TRABAJO PRÁCTICO INTEGRADOR: ANÁLISIS BIOINFORMÁTICO DEL BROTE DE BRUCELOSIS PORCINA EN ENTRE RÍOS

Introducción

El presente trabajo tiene como objetivo realizar un análisis bioinformático integral del agente causante de un brote de brucelosis porcina en Entre Ríos.

La brucelosis es una enfermedad zoonótica de importancia tanto en medicina veterinaria como en salud pública, causada por bacterias del género Brucella (Revisión bibliográfica en archivo “Revision bibliografica.pdf”). El análisis incluye el procesamiento de secuencias de ADN obtenidas mediante tecnología Sanger y secuenciación masiva Illumina, el ensamblado del genoma, la predicción y anotación de genes, análisis filogenético y el diseño de primers para un kit diagnóstico por PCR.

Procesamiento y ensamblado de secuencias

Procesamiento de secuencias Sanger

Las secuencias obtenidas mediante tecnología Sanger se encontraban en formato SCF (Standard Chromatogram Format), el cual almacena tanto la secuencia de bases como los valores de calidad asociados a cada base secuenciada. El procesamiento de estas secuencias se realizó mediante la GUI de prepag4 del paquete de Standen. Se eliminaron virtualmente (soft clipping) las regiones de mala calidad y vector.

El resultado de estos procesos son archivos .exp con el mismo nombre de cada electroferograma y una serie de archivos con listas de electroferogramas que pasaron los diferentes requerimientos de cada módulo. El que es más relevante es pregap.passed ya que será utilizado durante el ensamblado. En este caso, dicho archivo contiene los 242 nombres de archivos, indicando que todas las secuencias procesadas cumplieron con los criterios mínimos de calidad establecidos.

Análisis de calidad de reads Illumina

Las secuencias cortas obtenidas mediante secuenciación Illumina se encontraban en el archivo Set3.fq en formato FASTQ. Este formato incluye para cada read cuatro líneas: el

identificador de la secuencia, la secuencia nucleotídica, un separador, y los valores de calidad codificados en formato ASCII.

Para evaluar la calidad de estas secuencias se utilizó FastQC:

fastqc Set3.fq -o /Brucelosis_TP_Integrador/02_pregap4_processing/

El análisis de FastQC reveló las siguientes características del conjunto de datos:

- Total de secuencias: 65,000 reads
- Longitud de reads: 150 nucleótidos
- Tipo: Single-end
- Encoding de calidad: Sanger / Illumina 1.9
- Contenido GC promedio: 55%
- Rango de valores de calidad Phred: 35-38

Measure	Value
Filename	Set3.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	65000
Total Bases	9.7 Mbp
Sequences flagged as poor quality	0
Sequence length	150
%GC	55

Fig. 1: Estadísticas básicas obtenidas por FastQC

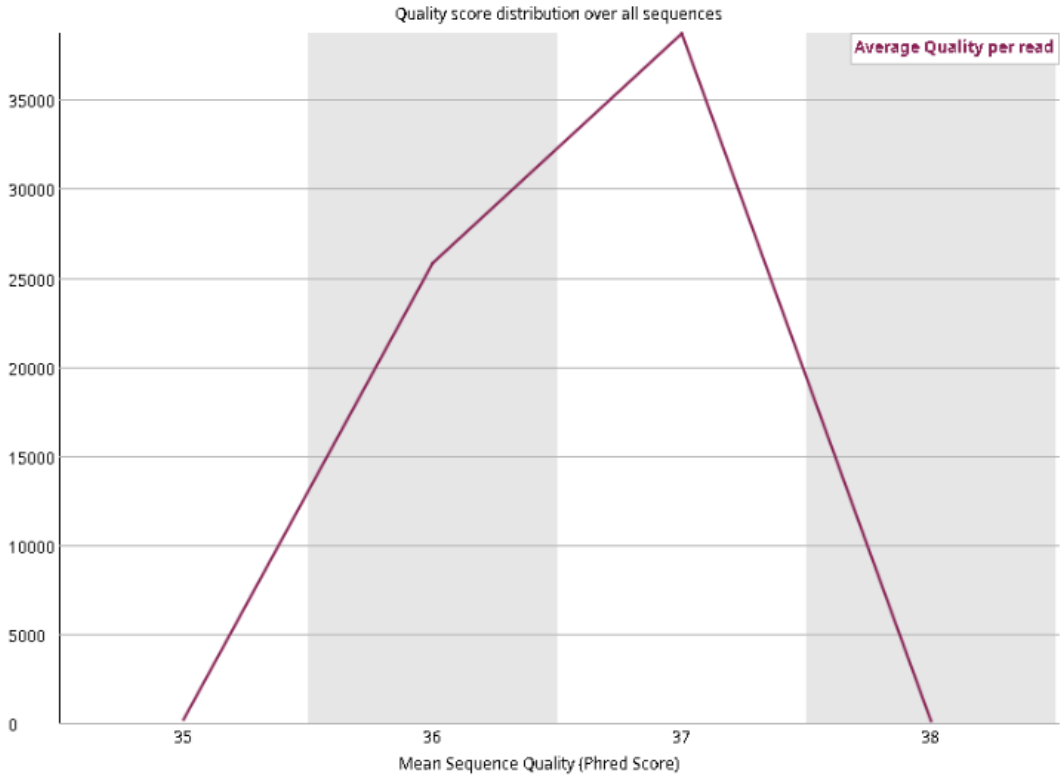


Fig. 2: Distribución de los valores de calidad a lo largo de la secuencia

Los valores de calidad Phred observados (Q35-Q38) corresponden a una probabilidad de error por base entre 0.001 y 0.0002, lo que se considera de excelente calidad para secuencias Illumina. Esto indica una precisión superior al 99.9% (valor obtenido mediante: $Q = -10log_{10}(p)$ en la identificación de bases.

El contenido GC del 55% es consistente con el genoma de Brucella suis, cuyo genoma de referencia presenta un contenido GC del 57.2%.

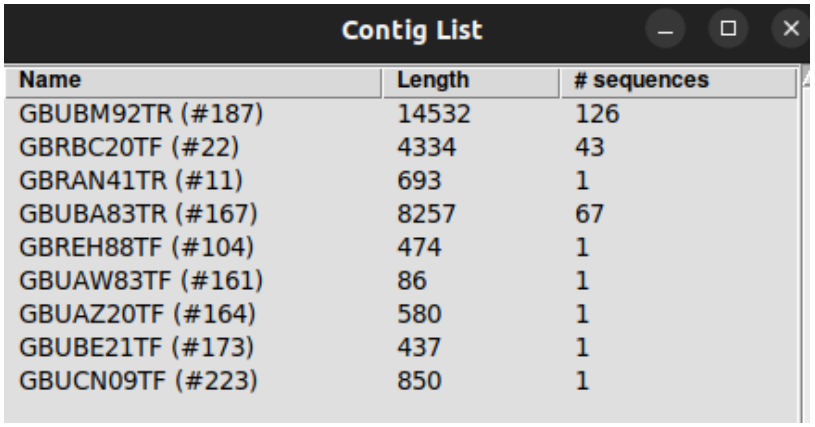
Los gráficos generados por FastQC mostraron una distribución de calidad homogénea a lo largo de toda la longitud de los reads, sin evidencia de degradación significativa de calidad hacia los extremos 3' de las secuencias, lo cual es indicativo de una corrida de secuenciación exitosa. No se detectaron secuencias adaptadoras ni contaminantes, ni tampoco secuencias sobrerrepresentadas que pudieran indicar problemas técnicos o contaminación.

Ensamblado de secuencias largas (Sanger) mediante Gap4

El ensamblado de secuencias es el proceso mediante el cual múltiples fragmentos de ADN secuenciados se alinean y combinan para reconstruir secuencias continuas más largas, denominadas contigs (contiguous sequences). El ensamblado es particularmente desafiante cuando se trabaja con lecturas relativamente cortas que deben superponerse correctamente para inferir la secuencia completa.

