UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

FACULTAD DE CONTADURÍA, ADMINISTRACIÓN E INFORMÁTICA

MAESTRÍA EN OPTIMIZACIÓN Y CÓMPUTO APLICADO

Herramienta computacional para la caracterización de matrices CRISPR

T E S I S

Que para obtener el Grado de

Maestra en Optimización y Cómputo Aplicado

Presenta

EDNA CRUZ FLORES

Director de Tesis:

Dra. Lorena Díaz González

Co-Director:

Dra. Blanca Itzelt Taboada Ramírez

Revisores:

Dr. José Alberto Hernández Aguilar

Dr. Luis Manuel Gaggero Sager

Dr. Mauricio Rosales Rivera

Dra. Lorena Díaz González

Dra. Blanca Itzelt Taboada Ramírez

CUERNAVACA, MORELOS ABRIL, 2022

Cuernavaca, Morelos a 1 de mayo del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, de la estudiante Edna Cruz Flores, con matrícula 10034119, con el título **Herramienta computacional para la caracterización de matrices CRISPR** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Lorena Díaz González
Profesor- investigador
Centro de Investigación en Ciencias

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LORENA DIAZ GONZALEZ | Fecha:2022-05-01 13:24:37 | Firmante

JSOzE7WSebJa4/U2WjloueB+Cfdb1Zf9/dsGE4vCLckcsdoUNJtnB75t3EBtrcCdbAzi6yqKsT5iN3rwI0Q5w/11JfIOXbHp1ZFoHYU5MHtQepO8lmRBD+l++rhXTFDJL+oPMbeBqKP6IwiodSf1s4cyUbRrDeMJAPpQrCCzf3yTa9INTRzOt8CxgxoHyLz302bj45G2BsrVcuCtgUBl+IuOWstr2bPK3b7J9QfxsDw72Rw9p1a/PqlbhDiSk2ksqWcN0lA7K8kJCR1/WSk6bclAOWC3t509eZLGBfzG/jCjtm3aJJVzkUnn9E9kAfiacLC3IokqC1NXrk1vGVuphg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

pEVIDyd6m

https://efirma.uaem.mx/noRepudio/2YCCPQXPYKkw0Mi7SPlitHbde5rghY3P

Cuernavaca, Morelos a 1 de mayo del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, de la estudiante Edna Cruz Flores, con matrícula 10034119, con el título **Herramienta computacional para la caracterización de matrices CRISPR** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dra. Blanca Itzelt Taboada Ramírez
Profesora- investigadora
Instituto de Biotecnología, UNAM

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

BLANCA ITZELT TABOADA RAMÍREZ | Fecha:2022-05-02 11:22:49 | Firmante

c0QpGGOnvvxU1GowafV6qANIKh4lHtErZ++bi7+D95Nr4TMbr0jscC+GMVMZIaepz+oMw9fXOZH8cuBKAYJGX/zDEoJFW0T6DGN6DlA+pPZTEnWKTtBfOnoS/vbdoxu33XXD8vIkfsgGmRqAG3MUoAC1pS8+8fZxmiQGjJFFZ7uXekMZ1ouIoc+bLSwiqGB1kXxbjP4HbElNPZpRuzANQw6/WFtSL3smQPfP0KggcQyax1b3obEKCmQDPxDn9+/5D8wEVtsrB6ZGp6Q+jFhQgeZ3SuvAf6Ar1CgMWVuCyGJkSOcyzCUkF2JB5+jnyiKgUBMf/hGGCeyTwxOBydmNCg==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

6kANMoPGe

https://efirma.uaem.mx/noRepudio/fluWIh5rxoCwqydVnqUBhs5LSg2kKOdP

Cuernavaca, Morelos a 29 de Abril del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, de la estudiante Edna Cruz Flores, con matrícula 10034119, con el título **Herramienta computacional para la caracterización de matrices CRISPR** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. José Alberto Hernández Aguilar
Profesor- investigador
Facultad de Contaduría, Administración e Informática

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

JOSE ALBERTO HERNANDEZ AGUILAR | Fecha:2022-04-29 11:47:29 | Firmante

Jnso3tlyRFnfDVDgdh9DqgfDXGeTNekL4tHPLTYPO2J9QxmCrGZ/GAJTphlQ/8Ed0/8EY/a4zRN3nue4siJ9svgJayi0kSXxw/hl0U5sgW9Ezge9fXpJ3MYL3Z7xchZxjysVmJR81AzeRR9OMe64yDI04NNE2D1lAi1glqWcjYgQfSjqd/FAegiP6+S9eGqPlJumg8Njb2O6rAQRolff9SxGmowb8F6wigTEB0oywLu+KwadckD/Nzqtwn6c2p+F8UVwLYRCSXPuOfpjbj6k96oCl3aUJVzV8ZvF3lft0WGFvsNtEPvwpysbri/JbXsJoikrrth53B4bGzKzOyIfbQ==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

QVvWYcU4d

https://efirma.uaem.mx/noRepudio/QsrKD8wRKqipcpszOmsJCjKknTBBj6c8

Cuernavaca, Morelos a 1 de mayo del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, de la estudiante Edna Cruz Flores, con matrícula 10034119, con el título **Herramienta computacional para la caracterización de matrices CRISPR** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Luis Manuel Gaggero Sager
Profesor- investigador
Facultad de Contaduría, Administración e Informática

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

LUIS MANUEL GAGGERO SAGER | Fecha:2022-05-01 10:32:27 | Firmante

Rwez/J5M0YOt9DuNuBnjufjMIS7CPI/km1mMwn1MTjXfMfD695q9jtscJiNhoRIQ70lZKk1oqX3iGfyy+Dh0nXmIxAHCDSgSwBIXDTmfg1lAX+oDvYFwQX2zAmefv OysElJvBL5NCl7 aA/tlj7QmH4D5D+/BncBmKi1AaGuLzt4A4e1YsuVpMAR55DpfKX+mE21F62ZVU0HJS2Mwsx2wyqFYfAQPc4aVG1pCx8zUzYxw9/Ti7qcuofL8UrLSZ3WKj0tPCH/Va0q5l4KNE1+ KkZV1iXQcKDeXBgsnH2QbkRWqxicHVFWjqUIcLcm8bLVidmymR30nCyD+NgtaUGpapw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

GX4YdJN8B

https://efirma.uaem.mx/noRepudio/ehkbYiEg6NZOKsHgLytF0p8LsGvelDoc

Cuernavaca, Morelos a 1 de mayo del 2022.

DR. AUGUSTO RENATO PÉREZ MAYO
SECRETARIO DE INVESTIGACIÓN DE LA F.C.A.e I.
PRESENTE

En mi carácter de revisor de Tesis, hago de su conocimiento que he leído con interés la tesis para obtener el grado de la Maestría en Optimización y Cómputo Aplicado, de la estudiante Edna Cruz Flores, con matrícula 10034119, con el título **Herramienta computacional para la caracterización de matrices CRISPR** por lo cual, me permito informarle que después de una revisión cuidadosa de dicha tesis, concluyo que el trabajo se caracteriza por el establecimiento de objetivos académicos pertinentes y una metodología adecuada para su logro. Además construye una estructura coherente y bien documentada, por lo cual considero que los resultados obtenidos contribuyen al conocimiento del tema tratado.

