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

www.editada.org

Method for the Unification and Reduction of the Search Space of V Gene Segments in Sequence Alignments

Juan Miguel-Ruiz[1], *Javier Ortiz-Hernández*[1], *Jesús Martínez-Barnetche*[2], *Yasmín Hernández*[1],
Juan Téllez-Sosa[2], *Manuel Erazo-Valadez*[1]
[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.

This paper proposes a novel method for effectively reducing the search space of the start and end positions of V gene segments in sequence alignments. The proposed method is evaluated by comparing it with the manual references generated by CISEI experts, using metrics to measure the error and its efficiency. This method not only seeks to improve accuracy in the identification stage, but also sets a benchmark for more efficient annotation, contributing to the knowledge and study of vertebrate genomes.

1.1 Structure of the V gene

Each of the V gene sections has the following characteristics: (a) signal peptide or *SP* is a short amino acid sequence that indicates where the V gene begins and has a length between 40 to 46 characters; (b) *Exon* is the section that contains the essential information of the V gene and has a length between 300 to 350 characters; and (c) the recombination signal or *RSS* which is responsible for the binding between the V and D gene, the length of the *RSS* and has 38 to 40 characters (Lefranc & Lefranc, 2001). Fig. 1 shows the structure of the V gene.

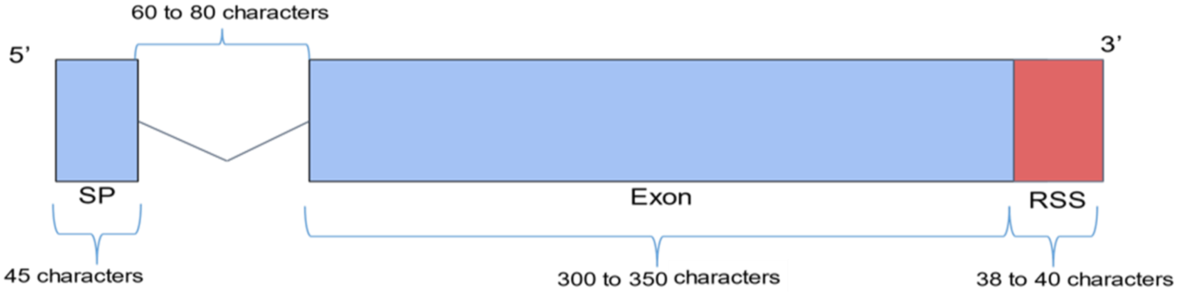


Fig. 1. V gene structure obtained from (Lefranc & Lefranc, 2001).

The identification of the V gene consists of determining the start and end positions of each of its sections. Correct identification improves the annotation and thus the representation of gene functionality (Bergman, 2007; Megrian, 2014). The use of sequences manually validated by experts to find new genes is called homology and is used for V gene identification (Mount, 2004; Olivieri & Gambón-Deza, 2019). These validated sequences are used in alignment systems to obtain data that are used in the identification of V genes (Amin et al., 2018; Ejigu & Jung, 2020; Mount, 2004). The following studies address genome annotation more extensively (Ejigu & Jung, 2020; Miguel-Ruiz et al., 2024; Serret et al., 2023).

Some of the most used alignment systems are *Blast*, *Hmmer* and *Exonerate* (Altschul et al., 1990; Eddy, 1998; Slater & Birney, 2005). Although they have the disadvantage of performing sequence alignment estimates of only some of the V gene sections (EMBL-EBI, 2022). They also have low precision and generate many approximations of the start and end positions of the V gene sections (Mount, 2004).

Currently, CISEI experts perform V gene with the *Integrative Genomics Viewer* (*IGV*) system (Robinson et al., 2023). Fig. 2 shows an identified and annotated V gene. Below the expert's annotation are four alignments that are used to identify the gene. The identification and annotation take 15-20 minutes per gene and up to one month to complete a genome of approximately 1600 genes (V, D and J).

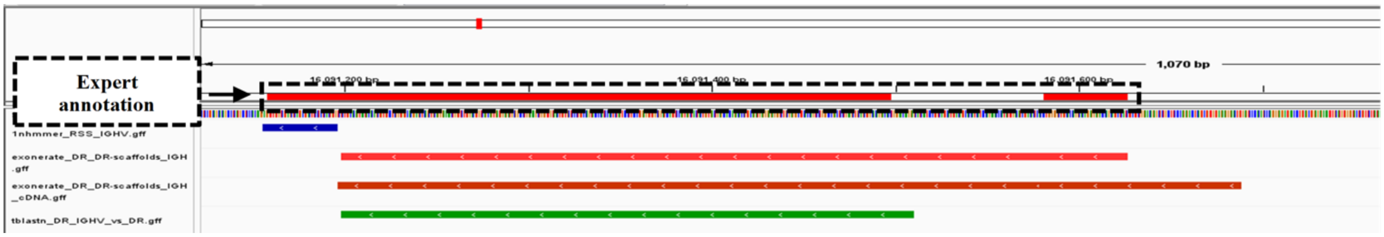

Expert annotation
1,070 bp
16,091,200 bp
16,091,400 bp
16,091,600 bp
1nhmmer_RSS_IGHV.gff
exonerate_DR_DR-scaffolds_IGH.gff
exonerate_DR_DR-scaffolds_IGH_cDNA.gff
tblastn_DR_IGHV_vs_DR.gff

Fig. 2. V gene identification and alignments with the *IGV* system.

In Fig. 2, the alignments appear collapsed, which facilitates the general location of a gene but makes precise identification of its sections difficult. Each alignment groups between 50 and 200 sequences, which improves detection, but complicates the identification of the V gene. For a more detailed analysis, Fig. 3 shows the alignments in their expanded version, allowing CISEI experts to manually determine the start and end positions. Because of this, experts must apply their knowledge to select the best solution and precisely define the boundaries of each gene section.

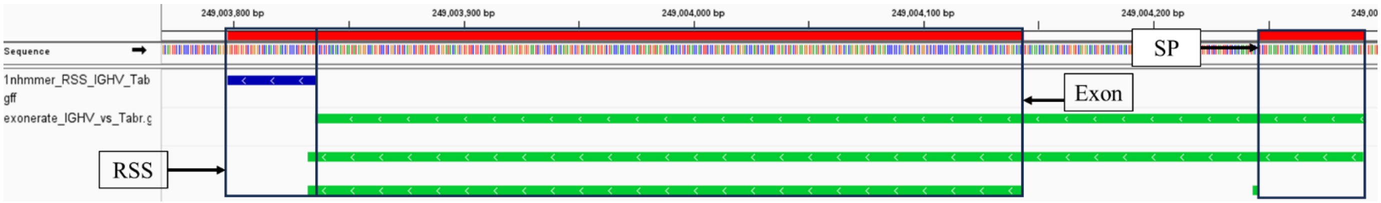


Fig. 3. Expanded alignments.

Each alignment system generates files with key information for V gene identification. *Hmmerscan* identifies the recombination signal (RSS); *Exonerate:protein2genome* detects the *Exon* and Signal Peptide (*SP*); *Exonerate:est2genome*, *Tblastx* and *Tblastn* also identify these sections, operating at different levels of the genome. Together, these systems allow CISEI experts to locate V genes, but the wide search space generates a large effort and time for gene identification. In this research, we propose a method that reduces the search space generated by sequence alignments, improving the accuracy and efficiency of V gene identification.

1.2 Literature review

In the literature review based on the *PRISMA* methodology (Urrutia & Bonfill, 2010), most relevant works for gene identification and annotation were identified. The research question posed was: How has the systematization of V gene characterization been approached and what techniques are used? The purpose of this review was to identify systems, architectures, methods, techniques and algorithms for the systematization of gene characterization. The following search string was used: ("gene annotation" OR "gene identification") AND ("machine learning" OR "computational modeling") AND ("V(D)J" OR "antibodies" OR "V genes") and "vertebrates".

The review began on May 7, 2024, and was completed on November 30, 2024. The review started on May 7, 2024, and was completed on November 30, 2024. Articles indexed in journals and presented at conferences were searched in the databases PubMed, MDPI, NCBI, SpringerLink, and ScienceDirect, in combination with the Google Scholar search engine. Only articles written in English and published between 2019 and 2024 were considered. The search retrieved a total of 160 articles, of which 9 were identified as relevant to this study (see Table 1).

Table 1. Literature review

Brief description	Method, technique or algorithm	Reference
GeMoMa homology-based web system using *GFF* and *Fasta* files	Based on homology with *GFF* files	(Keilwagen et al., 2019)
VgeneFinder system based on the homology principle with reference sequences	It is based on the use of motifs and genome sequence	(Olivieri & Gambón-Deza, 2019)
DeepGSR deep learning-based system for signal and genomic pattern detection	It is based on a convolutional neural network	(Kalkatawi et al., 2019)
Helixer is a system based on deep learning	Based on BLSTM to annotate sequences with Keras	(Stiehler et al., 2020)
GOODORFS is a system based on the ab initio method.	Based on Kmeans	(McNair et al., 2021)
RAPID is a web-based system that analyzes antibodies and annotates clonotypes	Not specified	(Y. Zhang et al., 2021)
TOGA is a sequence alignment pipeline	It is based on homology and machine learning	(Zhang et al., 2021)
IGDetective is a homology-based system	The system is based on graph theory and homology	(Sirupurapu et al., 2022)
Mgcod identifies the zones of a gene	Based on homology	(Pfennig et al., 2022)
GeMoMa homology-based web system using *GFF* and *Fasta* files	Based on homology with *GFF* files	(Keilwagen et al., 2019)

The literature review evidenced the importance of systematization and the use of artificial intelligence and optimization techniques for V gene identification, highlighting works such as Olivieri & Gambón-Deza (2019) and Sirupurapu et al. (2022a). Although there are few specific studies, most genomic annotation systems rely on prior identification of genes, creating a large area of opportunity. The most widely used and proven effective techniques for gene annotation are based on homology methods and *ab* *initio* approaches in conjunction with LSTM models, convolutional neural networks, and deep learning (Kalkatawi, Magana-Mora, et al., 2019). Other techniques still rely on sequence alignments with reference data.

2 Method for reducing the search space in V-gene sequence alignments

Several proposals to systematize the characterization of V genes have been reported in the scientific literature, including the studies of Olivieri and Gambón-Deza (2019) and Sirupurapu et al. (2022). Nevertheless, important challenges remain, particularly regarding the low accuracy in predicting the precise location of genes. To address this, a distance-based method is proposed to reduce the search space in V gene alignments. This approach relies on the structural organization of V genes and the defined distances between their three fundamental sections to generate a new dataset used for search space reduction. The method is structured into three main stages: preprocessing, processing, and postprocessing, as summarized in Table 2. An overview of the method is illustrated in Figure 4.