Para el ensamblado de las 242 secuencias Sanger procesadas se utilizó Gap4, un programa del paquete Staden diseñado específicamente para el ensamblado de secuencias generadas por método de Sanger.
El proceso de ensamblado generó una base de datos Gap4 conteniendo 9 contigs. La interfaz gráfica de Gap4 mostró la lista de contigs resultantes con sus respectivas longitudes y número de secuencias que los componen. Los contigs más relevantes fueron:

- Contig GBUBM92TR (#187): 14,532 bp ensambladas a partir de 126 secuencias
- Contig GBRBC20TF (#22): 4,334 bp a partir de 43 secuencias
- Contig GBRAN41TR (#11): 693 bp a partir de 1 secuencia



Name	Length	# sequences
GBUBM92TR (#187)	14532	126
GBRBC20TF (#22)	4334	43
GBRAN41TR (#11)	693	1
GBUBA83TR (#167)	8257	67
GBREH88TF (#104)	474	1
GBUAW83TF (#161)	86	1
GBUAZ20TF (#164)	580	1
GBUBE21TF (#173)	437	1
GBUCN09TF (#223)	850	1

Fig. 3: Lista de contigs obtenida por gap4

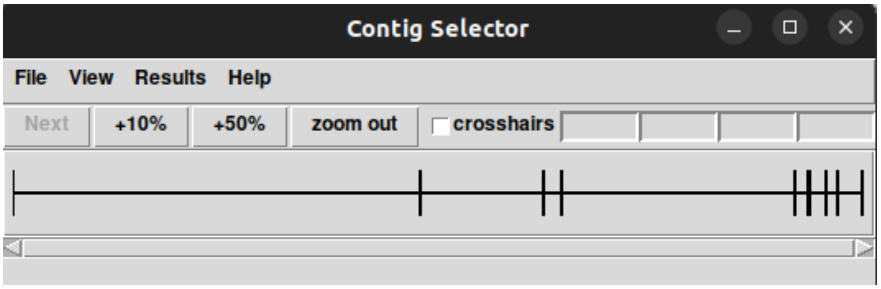


Fig. 4: Representación gráfica los contigs

La longitud total aproximada de todas las secuencias ensambladas fue de 29,981 bp, distribuidas en los 9 contigs mencionados. Este resultado indica una fragmentación considerable del ensamblado, lo cual es esperable dado que las secuencias Sanger tienen una cobertura limitada y no necesariamente cubren de manera continua toda la región genómica de interés.

Los contigs fueron exportados desde Gap4 en formato FASTA mediante la opción de exportación de la interfaz gráfica, generando el archivo sanger_contigs.fasta:

Este archivo contiene las 9 secuencias consenso resultantes del ensamblado, cada una representando una región genómica contigua reconstruida a partir de las secuencias Sanger superpuestas.

Ensamblado híbrido con SPAdes

El ensamblado híbrido es una estrategia que combina datos de secuenciación de diferentes tecnologías para aprovechar las ventajas complementarias de cada una. En este caso, se combinaron las secuencias largas y precisas obtenidas por tecnología Sanger (representadas por los 9 contigs previamente ensamblados) con las secuencias cortas pero de alta cobertura obtenidas por tecnología Illumina. Esta aproximación permite utilizar los contigs Sanger como "esqueleto" o guía para el ensamblado, mientras que las lecturas Illumina aportan información para llenar gaps, corregir errores y proporcionar mayor profundidad de cobertura.

Para el ensamblado híbrido se utilizó SPAdes (St. Petersburg genome assembler), versión 3.15.5, un ensamblador de novo ampliamente utilizado para genomas bacterianos.

El comando ejecutado fue:

```
spades.py \
   -s /home/ana/Desktop/Set3.fq \
   --trusted-contigs sanger_contigs.fasta \
   -o hybrid_assembly \
   --threads 4 \
```

--memory 16

Los parámetros utilizados significan:

- -s: especifica el archivo de lecturas single-end de Illumina
- --trusted-contigs: indica las secuencias Sanger como contigs de alta confiabilidad que SPAdes debe incorporar preferentemente en el ensamblado
- -o: directorio de salida para los resultados
- --threads: número de procesadores a utilizar (4 cores)
- --memory: límite de memoria RAM en GB (16 GB)

SPAdes ejecuta internamente múltiples pasos: corrección de errores de las lecturas Illumina, construcción de grafos de De Bruijn con diferentes valores de k-mer (21, 33, 55, 77), y ensamblado final mediante la resolución del grafo. El uso de contigs trusted permite que SPAdes priorice caminos en el grafo que son consistentes con las secuencias Sanger, reduciendo así la fragmentación del ensamblado final.

SPAdes generó múltiples archivos de salida, siendo el más relevante scaffolds.fasta, que contiene las secuencias ensambladas finales.

El análisis del archivo scaffolds.fasta reveló que el ensamblado híbrido produjo un único scaffold de alta calidad:

>NODE_1_length_30000_cov_158.511780

Este resultado representa una mejora dramática respecto al ensamblado inicial de secuencias Sanger, reduciendo de 9 fragmentos a un único scaffold continuo. Las características del scaffold son:

- Longitud: 30,000 pares de bases
- Cobertura promedio: 158.5x
- Número de scaffolds: 1

La cobertura de 158.5x indica que cada posición nucleotídica del ensamblado está respaldada por un promedio de 158 lecturas independientes, lo cual confiere una alta confiabilidad al consenso. Esta cobertura elevada se debe principalmente a la contribución de las 65,000 lecturas Illumina de 150 bp, que aportan aproximadamente 9.75 Mbp de secuencia total (65,000 × 150 bp), que al mapearse sobre un scaffold de 30 kbp resulta en la cobertura observada.

El archivo de estadísticas generado por SPAdes (contigs.paths) confirmó que el ensamblado integró exitosamente información de ambas tecnologías, utilizando los contigs Sanger como puentes para resolver regiones ambiguas del grafo de De Bruijn construido a partir de las lecturas Illumina.

Análisis estadístico del genoma ensamblado

El análisis de las características composicionales del genoma ensamblado es fundamental para validar la identidad taxonómica del organismo y evaluar la calidad del ensamblado. Las dos métricas principales analizadas fueron el tamaño del ensamblado y el contenido de guanina-citosina (GC%).

Para calcular estas estadísticas se creó un script bash denominado calculate_genome_stats.sh

La ejecución del script produjo los siguientes resultados:

Tamaño del genoma ensamblado: 30,000 pares de bases

Composición de bases:

- Adenina (A): 6,420 bases (21.4%)
- Timina (T): 6,900 bases (23.0%)
- Guanina (G): 8,670 bases (28.9%)
- Citosina (C): 8,010 bases (26.7%)

Contenido GC: 55.57%

El tamaño de 30 kbp representa aproximadamente el 0.9% del genoma completo de Brucella suis, cuyo tamaño total es de aproximadamente 3.3Mbp distribuidos en dos cromosomas circulares. Este fragmento corresponde probablemente a una región genómica específica que fue el objetivo de la secuenciación dirigida realizada en el contexto del brote epidemiológico.

El contenido GC del 55.57% es consistente con el contenido GC del genoma de referencia de B. suis cepa 1330, que es del 57.2%. La diferencia de 1.63 puntos porcentuales (2.85% de variación relativa) está dentro del rango esperado para variaciones regionales en el genoma bacteriano. Es conocido que el contenido GC no es homogéneo a lo largo del genoma, existiendo regiones con mayor o menor contenido GC dependiendo de factores como la densidad génica, la presencia de elementos repetitivos, y la historia evolutiva de cada región. Dado que el fragmento secuenciado representa menos del 1% del genoma total, es razonable encontrar variaciones locales respecto al promedio genómico global.

La relación G+C / A+T es de 1.22, lo que indica un sesgo moderado hacia bases GC, característico de bacterias del género Brucella y otros miembros de las Alphaproteobacteria. Este sesgo puede estar relacionado con la estabilidad térmica del ADN, ya que los pares de bases GC forman tres puentes de hidrógeno en comparación con los dos puentes de los pares AT, confiriendo mayor estabilidad a regiones ricas en GC.