Con base en los argumentos precedentes me permito expresar mi **VOTO APROBATORIO** por lo que de mi parte no existe inconveniente para que el estudiante continúe con los trámites que esta Secretaría de Investigación tenga establecidos para obtener el grado mencionado.

Atentamente
Por una humanidad culta
Una universidad de excelencia

Dr. Mauricio Rosales Rivera
Profesor- investigador
Centro de Investigación en Ciencias

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

Se expide el presente documento firmado electrónicamente de conformidad con el ACUERDO GENERAL PARA LA CONTINUIDAD DEL FUNCIONAMIENTO DE LA UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS DURANTE LA EMERGENCIA SANITARIA PROVOCADA POR EL VIRUS SARS-COV2 (COVID-19) emitido el 27 de abril del 2020.

El presente documento cuenta con la firma electrónica UAEM del funcionario universitario competente, amparada por un certificado vigente a la fecha de su elaboración y es válido de conformidad con los LINEAMIENTOS EN MATERIA DE FIRMA ELECTRÓNICA PARA LA UNIVERSIDAD AUTÓNOMA DE ESTADO DE MORELOS emitidos el 13 de noviembre del 2019 mediante circular No. 32.

Sello electrónico

MAURICIO ROSALES RIVERA | Fecha:2022-05-01 11:20:46 | Firmante

tnaACIz9/ISY+vsm9YNYiYO9VZxzR5Pbb0agtAYB/OCuByPZ0RA3p3UMuRpSTKtET+OllWdNiDKLGxb1y9pqW3V1dwqEH5THvI7UmfXXOPKcIGfkvuh3DA+dMH96Z2O87hA0iL
wAKetAJDsS6mbwKdNpRit6tKKAgbnQ4GY714Nadtb/3Owc5is662JVPib6VWTVUqj+kai49oYmCJWQWgtf/hxfjvzxq+e6u3sTIL4/uniWO8hnq3hXmYoUNF8I01OF2Sroy+FjHLskv
CcfXwQFUbHmGY9Jy48kG0jOlNCb4eZHJ5UP+10a6zRnBhLwGt+nwxcC1VphvTnFbtrvFw==

Puede verificar la autenticidad del documento en la siguiente dirección electrónica o escaneando el código QR ingresando la siguiente clave:

jd6QtkYAS

https://efirma.uaem.mx/noRepudio/UOElKQdTAFmiNmNWvu8ZINkrWt9u9ZN9

Agradecimientos

Doy gracias a Dios, por su amor inagotable, por ser mi guía y mi fortaleza.

Les agradezco a mis padres, hermanos y familia por su constante apoyo, por sus ánimos y cariño. Por ser mi más grande motivación.

Agradezco a la Dra. Lorena Díaz González y a la Dra. Blanca Itzelt Taboada Ramírez por haberme brindado la oportunidad de desarrollar el presente proyecto bajo su asesoría, por el conocimiento transmitido que fue imprescindible para la culminación del mismo; por su tiempo y confianza. Qué orgullo haber trabajado junto a dos grandes investigadoras. Las admiro.

Asimismo, agradezco a mi comité evaluador, a los Drs. José Alberto Hernández Aguilar, Luis Manuel Gaggero Sager y Mauricio Rosales Rivera, por la disposición en brindar sus puntuales observaciones y comentarios que contribuyeron al presente proyecto.

Agradezco al Grupo Arias-López del IBT-UNAM por el apoyo y la confianza brindada para trabajar en su laboratorio, otorgándome acceso al Clúster Teopanzolco de la UNAM.

Agradezco al cuerpo docente que integra la Maestría en Optimización y Cómputo Aplicado en la FCAeI-UAEM y al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de posgrado otorgada.

Agradezco a mis compañeros de posgrado Elizabeth Cadenas, Lizbeth Montiel, Víctor Saucedo y Víctor Pacheco por compartirme sus experiencias y conocimientos.

.

“Pues Dios no nos ha dado un espíritu de temor
y timidez sino de poder, amor y autodisciplina”
2 Timoteo 1:7 NTV

Resumen

Los bacteriófagos o fagos, virus que infectan a las bacterias, son considerados como una potencial herramienta para modular a las comunidades de bacterias, ya sea confiriéndoles ventajas de supervivencia o disminuyendo su población e incluso para erradicarlas. Una de las limitantes para estas aplicaciones es el poco conocimiento de las interacciones infecciosas entre los bacteriófagos y sus bacterias hospederas.

Algunos de los enfoques hacen uso de los espaciadores CRISPR como señales de interacción para establecer las relaciones fago – bacteria hospedera, siendo un método muy prometedor. Sin embargo, las predicciones de interacciones a través de los espaciadores no han sido tan favorables (Dion et al., 2021; Edwards et al., 2016; Zhang et al., 2019) dado que los resultados están en función de las herramientas que se utilicen para la identificación del sistema CRISPR-Cas en las secuencias genómicas bacterianas. Así como también, en la homología encontrada entre los espaciadores con los bacteriófagos registrados en las bases de datos de referencia.

En el presente trabajo, se planteó como objetivo desarrollar un modelo computacional basado en técnicas de aprendizaje profundo para la identificación de matrices CRISPR en secuencias de nucleótidos bacterianas, y que sea aplicado a la definición de interacciones bacteriófago – bacteria hospedera.

El modelo computacional propuesto para la identificación de matrices CRISPR, obtuvo un 96.29% de exactitud, 96.00% de Precisión, Exhaustividad y Puntuación-F1 y 92.04% de Coeficiente de correlación de Mathews sobre el conjunto de prueba.

Asimismo, el modelo fue usado en un caso real, caracterizando las matrices CRISPR de la bacteria *Actinoalloteichus sp. AHMU CJ021*. Los resultados mostraron que se identificaron 1,135 matrices CRISPR con un umbral definido como de mayor certeza, además de otras 1,198 matrices con un segundo umbral, el cual es menos restrictivo.

Finalmente, se realizó la extracción de las subsecuencias espaciadoras de las matrices y se les buscó similitud mediante alineamiento con los genomas de bacteriófagos anotados en la base de datos de NCBI Virus. El 47.52% y el 30.23% de los espaciadores en el umbral de “estructuras CRISPR confiables” y en el umbral de “estructuras CRISPR potenciales” respectivamente, encontraron su homólogo con su potencial protoespaciador en los bacteriófagos de la BD. Siendo esta información la que permitió la identificación de las posibles relaciones de interacción. El resto de los espaciadores sin homólogos podrían pertenecer a fragmentos de secuencias de ADN de nuevos fagos aún no han sido identificados.

Contenido

Índice de tablas

Índice de figuras

CAPÍTULO 1.
INTRODUCCIÓN

La metagenómica es el campo que ha permitido la exploración del material genético (genomas) de comunidades de microorganismos como bacterias, arqueas, protozoos, hongos, virus, entre otros, que se encuentran presentes en muestras de diferente índole, mediante el uso de la secuenciación de nueva generación, que genera millones de secuencias de ADN. Esta información permite analizar la diversidad taxonómica y funcional de dichas comunidades. El microbioma intestinal humano ha sido uno de los ecosistemas más estudiados, dado que se busca conocer la diversidad de microorganismos que lo conforman y comprender su papel en la salud y la enfermedad. Estudios recientes (p. ej. Azam et al., 2019; Beller et al., 2019; Manrique et al., 2021) han concluido que existe una relación entre el desequilibrio dentro de las comunidades microbianas intestinales, específicamente de bacterias y de los virus que las infectan (bacteriófagos), con ciertas enfermedades.