Table 2. Structure of method with its steps and activities

Step	Activity	Description
Preprocessing	Data unification	Different alignment sources (e.g., *Hmmer, Exonerate, Tblastn*) are integrated into a single representative dataset. This ensures that each record includes all sections of the V gene
	Data normalization and splitting	Data are adjusted using techniques such as Normalizer to improve clustering performance and then split into training and testing sets
Processing	Training and use of a clustering model	A model based on the Gaussian Mixture algorithm is implemented
	Group unification	Records within each cluster are consolidated into a single record per gene, reducing the search space and enabling fairer measurement of results
Postprocessing	Error measurement and best-group selection	Results are compared against expert reference using metrics such as MAE, MSE and RMSE to select the best-performing group.
	Transformation to *GFF* format	The best group is transformed back into *GFF* format for compatibility with genomic analysis tools (e.g., *IGV*)

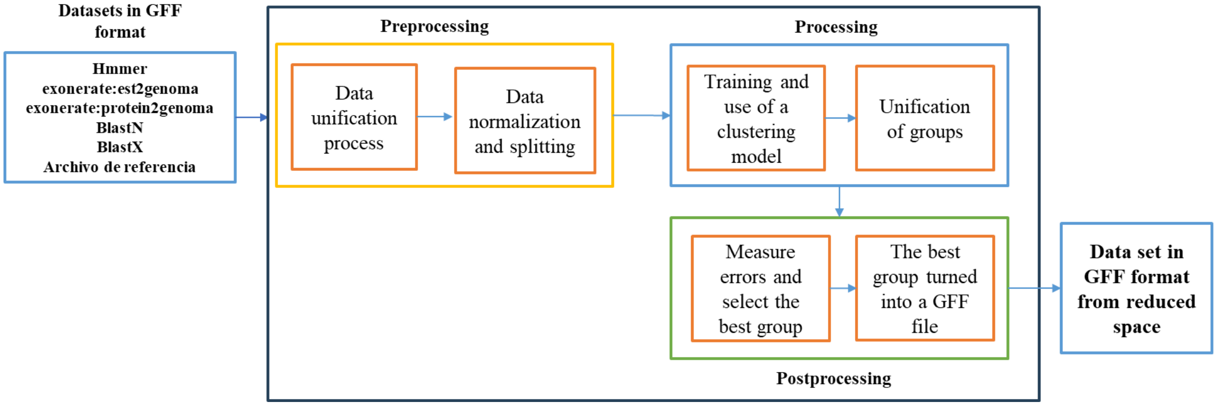

Datasets in GFF format
Hmmer
exonerate:est2genoma
exonerate:protein2genoma
BlastN
BlastX
Archivo de referencia
Preprocessing
Data unification process
Data normalization and splitting
Processing
Training and use of a clustering model
Unification of groups
Measure errors and select the best group
The best group turned into a GFF file
Postprocessing
Data set in GFF format from reduced space

Fig. 4. Method for reducing the search space of V gene(s).

The proposed method allows a significant reduction of the search space for V gene identification, achieving efficient performance on both large and small genomes, with processing times of less than 60 minutes. The following sections of the paper address key aspects such as (a) data sets in *GFF* format, (b) activities associated with preprocessing, (c) activities associated with data processing, and (d) results of the postprocessing stage.

2.1 Dataset in *GFF* format

GFF files are widely used in biological and bioinformatics projects due to their ability to organize and make sense of fragments of a genome. This plain text format consists of nine tab-delimited columns, allowing for a structured and detailed representation of genomic features (EMBL-EBI, 2022; García Simón, 2018). *GFF* files are comparable to datasets used in data science, because both contain attributes associated with specific features.

The data used in this study come from different systems specialized in sequence alignments, which identify areas of interest in a genome and in specific V genes. The systems used by the CISEI experts are *Blast*, *Hmmer* and *Exonerate*. Each of these systems employs a different principle for the execution of the alignments (EMBL-EBI, 2022) and is only interested in one or several sections of the gene. An example of this is *Exonerate:protein2genome* and *Exonerate:est2genome*, which identify the *Exon* and signal peptide. Each works at different levels of the genome and complement each other. If protein2genome does not identify the signal peptide, est2genome can complete the identification. Despite their effectiveness, these systems have important limitations, such as their low accuracy in determining the location of the complete structure of a gene (Mount, 2004). For this reason, in this work we propose to combine multiple results from these different systems.

The *Blast*, *Hmmer* and *Exonerate* systems generate key data in the *GFF* format, which represent sequence alignments representing possible gene locations. This format provides a structured framework for analyzing genetic features, facilitating their integration and analysis. An excerpt of a *GFF* file used in this research is presented in Table 3. The *GFF* files contain nine attributes that allow for the identification of key information, as presented in Table 4. For this study, the Start, End, and Strand columns are particularly important because they describe the structure of the V gene.

Table 3. Dataset in *GFF* format

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

Table 4. Nine attributes of *GFF* files

Field	Description
Seqname	Unique identifier of the sequence, such as the chromosome or scaffold where the feature is located
Source	Tool or method that generated the alignment, e.g., *Hmmer*, *Exonerate* or *Blast*
Feature	Types of aligned features, such as *Gene*, *Exon*, *CDS* or *mRNA*
Start	Initial position of the feature in the sequence
End	Final position of the feature in the sequence
Score	Confidence value associated with the annotation. It can be a number or a dot (.) if not available
Strand	Direction of the feature in the DNA, indicated by + or –
Frame	Specifies how the codons are to be read. Possible values are 0, 1 or 2
Attributes	Additional information about the features, such as unique identifiers, gene names or notes

In the context of genome characterization, bats (chiropters) have been widely studied due to their genetic diversity and unique adaptations. Specifically, at CISEI, we have worked with the species *Tadarida brasiliensis*, a bat native to the western and southern United States, Mexico, Central America and other regions (Webster et al., 2024). This species is notable for its extensive genome, containing more than 800 V genes, while one of its GFF files alone exceeds 167,165 records, representing an intermediate number for data used in genomic characterization. The specimen used in this study has as reference chromosome CM061257.1, which allows a detailed analysis of its genetic organization and its reference in the National Center for Biotechnology Information (NCBI).

When performing a detailed analysis of the GFF file characteristics, it was decided to use three key columns: Start, End and Strand. These columns concentrate the most relevant information for describing genomic structure and features, as seen in Table 5. Although the start and end positions have a scale that can exceed millions, this granularity is essential for identifying each of the V gene sections. A major challenge in working with these files is that each file identifies only one section of the gene, rendering any attempt at independent analysis ineffective. For this reason, it is essential to unify the dispersed information.

Moreover, the study of *Tadarida brasiliensis* provides an opportunity to explore not only the complexity of its immunogenetic repertoire but also the methodological challenges inherent in large-scale genomic projects. The high number of V genes, coupled with the fragmented representation in GFF files, highlights the necessity of implementing structured data reduction and integration strategies. Such approaches allow researchers to move beyond raw sequence data, enabling the extraction of biologically meaningful insights and supporting comparative studies with other vertebrate genomes.

Table 5. Basic statistics of the exonerate:est2genome set

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

2.2 Data preprocessing stage

Analysis of the data in *GFF* format shows that its sparse structure makes it difficult to accurately represent the V gene, limiting the application of machine learning algorithms. Each record captures only a part of the information, preventing a complete identification of the gene. To solve this problem, a preprocessing for unification has been developed that integrates the fragmented data into new records, improving their representation. This approach is based on the relative distances between V gene sections, allowing them to temporarily omit the character sequence and focus on the structural information in the *GFF* files.

Data unification. In this research, it is essential to implement a new preprocessing technique that focuses on the unification and integration of data from different sources into complete and accurate records. This unification enables the representation of the V gene's structural composition, incorporating the relative distances between its three sections (*SP*, *Exon*, and *RSS*). Consequently, the dependence on absolute positions is mitigated, enhancing the flexibility and accuracy of the analysis, even when tools vary in their start and end values.

This approach optimizes time and resources in V gene identification by consolidating scattered data, making annotation more efficient. Systematic integration of these data not only improves V gene identification but also creates a reusable framework for consolidating data from other genomes. The subsequent section delineates data unification with a five-stage flowchart, as shown in Table 6.

Table 6. Five stages of data unification

Stage		Description
1	*GFF* data import	Recovery and adaptation of *GFF* files, transforming them into a tabular format that facilitates subsequent manipulation and analysis
2	Identification of file with least redundancy	Comparison of all available files to determine which has the fewest redundant records; this file is chosen as the starting point because it offers the most specific and reliable information for constructing V-gene records
3	Data extraction and preparation	Using the positions from the least-redundant file, missing data are extracted from the other files to improve the representation of the V gene
4	Construction and validation of V gene records	Assembly of V-gene records according to biological constraints, ensuring that generated records do not exceed biologically consistent distances between feature points
5	Adjacency matrix construction and vectorization	Generation of an adjacent matrix for each gene to represent relational data; new distance-based features are derived from this matrix and then vectorized to facilitate their use by machine-learning algorithms

The five stages are articulated to ensure that the processed data are of high quality and ready for analysis at advanced stages. The combination of these activities not only improves the accuracy of V gene identification, but also simplifies the handling of complex data, promoting a more efficient and reliable workflow. Fig. 5 presents the general structure of data unification and its five stages.

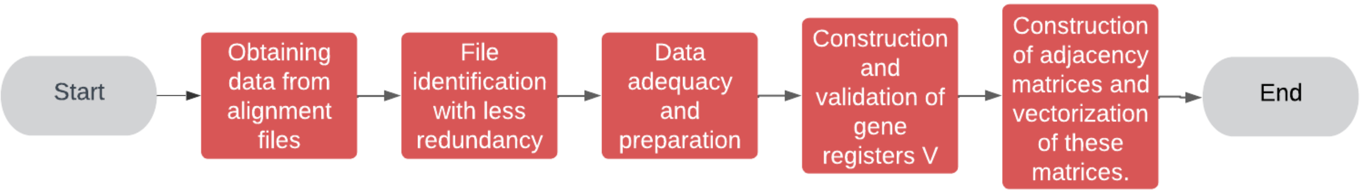

Start
Obtaining data from alignment files
File identification with less redundancy
Data adequacy and preparation
Construction and validation of gene registers V
Construction of adjacency matrices and vectorization of these matrices.
End

Fig. 5. Data unification process.