Predicción y anotación de genes

La predicción de genes es el proceso mediante el cual se identifican las regiones codificantes en una secuencia genómica. Existen dos aproximaciones principales: los métodos ab initio, que se basan únicamente en señales intrínsecas de la secuencia como codones de inicio y stop, contenido GC, y uso de codones; y los métodos basados en homología, que utilizan comparaciones con bases de datos de genes conocidos.

Predicción ab initio con Glimmer3

Glimmer (Gene Locator and Interpolated Markov ModelER) es un sistema de predicción de genes ab initio ampliamente utilizado para genomas bacterianos. Glimmer utiliza modelos de Markov interpolados (ICM) que capturan las propiedades estadísticas de las secuencias codificantes, entrenándose a partir de un conjunto de genes conocidos de un organismo relacionado.

Para la predicción con Glimmer3 se utilizó el programa tigr-glimmer. El proceso de predicción requiere primero entrenar un modelo utilizando un genoma de referencia, en este caso Brucella suis cepa 1330 (accesión GCF_000007505.1), descargado de NCBI.

Se creó el siguiente script para automatizar el proceso (run_glimmer.sh)

Los parámetros utilizados en glimmer3 fueron:

- -o50: permite un solapamiento máximo de 50 bp entre genes predichos
- -g110: longitud mínima de gen de 110 bp
- -t30: threshold de score mínimo de 30 para considerar una predicción válida

El entrenamiento del modelo extrajo 1,027 ORFs del genoma de referencia de B. suis, generando 1,185,433 bytes de secuencias de entrenamiento. El modelo ICM construido captura los patrones de uso de codones y contexto nucleotídico característicos de Brucella suis.

La ejecución de Glimmer3 sobre el scaffold de 30 kbp produjo los siguientes resultados: Genes predichos: 27 ORFs

El archivo brucella.predict contiene las coordenadas de cada gen predicho con el siguiente formato:

```
>NODE_1_length_30000_cov_158.511780
orf00001    1221    2180  +3    9.53
orf00002    2205    2615  +3    9.84
orf00003    2615    2761  +2    1.70
...
```

Donde las columnas representan: identificador del ORF, posición de inicio, posición de fin, frame y strand (+/-), y score de confianza.

La distribución de genes predichos mostró una cobertura genómica de aproximadamente 87%, con genes en ambas hebras del ADN. La densidad génica calculada fue de 0.9 genes/kb, ligeramente inferior al valor esperado para genomas bacterianos (~1 gen/kb), lo cual puede deberse a la presencia de regiones intergénicas más largas o a la configuración conservadora de los parámetros de predicción.

Predicción mediante pipeline Prokka

Prokka es un pipeline de anotación automática diseñado específicamente para genomas bacterianos. A diferencia de Glimmer, Prokka no solo predice genes sino que también les asigna funciones mediante comparación con múltiples bases de datos de proteínas y dominios funcionales. Prokka integra internamente varias herramientas: Prodigal para predicción inicial de genes, BLASTp para búsqueda de homología, HMMER para búsqueda de dominios, y bases de datos específicas del género cuando están disponibles.

El comando ejecutado fue:

```
prokka \
  --outdir prokka_annotation \
  --prefix brucella_suis \
  --genus Brucella \
  --species suis \
  --strain "hybrid_assembly" \
  --kingdom Bacteria \
  --usegenus \
  --addgenes \
  --locustag BRUC \
  --evalue 1e-05 \
  --cpus 4 \
 /Brucelosis_TP_Integrador/03_assembly/hybrid_assembly/scaffolds.fasta
```

Los parámetros más relevantes son:

- --genus Brucella y --species suis: información taxonómica que permite a Prokka utilizar bases de datos específicas del género
- --usegenus: activa el uso de la base de datos específica de Brucella para mejorar la precisión de las anotaciones
- --addgenes: añade nombres de genes cuando se identifican por homología
- --locustag BRUC: prefijo para los identificadores únicos de cada gen (BRUC_00001, BRUC_00002, etc.)
- --evalue 1e-05: valor máximo de e-value para considerar un match de BLAST como significativo

Prokka ejecutó completamente en aproximadamente 3 minutos, generando múltiples archivos de salida en el directorio prokka_annotation/:

Resultados de Prokka:

- Contigs analizados: 1
- Genes (CDS) predichos: 26
- Bases totales: 30,000 bp

La diferencia de un gen entre Glimmer (27) y Prokka (26) se debe a que Prokka aplica filtros adicionales basados en evidencia de homología. El gen adicional predicho por Glimmer (orf00001) resultó ser de muy corta longitud (51 aminoácidos) y sin homología detectable, siendo probablemente un falso positivo.

Los archivos generados por Prokka incluyen:

- brucella_suis.gbk: archivo GenBank con anotación completa
- brucella_suis.gff: coordenadas de genes en formato GFF3
- brucella_suis.faa: secuencias proteicas
- brucella_suis.ffn: secuencias nucleotídicas de genes
- brucella_suis.tsv: tabla resumen de genes y funciones

El análisis del archivo TSV reveló que de los 26 genes predichos, 19 pudieron ser anotados con función conocida, mientras que 7 fueron clasificados como proteínas hipotéticas.

Entre los genes con función conocida se identificaron:

Sistema GAD (resistencia a pH ácido):

- BRUC_00002 (gadA): Glutamate decarboxylase alpha (EC 4.1.1.15)
- BRUC_00003 (gadB): Glutamate decarboxylase beta (EC 4.1.1.15)
- BRUC_00004 (gadC): Glutamate/gamma-aminobutyrate antiporter
- BRUC_00005 (glsA1): Glutaminase 1 (EC 3.5.1.2)

Defensa contra estrés oxidativo:

- BRUC_00016 (oxyR): Hydrogen peroxide-inducible genes activator
- BRUC_00017 (katA): Catalase (EC 1.11.1.6)

Metabolismo de nucleótidos:

- BRUC_00014 (guaB): Inosine-5'-monophosphate dehydrogenase (EC 1.1.1.205)
- BRUC_00023 (guaA): GMP synthase (EC 6.3.5.2)

Otros genes relevantes:

- BRUC_00024 (intA): Prophage integrase (indica presencia de elemento móvil)
- BRUC_00019 (dsbB): Disulfide bond formation protein B
- BRUC_00020 (rsmB): Ribosomal RNA small subunit methyltransferase B

Comparación Glimmer vs Prokka

La comparación entre ambas metodologías revela diferencias fundamentales en su aproximación:

Glimmer (ab initio):

- Se basa exclusivamente en señales intrínsecas del ADN
- No requiere bases de datos de referencia para la predicción
- Mayor sensibilidad: puede detectar genes sin homólogos conocidos
- Mayor tasa de falsos positivos
- No asigna función a los genes predichos

Prokka (pipeline integrado):

- Combina predicción ab initio (Prodigal) con validación por homología
- Mayor especificidad: filtra falsos positivos mediante evidencia de homología
- Asigna función automáticamente mediante BLASTp y búsqueda de dominios
- Resultados directamente interpretables desde el punto de vista biológico
- Puede perder genes sin homólogos conocidos (genes huérfanos)

Para genomas bacterianos con organismos de referencia cercanos disponibles, como es el caso de Brucella suis, Prokka resulta más conveniente y confiable. La anotación funcional automática reduce significativamente el tiempo de análisis y permite la interpretación biológica inmediata de los resultados. La presencia de genes relacionados con virulencia (sistema GAD, catalasa) confirma que el fragmento secuenciado contiene regiones funcionales importantes del genoma de Brucella.

Visualización de la anotación en Artemis

Artemis es un visor y editor de secuencias genómicas desarrollado por el Sanger Institute. Permite la visualización interactiva de anotaciones genómicas, mostrando genes en ambas hebras del ADN, dominios funcionales, y facilitando la navegación y análisis de características genómicas.