Los bacteriófagos o fagos, son virus bacterianos que son considerados como una potencial herramienta para modular a las comunidades de bacterias, ya sea confiriéndoles ventajas de supervivencia o para disminuir su población o incluso para erradicarla. Una de las limitantes para estas aplicaciones es el poco conocimiento de las interacciones infecciosas entre los bacteriófagos y sus bacterias hospederas. Para detectar estas relaciones, existen diferentes métodos que analizan las señales ecológicas de interacción. Sin embargo, los métodos que muestran mejores resultados dependen de bases de datos de referencia para la identificación de estas señales.

Un método prometedor es analizar las señales en el sistema de Repeticiones Palíndromas Cortas Agrupadas y Regularmente Inter espaciadas (CRISPR por sus siglas en inglés, *Clustered Regularly Interespaced Short Palindromic Repeats*) encontrado en bacterias y arqueas. Este sistema es considerado como un mecanismo inmune adaptativo en donde las proteínas generadas, por los genes asociados al sistema CRISPR (genes Cas), cortan fragmentos de la secuencia de ADN de los bacteriófagos invasores y los integran a su propio genoma como espaciadores (en inglés *spacers*). Estos a su vez están delimitados por subsecuencias de repeticiones directas (en inglés *direct repeats - DR*), también llamadas repetidores CRISPR, formando lo que se conoce como una matriz CRISPR. Posteriormente, las bacterias utilizan los espaciadores para defenderse ante una nueva reinfección, haciendo a estas subsecuencias una fuerte señal de interacción entre los bacteriófagos y las bacterias.

Actualmente, la identificación de interacciones bacteriófago – bacteria hospedera mediante este método de los espaciadores CRISPR, ha presentado bajos porcentajes de predicción (Edwards et al., 2016; Li et al., 2021). Esto se debe a la manera en la que se han extraído los espaciadores CRISPR y a su poca o nula homología encontrada en bases de datos de bacteriófagos conocidos.

Las repeticiones directas o repetidores en las matrices CRISPR, tienden a conservarse por especie y se presentan como patrones repetidos separados por una subsecuencia no repetida de longitud similar. Esto hace a la matriz CRISPR una región susceptible a ser identificada con modelos de aprendizaje

profundo, que son ampliamente utilizados en el área de la inteligencia artificial para el reconocimiento de patrones complejos en grandes conjuntos de datos (Nasko et al., 2019).

En el presente trabajo de investigación se realizó la identificación de interacciones bacteriófago – bacteria. Para ello se desarrolló una herramienta computacional basada en un modelo de aprendizaje profundo, para caracterizar matrices CRISPR en secuencias bacterianas e identificar las subsecuencias espaciadoras en las matrices predichas. Estas subsecuencias representan los potenciales fragmentos de ADN de los bacteriófagos que han interactuado con una determinada bacteria hospedera.

1.1. Estructura de la tesis

En este capítulo, se presentan los antecedentes principales de la investigación y el estado del arte de las técnicas que motivaron al desarrollo del presente proyecto. Así como también, la definición del problema, la hipótesis y los objetivos de la investigación.

En el capítulo dos, se presenta el marco teórico, el cuál describe brevemente los principales conceptos tanto biológicos como computacionales utilizados en el desarrollo de la presente investigación.

En el capítulo tres, se desglosa la metodología propuesta que va desde la extracción y depuración del conjunto de datos de entrenamiento y prueba, la exploración de los mismos, la formulación del modelo computacional propuesto, el posprocesamiento de resultados y la aplicación del modelo propuesto para caracterizar matrices CRISPR en secuencias bacterianas.

En el capítulo cuatro se presenta el análisis de los resultados de la evaluación del modelo computacional desarrollado y los resultados obtenidos en la predicción de las matrices CRISPR en un genoma bacteriano real.

Finalmente, en el capítulo cinco se realiza una discusión de los resultados, se presentan las principales conclusiones obtenidas y se plantean los trabajos futuros para la mejora del rendimiento del modelo computacional propuesto.

1.2. Antecedentes de la investigación

Los estudios metagenómicos hacen uso de las tecnologías de secuenciación de nueva generación (NGS por sus siglas en inglés, *Next Generation Sequencing*) o también conocida como secuenciación masiva, para obtener todos los ácidos nucleicos de comunidades microbianas cultivables o no cultivables en laboratorios, que se encuentran presentes en muestras ambientales o de un nicho de interés. Este tipo de análisis metagenómicos ha favorecido la generación de información que puede ser utilizada para la identificación y el descubrimiento de microorganismos. Además, permite comprender la diversidad taxonómica que conforma un ecosistema en específico (Nooij et al., 2018).

En sus inicios, la metagenómica fue mayormente utilizada para el análisis de la diversidad bacteriana, sin embargo, en la actualidad, también se usa para la caracterización de las comunidades víricas (Nooij et al., 2018). Un ecosistema que ha captado gran interés, para su análisis a nivel estructural y

funcional, es la microbiota intestinal de humanos, debido a que su diversidad taxonómica desempeña un papel importante en la salud y en la enfermedad.

La microbiota es el conjunto de microorganismos (bacterias, hongos, levaduras, virus, entre otros) que ocupan un hábitat en específico. El término microbioma engloba a la microbiota y a las funciones que ésta cumple dentro de un determinado ecosistema. La estructura de la microbiota intestinal "normal" en humanos influye en funciones como la maduración del sistema inmune, inhibición de patógenos, síntesis de vitaminas, modulación de fármacos, entre otras. Es adquirida al momento de nacer y se ve influenciada por diversos factores externos y genéticos, desde la vía de nacimiento (cesárea o parto), la dieta alimentaria, enfermedades, suministro de antibióticos, entorno de desarrollo, actividad física, etcétera (Moreno del Castillo et al., 2018; Rowan-Nash et al., 2019).

Por otra parte, la disbiosis intestinal es la afectación que produce cambios en las comunidades microbianas y se encuentra correlacionada a enfermedades como las gastrointestinales, autoinmunes, diabetes, obesidad, diarrea en lactantes e incluso enfermedades neurodegenerativas y del sistema nervioso central (Beller et al., 2019; Moreno del Castillo et al., 2018).

El viroma intestinal hace referencia a los virus eucariontes y a los bacteriófagos (fagos). Según estudios, se ha determinado que los virus que predominan el cuerpo humano son los bacteriófagos (Manrique et al., 2021; Rowan-Nash et al., 2019). Estos son virus infecciosos intracelulares específicos de bacterias, que se multiplican dentro de las mismas usando sus recursos biosintéticos.

Dependiendo del tipo de fago, puede estar constituido por ácido desoxirribonucleico (ADN) o ácido ribonucleico (ARN). Los fagos pueden seguir dos tipos diferentes de ciclo de vida para su replicación:

- Ciclo lítico (o virulento). Los fagos inicialmente realizan la identificación y la fijación a la bacteria, a través de la adsorción con los receptores específicos en la membrana celular del huésped a infectar. El fago inyecta su ácido nucleico (ADN o ARN) hacia el citoplasma de la célula bacteriana, donde se comienzan a sintetizar proteínas víricas para posteriormente ensamblar nuevos fagos hasta ocasionar lisis en la bacteria provocando su muerte.