The first stage consists of importing *GFF* data: this stage involves retrieving and adapting the *GFF* files, transforming them into a tabular format that facilitates their manipulation and subsequent analysis. The second stage, shown in Fig. 6, consists of identifying the file with the lowest redundancy, ensuring that the selected set contains only unique information. To do this, the redundancy of each *GFF* set is calculated, verifying whether at least one of the Start or End columns contains unique values and whether these represent the totality of the records within the analyzed file. Once the appropriate option is identified, the set with the lowest redundancy is assigned to the *DUMrev* or *DUMforw* (Data Unique Matrix) variable, depending on the direction of the alignment. At this stage, it is considered whether the alignment is Reverse or Forward. Finally, the *DUMrev* and *DUMforw* matrices are returned for use in the following stages of the unification.

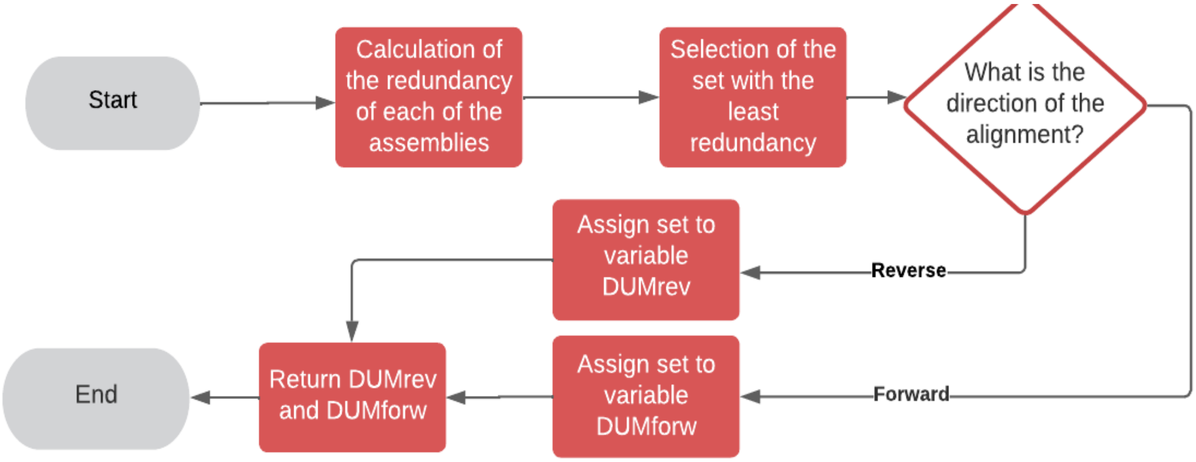

Start
Calculation of the redundancy of each of the assemblies
Selection of the set with the least redundancy
What is the direction of the alignment?
Assign set to variable DUMrev
Reverse
Return DUMrev and DUMforw
Assign set to variable DUMforw
Forward
End

Fig. 6. Second stage, identification of the file with less redundancy.

The third stage, shown in Fig. 7, consists of data matching and preparation, ensuring that the records are structured correctly. Therefore, the starting positions of the *DUMrev* and *DUMforw* matrices are extracted in pairs and used as ranges. With these ranges, data is retrieved from the other matching *GFF* files within the intervals set by these ranges. Subsequently, the extracted records are assigned to lists according to their direction: *LISTMrev* or LISTMforw. These lists store the records in a structured manner for further analysis. The process continues iteratively until there is no more outstanding data in *DUMrev* or *DUMforw*.

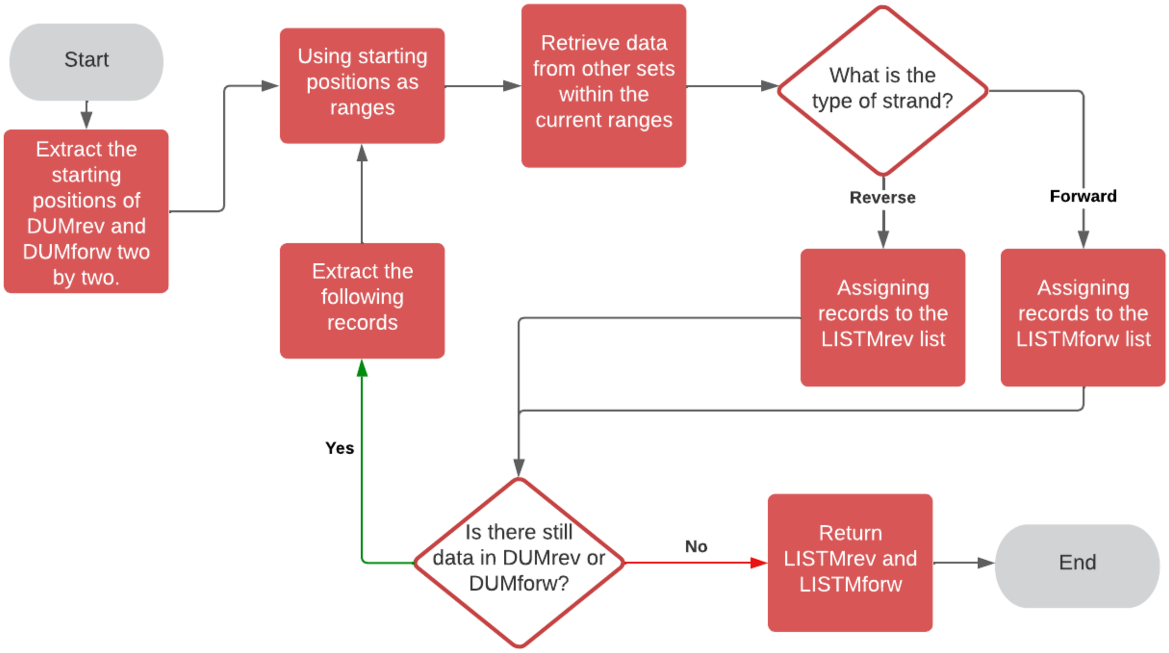

Start
Extract the starting positions of DUMrev and DUMforw two by two.
Using starting positions as ranges
Retrieve data from other sets within the current ranges
What is the type of strand?
Reverse
Forward
Assigning records to the LISTMrev list
Assigning records to the LISTMforw list
Extract the following records
Yes
Is there still data in DUMrev or DUMforw?
No
Return LISTMrev and LISTMforw
End

Fig. 7. Third stage, data adequacy and data preparation.

The fourth stage, shown in Fig. 8, consists of the construction and validation of V-gene records. The objective is to ensure that the generated records comply with biologically consistent distances between sections, consolidating the information from the previous subprocesses. The subprocess starts by extracting a record from *DUMrev* or *DUMforw* and assigning it to the *RSS* variable. The next step is to process the data from *LISTMrev* or *LISTMforw*, depending on the *DUM* from which it was extracted for the RSS. The following mappings follow the structure of the first *LISTM* record (depending on the strand) mapped to *Exon*, while the next one is associated with SP.

Record construction is based on a generic V-gene structure, ensuring that the three sections are correctly represented and that each record is biologically valid. To maintain this coherence, four fundamental rules are established to delimit the gene structure, which are presented in Table 7.

Table 7. Rules for data unification

Rule		Description
1	Distance between *RSS* and *Exon*	The absolute difference between the start position of the *RSS* and the start of the *Exon* must be ≥ 0 and ≤ 15 characters. Ensure sections are close without overlaps or excessive gaps; otherwise, the next record is extracted as an *Exon* and reevaluated
2	Distance within the *Exon*	The absolute difference between the *Exon's* start and end must lie between 280 and 350 characters. Absolute values are used because reverse alignments increase and forward alignments decrease; if not met, the next record is analyzed and the entire process restarts
3	Distance within the *SP*	The absolute difference between the SP's start and end must be > 45 but ≤ 50 characters. If this condition fails, the record is extracted as *SP* and the full evaluation process restarts
4	Distance between *SP* and *Exon*	The absolute distance between the *SP's* start and the *Exon's* end must be between 60 and 150 characters. This ensures biologically consistent separation; if the criterion is not met, another record is treated as a potential *SP* and the whole process restarts

If all validation criteria are met, the *RSS*, *SP* and *Exon* records are stored in the *Mat_Gen* array according to the direction of alignment and these arrays are stored in *ListMat_Genrev* or *ListMat_Genforw* depending on the direction. Once there is no more data left in *DUMrev* or *DUMforw*, the lists are returned, containing the final records ready for analysis in the next phase of the unification method. This subprocess is critical for structuring the data in a consistent manner, ensuring that the V gene sections are represented with biological accuracy.

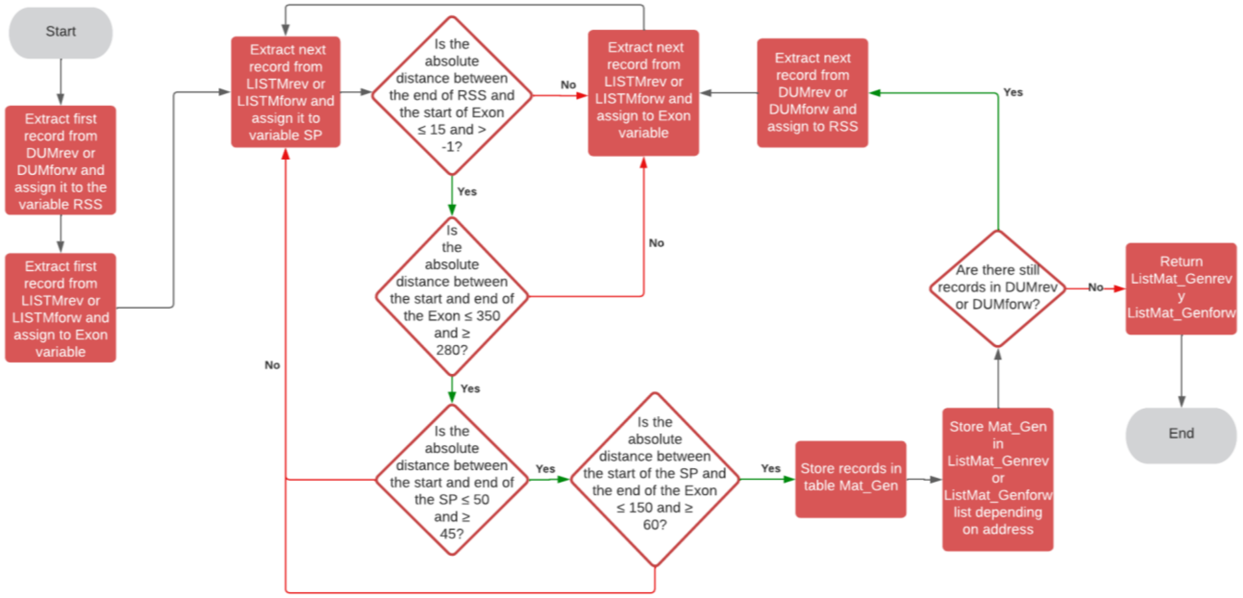

Start
Extract first record from DUMrev or DUMforw and assign it to the variable RSS
Extract first record from LISTMrev or LISTMforw and assign to Exon variable
Extract next record from LISTMrev or LISTMforw and assign it to variable SP
Is the absolute distance between the end of RSS and the start of Exon ≤ 15 and > -1?
No
Yes
Extract next record from LISTMrev or LISTMforw and assign to Exon variable
Extract next record from DUMrev or DUMforw and assign to RSS
Yes
Is the absolute distance between the start and end of the Exon ≤ 350 and ≥ 280?
No
Yes
No
Is the absolute distance between the start and end of the SP ≤ 50 and ≥ 45?
Yes
Is the absolute distance between the start of the SP and the end of the Exon ≤ 150 and ≥ 60?
Yes
Store records in table Mat_Gen
Store Mat_Gen in ListMat_Genrev or ListMat_Genforw list depending on address
Are there still records in DUMrev or DUMforw?
No
Return ListMat_Genrev y ListMat_Genforw
End

Fig. 8. Fourth stage, construction and validation of V gene registers.