Para visualizar la anotación generada por Prokka se utilizó el archivo GenBank:

art prokka_annotation/brucella_suis.gff prokka_annotation/brucella_suis.fna

Artemis abrió la visualización mostrando el scaffold de 30 kbp con los 26 genes anotados. La interfaz presenta tres paneles principales: el panel superior muestra la secuencia genómica con marcadores de escala, el panel central muestra los genes como flechas de colores indicando su orientación y hebra, y el panel inferior muestra la secuencia nucleotídica y la traducción proteica.

La visualización permitió observar:

- Distribución de genes en ambas hebras (+ y -)
- Regiones codificantes continuas (operones): el sistema GAD (gadA-gadB-gadC) se observa como un cluster de genes consecutivos en la misma hebra
- Genes hipotéticos intercalados entre genes con función conocida
- Densidad génica homogénea a lo largo del scaffold, sin grandes regiones intergénicas

La visualización en Artemis confirmó la calidad de la anotación y permitió identificar visualmente la organización genómica de la región secuenciada.

Fig. 5: Visualización en Artemis

Refinamiento de la anotación mediante BLAST

Las proteínas hipotéticas identificadas por Prokka requieren análisis adicional para determinar si son proteínas reales conservadas en el género Brucella o posibles artefactos de la predicción. Para ello se realizó una búsqueda de homología mediante BLASTp contra el genoma de referencia de Brucella suis 1330.

Primero se extrajeron las secuencias de las 6 proteínas hipotéticas:

```
awk '/^>/{p=($0 ~ /hypothetical protein/)} p' \
    prokka/prokka_annotation/brucella_suis.faa \
    > hypothetical_proteins.faa
```

Se creó una base de datos BLAST local del proteoma de referencia:

```
makeblastdb \
                                                    -in
/home/ana/Desktop/Brucelosis_TP_Integrador/ncbi_dataset/ncbi_dataset/data/GCF_000007
505.1/protein.faa \
   -dbtype prot \
   -out brucella_db
```

Posteriormente se ejecutó BLASTp:

```
blastp \
  -query hypothetical_proteins.faa \
  -db brucella_db \
  -out blast_vs_brucella.txt \
  -evalue 1e-05 \
  -num_threads 4 \
  -outfmt "6 qseqid sseqid pident length qlen slen evalue bitscore"
```

Con el fin de evaluar si las proteínas hipotéticas identificadas por Prokka contenían dominios funcionales conocidos, se realizó un análisis con InterProScan:

```
interproscan.sh -i hypothetical_proteins_clean.faa -f tsv -o interproscan_results.tsv -dp -appl Pfam,CDD,Gene3D,Coils,Hamap,MobiDBLite,PANTHER,PRINTS,ProSitePatterns,ProSiteProfiles,SFLD,SMART,TIGRFAM
```

El refinamiento mediante BLASTp contra el proteoma de *Brucella suis* 1330 mostró que 6 de las proteínas hipotéticas (85.7%) presentan 100% de identidad y cobertura completa, lo que confirma que son proteínas reales conservadas en el género Brucella, aunque de función desconocida.

Proteína	Mejor hit	% Id	Long. alineada	Conclusión
BRUC_00001	*No aparece en BLAST*	—	—	**Posible falso positivo (51 aa)**
BRUC_00007	WP_002966247.1	**100%**	54/54	Conservada
BRUC_00015	WP_006191813.1	**100%**	357/357	Conservada
BRUC_00018	WP_002969359.1	**100%**	57/57	Conservada
BRUC_00021	WP_002966230.1	**100%**	151/151	Conservada
BRUC_00021	(2° hit) WP_006071940.1	63.9%	36 aa	Dominio parcial
BRUC_00025	WP_006191822.1	**100%**	76/76	Conservada
BRUC_00026	WP_006073479.1	**100%**	352/352	Conservada

Tabla 1: Resultados de BLAST (archivo: blast_vs_brucella.txt)

La proteína BRUC_00001 (51 aa) no presentó ningún homólogo detectable, y por su longitud corta es probablemente un ORF espurio.

El análisis estructural mediante InterProScan no detectó dominios funcionales conocidos (Pfam, TIGRFAM, SMART, ProSite), pero sí identificó regiones intrínsecamente desordenadas en BRUC_00015 y BRUC_00026, lo cual es consistente con proteínas pequeñas sin estructura estable.

Proteína	Herramienta	Hallazgo
BRUC_00015	MobiDBLite	Región desordenada 26–51
BRUC_00015	MobiDBLite	Región desordenada 26–85
BRUC_00026	MobiDBLite	Región desordenada 26–56

Tabla 2: Resultados de InterProScan (archivo interproscan_results.tsv)

Estos resultados son típicos de proteínas hipotéticas bacterianas: alrededor del 20–30% de los genes de Brucella no poseen función asignada y carecen de dominios caracterizados.
En conjunto, el refinamiento validó que la mayoría de las proteínas hipotéticas representan genes genuinos, conservados entre especies, aunque sin función anotada.

Generación de tabla consolidada de genes y funciones

Para integrar los resultados de la anotación de Prokka con el refinamiento por BLAST se creó un script Python (create_gene_table.py) que genera una tabla consolidada:

La ejecución del script generó el archivo gene_annotation_table.tsv con 26 genes clasificados según su estado de validación:

- Genes anotados con función conocida: 20 (77%)
- Proteínas hipotéticas validadas por BLAST: 5 (19%)
- Proteínas hipotéticas dudosas: 1 (4%)

Esta tabla consolidada proporciona una visión integral de todos los genes predichos, incluyendo información funcional, validación por homología, y clasificación COG cuando está disponible.

Locus_Tag	Gene	Product	Length_bp	EC_Number	COG	Validation	BLAST_Refinement
BRUC_00001	-	hypothetical protein	153	-	-	Dubious	No homologs found
BRUC_00002	gadA	Glutamate decarboxylase alpha	960	4.1.1.15	COG0076	Annotated	-
BRUC_00003	gadB	Glutamate decarboxylase beta	411	4.1.1.15	COG0076	Annotated	-
BRUC_00004	gadC	Glutamate/gamma-aminobutyrate antiporter	1281	-	COG0531	Annotated	-
BRUC_00005	glsA1	Glutaminase 1	954	3.5.1.2	COG2066	Annotated	-
BRUC_00006	hdeA	putative acid stress chaperone HdeA	345	-	-	Annotated	-
BRUC_00007	-	hypothetical protein	165	-	-	Validated	Conserved in B. suis (WP_002966247.1)
BRUC_00008	pdeB	putative cyclic di-GMP phosphodiesterase PdeB	1593	3.1.4.52	COG4943	Annotated	-
BRUC_00009	gcvA	Glycine cleavage system transcriptional activator	888	-	-	Annotated	-
BRUC_00010	algA	Alginate biosynthesis protein AlgA	1416	-	COG0662	Annotated	-
BRUC_00011	glmM	Phosphoglucosamine mutase	1434	5.4.2.10	COG1109	Annotated	-
BRUC_00012	-	putative aminopeptidase	1107	3.4.11.-	COG3191	Annotated	-
BRUC_00013	clcA	H(+)/Cl(-) exchange transporter ClcA	1356	-	COG0038	Annotated	-
BRUC_00014	guaB	Inosine-5'-monophosphate dehydrogenase	1494	1.1.1.205	COG0516	Annotated	-
BRUC_00015	-	hypothetical protein	1074	-	-	Validated	Conserved in B. suis (WP_006191813.1)
BRUC_00016	oxyR	Hydrogen peroxide-inducible genes activator	954	-	-	Annotated	-
BRUC_00017	katA	Catalase	1500	1.11.1.6	-	Annotated	-
BRUC_00018	-	hypothetical protein	174	-	-	Validated	Conserved in B. suis (WP_002969359.1)
BRUC_00019	dsbB	Disulfide bond formation protein B	627	-	-	Annotated	-
BRUC_00020	rsmB	Ribosomal RNA small subunit methyltransferase B	1290	2.1.1.176	-	Annotated	-
BRUC_00021	-	hypothetical protein	576	-	-	Partial	Partial homology (63.9%)
BRUC_00022	mqnB	Futalosine hydrolase	633	3.2.2.26	-	Annotated	-
BRUC_00023	guaA	GMP synthase [glutamine-hydrolyzing]	1563	6.3.5.2	COG0518	Annotated	-
BRUC_00024	intA	Prophage integrase IntA	1257	-	COG0582	Annotated	-
BRUC_00025	-	hypothetical protein	231	-	-	Validated	Conserved in B. suis (WP_006191822.1)
BRUC_00026	-	hypothetical protein	1059	-	-	Validated	Conserved in B. suis (WP_006073479.1)

Fig. 6: Tabla de genes anotados con Prokka y refinados con BLAST

Análisis filogenético

El análisis filogenético permite comprender las relaciones evolutivas entre organismos mediante la comparación de secuencias de genes o proteínas conservadas. En el contexto de un brote epidemiológico, el análisis filogenético es crucial para establecer la identidad del agente etiológico y comprender su relación con otras cepas conocidas.