- Ciclo lisogénico (o templado). Los fagos realizan la inyección de su ácido nucleico de manera similar a los líticos, pero la diferencia es que buscan unirse al material genético de la bacteria para permanecer dentro ella en un estado reprimido, al que también se le conoce como profago. Al encontrarse el profago integrado en el cromosoma de la bacteria, es heredado a las bacterianas de las siguientes generaciones en el proceso de replicación.

Los fagos intestinales pueden influir en las comunidades de sus huéspedes bacterianos en términos de composición y función mediante los procesos líticos o lisogénicos. Estos pueden, desde interferir fuertemente en la modulación de la expresión génica de las comunidades bacterianas, hasta conferir ventaja competitiva o resistencia a las bacterias, a través de los profagos o la transferencia horizontal de genes en los procesos de infección e integración (Beller et al., 2019; Manrique et al., 2021).

Los fagos que habitan el intestino humano son aproximadamente 10^{15}. Los más predominantes son del orden *Caudovirales* con ADN de cadena doble como *Myoviridae*, *Podoviridae* y *Siphoviridae*, y los de ADN de cadena simple son de la familia *Microviridae*. Los *Caudovirales* y *Microviridae* tienen un tamaño genómico de 18 a 30 kilo pares-bases (kb) y exhiben ambos ciclos de vida, siendo los fagos templados quienes dominan el intestino humano (Rowan-Nash et al., 2019).

En las secuencias de estos microorganismos, es posible encontrar diversas señales de interacción resultantes de los procesos ecológicos de infección que existen o han existido entre ellos y sus bacterias hospederas. En la actualidad, aún nos enfrentamos a la ausencia de algún método universal, ya sea experimental o computacional, que esté establecido para realizar las anotaciones de relaciones fago – huésped. Por tal motivo, en la literatura se siguen proponiendo herramientas y enfoques bioinformáticos para realizar este tipo de predicciones. Coclet & Roux (2021) realizaron una extensa revisión de los principales métodos para anotar dichas interacciones clasificándolos en las siguientes tres categorías:

- **Enfoques dependientes de alineamientos.** Consisten en encontrar la similitud de secuencias genómicas (ADN o ARN) entre los fagos y las bacterias, y también, la similitud entre los fagos y genes marcadores exclusivos de virus con huéspedes conocidos. Para detectar las similitudes, este enfoque comúnmente utiliza bases de datos de referencia y la Herramienta Básica de Búsqueda de Alineación Local (BLAST, por sus siglas en inglés) a nivel de nucleótidos o aminoácidos. Algunas de las regiones con similitud, son los sitios con profagos integrados en las bacterias, espaciadores CRISPR, genes transferidos horizontalmente (como los genes metabólicos auxiliares (AMG)), sitios de inserción (como los ARN de transferencia (ARNt)), por mencionar algunos. Las principales herramientas utilizadas con este enfoque son BLAST, SpacePHARER, RaFAH, vHULK, VPF-Class (Coclet et al., 2021) y con técnicas de aprendizaje profundo, el enfoque PHIAF (Li et al., 2021).

- **Enfoques sin alineamientos.** Determinan la similitud de la composición de las secuencias genómicas, ya sea a nivel nucleótido o aminoácido, de los fagos y de sus potenciales bacterias hospederas. Esto es posible por los procesos ecológicos entre los fagos y las bacterias, como los cambios de adaptación que realizan los fagos en sus secuencias genómicas para tener una mayor similitud con la maquinaria de replicación, transcripción y traducción de la bacteria huésped. Comúnmente estos métodos analizan los perfiles de oligonucleótidos, subsecuencias cortas extraídas de un genoma, con una longitud *k* específica a las cuales se les conoce como *k*-meros. Las principales herramientas con este enfoque son PHIST (Zielezinski et al., 2022), HostPhinder, ILMF-VH, VirHostMatcher, WIsH y PHP (Coclet et al., 2021) y con técnicas de aprendizaje profundo, el enfoque DeepHost (Ruohan et al., 2022).

- **Enfoques integradores.** Son tuberías computacionales que integran múltiples métodos para la detección o predicción de las señales de interacción fago - bacteria hospedera, con la finalidad de mejorar el rendimiento en las métricas, tales como precisión y exhaustividad. Estos métodos integran enfoques basados en alineamiento y sin alineamiento, como PHISDetector (Zhang et al., 2019) y VirHostMatcher-Net (Wang et al., 2020), que analizan características como la similitud de secuencia libre de alineación en función de las

frecuencias de *k*-meros, análisis de espaciadores CRISPR, interacciones de proteínas y coincidencias exactas entre los fagos y sus huéspedes (Coclet et al., 2021).

Por lo que concluyen que el rendimiento de los enfoques de alineamiento está limitado a la poca anotación de fagos conocidos en las bases de datos de referencia, en consecuencia, estos enfoques obtienen pocas o nulas predicciones cuando se enfrentan a fagos desconocidos. Por su parte, los enfoques sin alineamientos tienden a mostrar menor precisión en sus predicciones que los enfoques de alineamiento, pero alcanzan mejores resultados en la métrica de exhaustividad (en inglés *recall*). Finalmente, los enfoques integradores han mostrado, hasta el momento, buenos resultados en sus predicciones, pero requieren más recursos computacionales debido a que integran distintas herramientas para obtener sus resultados.

Diferentes tuberías computaciones con enfoques basados en alineamientos y enfoques integrados, hacen uso de los espaciadores CRISPR. Esto debido a que son señales muy prometedoras para identificar las interacciones entre los fagos y sus bacterias hospederas (Nasko et al., 2019). Los espaciadores en las matrices CRISPR de las bacterias tienen similitud con la secuencia del fago invasor, en las regiones objetivo a escindir, que son conocidas como protoespaciadores.

En el estudio de Edwards y colaboradores (Edwards et al., 2016), extrajeron los espaciadores de las matrices CRISPR identificadas en el 39.5% (1,066) de los 2,698 genomas bacterianos analizados con la herramienta PILER-CR v1.06 (Edgar, 2007). Posteriormente, realizaron el alineamiento de los espaciadores con la herramienta BLAST para identificar los protoespaciadores objetivo en un conjunto de datos conformado por 820 genomas de bacteriófagos. Como resultado, las predicciones de bacterias con coincidencias de similitud a partir de 1 espaciador, fueron correctas para el 15.1% de los fagos y, las bacterias con coincidencias a partir de 2 espaciadores, fueron correctas para el 21.3% de los fagos.

En el enfoque integrador PHISDetector (Zhang et al., 2019), también analizaron las señales de interacción a partir de diferentes características tales como espaciadores CRISPR (extraídos con PILER-CR v1.06), profagos integrados, similitud en la composición de *k*-meros, homología genética e interacciones de proteínas. La combinación de estas características fue utilizada para entrenar un modelo de aprendizaje máquina para predecir la interacción bacteriófago – bacteria huésped. Obtuvieron que, con el uso de una sola característica sólo el 16.4% al 41.25% de las predicciones de interacción fueron correctas y con la combinación de las diferentes características el 70.13% al 89.84% de las predicciones de interacción a nivel de especie y familia, fueron correctas en su conjunto de entrenamiento.

