588198	54.92957746
	2	300	28	71	14.869048	24.284279	39.43661972
	3	448	51	71	3.526144	7.4173	71.83098592
	4	634	41	71	3.296748	5.416401	57.74647887
	5	211	29	71	14.189655	21.65999	40.84507042
	6	361	33	71	13.919192	24.1297	46.47887324
	7	556	60	71	4.105556	8.144528	84.50704225
Gaussian mixtures	1	96	11	71	7.984848	13.928932	15.49295775
	2	1096	52	71	3.112179	5.330428	73.23943662
	3	423	35	71	14.409524	24.453649	49.29577465
	4	665	42	71	4.198413	8.922061	59.15492958
	5	268	28	71	14.690476	24.071323	39.43661972
	6	176	23	71	6.15942	11.127717	32.3943662
	7	179	27	71	14.62963	22.282418	38.02816901
OPTICS	1	2612	71	71	4.910798	10.38304	100
	2	28	9	71	1.111111	1.527525	12.67605634
	3	28	16	71	3.385417	6.512008	22.53521127
	4	31	17	71	0.882353	1.371989	23.94366197
	5	27	11	71	3.393939	7.904736	15.49295775
	6	69	11	71	12.848485	21.589139	15.49295775
	7	30	15	71	15.4	22.24485	21.12676056
Spectral clustering	1	639	43	71	3.248062	7.419334	60.56338028
	2	493	51	71	3.447712	6.200095	71.83098592
	3	240	11	71	9.348485	14.432182	15.49295775
	4	504	37	71	15.977477	26.092093	52.11267606
	5	1	1	71	1.166667	1.870829	1.408450704
	6	491	53	71	4.333333	10.607343	74.64788732
	7	535	53	71	3.437107	5.848453	74.64788732

Anexo C producción científica asociada a la investigación

Este anexo presenta los artículos científicos publicados y derivados del trabajo de investigación descrito en esta tesis. Las publicaciones incluidas documentan etapas específicas del desarrollo del método RIVAM, desde el modelo conceptual del proceso de caracterización hasta la propuesta metodológica de reducción del espacio de búsqueda. Su inclusión tiene como propósito evidenciar que los aportes de esta investigación han sido sometidos a revisión por pares y aceptados en revistas indexadas, constituyendo una validación externa adicional a la presentada en los capítulos experimentales.

Conceptual modeling of the V gene annotation process in antibodies

Noemi Serret

Departamento de ciencias computacionales
Tecnológico Nacional de México/CENIDET
Cuernavaca, México
m22ce046@cenidet.tecnm.mx

Juan Miguel-Ruiz

Departamento de ciencias computacionales
Tecnológico Nacional de México/CENIDET
Cuernavaca, México
d20ce065@cenidet.tecnm.mx

Javier Ortiz-Hernandez

Departamento de ciencias computacionales
Tecnológico Nacional de México/CENIDET
Cuernavaca, México
javier.oh@cenidet.tecnm.mx

Jesús Martínez Barnetche

Centro de investigación sobre enfermedades infecciosas
Instituto Nacional de Salud Pública
Cuernavaca, México
jmbarnet@insp.mx

Yasmin Hernández

Departamento de ciencias computacionales
Tecnológico Nacional de México/CENIDET
Cuernavaca, México
yasmin.hp@cenidet.tecnm.mx

Abstract— The genes present in the V(D)J segments are components for the generation of antibodies. The function of antibodies is to detect foreign agents that are harmful to the organism. The location of the V, D and J genes is of great interest for the diagnosis of diseases and genomic disorders. The gene recognition or genome annotation process is of high complexity due to the integration of large volumes and diverse data sources. Genome annotation is the process of assigning information or labels to sections of the genome of a species. Advances have been made in the systematization of some steps of the annotation process, but activities that are performed manually by molecular biology specialists still prevail. Due to the diversity of rules that must be taken into account to determine the location of the V, D and J genes in the annotation process, a great amount of time and computational resources are consumed. In this work we propose the use of design science methodology for knowledge representation by means of conceptual modeling. The model represents the annotation process, the rules to determine its functionality and the parts that compose the V genes.

Keywords—component; *conceptual model*; *annotation process*; *V(D)J segments*; *V gene*; *design science*; *knowledge engineering*

I. INTRODUCTION

Genome annotation is the process of assigning information or labels to sections of the genome in a species [1]. Although there are some automatic annotation systems for gene segments as mentioned in [1][2], most of them focus on annotating functional genes and data presentation, without explaining the process that is carried out for annotation. The methodology followed for interaction and

elicitation of knowledge by genomics specialists and molecular biologists is generally not described. Throughout this article we will refer to them only as "specialists". Genome annotation is still performed manually. This is an iterative process that uses bioinformatics tools to identify and classify sections of the genome.