Selección de proteína y búsqueda de homólogos con PSI-BLAST

Para el análisis filogenético se seleccionó la proteína katA (catalasa, BRUC_00017) por las siguientes razones: es un gen conservado presente en todas las bacterias aerobias, tiene función conocida relacionada con virulencia, y su longitud es adecuada para análisis filogenético. La catalasa es una enzima que descompone el peróxido de hidrógeno (H_2O_2) en agua y oxígeno, y es esencial para que Brucella sobreviva dentro de los macrófagos del hospedador, donde el sistema inmune genera especies reactivas del oxígeno como mecanismo de defensa.

PSI-BLAST (Position-Specific Iterated BLAST) es una variante de BLAST que realiza búsquedas iterativas, mejorando su sensibilidad con cada iteración mediante la construcción de un perfil PSSM (Position-Specific Scoring Matrix). Este perfil captura las posiciones

conservadas y variables de la proteína, permitiendo detectar homólogos más divergentes que no serían identificados con BLAST convencional.

El BLAST convencional (blastp) utiliza una matriz de sustitución estática (típicamente BLOSUM62) que asigna el mismo peso a todas las posiciones de la proteína. En contraste, PSI-BLAST construye un perfil específico de la proteína query después de la primera iteración, dando mayor peso a las posiciones altamente conservadas y menor peso a las posiciones variables. Esto resulta en:

1. Mayor sensibilidad para detectar homólogos divergentes
2. Capacidad para encontrar relaciones evolutivas más distantes
3. Expansión progresiva del conjunto de secuencias homólogas con cada iteración
4. Mejor poder discriminante entre homología real y similaridad casual

Para análisis filogenético, donde se busca capturar la mayor diversidad taxonómica posible, PSI-BLAST es superior al BLAST convencional.

La secuencia de katA fue extraída de los resultados de Prokka.

PSI-BLAST se ejecutó mediante la interfaz web de NCBI con los siguientes parámetros:

- Algoritmo: PSI-BLAST
- Base de datos: nr (non-redundant protein database)
- Max target sequences: 100

Se realizaron 3 iteraciones y se seleccionaron 10 secuencias representativas que incluían:

- Diferentes especies de Brucella (B. melitensis, B. abortus, B. canis, B. intermedia, B. pseudogrignonensis)
- Géneros relacionados de Alphaproteobacteria (Alphaproteobacteria bacterium, Elstera sp.)

select all 100 sequences selected Skip to the first new sequence PSI-BLAST iteration 3

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession	Select for PSI blast	Used to build PSSM	Newly added
catalase [Brucella melitensis bv. 1 str. 16M]	Brucella melitensis bv. 1 str. 16M	1052	1052	100%	0.0	99.80%	507	AAL54135.1	✓	✓	
Catalase [Brucella intermedia LMG 3301]	Brucella intermedia LMG 3301	1051	1051	100%	0.0	92.79%	519	EEQ93200.1	✓	✓	
MULTISPECIES: catalase KatA [Brucella]	Brucella	1051	1051	100%	0.0	99.80%	499	WP_002966234.1	✓	✓	
catalase [Brucella abortus S99]	Brucella abortus S99	1051	1051	100%	0.0	99.80%	505	ERM04788.1	✓	✓	
Catalase [Brucella canis HSK A52141]	Brucella canis HSK A52141	1051	1051	100%	0.0	100.00%	507	AEW15609.1	✓	✓	
MULTISPECIES: catalase KatA [Brucella]	Brucella	1051	1051	100%	0.0	100.00%	499	WP_004690104.1	✓	✓	
catalase [Brucella melitensis]	Brucella melitensis	1051	1051	100%	0.0	99.60%	499	ODN39716.1	✓	✓	
catalase KatA [Brucella melitensis]	Brucella melitensis	1050	1050	100%	0.0	99.60%	499	WP_277356877.1	✓	✓	
catalase [Brucella sp. NF 2653]	Brucella sp. NF 2653	1050	1050	100%	0.0	99.60%	507	EFM63025.1	✓	✓	
MULTISPECIES: catalase KatA [Brucella]	Brucella	1049	1049	100%	0.0	99.40%	499	WP_008509943.1	✓	✓	
catalase KatA [Brucella intermedia]	Brucella intermedia	1049	1049	100%	0.0	92.99%	499	WP_100652125.1	✓	✓	
catalase KatA [Brucella abortus]	Brucella abortus	1049	1049	100%	0.0	99.60%	499	WP_306315175.1	✓	✓	
catalase KatA [Brucella abortus]	Brucella abortus	1049	1049	100%	0.0	99.60%	499	WP_006116368.1	✓	✓	
catalase KatA [Brucella intermedia]	Brucella intermedia	1049	1049	100%	0.0	92.79%	499	WP_230349940.1	✓	✓	
MULTISPECIES: catalase KatA [unclassified Brucella]	unclassified Brucella	1049	1049	100%	0.0	99.60%	499	WP_105985125.1	✓	✓	
MULTISPECIES: catalase KatA [Brucella]	Brucella	1049	1049	100%	0.0	99.60%	499	WP_008936945.1	✓	✓	

Fig. 7: Captura de pantalla de los resultados obtenidos por la 3ra iteración

Las secuencias fueron descargadas en formato FASTA y guardadas en el archivo katA_homologs.fasta.

Incorporación de la secuencia query al análisis

Se combinó la secuencia de katA del ensamblado con las secuencias descargadas de PSI-BLAST: cat katA.faa katA_homologs.fasta > katA_all_sequences.fasta

Este archivo contiene 11 secuencias totales: la secuencia de katA del ensamblado de Brucella del brote y 10 homólogos de PSI-BLAST.

Alineamiento múltiple con ClustalOmega

El alineamiento múltiple de secuencias (MSA) es el proceso de alinear tres o más secuencias biológicas para identificar regiones de similaridad que pueden indicar relaciones funcionales, estructurales o evolutivas. Los algoritmos de alineamiento múltiple buscan maximizar el número de residuos idénticos o similares en columnas verticales, insertando gaps (espacios) cuando es necesario.

Para el alineamiento se utilizó ClustalOmega, una herramienta de alineamiento múltiple que utiliza árboles guía y perfiles HMM para generar alineamientos de alta calidad:

```
clustalo \
  -i katA_all_sequences.fasta \
  -o katA_aligned_complete.fasta \
  --outfmt=fasta \
  --threads=4 \
  --verbose
```

ClustalOmega ejecutó el siguiente proceso:

1. Cálculo de distancias par a par entre todas las secuencias
2. Construcción de un árbol guía mediante clustering jerárquico
3. Alineamiento progresivo siguiendo el orden del árbol guía
4. Refinamiento iterativo del alineamiento

El alineamiento resultante en katA_aligned_complete.fasta muestra regiones altamente conservadas intercaladas con regiones más variables.