Recientemente, Dion y colaboradores (Dion et al., 2021) desarrollaron un enfoque integrador para la predicción de interacciones. Extrajeron más de 11 millones de espaciadores CRISPR de 367,446 genomas bacterianos, con la herramienta CRISPRDetect v2 (Biswas et al., 2016). Definieron diferentes criterios en el total de desajustes a permitir entre la similitud de protoespaciadores de los fagos a predecir su huésped y los espaciadores en su base de datos, aplicando un alineamiento con BLAST. El mejor criterio fue permitir 2 desajustes para permitir errores de secuenciación o mutaciones. Como resultados obtuvieron 69% de precisión y 49% de recuperación en la predicción de la bacteria huésped para 9,484 fagos.

En vista de lo anterior, se interpreta que los resultados en la precisión de los diferentes enfoques que analizan los espaciadores CRISPR para la predicción de interacciones bacteriófago – bacteria huésped, son dependientes de las herramientas que utilizan para la detección de las matrices CRISPR y sus subsecuencias. Aunado a ello, la poca homología encontrada de los espaciadores con los genomas de bacteriófagos conocidos y registrados en bases de datos. De manera general, los distintos enfoques han reportado porcentajes alrededor del 15.1% al 89.84% en la exactitud de las predicciones de interacción (Coclet et al., 2021; Dion et al., 2021; Zhang et al., 2019; Edwards et al., 2016).

1.3. Estado del arte sobre la caracterización de matrices CRISPR

Desde el descubrimiento del sistema CRISPR-Cas han ido en aumento los estudios de su estructura y funcionamiento, así como el desarrollo de herramientas bioinformáticas para caracterizar a estos sistemas completos o partes específicas de ellos en los genomas bacterianos y de arqueas. A continuación, se presenta una breve descripción de algunas de las herramientas bioinformáticas más sobresalientes para esta tarea.

1.3.1. PILER-CR

PILER-CR (Edgar, 2007) es una herramienta diseñada para la identificación y análisis de repetidores CRISPR. De manera general, el algoritmo consiste en encontrar alineamientos locales de un genoma contra sí mismo e identificar subsecuencias repetidas separadas por una distancia corta. Para esto, construye pilotes, a lo largo del genoma, utilizando el concepto de pila para hacer referencia a las subsecuencias que se encuentran repetidas al menos una vez y que están separadas por otra subsecuencia única, pudiendo vincularlas. De esta manera, los pilotes permiten extraer matrices CRISPR completas a partir de la generación de gráficos de conectividad o de autosimilitud (gráfico de puntos) de los pilotes.

Por último, en el posprocesamiento se evalúa que las repeticiones directas en las matrices CRISPR candidatas tengan una conservación mayor al 90% de similitud. Después se extrae una repetición directa consenso y se fusionan las matrices adyacentes con subsecuencias consenso similares. Este método fue aplicado en 346 genomas procariotas, comparando sus resultados se compararon con los repetidores CRISPR reportados por Jansen et al., 2002 y contra las matrices CRISPR asociadas con sus genes Cas reportadas por Godde & Bickerton (2006), se estimó una sensibilidad cercana al 100% porque se encontraron todas las instancias de los estudios, y una especificidad del 94% aproximadamente debido al obtuvieron 8 repeticiones directas cuestionables y 11 como falsos positivos de un total de 319 repeticiones directas.

1.3.2.CRISPR Recognition Tool – CRT

CRISPR Recognition Tool – CRT (Bland et al., 2007) incorporó una técnica de búsqueda secuencial que detecta repeticiones en la secuencia de ADN. Para esto, realiza búsquedas de coincidencias

exactas de repeticiones mediante una ventana deslizante de longitud *k* en un determinado rango de búsqueda, tomando en cuenta longitudes mínimas y máximas de las repeticiones directas y espaciadores en las matrices CRISPR.

Posteriormente, se aplican filtros para descartar repeticiones contiguas o con posiciones iniciales y/o finales incorrectas. Después se comprueba que las repeticiones cumplan con las longitudes mínimas y máximas definidas y que los espaciadores no se repitan y tengan longitudes similares a las reportadas. Finalmente, se verifican los desajustes entre los repetidores en el extremo izquierdo y el derecho, usando la distancia de Hamming. Esto, ya que la discrepancia de repeticiones directas aumenta más en el extremo derecho de la matriz, debido a que los últimos repetidores CRISPR contienen mayor número de mutaciones a lo largo del tiempo.

La herramienta CRT fue comparada con las herramientas PILER-CR y Patscan en función de la velocidad de ejecución y la capacidad para identificar correctamente los CRISPR, usando las métricas de calidad, precisión y exhaustividad (*recall*). Cabe señalar que la herramienta Patscan es de uso general para la detección de patrones genéricos en secuencias genómicas (Dsouza et al., 1997). Se usaron los siguientes dos conjuntos de datos: (i) 27 especies bacterianas seleccionadas aleatoriamente de 101 con CRISPR adyacentes a genes Cas, detectados con la herramienta Patscan por Godde & Bickerton (2006) y (ii) 80 genomas seleccionados aleatoriamente de la base de datos IMG versión 1.5 (Markowitz et al., 2006).

La herramienta CRT mostró mayor velocidad de ejecución con un tamaño de $k = 6$, considerando repetidores de 19 – 50 nucleótidos y espaciadores de 19 – 60 nucleótidos de longitud. Una secuencia de casi 6 millones de nucleótidos se procesó en 3 segundos con un tamaño de $k = 8$ y en 2 segundos con el tamaño $k = 6$.

En cuanto al rendimiento, los resultados del análisis comparativo, muestran que CRT fue superior en todas las métricas (Tabla 1).

Tabla 1. Resultados de las herramientas comparadas contra CRT (Bland et al., 2007).

Métricas	**Conjunto Godde & Bickerton, (2006)**			**Conjunto de la base de datos IMG**	
	CRT	**PILER-CR**	**PATSCAN**	**CRT**	**PILER-CR**
Calidad	0.95	0.77	0.74	0.90	0.75
Precisión	0.99	1	0.89	0.89	1
Exhaustividad	0.99	0.95	N/A	1	0.86

1.3.3. CRISPRDetect

CRISPRDetect (Biswas et al., 2016) es una herramienta disponible en versión web y en línea de comandos. Inicialmente, el algoritmo realiza una búsqueda de repeticiones directas a lo largo del genoma con un tamaño de *k*-mero de 11 nucleótidos. Después los repetidores CRISPR putativos y los espaciadores se comparan mediante alineamientos para descartar repeticiones contiguas.

El refinamiento de los repetidores CRISPR consiste en extender las repeticiones putativas hacia ambos lados de la secuencia, comparando los nucleótidos adyacentes e identificando repeticiones con un 75% de identidad similar mínima. Para el análisis de los límites de las matrices, se selecciona una repetidor representativo o consenso, permitiendo más desajustes, de hasta 66% de identidad, en matrices más extensas. Adicionalmente, se implementa la herramienta CRISPRDirection (Biswas et al., 2014) para analizar las matrices predichas y obtener la dirección u orientación transcripcional correcta de la región CRISPR.