This activity is crucial for the study of immunology, as well as for the diagnosis and treatment of genomic diseases. The genes present in the V(D)J segments are the components for the generation of antibodies. The function of antibodies is to detect external agents that are harmful to the organism. The annotation of the V, D and J genes requires the integration of large volumes of data, different sources of information and the knowledge of specialists in molecular biology to determine the functionality of the genes [1]–[4].

Conceptual modeling makes it possible to represent the fundamental elements and relationships of a system or process. Its purpose is to present a general view of the domain, facilitating its understanding. The use of conceptual models has increased due to the need to represent complex problems in various domains [5].

This paper proposes the use of design science methodology for the development of conceptual modeling of the V gene annotation process. The domain was represented by means of conceptual models, one of them is the SADT model that was used for the representation of the genome annotation process performed by specialists in a general way. The second model developed was an activity

A Design Science Approach to Modeling the V Gene Annotation Process

Juan Miguel-Ruiz^{a,*}, Noemi Serret^{a,**}, Javier Ortiz-Hernandez^{a,***},
Jesus Martinez Barnetche^{b,****}, and Yasmín Hernández^{a,*****}

^a*Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), Cuernavaca, Mexico*

^b*Instituto Nacional de Salud Pública., Cuernavaca, Mexico*

*e-mail: d20ce065@cenidet.tecnm.mx

**e-mail: m22ce046@cenidet.tecnm.mx

***e-mail: javier.oh@cenidet.tecnm.mx

****e-mail: jmbarnet@insp.mx

*****e-mail: yasmín.hp@cenidet.tecnm.mx

Received May 25, 2024; revised August 20, 2024; accepted September 12, 2024

Abstract—Design science is a methodology that addresses complex and evolving problems requiring a multi-disciplinary 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: 10.1134/S0361768824700798

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 [1–3].

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 objectives of the resulting products [7]. This methodology addresses real problems using consolidated knowledge in a specific area and designs artifacts that solve needs and take advantage of existing opportunities [8, 9].

In the field of information systems (IS), design science is a relevant and crucial area of study. It examines the role of information technology (IT) artifacts and the professional justification for research in this domain. In the early 1990s, the IS community came to recognize the importance of design science for improving the effectiveness of IT artifacts in solving



www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(4), Sep-Dec 2025, 156-175. ISSN: 2007-1558.
<https://doi.org/10.61467/2007.1558.2025.v16i4.1028>

Method for the Unification and Reduction of the Search Space of V Gene Segments in Sequence Alignments

Juan Miguel-Ruiz¹, Javier Ortiz-Hernández¹, Jesús Martínez-Barnetche², Yasmin Hernández¹,
Juan Téllez-Sosa², Manuel Erazo-Valadez¹

¹ TecNM Centro Nacional de Investigación y Desarrollo Tecnológico, México

² Centro de Investigación Sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, México
{d20ce065, javier.oh, yasmin.hp, d20ce059}@cenidet.tecnm.mx, {jmbarnet, jmtellez}@insp.mx

Abstract. The identification and characterisation of V genes poses a significant challenge due to the substantial number of alignments generated by diverse sequencing systems. This study proposes a method for the unification and reduction of the search space, with the objective of optimising the identification of V genes. The method integrates preprocessing, normalisation, and clustering using Gaussian Mixture Models. This approach facilitates data consolidation and reduces redundancy, thereby enhancing the efficiency and accuracy of the subsequent analysis. The elbow method was employed to determine the optimal number of groups, achieving a 98% reduction in the search space. The findings were validated through the use of metrics such as mean absolute error, mean squared error, and root mean squared error, thereby confirming the effectiveness of the method in improving the precision of gene identification.

Keywords: Clustering, data integration, reduction of search space, sequence alignment, V gene.

Article Info

Received March 15, 2025

Accepted July 6, 2025

1 Introduction

The technological development for massive sequencing has made it easier to obtain genomes, however, the annotation of these genomes has become a great challenge due to the variety of sequenced genomes and the large amount of data obtained. The purpose of annotation is the localization of genes in a genome, as well as the determination of their structure and the proteins they produce through computational analysis and experimental approaches. Most annotation methods share common features, but in the end, the best approach depends on the time and resources available.

The first part of the annotation is the identification of the start and end positions of the genes to be analyzed. For gene identification, molecular biologists use various methods of sequence alignments to estimate the start and end positions of the targeted genes, or the sections that constitute these genes. In other words, most of the current methods cannot detect the entire structure of the genes simultaneously, which makes it necessary to analyze their sections separately, thus increasing the complexity of genes.

At Centro de Investigación Sobre Enfermedades Infecciosas (CISEI) of Instituto Nacional de Salud Pública (INSP), located in Cuernavaca, México, the Variable, Diversity and Union genes (V, D and J) are currently being annotated. These genes play a key role in the formation of B-cell antibodies, which are responsible for detecting and neutralizing harmful agents such as viruses and bacteria (Kindt et al., 2007; Pieper et al., 2013). Within this context, the focus is placed on V genes because of their high frequency in vertebrate genomes and their complex structure. The aim is to reduce the time and effort required by molecular biologists for their identification.