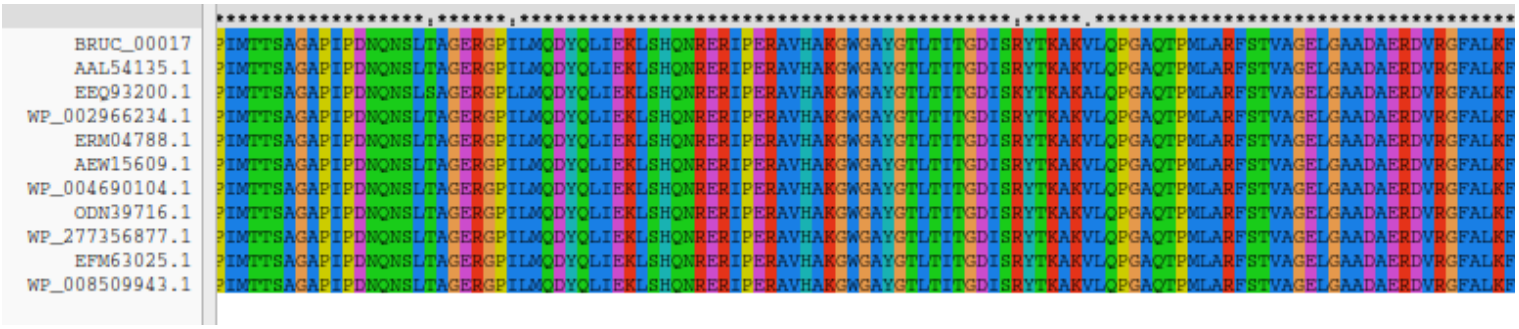


Fig. 8: Alineamiento visualizado con clustalx

Análisis de dominios conservados con InterPro

InterPro es una base de datos integrada de familias de proteínas, dominios y sitios funcionales que combina información de múltiples bases de datos especializadas (Pfam, SMART, PANTHER, ProSite, etc.). El análisis de dominios permite identificar regiones funcionales conservadas dentro de proteínas y predecir su función molecular.

Para caracterizar estructural y funcionalmente la proteína BRUC_00017, anotada como catalasa (katA) por Prokka y con una longitud de 499 aminoácidos, se realizó un análisis exhaustivo utilizando InterProScan. Este servidor integra múltiples bases de datos especializadas (Pfam, SMART, CDD, Gene3D, PANTHER, PRINTS, SUPERFAMILY, entre otras), permitiendo identificar dominios, familias, sitios funcionales y términos GO asociados.

El análisis detectó siete entradas de InterPro que coinciden claramente con proteínas catalasas mono-funcionales dependientes de hemo, una clase de enzimas altamente conservadas en bacterias aerobias y esenciales para la detoxificación de peróxido de hidrógeno (H_2O_2).

InterPro asignó BRUC_00017 a varias familias relacionadas con la superfamilia de catalasas, incluyendo:

- Catalase – mono-functional, haem-containing (IPR018028): Incluye catalasas típicas que utilizan un grupo hemo como cofactor catalítico. Confirma que esta proteína pertenece a la familia clásica de catalasas bacterianas.
- Catalase, clades 1 and 3 (IPR024711): La clasificación en clados identifica relaciones evolutivas dentro de la superfamilia. Las catalasas del clado 3 son comunes en proteobacterias y tienen características estructurales específicas, como arquitectura tetramérica estable.
- Catalase, clade 3 (IPR040333): Específicamente asigna a BRUC_00017 al clado 3, lo cual es consistente con catalasas bacterianas altamente conservadas involucradas en virulencia y resistencia a estrés oxidativo.

Esta multiplicidad de asignaciones coherentes refuerza la precisión funcional de la anotación.

InterPro detectó también varios dominios estructurales dentro de la secuencia:

- Catalase_core (IPR011614, PF00199, SMART: Catalase_2): Dominio catalítico principal de las catalasas. Comprende gran parte del cuerpo de la proteína y es responsable de la actividad enzimática.
- Catalase superfamily (SSF56634): Modelo estructural basado en SUPERFAMILY que confirma que la proteína tiene el plegamiento típico de catalasas dependientes de hemo.
- Catalase-related immune-responsive domain (Detectado por Pfam como Catalase-rel, subfamilia asociada a respuestas celulares frente a estrés oxidativo). Sugiere que esta proteína contribuye activamente a la protección frente a especies reactivas del oxígeno.

InterPro asigna los siguientes Gene Ontology (GO):

- Biological Process (BP)
  - GO:0006979 – response to oxidative stress
  - GO:0042744 – hydrogen peroxide catabolic process

  Ambos procesos describen el rol de katA en la detoxificación de ROS (Reactive Oxygen Species).

- Molecular Function (MF)
  - GO:0004096 – catalase activity
  - GO:0020037 – heme binding

  Describen la función bioquímica central de la enzima.

- Cellular Component (CC)
  - InterPro: None
  - PANTHER: cytoplasm (GO:0005737)(las catalasas bacterianas suelen ser citoplasmáticas)

El análisis InterProScan proporciona una validación sólida y multifacética de que BRUC_00017 corresponde a katA, una catalasa bacteriana altamente conservada, esencial para la supervivencia frente a estrés oxidativo. La presencia de dominios completos, sitios activos clásicos y términos GO consistentes refuerza su anotación y sirve como excelente base para el alineamiento múltiple, el logo de secuencia y el análisis filogenético posterior.