Finalmente, predicen el tipo de sistema CRISPR-Cas de las matrices predichas, comparándolas con un conjunto de referencia tipificado. Al mismo tiempo, corrigen los límites de repetidores, y validan las matrices CRISPR. Esta herramienta fue evaluada en 2,806 genomas bacterianos y arqueales completos. En donde obtuvieron 3,901 matrices CRISPR, de las cuales el 97% (3,870 matrices) fueron matrices CRISPR verdaderas, 16,607 cuestionables y 160 posibles repeticiones en tándem. En los refinamientos iterativos, el 12% de las repeticiones no fueron idénticas al repetidor CRISPR representativo, ya que identificaron 50 por debajo del 70% de identidad y 399 por debajo del 80%. Los autores realizaron una comparación analizando las 3,870 matrices correctas contra otras tres herramientas, en donde se obtuvo los siguientes resultados: (i) CRT predijo 3,681 (95%) matrices CRISPR; (ii) PILER-CR predijo 3,743 (96%); y, (iii) CRISPRFinder (Grissa et al., 2007a) predijo 2,750 (71%). Todas las herramientas predijeron 1,782 (46%) matrices en común.

1.3.4. CRISPR Finder Random Forest – CRF

CRISPR Finder Random Forest – CRF (Wang et al., 2017) es una herramienta basada en web que implementa un clasificador de bosque aleatorio de aprendizaje automático, para filtrar matrices CRISPR putativas en función de los repetidores CRISPR y sus características estructurales. Inicialmente, esta herramienta detecta las matrices CRISPR candidatas con la herramienta CRT disminuyendo el tamaño de la palabra semilla predeterminada de $k = 8$ nt a $k = 5$ nt para detectar más repetidores con desajustes.

El conjunto de datos positivo para entrenar el clasificador, se descargó de la base de datos CRISPRdb (Grissa et al., 2007b). Tras la eliminación de redundancia en los datos, obtuvieron 11,407 repetidores positivos o verdaderos. El conjunto de datos negativo fue generado mediante un modelo de Markov de primer orden, obteniendo 12,000 repetidores negativos o falsos.

El clasificador fue implementado con el paquete R randomForest v4.6-10, y se entrenó con el 80% de los datos y el 20% de prueba. Con la herramienta Phobos (Mayer, 2008) detectaron las repeticiones en tándem y finalmente calcularon la distancia de Hamming para excluir matrices CRISPR no válidas, que contenían espaciadores con una similitud superior al 50% a los repetidores.

El modelo alcanzó una precisión de 94.42%, una sensibilidad de 93.99% y el área bajo la curva (AUC) de 0.9837 en el conjunto de prueba. El modelo se comparó con las herramientas PILER-CR, CRT y CRISPRDetect utilizando su mismo conjunto de prueba (1,139 genomas bacterianos), que contenían 3,689 matrices CRISPR. Como resultado, se obtuvo que PILER-CR detectó matrices CRISPR sólo en 1,033 (90%) genomas, CRT en 1,067 (93%), CRISPRDetect en 948 (83%) y la herramienta

propuesta CRF en el 100% de los genomas. Asimismo, CRF obtuvo una mayor sensibilidad y precisión (Tabla 2).

Tabla 2. Resultados de las herramientas comparadas contra CRF en el conjunto de prueba de 1,139 genomas bacterianos.

Métricas	CRF	CRT	CRISPRDetect	PILER-CR
Sensibilidad (%)	61.62	66.28	34.46	69.75
Precisión (%)	83.29	73.80	51.93	75.21

1.3.5. CRISPRCasFinder

CRISPRCasFinder (Couvin et al., 2018) es una actualización de CRISPRFinder (Grissa et al., 2007a), que realiza la predicción de matrices CRISPR. La actualización agrega la función de predicción de la orientación CRISPR, así como la tipificación de proteínas Cas. Inicialmente, para detectar las repeticiones directas de las matrices, hace uso de la herramienta Vmatch versión 2.3 implementada en CRISPRFinder.

Implementan un sistema para la discriminación de elementos similares a CRISPR falsos, estableciendo cuatro niveles de evidencia, midiendo la conservación de los repetidores CRISPR con base a la entropía de Shannon: i) las matrices que contienen de uno a tres espaciadores, son poco confiables y se asignan al nivel de evidencia 1; ii) las matrices con índices de conservación <70 en el nivel 2; iii) las matrices con índice ≥70 y un porcentaje de identidad entre sus espaciadores >8% en el nivel 3 y, iv) las matrices con índice ≥70 y un porcentaje de identidad entre sus espaciadores ≤8%, son las de mayor confiabilidad, en el nivel 4.

Así también, proporciona dos indicadores de orientación usando la herramienta CRISPRDirection, la cual busca la región rica en AT que flanquea la matriz para detectar la orientación. Con la herramienta Prodigal v2.6.3 identifica y tipifica los genes Cas a partir de lo reportado en la literatura. Con la herramienta CasFinder analiza las secuencias codificantes putativas para identificar los sistemas y sus componentes. Posteriormente, las proteínas Cas putativas se analizan por similitud de secuencia aplicando perfiles de proteínas con Modelos Ocultos de Markov (HMM).

La evaluación de CRISPRCasFinder se realizó sobre un conjunto de 1,263 matrices CRISPR de 400 genomas y se compararon los resultados con las herramientas PILER-CR, CRT y CRISPRDetect. CRISPRCasFinder obtuvo una la precisión, exhaustividad y valor-F iguales a PILER-CR, siendo ambas las de mayor valor-F (Tabla 3).

Tabla 3. Resultados de las herramientas comparadas contra CRISPRCasFinder (Couvin et al., 2018).

Métricas	CRISPRCasFinder	CRT	PILER-CR	CRISPRDetect
Precisión	0.96	0.92	0.96	0.98
Exhaustividad	0.98	0.94	0.98	0.93
Valor - F	0.97	0.93	0.97	0.95

1.3.6. CRISPRidentify

CRISPRidentify (Mitrofanov et al., 2021) es una tubería computacional que integra un enfoque de aprendizaje automático para detectar las matrices CRISPR en función de 13 características derivadas de un conjunto de ejemplos positivos y negativos de matrices CRISPR. Entre las principales se encuentra la similitud entre los repetidores conocidos, el contenido de AT y la estabilidad de la horquilla del repetidor en base a la energía libre mínima de la estructura tallo-bucle. A partir de esto, anota la matriz CRISPR con información adicional como su orientación, secuencia líder y genes Cas.

Utilizaron cuatro conjuntos de datos para entrenar y validar el modelo. El primero de 400 matrices CRISPR positivas y negativas de arqueas. El segundo de 600 matrices CRISPR positivas y negativas de bacterias. Ambos conjuntos fueron usados en el entrenamiento. El tercer conjunto, de 300 matrices falsas, generado a partir de repeticiones en tándem, con supuestos espaciadores idénticos a los repetidores CRISPR y con la distorsión de matrices reales, usando 100 en el entrenamiento y 200 para la validación. El cuarto fue de prueba como positivo de alta calidad, constituido por 550 matrices de arqueas y bacterias con los siguientes criterios: tres repeticiones directas mínimas por matriz y que la distancia máxima entre la matriz CRISPR y el operón Cas fuera de 500 nucleótidos.