Research at CISEI also seeks to explore the structural diversity of DNA in different bat species. This is of fundamental epidemiological importance since bats serve as reservoirs of potentially zoonotic viruses. The annotation of V(D)J segments is expected to deepen the understanding of the immune system and contribute to the development of new strategies for addressing infectious diseases.



OPEN ACCESS

EDITED BY
Anna E. Savage,
University of Central Florida, United States

REVIEWED BY
Shamsa Batool,
Université de Montpellier, France
John Eimes,
Sungkyunkwan University, Republic of
Korea

*CORRESPONDENCE
Jesús Martínez-Barnetche
✉ jmbarnet@insp.mx
Constantino López-Macías
✉ constantino.lopez135@gmail.com
Stephanie Saint Remy-Hernández
✉ stephstremy2691@gmail.com

RECEIVED 31 October 2025
REVISED 23 December 2025
ACCEPTED 04 February 2026
PUBLISHED 23 February 2026

CITATION
Saint Remy-Hernández S,
Pacheco-Olvera DL, Godoy-Lozano EE,
Miguel-Ruiz J, Téllez-Sosa J,
Valdovinos-Torres H, Rivas N,
López-Macías C and
Martínez-Barnetche J (2026)
An update on the characterization
of immunoglobulin loci in
Ambystoma mexicanum.
Front. Immunol. 17:1736245.
doi: 10.3389/fimmu.2026.1736245

COPYRIGHT
© 2026 Saint Remy-Hernández,
Pacheco-Olvera, Godoy-Lozano,
Miguel-Ruiz, Téllez-Sosa,
Valdovinos-Torres, Rivas, López-Macías
and Martínez-Barnetche. This is an open-
access article distributed under the terms
of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication
in this journal is cited, in accordance
with accepted academic practice. No
use, distribution or reproduction is
permitted which does not comply with
these terms.

An update on the characterization of immunoglobulin loci in *Ambystoma mexicanum*

Stephanie Saint Remy-Hernández^{1,2*},
Diana Laura Pacheco-Olvera^{2,3},
Elizabeth Ernestina Godoy-Lozano⁴,
Juan Miguel-Ruiz⁵, Juan Téllez-Sosa⁴,
Humberto Valdovinos-Torres⁴, Nancy Rivas¹,
Constantino López-Macías^{2*} and Jesús Martínez-Barnetche^{4*}

¹Departamento de Parasitología, Posgrado en Ciencias Químico-biológicas, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, México City, Mexico, ²Unidad de Investigación Médica en Inmunología, UMAE Hospital de Especialidades, Centro Médico Nacional Siglo XXI, Instituto Mexicano del Seguro Social, México City, Mexico, ³Departamento de Bioquímica, Posgrado en Ciencias Químico-biológicas, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, México City, Mexico, ⁴Centro de Investigación sobre Enfermedades Infecciosas, Instituto Nacional de Salud Pública, Cuernavaca, Morelos, Mexico, ⁵Departamento de Ciencias Computacionales, Tecnológico Nacional de México/CENIDET, Cuernavaca, Morelos, Mexico

Background: We previously reported the genomic characterization of immunoglobulin loci in *Ambystoma mexicanum* of the laboratory d/d white strain, where the IGH locus gene orientation was incompatible with VDJ recombination, suggesting scaffold orientation errors. A novel 29.1 Gbp *A. mexicanum* genome derived from an F1 cross between *A. mexicanum* and *A. tigrinum* (UKY_AmexF1_1) has recently been released, containing only 220 unmapped scaffolds. Here, we present an updated description of the immunoglobulin loci based on this improved genome assembly.

Methods: Using our previously annotated AmbMex60DD sequences, we performed an alignment-based annotation followed by manual curation on the UKY_AmexF1_1 assembly. Gene models were further refined using RNA-Seq datasets from axolotl spleen, liver, lung, heart, and gill tissues.

Results: The IGH locus was mapped at 350–364 Mbp on chromosome 13q. Synteny is preserved between both genome versions, but in UKY_AmexF1_1, all IGH genes share the same transcriptional orientation. A major difference was observed in the lambda locus, which in the UKY_AmexF1_1 genome contains 13 IGLC–IGLJ clusters comprising 60 IGLJ genes, compared with only three clusters in the AmbMex60DD genome. No kappa locus was detected in either assembly.

Conclusion: This study confirms our previous findings and provides an example of intraspecies structural variation in adaptive immune receptor loci. It underscores the importance of well-assembled genomes and establishes the current *A. mexicanum* reference as a valuable resource for investigating immune evolution and function in axolotls and other vertebrates.

KEYWORDS

Ambystoma, comparative genomics, copy number variation, Ig loci, immunoglobulin genes