The fifth stage, presented in Fig. 9, consists of the construction of adjacent matrices and their vectorization, a critical step in transforming the V gene records into a matrix and vector representation. This process optimizes downstream analysis by ensuring an accurate relationship between gene sections. The procedure begins with the combination of *ListMat_Genrev* and *ListMat_Genforw*, generating a new list called *ListMat*, which contains all the unified records. From *ListMat*, the first record is extracted and stored in *MatrizGen*, establishing the basis for the construction of relationships between the data.

The adjacency matrix (*AdMatrix*) is constructed to represent all possible distances between the data of the current record and specifically its start and end fields. To avoid redundancies, all values below the diagonal are set to zero, ensuring a coherent structure. The records are then vectorized for storage. The start and end positions are stored in the Gen array, while the adjacency matrix is vectorized and stored in *Distans*. These two lists are combined to form the *Totalgen* array, facilitating their manipulation and structuring into an ordered data set. Subsequently, the direction of the records is determined from *MatrixGen*, classifying them as Forward or Reverse. Depending on their direction, a new record is added to *Totalgen*, assigning values 0 or 1 accordingly. All records generated for *Totalgen* are stored in *MatFinal*, which will be the new dataset, since it will contain the gene structure, its direction and distances. The subprocess is repeated until there are no more records in *ListMat*.

Data unification is essential to generate complete and structured records that comprehensively represent a V gene. Each record must cover the entire gene structure, avoiding partial fragmentations. The construction of records is based on the formation of adjacent matrices and their vectorization, processes that will be detailed in the construction of adjacent matrices and their vectorization, because they depend on the application of specific distances and methods of data structuring. The data unification method is highly adaptable to any genome containing V genes and using sequence alignments for their identification. Its versatility allows the method to be reused in different genomic studies, ensuring accuracy, efficiency and applicability in genetic characterization.

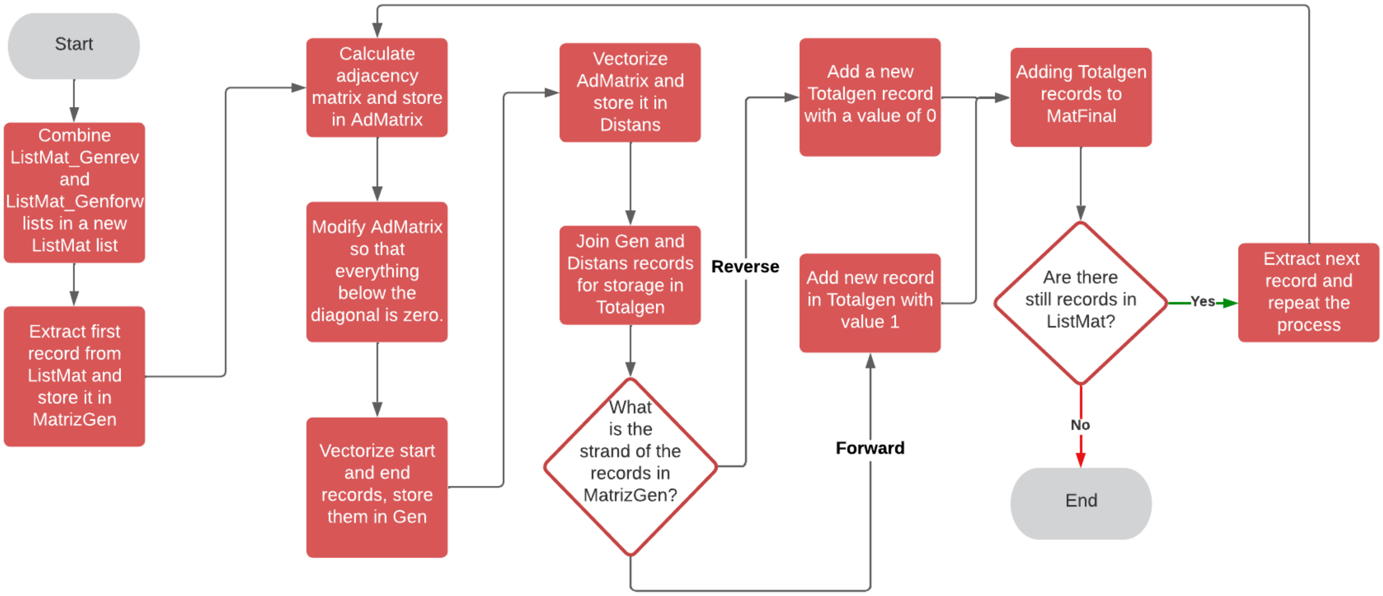

Start
Combine ListMat_Genrev and ListMat_Genforw lists in a new ListMat list
Extract first record from ListMat and store it in MatrizGen
Calculate adjacency matrix and store in AdMatrix
Modify AdMatrix so that everything below the diagonal is zero.
Vectorize start and end records, store them in Gen
Vectorize AdMatrix and store it in Distans
Join Gen and Distans records for storage in Totalgen
What is the strand of the records in MatrizGen?
Reverse
Forward
Add a new Totalgen record with a value of 0
Add new record in Totalgen with value 1
Adding Totalgen records to MatFinal
Are there still records in ListMat?
Yes
No
Extract next record and repeat the process
End

Fig. 9. Fifth stage, construction of adjacent matrices and their vectorization.

Construction of adjacent matrices and their vectorization. One of the fundamental steps of the data unification method is the formation of functional records from structurally valid arrays. The section will focus on the construction of these records using an adjacency matrix, with the goal of ensuring that each record within the array represents validated distances between V gene sections. Array formation depends on the records generated by sequence alignments, organized in tables that group three records to represent the sections of a gene. The example presented is based on data from the individual Tadarida brasiliensis, which has six sequence alignment files: (a) *Hmmerscan* for *RSS*, (b) *Exonerate:protein2genome*, (c) *Exonerate:est2genome*, (d) *Tblastx*, (e) *Tblastn* and (f) *Hmmerscan* for *SP* and *Exon*.

As shown in Table 8, there are examples of records that can be found in *ListMat_Genforw* or *ListMat_Genrev*, which are stored in *MatrixGen*. Each group of three records represents a V gene, along with its respective sections and associated strands. Although the Feature column specifies the section of the gene that has been identified, it cannot always be used as an absolute criterion when forming the records. In this example, the first record corresponds to the RSS, the second to the Exon, and the third to the SP. The same sequence is repeated in the fourth, fifth, and sixth records, thus representing a second V gene.

Table 8. Example of records for a V gene with all its sections

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
CM040297.1	*hmmerscan*	*RSS*	3712219	3712180	NA	-	NA	NA
CM040297.1	*exonerate:est2genome*	*Exon*	3712181	3711873	NA	-	NA	NA
CM040297.1	*exonerate:est2genome*	*gene*	3711792	3711746	NA	+	NA	NA
CM040297.1	*hmmerscan*	*RSS*	2168368	2168403	NA	-	NA	NA
CM040297.1	*exonerate:est2genome*	*Exon*	2168406	2168717	NA	-	NA	NA
CM040297.1	*exonerate:est2genome*	*gene*	2168796	2168837	NA	+	NA	NA

The first step in record construction is the assignment of data to the Gen variable, using the information in Table 8. In this process, only the absolute positions in the genome are extracted and vectorized. The Gen variable defines the positions to be used for the adjacency matrix. Each of the records represents one of the sections of the V gene, the first 3 records in the Start column represent the start of the RSS, the Exon and the SP while the last three are the end positions of these, the record is presented in Table 9. These two rows represent the two possible genes obtained from the records in Table 8.

Table 9. Examples of V genes organized into Their six sections

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

The calculation of the distance is based on the progressive subtraction of each position of a gene with respect to all other positions, repeating this procedure for each position. As shown in Equation 1, the subtraction of the first position of the gene is presented, which corresponds to the start of the RSS with a position of 3712219. The result of all these subtractions corresponds to the first row of the adjacency matrix, which represents the absolute distances between the V gene sections. This process is repeated for each of the positions representing the gene and is performed for all other positions representing the V gene.

$$\begin{bmatrix} 3712219 - 3712219 & 3711873 - 3712219 \\ 3712180 - 3712219 & 3711792 - 3712219 \\ 3712181 - 3712219 & 3711746 - 3712219 \end{bmatrix} = \{0 \quad 39 \quad 38 \quad 346 \quad 427 \quad 473\} \tag{1}$$

Table 10 presents the new adjacency matrix obtained from the process of Fig. 10, in which there are absolute values between the different distances. It is noteworthy that, at first glance, this matrix assumes a mirror-like configuration. Given the occurrence of the mirror in all values below the diagonal, it is determined that all values below the diagonal should be converted to zero. The new matrix could be used in its current form; however, it would not be entirely suitable for a machine learning algorithm. Therefore, it must be vectorized to generate a single record. This process is repeated for each record.

Table 10. Example of adjacency matrix

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

With the adjacency matrix generated, the values of the distances are vectorized and added to the *Totalgen*. Subsequently, the direction of gene V is verified using the first record of the *MatrixGen* variable, which allows this information to be incorporated into the vector. Finally, the *Totalgen* variable is integrated into the *MatFinal* Table, where all the V gene records are stored.