Con Vmatch v2.3.0 analizaron los genomas para identificar matrices CRISPR putativas (subsecuencias de 21 a 55 nucleótidos repetidos, separados por un rango de 18 a 78 nucleótidos). Después extrajeron los vectores de características a evaluar mediante uno o los tres modelos de aprendizaje automático propuestos, cada uno con diferentes combinaciones de las características. Se usó árboles extremadamente aleatorios (en inglés *extra-trees classifier*) entrenado con ejemplos positivos y negativos. Las matrices con puntuaciones ≥ 0.75 son etiquetadas como candidatas, las de 0.40 a 0.75 como candidatas potenciales y las matrices con puntuaciones inferiores se clasifican como candidatas de puntuación baja.

Una vez clasificadas las matrices, su orientación fue determinada con la herramienta CRISPRstrand (Alkhnbashi et al., 2014) y se anotó la secuencia líder usando CRISPRleader (Alkhnbashi et al., 2016). Utilizaron la herramienta Prodigal (Hyatt et al., 2010) para la predicción de proteínas Cas mediante un Modelo oculto de Markov.

El conjunto de prueba se usó para comparar su rendimiento contra las herramientas CRT, CRISPRCasFinder y CRISPRDetect. Los resultados mostraron que CRISPRidentify obtuvo una mayor la tasa de verdaderos positivos (TPR), de verdaderos negativos (TNR), exactitud (ACC), exactitud equilibrada (BACC) y coeficiente de correlación de Mathews (MCC) (Tabla 4).

Tabla 4. Resultados de las herramientas comparadas contra CRISPRidentify (Mitrofanov et al., 2021).

Herramientas	TPR	TNR	ACC	BACC	MCC
CRISPRidentify	0.99	1	0.99	0.99	0.99
CRT	0.93	0.42	0.79	0.67	0.42
CRISPRCasFinder	0.96	0.64	0.87	0.80	0.67
CRISPRDetect	0.96	0.74	0.89	0.85	0.73

Por otra parte, del conjunto 2 de prueba, realizaron la evaluación de las 200 matrices falsas, identificando CRISPRidentify correctamente a todas estas matrices, mientras, la herramienta CRT obtuvo la mayor tasa de falsos positivos (58.5%), seguida por CRISPRCasFinder (36%) y CRISPRDetect (26.5%). En una evaluación más, los autores emplearon otro conjunto de datos de 987 arqueas y 27,028 genomas bacterianos extraídos del Centro Nacional para la Información Biotecnológica (NCBI, por sus siglas en inglés) con un total de 35,310 matrices CRISPR. En esta prueba CRISPRidentify tuvo un desempeño del 99.8%, CRISPRCasFinder de 98.9%, CRISPRDetect de 94.5% y CRT de 93.5%.

Finalmente, aplicaron el algoritmo ReBATE (Urbanowicz et al., 2018) para definir la contribución de las características utilizadas en la identificación de matrices CRISPR. Se encontró que la característica más significativa es el análisis de la similitud de los repetidores con lo anotado.

1.4. Estado del arte de herramientas para el análisis completo o de regiones del sistema CRISPR-Cas

En la literatura, se puede encontrar diversas tuberías computacionales que integran alguna de las herramientas anteriormente descritas para realizar análisis completos o de regiones específicas del sistema CRISPR-Cas. A continuación, se describen brevemente las más relevantes.

- CRISPRleader (Alkhnbashi et al., 2016) implementa la herramienta CRISPRstrand que a su vez usa las herramientas CRISPRFinder y CRT para extraer las matrices CRISPR. Posteriormente, se analizan los nucleótidos *aguas arriba* de la primera repetición directa en la matriz para identificar a la secuencia promotora o secuencia líder del sistema CRISPR-Cas.

- CRISPRminer integra funciones para la predicción, usando las herramientas PILER-CR, HmmScan y CRISPRCasFinder; la clasificación, usando las herramientas FragGeneScan, RPS-BLAST, Prodigal v2.6.3 y MacSyFinder (Zhang et al., 2018); la anotación y la visualización de los sistemas CRISPR-Cas.

- CRISPRdisco (Crawley, Henriksen, & Barrangou, 2018) realiza la predicción de las matrices CRISPR con minCED (http://github.com/ctSkennerton/minced) que es una herramienta derivada de CRT (Bland et al., 2007), la identificación de los genes Cas usando la herramienta BLAST.

- CRISPRmap (Alkhnbashi et. al, 2019) utiliza herramientas externas (p. ej. CRISPRFinder y CRT) para extraer los sistemas CRISPR-Cas, y después los clasifica mediante el uso de los repetidores, los cuales pueden representar una fuente de información para investigar la evolución de estos sistemas.

- CRISPRstrand (Alkhnbashi et al., 2014) usa herramientas externas (p. ej. CRISPRFinder y CRT) para extraer sistemas CRISPR-Cas y, aplica modelos discriminativos basados en núcleos de gráficos para predecir la orientación de las subsecuencias CRISPR a transcribirse en ácidos ribonucleicos CRISPR (crARN).

- CRISPRCasTyper (Russel et al., 2020) primero detecta operones de genes Cas usando la herramienta HMMER3 (Eddy, 2009) con 680 Modelos Ocultos de Markov de operones Cas tipificados. Después detecta las matrices CRISPR adyacentes para asignar una clasificación de sistema CRISPR-Cas. Finalmente, usa minCED, para determinar la orientación de los repetidores, generar subsecuencias consenso y contabilizar el número de repetidores para indicar el tamaño de las matrices CRISPR detectadas.

- CRISPRCasIdentifier (Padilha et al., 2020) predice y clasifica los genes Cas, con base a reglas y evidencia de asociación de los genes característicos conocidos, usando técnicas de clasificación y regresión, como árboles extremadamente aleatorios y máquinas de soporte vectorial.

- CRISPRloci (Alkhnbashi et al., 2021) es un servidor web que integra CRISPRIdentify (Mitrofanov et al., 2021) para la detección de matrices CRISPR, su orientación y secuencia líder. Mide la energía libre mínima de las estructuras secundarias tallo-bucle de los repetidores CRISPR consenso en las matrices con la herramienta RNAfold.

- Casboundary permite definir los límites de la región usando los genes Cas, mediante árboles extremadamente aleatorios y redes neuronales artificiales profundas. Para ello, definen una región como intervalo para detectar genes y predice una etiqueta de la relación de dichos genes representativos de los operones Cas tipificados. Después realiza la clasificación de los genes Cas de acuerdo con las familias de proteínas conocidas, lo cual es utilizado como entrada para CRISPRCasIdentifier que clasifica a un subtipo de sistema CRISPR-Cas con base a la combinación de las proteínas Cas y, realiza inferencias de proteínas potencialmente faltantes. Finalmente, detecta coincidencias entre los espaciadores CRISPR extraídos por la herramienta CRISPRidentify contra genomas completos o parciales de fagos para detectar las interacciones fago – huésped (Alkhnbashi et al., 2021).

1.5. Estado del arte de las bases de datos disponibles relacionadas con el sistema CRISPR-Cas

Existen bases de datos (BD) donde se almacenan los resultados predictivos y algunos con validación experimental de los sistemas CRISPR-Cas obtenidos por las tuberías computacionales. A continuación, se describen algunas BDs disponibles.

- CRISPRCasdb (Pourcel et al., 2020) (Grissa et al., 2007b) fue la primer BD dedicada al almacenamiento exclusivo de matrices CRISPR de genomas arqueales y bacterianos extraídas con la herramienta CRISPRCasFinder. Esta BD se actualiza periódicamente con metadatos, taxonomía procariota y los resultados del análisis de secuencias de genomas contenidos en NCBI.