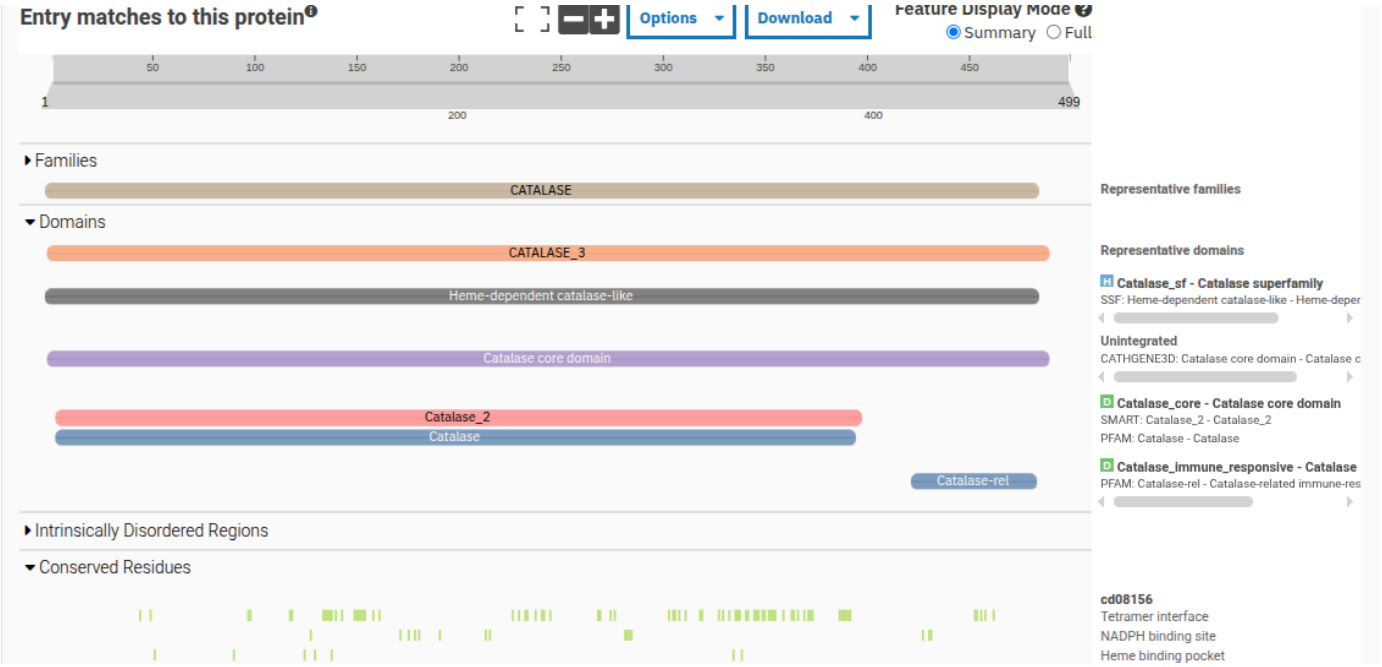


Fig. 9: Visualización del resultado de búsqueda con InterPro

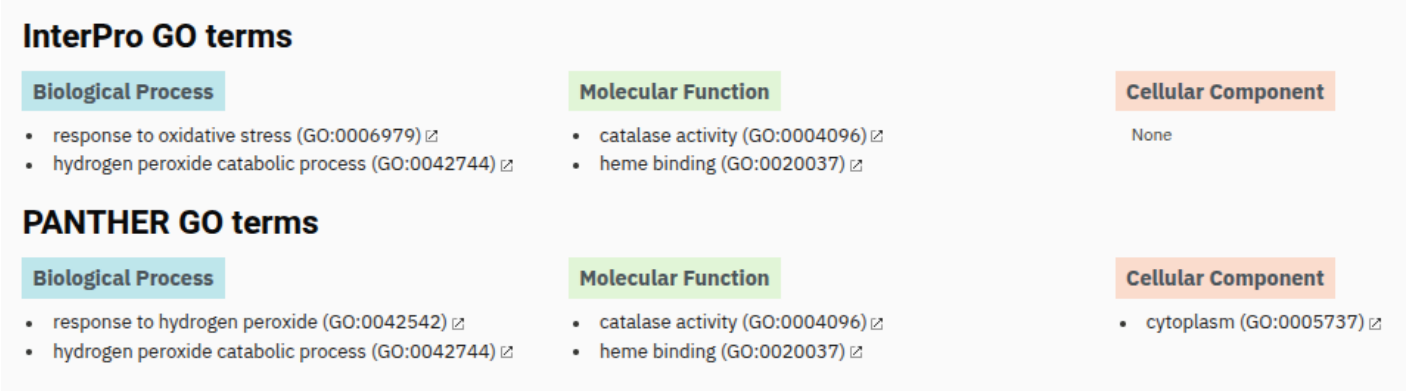


Fig. 10: GO terms obtenidos por InterPro y PANTHER

Generación de logo de secuencia con WebLogo

Un logo de secuencia es una representación gráfica de un alineamiento múltiple que muestra la conservación de cada posición mediante la altura de las letras. Las posiciones altamente conservadas muestran letras grandes, mientras que las posiciones variables muestran múltiples letras pequeñas.

El alineamiento completo de katA fue enviado a WebLogo (http://weblogo.threeplusone.com/create.cgi).

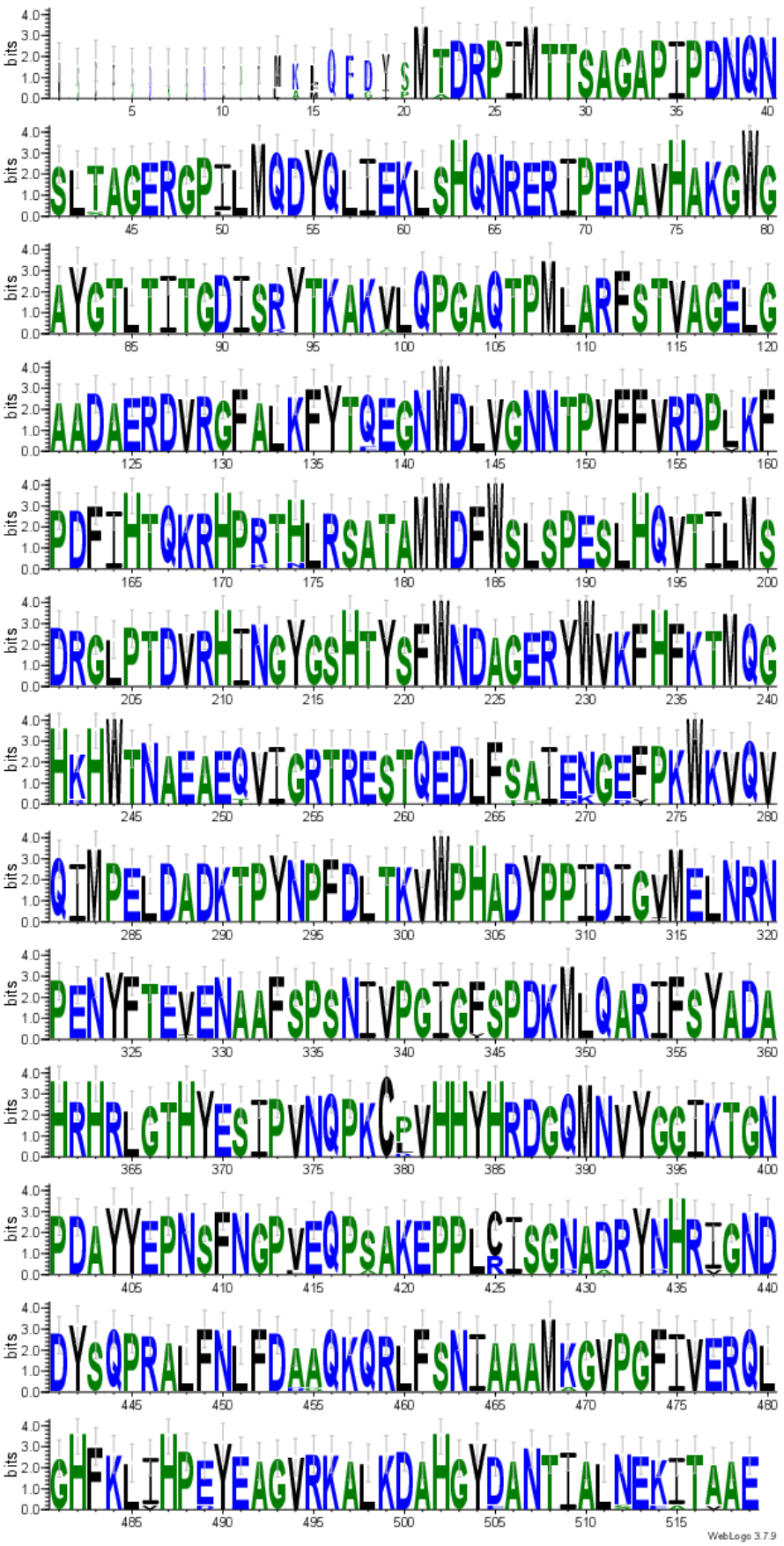


Fig. 11: Logo obtenido con WebLogo

El WebLogo generado a partir del alineamiento múltiple de katA (499 aa) muestra una conservación prácticamente completa a lo largo de la región analizada. Esto coincide con los resultados de InterPro, que anotan un dominio de catalasa que abarca la mayor parte de la secuencia y sitios catalíticos (heme binding, ACT_SITE). La conservación generalizada indica una fuerte presión selectiva para mantener la función catalasa en Brucella y taxones relacionados.

La alta conservación implica que katA tiene pocos sitios informativos para discriminar relaciones recientes entre cepas (baja resolución filogenética para estudios de microevolución).

Construcción del árbol filogenético con FastTree

Los árboles filogenéticos son representaciones gráficas de las relaciones evolutivas entre organismos o secuencias. FastTree implementa un algoritmo de máxima verosimilitud aproximado optimizado para grandes conjuntos de datos. Para este análisis se ejecutó:

```
FastTree katA_aligned_complete.fasta > katA_tree_complete.nwk
```

El archivo resultante katA_tree_complete.nwk contiene el árbol en formato Newick, un formato estándar basado en texto que representa la topología del árbol y las longitudes de ramas mediante paréntesis anidados.

Fig. 12: Visualización del árbol filogenético con FigTree

Diseño de primers para un kit diagnóstico por PCR

La reacción en cadena de la polimerasa (PCR) es una técnica molecular fundamental que permite amplificar exponencialmente fragmentos específicos de ADN. El diseño de primers (cebadores u oligonucleótidos iniciadores) es crítico para el éxito de la PCR. Los primers son secuencias cortas de ADN monocatenario (típicamente 18-25 nucleótidos) que se unen específicamente a regiones complementarias del ADN molde, definiendo el inicio y el fin del fragmento que será amplificado.