The new dataset is composed of 43 features of which the first seven are the positions of the alignments in the genome. These first features are maintained by preserving the relationship between the records and their location in the genome. The first seven attributes correspond to 1) *Strand*, 2) *Start_RSS*, 3) *Start_Exon*, 4) *Start_SP*, 5) *End_RSS*, 6) *End_Exon* and 7) *End_SP*.

Table 11 presents the 36 attributes representing the relationships between each of the V gene sections and the distances between them. As previously mentioned, all values below the diagonal of the adjacency matrix are converted to zeros. With the integration of these data, the final set is obtained, ready to be used in some machine learning algorithm.

Table 11. Attributes resulting from data unification

Attributes 8 to 13	Attributes 14 to 19	Attributes 20 to 25	Attributes 26 to 31	Attributes 32 to 37	Attributes 38 to 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

It is essential to understand both the structure of the original data and the structure of the new set, because they follow an unsupervised learning approach. To avoid the loss of relevant information, traceability of records must be preserved during data unification. Furthermore, the selection of suitable algorithms is crucial to processing the data in an optimal way and to ensure a correct interpretation of the spatial relationships between the V-gene sections. Therefore, the next activity consists of performing column filtering and data normalization.

Data normalization and data splitting. The evaluation of the effectiveness of data unification was based on its use in the *GFF* datasets of the species *Tadarida brasiliensis*. The analyzed set consists of 1,006,797 records, distributed among its six *GFF* files, representing the magnitude of data that experts must work with daily in genome characterization and V gene identification.

The total number of records after the application of the unification is 46,787 records. The percentage reduction in search space is calculated using the following Equation 2. This Equation enables the identification of the initial percentage reduction in the search space, thereby providing a quantitative metric for assessing data unification. In this case, the reduction achieved is 95.35%, which represents a significant reduction in the number of records to be analyzed. Although there is a significant reduction in the search space, many records with similar values remain, as shown by the fraction in the total space reduction.

$$((Initial\ total - Final\ total)/Initial\ total)\ \times 100 \tag{2}$$

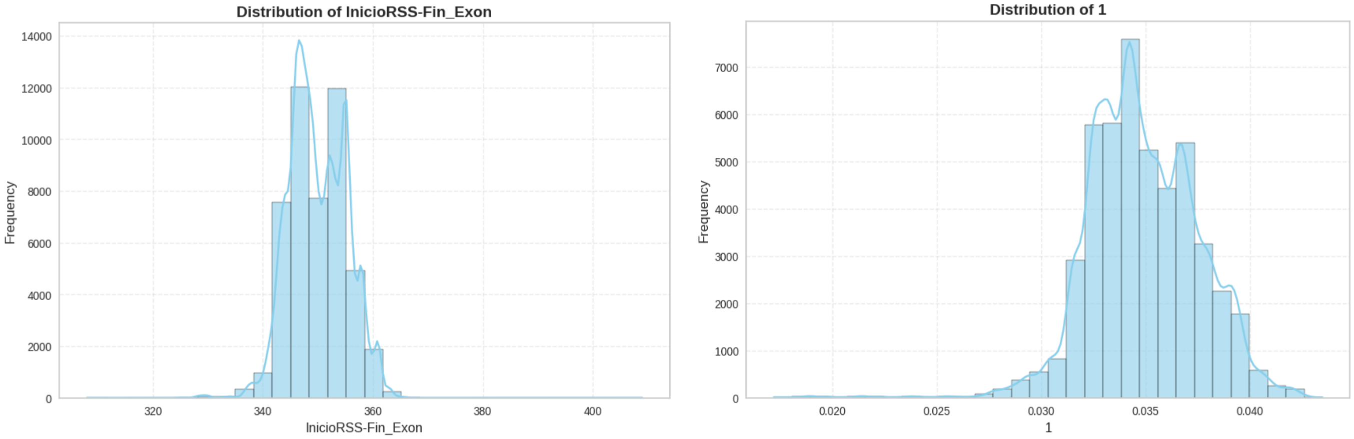
One of the first points that are addressed on the data set are the columns composed entirely of zeros. Table 12 presents a fraction of the data set resulting from data unification. Examples of columns with no relevance are a) *StartRSS-StartRSS* or b) *StartRSS-EndRSS*, these attributes are made up entirely of zeroes, so they are filtered and removed. After this filtering, the dataset is reduced to 22 columns, which contain information about the distances between V gene sections and their original positions in the genome. The final step in the process of splitting the data set consists of extracting the initial seven columns, resulting in a total of 15 attributes used. The selection of these 15 attributes is based on their ability to encompass the complete set of distances between each section of the V gene. This approach allows the most relevant information to be used for model training.

Table 12. New data set obtained from data unification

Strand	Start RSS	Start Exon	Start SP	End RSS	End Exon	End SP	StartRSS-StartRSS	StartRSS-EndRSS	StartRSS-StartExon	StartRSS-EndExon	StartRSS-StartSP	StartRSS-EndSP
0	15507	15543	15931	15542	15857	15975	0	35	36	350	424	468
0	15507	15543	15929	15542	15851	15970	0	35	36	344	422	463

When starting the analysis of the data, the first step is to verify its normality. The distribution plots in Fig. 10 represent exclusively the distances between the gene sections and in particular, the relationship between the start of the *RSS* and the end of the *Exon*. As shown in the non-standardized results, as shown in Fig. 10 (a) plot, the distances range from 300 to 38 characters, which is a common range for the distance between the two gene sections.

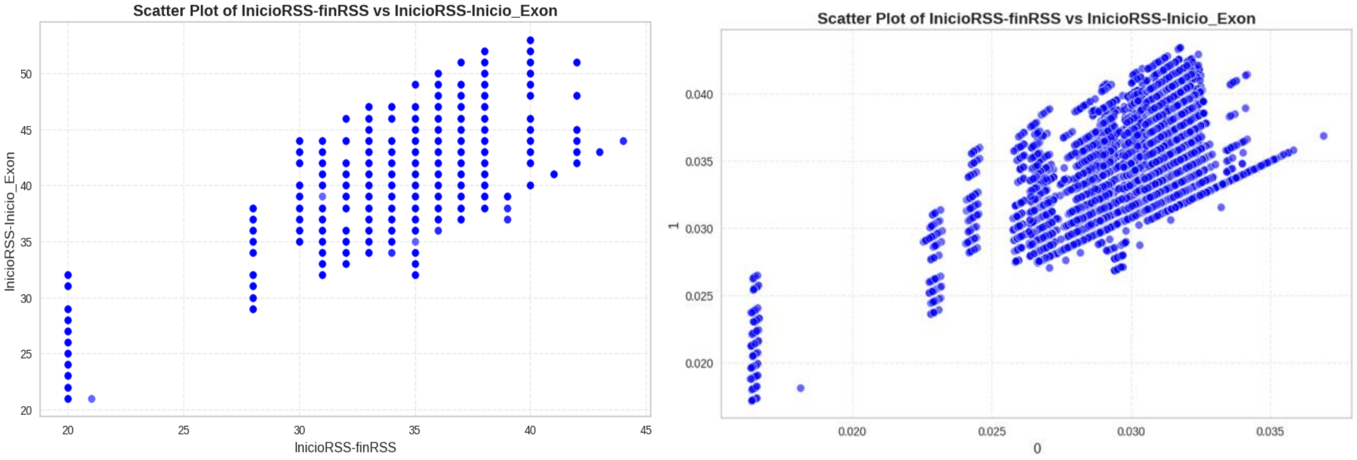
However, the data still does not present a uniform scale or a clear approximation to normality, highlighting the need to normalize the data. For this study, the normalized function of scikit-learn was used, based on the L2 or Euclidean norm (Pedregosa et al., 2011), fitting the values to a vector of length 1. The second plot in Fig. 10 (b) shows the data after normalization, evidencing a more uniform distribution and better adaptation to the processing algorithms, which optimizes their performance in the following stages of the analysis.



(a) (b)

Fig. 10. Distribution plots of the start of the *RSS* at the end of the *Exon*. (a) distribution of the distance from the start of the *RSS* to the end of the *Exon* (b) distribution of the distance from the start of the *RSS* to the end of the normalized *Exon*.

With the normalization of the data, we proceeded to the elaboration of two new graphs that represent the dispersion of the data. Fig. 11 shows these graphs, where the state of the data before and after normalization can be observed. It is important to highlight the new distribution after normalization. Fig. 11 (a) presents the original and unprocessed data, while Fig. 11 (b) represents the data after normalization. This normalization will allow correct processing by the algorithms to be used, allowing the identification of unknown relationships and patterns.



(a) (b)

Fig. 11. Scatter plots of the first two attributes of the set (a) scatter plot of the original data (b) scatter plot of the data normalized with the Euclidean function.

Finally, it is worth mentioning that normalization makes use of the alignment data and reference data processed by the data unification. The reason for using all data at the same time is to use as much data as possible and to verify the effect of the reference data. But these reference data are extracted for data processing so as not to affect the model implementation. The algorithm will only make use of the 46,787 records for its implementation, bypassing a total of 918 reference records that will be used for the validation of the best group.

2.3 Data processing stage

The processing stage consists of the application of a machine learning algorithm suitable for the type of data used. In this case, the objective is the reduction of the search space and the identification of V genes, which implies working with an unclassified dataset, because there are no features available to clearly differentiate between defined classes. Because of this, the problem falls within the scope of unsupervised learning, which requires the selection of a clustering algorithm capable of identifying records with greater similarity to each other.

The selected algorithm is Gaussian Mixture Models (GMM) due to its ability to model complex data distributions and detect hidden patterns without the need for predefined labels. Unlike other clustering methods such as K-Means, GMM allows modeling data that does not necessarily follow a spherical structure, which makes it more suitable for data sets with variability in their distributions (Oscar Contreras Carrasco, 2024; Pedregosa et al., 2011).

With the definition of the algorithm, it is necessary to determine the optimal number of groups. Although this decision can be made arbitrarily, in the present study it has been decided to make use of one of the most used approaches for the definition of groups which is the Elbow Method, which identifies the point where the inertia (variability within groups) decreases (Shi et al., 2021).

Finally, once the number of groups has been defined, the last stage of data processing is carried out: unification by groups. This step consists of consolidating the data within each group, since similar records are expected to be found in the same cluster. The goal is to minimize the number of results per group. This final reduction is necessary because only one result is reported for each V gene in every group. The data processing performed for the *Tadarida brasiliensis* dataset is presented below.

Selection of groups and application of the Gaussian mixture model. The first step in this section was the selection of a clustering algorithm for the nature of the data. As mentioned earlier, due to the structure of the data, it is necessary to use an unsupervised learning technique, particularly a clustering algorithm. Although there are multiple approaches for data segmentation, in this study we have chosen to use the Gaussian mixture model (GMM). This algorithm assumes that the data comes from multiple overlapping normal distributions, which allows modeling uncertainty in the assignment of points to groups.

Unlike algorithms such as K-Means, which assigns each point strictly to a single group, GMM allows points to have a probability of belonging to different groups, which makes it more suitable for data with non-spherical structures (Oscar Contreras Carrasco, 2024; Pedregosa et al., 2011). The analysis of Fig. 12, which represents the dispersion of the data, allows for the identification of three key characteristics in its distribution. These characteristics are presented in Table 13.

Table 13. A comparison of K-Means and GMM for non-spherical structures

Observation	K-Means Limitation	GMM Advantage
Non-spherical structure	The K-Means algorithm assumes that clusters are isotropic (circular), which prevents the detection of groupings with elongated or elliptical geometries	The Gaussian mixture model represents each cluster as a multivariate Gaussian distribution, each with its own vector. This approach captures elliptic and other complex geometries
Overlapping of potential groups	Each observation is assigned exclusively to a single cluster, imposing rigid boundaries even in regions where groups overlap	It assigns to each observation a probability of belonging to each component, which facilitates smooth boundaries that adapt to regions of high density and allow the overlapping of subgroups
Heterogeneity in dispersion	It assumes that all clusters share the same variance, making it unsuitable for modeling structures with heterogeneous dispersion	As each Gaussian component has an independent covariance matrix, the model supports clusters with heterogeneous variances and irregular dispersions

Based on these observations, GMM was defined as the algorithm selected for this study, because it allows capturing complexities in the data distribution and generating clusters more representative of the underlying structure in the analyzed set. The need to obtain clusters with clear similarities in their data is one of the main reasons for its selection.

After preprocessing the data and defining the clustering algorithm, we proceeded to define the optimal number of clusters. Therefore, it was determined that the elbow method would be employed to ascertain the point at which augmenting the number of groups no longer engenders substantial enhancements in data compactness. Fig. 12 shows the relationship between the number of groups (k) and inertia, showing a drastic reduction in inertia between k = 2 and k = 4. However, from k = 4 onwards, the decrease in inertia becomes less pronounced, indicating that adding more groups does not significantly improve the quality of the clustering. In this study, 4 clusters will be used.

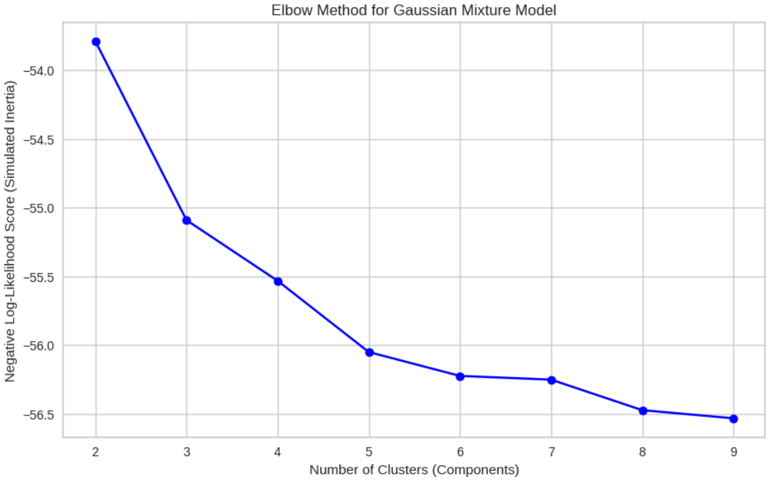
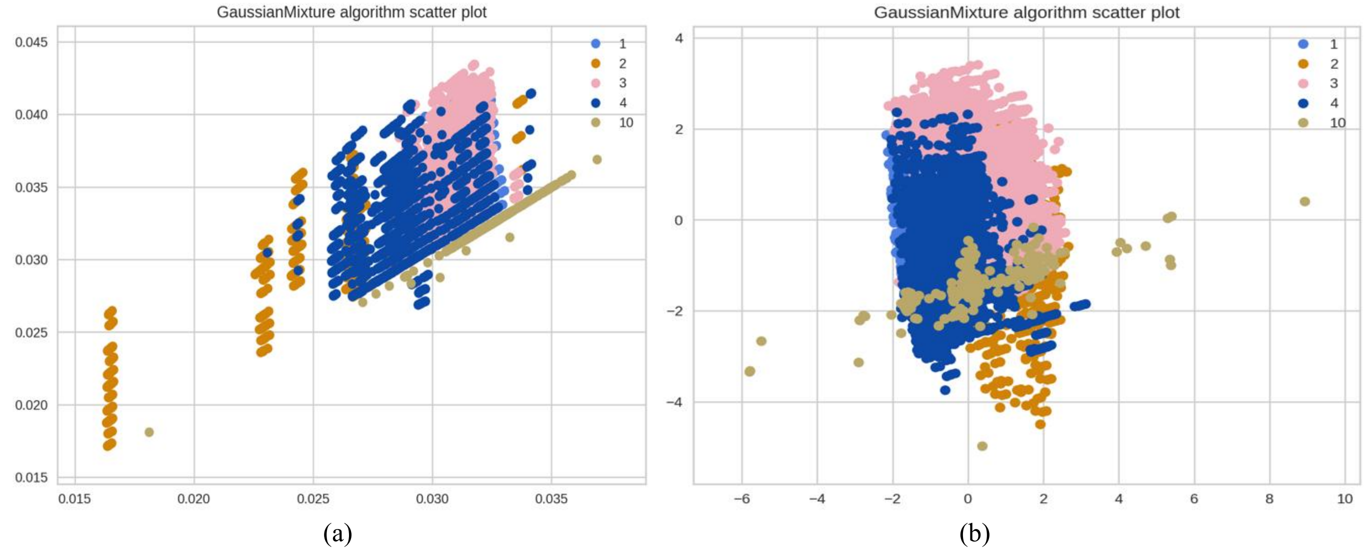


Fig. 12. Results of the Elbow Method.

After defining the number of groups and the algorithm to be used, the data were trained and grouped. Fig. 13 presents the scatter diagrams with the grouped data. The first diagram (Fig. 13 (a)) shows the original data from Fig. 11 (b). The four groups are represented by different colors, and the principal component analysis (PCA) technique was used to improve the graphs' representation (see Fig. 13 (b)). Note that a new group labeled 10 is presented in the graphs. This group represents the reference data that were not used to train the algorithm. The purpose of representing these data is to evaluate their proximity to the other groups and to analyze if they present any relevance within the segmentation performed.



(a) (b)

Fig. 13. Scatter plots after applying the GMM algorithm. (a) Scatter plot of normalized and pooled data (b) Two-component scatter plot with PCA.

Finally, the cluster analysis indicates that: a) cluster one has clustered 8219, b) cluster two has clustered 1117, c) cluster three has clustered 33968 and d) cluster four has clustered a total of 3482. Each of these clusters contains N number of V genes, but one of the problems with these clusters is that there is still repeatability so performing another data unification will solve this problem.

Unification of data by groups. The unification of data by group is the last point of processing, which allows minimizing the number of results within each grouping. This reduction is essential for the comparison of the quality of each group with reference. The new data unification is because each group identifies a number N of V genes. However, these genes may appear in several records with small variations, generating redundancies within a group. For example, position 15946089 in Table 14 corresponds to the start of one gene, while position 247913721 corresponds to the start of another gene. These two positions mark the start of the RSS in their respective genes. This pattern in the two genes will be repeated for each gene in the different groups.

Table 14. Positions grouped by the algorithm

Number	StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
1	15946089	15946130	15946516	15946129	15946438	15946566
2	15946089	15946130	15946516	15946129	15946441	15946566
3	15946089	15946130	15946516	15946129	15946444	15946566
4	247913721	247913760	247914148	247913759	247914065	247914189
5	247913721	247913762	247914142	247913759	247914070	247914183
6	247913721	247913762	247914143	247913759	247914067	247914190

In view of the high prevalence of repetition among the identified genes, it is proposed that the data for each group be consolidated by computing the column-wise mean. This will result in the production of a single representative row per gene. As demonstrated in Table 15, this procedure results in the derivation of two genes from the six rows presented in Table 14. This consolidation process serves to reduce redundancy and ensures that each gene is represented by a single, comprehensive record.

Table 15. Positions grouped by the algorithm

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

However, the process used to generate the records in Table 15 often produces decimal values because of averaging. To address this issue and obtain a single representative record per gene, the value closest to the average is selected within each group, prioritizing those that preserve biological coherence. Following this criterion, the second and fourth records from Table 14 are retrieved.

This procedure is repeated for each gene across the different groups, aiming to obtain unique V genes within each grouping, thereby enabling their evaluation against the reference standard. Given the nature of the problem, generic evaluation metrics cannot be applied, as a specific approach is required to ensure precise and consistent measurements. To this end, the final phase of the method incorporates a post-processing step.

3 Postprocessing

The final stage of the method corresponds to postprocessing, whose purpose is to evaluate and select the most appropriate cluster among those generated by the GMM algorithm. The evaluation is based on error-related metrics, as they quantify the quality of the clusters in relation to the reference. The metrics applied include a) Mean Absolute Error (MAE), b) Mean Squared Error (MSE), c) Root Mean Squared Error (RMSE), and d) the number of genes in each group. The use of these metrics allows for an effective assessment of the method's capacity to reduce the search space and accurately identify V genes. This ensures that the unified records faithfully represent the actual genomic structure.

Finally, a reduced file is generated by comparing the genomic positions of the best-performing group against the original files. This reduced file, containing the most accurate positions, is subsequently used in the *IGV* tool to demonstrate the effectiveness of the method in narrowing the search space for V genes in vertebrate genomes.

3.1 Evaluation and selection of the best group

In interpreting the results for the selection of the best group, it is first necessary to understand the relationship between the new data set and the reference used for validation. As mentioned above, the unified dataset contains 46,787 records, each representing a possible V gene. However, one of the key questions for the method is: How many actual V genes can be identified within these

46,787 records? To answer this question, a unique function is used on the first column of the dataset (*Start_RSS*), which allows to count the number of unique V genes being the total of 842 V genes that can be identified.

It is essential to know this number, because it sets a limit on the maximum number of V genes that can be identified in the dataset. From this, the second question arises: How many complete and validated V genes are there in the reference? The answer to this question is a total of 917 records. These records will be used for the evaluation of the method.

The evaluation follows the principle of regression measurement, where an expected value (expert reference) and a predicted value (data obtained from the search space reduction method) are compared. Table 16 shows this comparison, where: a) the first row represents the values scored by the experts (expected reference); b) the second row corresponds to the data generated by the unification method; c) The third record shows the difference between both values, used to calculate the error metrics.