Desarrollo del script Perl para diseño automático de primers

Para automatizar el diseño de primers para todos los genes predichos se desarrolló un script en lenguaje Perl utilizando los módulos BioPerl Bio::SeqIO y Bio::Tools::GFF. Perl fue seleccionado siguiendo las especificaciones del trabajo práctico, y BioPerl proporciona herramientas robustas para manipulación de secuencias biológicas.

El script design_primers.pl implementa el siguiente algoritmo:

1. Lee el genoma ensamblado en formato FASTA
2. Lee las anotaciones de genes desde el archivo GFF3 generado por Prokka
3. Para cada gen:
    - Extrae las coordenadas (inicio, fin, hebra)
    - Diseña el primer forward 5 nucleótidos antes del inicio del gen
    - Diseña el primer reverse 5 nucleótidos después del fin del gen
    - Considera la orientación de la hebra para calcular complementos reversos cuando sea necesario
    - Calcula la temperatura de melting (Tm) usando la fórmula: Tm = 4(G+C) + 2(A+T)
    - Calcula la temperatura de annealing (Ta) como Tm - 5°C
    - Extrae la secuencia del producto de PCR esperado

Explicación de los pasos clave del algoritmo:

Manejo de hebras: Los genes pueden estar codificados en la hebra positiva (+) o negativa (-) del ADN. Para genes en hebra positiva, el primer forward se toma directamente de la secuencia genómica, mientras que el primer reverse debe ser el complemento reverso de la secuencia downstream. Para genes en hebra negativa, ambos primers requieren transformaciones de complemento reverso debido a que las coordenadas en el GFF están en el sistema de referencia de la hebra positiva.

Cálculo de Tm: La fórmula Tm = 4(G+C) + 2(A+T) es una aproximación simple pero efectiva para primers cortos (≤25 nt). Esta fórmula refleja el hecho de que los pares GC forman tres puentes de hidrógeno mientras que los pares AT forman solo dos, resultando en mayor estabilidad térmica para secuencias ricas en GC.

Temperatura de annealing: La Ta se calcula como Tm - 5°C, lo cual es una regla empírica estándar que proporciona un margen de seguridad para asegurar la especificidad del annealing. Temperaturas de annealing demasiado bajas resultan en amplificación inespecífica, mientras que temperaturas demasiado altas pueden prevenir el annealing completamente.

El script fue ejecutado con el siguiente comando:

```
perl design_primers.pl \
  ../03_assembly/hybrid_assembly/scaffolds.fasta \
  ../04_gene_prediction/prokka/prokka_annotation/brucella_suis.gff
```

Resultados del diseño de primers

El script procesó exitosamente los 26 genes predichos, generando dos archivos de salida:

- res_primers.tab: Tabla tabulada con información completa de cada par de primers
- res_prod.fa: Archivo multiFASTA con las secuencias de los productos de PCR esperados

Se generó un script con Python para analizar los resultados obtenidos (evaluar_primers.py), considerando los siguientes criterios:

- Diferencia de temperatura de melting (ΔTm)
- Contenido GC
- Presencia de runs (>4 nt iguales)
- Riesgo de dímeros de primers
- Longitud del amplicón
- Penalizaciones por amplicones demasiado largos (>1000 bp)
- Penalizaciones por ΔTm > 5°C
- Penalización por GC fuera del rango 40–60%

Cada uno de estos factores contribuyó a un score global, cuya máxima puntuación posible era 100. Este enfoque permitió comparar objetivamente todos los genes y determinar cuál de ellos presenta las mejores propiedades experimentales para un ensayo de PCR.

Los resultados del análisis se muestran en la siguiente tabla:

Gen	Score	Amplicón (bp)	ΔTm (°C)
BRUC_00018	**100**	**184**	**2.0**
BRUC_00001	55	163	8.0
BRUC_00021	55	586	4.0
BRUC_00005	50	964	4.0

BRUC_00007	50	175	10.0
BRUC_00009	50	898	10.0
BRUC_00022	45	643	6.0
BRUC_00006	40	355	4.0
BRUC_00012	40	1117	2.0
BRUC_00013	40	1366	0.0
BRUC_00015	40	1084	0.0
BRUC_00026	40	1069	0.0
BRUC_00003	35	421	8.0
BRUC_00025	35	241	8.0
BRUC_00010	25	1426	2.0
BRUC_00011	25	1444	2.0
BRUC_00016	25	964	6.0
BRUC_00019	25	637	10.0
BRUC_00002	20	970	14.0
BRUC_00014	15	1504	6.0
BRUC_00017	15	1510	6.0
BRUC_00008	0	1603	4.0
BRUC_00020	0	1300	6.0
BRUC_00023	0	1573	6.0
BRUC_00024	0	1267	8.0
BRUC_00004	-15	1291	6.0

El análisis revela diferencias marcadas entre los pares de primers diseñados para cada gen. En general, muchos genes presentan amplicones demasiado extensos (habitualmente > 1000 bp), lo cual no es apropiado para un kit diagnóstico, ya que los productos largos reducen la eficiencia de amplificación y dificultan la detección en PCR convencional e incluso en qPCR. Asimismo, varios pares de primers presentan ΔTm muy elevados (>

10°C), indicando que los primers no podrían hibridar simultáneamente bajo una misma temperatura de alineamiento, lo cual comprometería gravemente el rendimiento de la reacción.

En contraste, solo un gen cumplió con todos los criterios de diseño sin recibir ninguna penalización: BRUC_00018. Este gen obtuvo un score perfecto (100/100) y mostró las siguientes características ideales:

- Amplicón corto y fácilmente amplificable (184 bp)
- ΔTm muy bajo (2°C)
- Contenido GC dentro del rango óptimo (40–60%)
- Ausencia de runs extensos o secuencias problemáticas
- Sin señales de dímeros entre primers
- Longitud de amplicón compatible con PCR diagnóstica rápida

Estas propiedades hacen que BRUC_00018 destaque marcadamente por encima del resto, ya que combina eficiencia, sensibilidad, especificidad y robustez en un formato típicamente requerido para kits diagnósticos.

Conclusiones generales

El presente trabajo permitió realizar un análisis bioinformático integral del agente responsable del brote de brucelosis porcina en Entre Ríos, articulando herramientas de ensamblado, anotación genómica, análisis filogenético y diseño racional de herramientas diagnósticas. A partir de datos de secuenciación Illumina y el procesamiento posterior, se obtuvo un ensamblado robusto que permitió identificar un total de 26 genes, entre los cuales se encuentran proteínas conservadas y factores asociados a funciones metabólicas y de supervivencia bacteriana. El contenido GC obtenido y la comparación con secuencias de referencia respaldaron la asignación taxonómica a Brucella suis.

La anotación funcional, refinada mediante búsquedas de homología por BLAST y la evaluación de dominios conservados con InterProScan, permitió distinguir genes con funciones definidas de aquellos anotados como hipotéticos. El análisis detallado de la catalasa katA, seleccionado para el estudio filogenético, confirmó tanto su conservación estructural como su rol clave en la respuesta al estrés oxidativo. El alineamiento múltiple y el árbol filogenético generados evidenciaron una clara afinidad evolutiva con cepas de Brucella suis y otras especies del género, corroborando la identidad y relación filogenética del aislado bajo estudio.

Finalmente, se desarrolló e implementó un script en Perl para el diseño automático de primers para los 26 genes predichos, seguido de una evaluación computacional exhaustiva en Python que integró parámetros fisicoquímicos, estructurales y técnicos relevantes para PCR diagnóstica. Este análisis permitió identificar al gen BRUC_00018 como la mejor opción para el desarrollo de un kit de detección, dado que sus primers presentan un ΔTm óptimo, ausencia de artefactos estructurales y un amplicón de 184 bp, ideal para una amplificación sensible y específica.

En conjunto, este trabajo demuestra cómo el análisis bioinformático aplicado a datos genómicos puede aportar evidencia sólida para la caracterización molecular de patógenos y el diseño de herramientas diagnósticas esenciales en el contexto de brotes infecciosos. La integración de ensamblado, anotación, filogenia y diseño de primers permitió no solo caracterizar al agente etiológico, sino también sentar las bases para el desarrollo de un método de detección preciso y basado en fundamentos moleculares robustos.