Table 16. Evaluation of the results of the method against the reference

	StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
Reference	88965	88926	88539	88930	88615	88495
Method	88967	88929	88536	88929	88621	88490
Difference	-2	-3	3	1	-8	5

The evaluation results are presented in Table 17, which summarizes the quality of each group generated by the Gaussian Mixture Model algorithm. This comparison reveals variability in the number of genes identified, with an approximate error of 3 in most cases. The best-performing group achieved an error of 3.166583, identifying 734 genes out of 842, which corresponds to 87.17 % of the total.

Although the total number of identified genes does not reach 842, this discrepancy is explained by structural alterations in certain genes, which hinder their proper grouping and identification by the proposed method. Nevertheless, Group 3 stands out as the top performer, yielding the lowest error and the highest number of V genes identified.

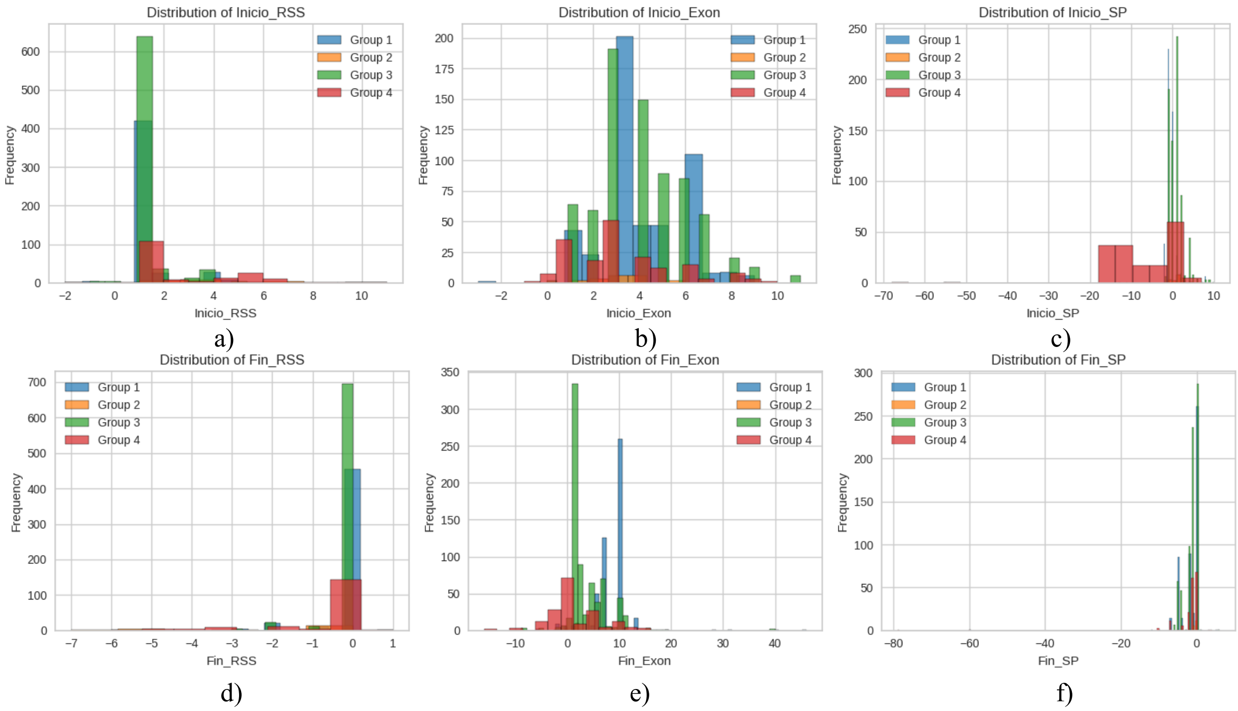
Table 17. Results of the test of the search space method

Group	No. of data	Number of genes	MAE	MSE	RMSE
1	8219	490	2.820408	22.254422	4.717459
2	1117	21	2.47619	15.793651	3.974123
3	33968	734	1.986376	10.027248	3.166583
4	3482	175	3.333333	40.95619	6.399702

Table 17 provides an overview of the error, while the distribution plots in Fig. 14 offer a more detailed perspective on the magnitude and variability present in each section of the V gene. These plots show how the displacement varies across the different sections of the V gene within the analyzed groups. For instance, in the *Start_RSS* position from Fig. 14 (a), Group 3 exhibits an error close to zero, with most values ranging between +1 and +4. This stability indicates that the records in this group align more closely with the expert reference for the *RSS* start section. As shown in Fig. 14 (a), most genes in Group 3 are positioned at +1, with a bar exceeding 600 records, followed by Group 2 with approximately 400 genes. This demonstrates that, at least for the gene start section, Group 3 delivers the best results.

The *Exon_start* and *SP_start*, shown in Fig. 14 (b) and 14 (c), greater variability is observed, even within Group 3. In the case of *Exon_start* from Figure 14 (b), values range from +1 to +10, with notable concentrations at positions +4 and +5, indicating a higher degree of dispersion and reduced precision in correctly identifying the gene. For *SP_start* Fig. 14 (c), values range from -1 to +10; however, in Group 3, most data points are concentrated at +1. This plot also reveals that certain groups exhibit variations of up to 80 positions, with minimal records of a single gene, but associated with an abnormal error.

Finally, Figs. 14 (d), 14 (e), and 14 (f) present the results for the final sections of the V gene. In Fig. 14 (d) (**End_RSS**), most values are concentrated at 0, with slight variations at -1, indicating the presence of genes with missing characters that prevented perfect identification. In Fig. 14 (e) (*End_Exon*), a pattern like Fig. 14 (b) is observed, where gene identification is more challenging due to greater positional variation, with displacements ranging from +1 to +10. For Group 3, the highest concentration occurs at +1, with nearly 300 genes, suggesting an error of at least one base. Lastly, Fig.14 (f) (**End_SP**) exhibits the largest displacement, with errors between -2 and -1. Although some error margin remains in this plot, the presence of values at 0 demonstrates that the method can accurately identify genes in this section.



a) b) c)

d) e) f)

Fig. 14. Error distribution plots for each group: a) *Start_RSS* position, b) *Start_Exon* position, c) *Start_SP* position, d) *End_RSS* position, e) *End_Exon* position, and f) *End_SP* position.

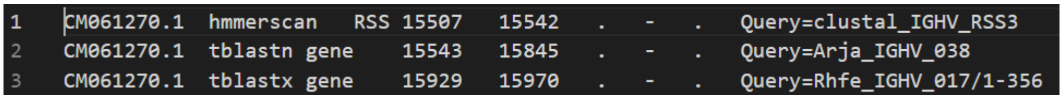
3.2 Reduced file and its use in *IGV*

Consequently, the reduced file is obtained from the dataset with the highest identification accuracy. In this final phase, the first seven columns of the optimal dataset are used to determine the positions of the V gene. As shown in Table 18, an example from dataset 3 is presented, where the first column indicates direction, the next three columns correspond to the initial sections, and the last three columns to the final sections of the gene. To construct the reduced file, the first three columns of the first row are merged into a single column, and the same procedure is then applied to the last three columns of that row. This process yields a structured representation of the V gene.

Table 18. Example of final records for reduced V gene(s)

Strand	StartRSS	StartExon	StartSP	EndRSS	EndExon	EndSP
Reverse	15507	15543	15928	15542	15851	15975
Forward	2107209	2107245	2107629	2107244	2107553	2107676

Once a Table with only two records is obtained, the original files are searched for records that generate the same positions as the table, for example, *StartRSS* with 15507 and *EndRSS* with 15542, and the record that is identical to these positions is extracted. Finally, these results are transformed into *GFF* format and unified in a single file. Fig. 15 shows an example of a fraction of this file with the present example. In this format, every three records represent one gene, consolidating the information in a structured way and facilitating its analysis in genomic annotation tools.



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

Fig. 15. Example of a V gene in the reduced file.

To demonstrate the method's feasibility, the results were presented using the tool employed by experts. This approach shows how the method reduces the search space and facilitates the identification of V genes in the analyzed genome. Fig. 16 shows the clusters generated by the method, displaying the distribution and segmentation of the data. The *IGV* system facilitates the assessment of the magnitude of the *Tadarida brasiliensis* genome in relation to the V genes.

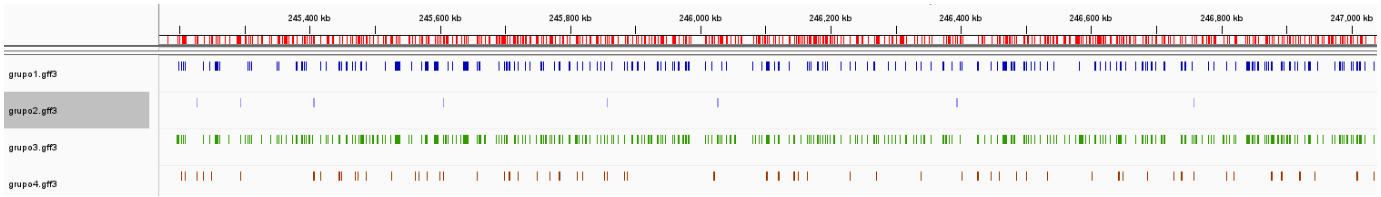


Fig. 16. Set of alignments reduced by the method.

When the analysis is extended over a set of alignments, a key phenomenon emerges: some groups identify genes that others do not. This highlights the importance of understanding that the method is not completely foolproof, because its accuracy is influenced by the nature of the V genes and the variability in their alignments. While certain groups may detect genes that others fail to identify, as shown in Fig. 17, the fundamental question is not only which genes are identified, but how accurately these results are achieved. Evaluating this aspect is essential to determine the reliability of the method, ensuring that the detections are consistent and reproducible within the genomic characterization process.

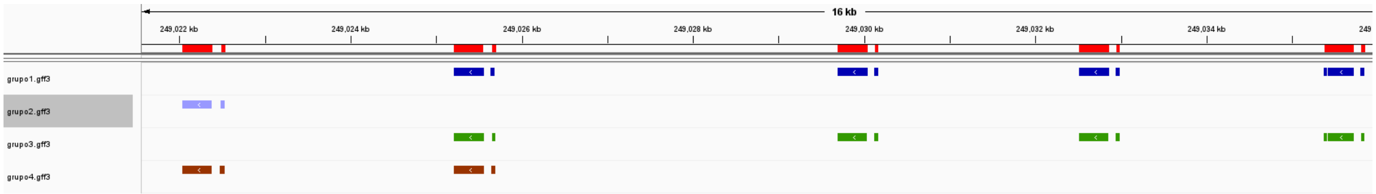


Fig. 17. Close-up of a set of alignments.

When selecting a gene, Fig. 18 shows the three sections of the gene; in this case the gene has a reverse direction, in red the reference of the expert is shown. It is relevant to mention that group three shows a great closeness to the reference, but it is not perfect, this is demonstrated by the missing characters at the start of the gene. But it compensates mostly with the accuracy in the SP, the last section of the alignment.

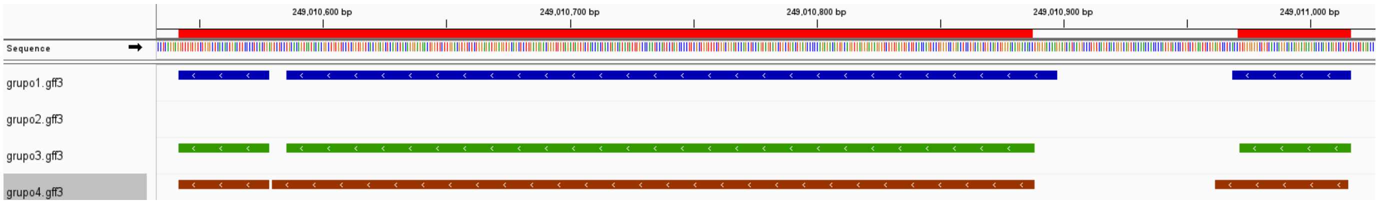


Fig. 18. Example of a gene with alignments.

Finally, when a closer approach is made to the start of the *Exon* and the alignment of the SP, Fig. 19 shows the closeness of each group to reference. As shown in group 3 and 4 present a great closeness to the reference in the case of the start of the *Exon* but in the alignment of the *SP* is where the difference of each group is demonstrated. In the case of the *SP* for group 3 is the one that presents a greater accuracy of all the groups, in the case of group 3 it only varies by one position being almost perfect.

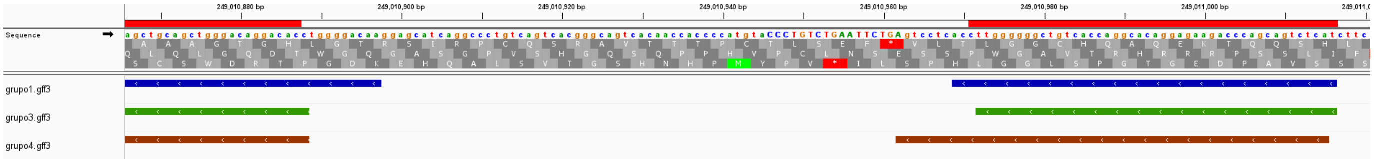


Fig. 19. Example of the alignments of a SP.

4 Discussion

The proposed method for data unification and search space reduction proved to be an effective tool for improving V gene identification in large volumes of genomic data. The integration of multiple alignment sources allowed for a more complete representation of the gene, reducing redundancy and minimizing the manual effort required by experts. However, despite the improvement in accuracy and efficiency, the method still presents challenges, especially in the identification of specific sections of the gene, such as the *Exon*, where higher variability was observed compared to the expert reference.

One of the most relevant aspects of the analysis was the evaluation of the performance of the method in terms of computational time. Running on a computer with moderate specifications (Asus TUF with 16 GB RAM, mechanical hard disk, Intel Core i7U and an NVIDIA RTX 3070 graphics card) allowed the method to be completed in approximately 50 minutes, which represents a significant improvement compared to the manual method. However, this time could be optimized with the use of more advanced computational architectures or by comparing other models or dimension reduction techniques that facilitate data processing.

Another key point in the discussion is the reliability of the Gaussian Mixture Model (GMM) used for clustering data. While GMM allowed modeling complex structures and capturing variations in the distribution of alignments, the probabilistic assignment of records to clusters suggests that some genes could have been classified into multiple clusters. Future research explores the combination of GMM with other machine learning approaches, such as dimension reduction to vary the structure and dispersion of the data. As well as analyzing other data normalization and cluster unification techniques and identifying if these produce any variation for method improvement.

Finally, the transformation of the results to *GFF* format was a crucial step to ensure the applicability of the method in genomic analysis tools such as *IGV*. However, it was identified that the accuracy of the alignments is highly dependent on the quality of the input data. In the future, more extensive evaluation with different datasets is recommended. Validating the generalization to different genomes will allow robustness of the method and verify that the method is indeed applicable to multiple species.

5 Conclusions

This study presents an innovative method for reducing the search space in the identification of V genes, achieving a substantial improvement in both the precision and efficiency of the process. The combination of preprocessing, normalization, and clustering techniques significantly reduced the number of records to be analyzed, facilitating the identification of complete genes with high agreement to expert annotations. This is evidenced by the reduction in the search space: starting with a total of 1,006,797 records for all genes and concluding with clusters containing, at most, 734 records—one per gene. Although some data loss occurs, the method's optimization continues to offer opportunities for performance enhancement.

The results demonstrate that the method can be applied to other genomes, provided that the input data meets the defined alignment and structuring criteria. The integration of multiple data sources into a unified reference set improves record quality and reduces redundancy, thus optimizing the time and computational resources required for genomic characterization.

A noteworthy aspect of this method is the lack of a similarly structured proposal in the current literature that direction search space reduction with this level of systematization. This establishes an important precedent in the optimization of biological data through computational techniques.

Despite its advantages, the method still presents areas for improvement—particularly in the identification of specific genomic sections, such as the signal peptide (*SP*) and *Exon* regions, as well as in computational efficiency. Future research will focus on refining clustering models, exploring new dimensionality reduction techniques, and assessing how genetic variability affects alignment precision.

In conclusion, the proposed methodology represents a significant advancement in the systematization of V gene identification, especially when working with distance-based data instead of complete genomic sequences. Although the use of sequences is not excluded, this approach provides a solid foundation for future research in bioinformatics, genomic characterization, and computational process optimization. Its implementation in genomic analysis tools holds the potential to enhance gene annotation and facilitate data interpretation in studies of immunogenomics and molecular evolution.

References

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.

Amin, M. R., Yurovsky, A., Tian, Y., & Skiena, S. (2018). DeepAnnotator: Genome annotation with deep learning. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 254–259).

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: Theory and algorithms*. John Wiley & Sons.

Bergman, N. H. (Ed.). (2007). *Comparative genomics: Volumes 1 and 2*. Humana Press.

Dagdia, Z. C., & Mirchev, M. (2020). Chapter 15 – When evolutionary computing meets astro- and geoinformatics. In P. Škoda & F. Adam (Eds.), *Knowledge discovery in big data from astronomy and Earth observation* (pp. 283–306). Elsevier. https://doi.org/10.1016/B978-0-12-819154-5.00026-6

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics, 14*(9), 755–763.

Ejigu, G. F., & Jung, J. (2020). Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology, 9*(9), 295.

EMBL-EBI. (2022). GFF/GTF file format: Definition and supported options. Ensembl. Retrieved from https://www.ensembl.org/info/website/upload/gff.html

García Simón, A. (2018). *Gestión de datos genómicos basada en modelos conceptuales* [Master's thesis, Universitat Politècnica de València]. RiuNet. https://riunet.upv.es/handle/10251/111666

Kalkatawi, M., Magaña-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. In M. Kollmar (Ed.), *Gene prediction* (*Methods in Molecular Biology*, Vol. 1962, pp. 161–177). Humana Press. https://doi.org/10.1007/978-1-4939-9173-0_9

Kindt, T. J., Goldsby, R. A., & Osborne, B. A. (2007). *Inmunología de Kuby*. McGraw-Hill.

Lefranc, M.-P., & Lefranc, G. (2001). *The immunoglobulin factsbook*. Academic Press.

McNair, K., Ecale Zhou, C. L., Souza, B., Malfatti, S., & Edwards, R. A. (2021). Utilizing amino acid composition and entropy of potential open reading frames to identify protein-coding genes. *Microorganisms, 9*(1), 129.

Megrian, D. (2014). *Identificación de genes de inmunoglobulinas en el genoma bovino* [Bachelor's thesis, institution unknown].

Miguel-Ruiz, J., Serret, N., Ortiz-Hernandez, J., Barnetche, J. M., & Hernández, Y. (2024). A design science approach to modeling the V gene annotation process. *Programming and Computer Software, 50*(8), 829–843. https://doi.org/10.1134/S0361768824700798

Mount, D. W. (2004). *Bioinformatics: Sequence and genome analysis*. Cold Spring Harbor Laboratory Press.

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

Contreras Carrasco, O. (2024). Gaussian mixture model explained. Retrieved from https://builtin.com/articles/gaussian-mixture-model

Pfennig, A., Lomsadze, A., & Borodovsky, M. (2022). Annotation of phage genomes with multiple genetic codes [Preprint]. *bioRxiv*.

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., Thorvaldsdóttir, H., Turner, D., & Mesirov, J. P. (2023). igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics, 39*(1). https://doi.org/10.1093/bioinformatics/btac830

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. *Journal of Machine Learning Research, 12*, 2825–2830.

Serret, N., Ortiz-Hernández, J., Miguel-Ruiz, J., Barnetche, J. M., & Hernández, Y. (2023). Conceptual modeling of the V gene annotation process in antibodies. In *Proceedings of the 11th International Conference on Software Engineering Research and Innovation (CONISOFT)* (pp. 256–264).

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking, 2021*(1), 31. https://doi.org/10.1186/s13638-021-01910-w

Sirupurapu, V., Safonova, Y., & Pevzner, P. A. (2022). Gene prediction in the immunoglobulin loci. *Genome Research, 32*(6), 1152–1169.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics, 6*, 1–11.

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.

Webster, C. F., Smotherman, M., Pippel, M., Brown, T., Winkler, S., Pieri, M., Mai, M., Myers, E. W., Teeling, E. C., & Vernes, S. C. (2024). The genome sequence of *Tadarida brasiliensis* I. Geoffroy Saint-Hilaire, 1824 [Molossidae; Tadarida]. *Wellcome Open Research, 9*.

Zhang, Y., Chen, T., Zeng, H., Yang, X., Xu, Q., Zhang, Y., Chen, Y., Wang, M., Zhu, Y., & Lan, C. (2021). RAPID: A rep-seq dataset analysis platform with an integrated antibody database. *Frontiers in Immunology, 12*, 717496.

Zhu, Y., Watson, C., Safonova, Y., Pennell, M., & Bankevich, A. (2024). Assessing assembly errors in immunoglobulin loci: A comprehensive evaluation of long-read genome assemblies across vertebrates [Preprint]. *bioRxiv*.