- CRISPRBank es una BD desarrollada a partir de los resultados de la herramienta CRISPRDetect.

- CRISPRone (Zhang et al., 2017) es una BD que fue creada con el uso de la herramienta derivada de CRT denominada MetaCRT (Rho et al., 2012) para extraer sistemas CRISPR-Cas de genomas bacterianos y arqueales de NCBI.

- CRISPI (Rousseau et al., 2009) es una BD que se construyó con el análisis de secuencias de bacterias y arqueas de NCBI con la herramienta CRISPRFinder.

- CRISPRminerDB es una BD creada a partir de los resultados de la tubería computacional CRISPRminer.

- CrisprOpenDB es una BD que contiene sólo subsecuencias espaciadoras y fue creada a partir del enfoque integrador de Dion y colaboradores (Dion et al., 2021).

El sistema CRISPR-Cas9 ha ganado gran interés, debido a que la proteína Cas9 y similares a esta, han demostrado beneficios para la edición genética de genomas y tipos de células de manera guiada como en humanos, monos, peces cebras, ratones, moscas de fruta y cerdos. Esta proteína debe ser guiada con un ARN guía (ARNg) el cual contiene la secuencia objetivo del genoma a editar ya sea para reemplazarla o eliminarlas.

Actualmente, las técnicas de aprendizaje profundo (en inglés *deep learning*, DL) han sido usadas para el desarrollo de modelos predictivos para el diseño y la actividad objetivo de los ARN guías CRISPR en la edición de genes. Sin embargo, aún no se proponen modelos de DL para identificar y predecir las regiones del sistema CRISPR-Cas mediante el análisis directo de secuencias de nucleótidos. La caracterización de las matrices CRISPR comienza en encontrar los patrones de las repeticiones directas parcialmente palíndromas e identificar el espaciador entre ellas. La implementación de una técnica de DL como las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés), podría representar mayor eficiencia para realizar este proceso.

Las CNN son modelos de aprendizaje profundo que utilizan capas convolucionales para extraer características específicas de los datos de entrada. Cada capa convolucional, mediante filtros extrae las características más significativas de la capa anterior. Finalmente, las CNN contienen una capa final de neuronas completamente conectadas que realizan clasificaciones binarias o multiclase (Nguyen et al., 2016).

Recientemente, las CNN han sido utilizadas para el análisis de secuencias de nucleótidos, que incluyen tareas de clasificación binaria de secuencias promotoras (Nguyen et al., 2016; Zuallaert et

al., 2018), clasificaciones taxonómicas de secuencias (Khawaldeh et al., 2017), predicción de fenotipos de muestras genéticas (Reiman et al., 2017), entre otras, mostrando resultados alentadores.

1.6. Definición del problema

Las interacciones entre los fagos y sus bacterias huéspedes, frecuentemente dejan señales en sus secuencias genómicas de ADN o ARN que pueden ser identificadas mediante métodos computacionales. Los enfoques que muestran mayor confiabilidad para la definición de interacciones fago – bacteria hospedera son los basados en alineamientos. Sin embargo, son dependientes de lo anotado en las bases de datos. Esto es una limitante debido a que gran parte de las secuencias de genomas virales obtenidos, mediante metagenómica, presentan escasa o nula homología con lo anotado en las BDs de referencia.

Algunos de los enfoques hacen uso de los espaciadores CRISPR como señales de interacción para establecer las relaciones fago – bacteria hospedera, siendo un método muy prometedor. Sin embargo, las predicciones de interacciones a través de los espaciadores no han sido tan favorables (Dion et al., 2021; Edwards et al., 2016; Zhang et al., 2019) dado que los resultados están en función de la herramienta que se utilice para la identificación del sistema CRISPR-Cas en las secuencias genómicas bacterianas y en que el bacteriófago infectante, se encuentre registrado en las bases de datos de lo conocido para encontrar su homólogo.

Con la revisión de las herramientas y tuberías computacionales para la identificación del sistema CRISPR-Cas, se encontró que aún no se han implementado técnicas de aprendizaje profundo (*deep Learning*, DL) para extraer estas regiones directamente de las secuencias de nucleótidos bacterianas. Este tipo de técnicas han demostrado ser eficientes en el análisis de datos metagenómicos obteniendo altas precisiones en predicciones y clasificaciones (Talukder et al., 2021).

Hasta el momento, aún existe mucho que explorar sobre las interacciones entre los fagos y sus bacterias huéspedes, más aún con relación a fagos desconocidos o identificados recientemente. Esta información podría representar un beneficio para comprender cómo éstas interacciones influyen en la dinámica de un determinado ecosistema, como por ejemplo en el microbioma intestinal, lo cual, puede ser aplicado en tratamientos que contribuyan a la homeóstasis o el equilibrio de la microbiota en beneficio de la salud humana.

1.7. Hipótesis

El uso de técnicas de aprendizaje profundo permitirá identificar las matrices CRISPR de genomas de bacterias que permitan establecer relaciones de interacción bacteriófago – bacteria hospedera, utilizando patrones de estas regiones, como las repeticiones directas y los espaciadores CRISPR.

1.8. Objetivos de la investigación

1.8.1. Objetivo general

Desarrollar un modelo computacional basado en técnicas de aprendizaje profundo para la identificación de matrices CRISPR en secuencias de nucleótidos bacterianas, y que sea aplicado a la definición de interacciones bacteriófago – bacteria hospedera.

1.8.2. Objetivos específicos

- Extracción y depuración de la base de datos de sistemas CRISPR-Cas para obtener un conjunto de datos de entrenamiento y prueba.

- Análisis de las subsecuencias de nucleótidos que conforman el sistema CRISPR-Cas de bacterias para la caracterización de las matrices CRISPR.

- Diseño y desarrollo de un modelo computacional binario para caracterizar las matrices CRISPR en secuencias bacterias.

- Aplicación del modelo computacional sobre el genoma de una bacteria real para predecir las matrices CRISPR.

- Extracción de las subsecuencias espaciadoras de las matrices CRISPR predichas de la secuencia analizada.

- Alineamiento de las subsecuencias espaciadoras contra bases de datos de bacteriófagos para identificar al virus invasor de la bacteria analizada. Si no existe homólogo con la base de datos, realizar anotación de interacción de la bacteria analizada con un potencial fago desconocido.

CAPÍTULO 2. MARCO TEÓRICO

En este capítulo se presenta la recopilación de los conceptos principales del área biológica, así como del área computacional, que son ampliamente utilizados en el desarrollo del presente trabajo.

2.1. Conceptos biológicos

2.1.1. Ácido desoxirribonucleico (ADN)

El ácido desoxirribonucleico (ADN) es el nombre químico de la molécula clave responsable del almacenamiento, la duplicación y la realización de la información genética en todos los seres vivos. El ADN es una molécula hetero-polimérica que consta de dos cadenas que se enrollan entre sí para formar una estructura de doble hélice; en la parte central se encuentran enlazadas por azúcares (desoxirribosas) y grupos de fosfatos. Las cadenas se mantienen unidas mediante los enlaces de cuatro bases, la adenina (A) se enlaza con la timina (T) y la citosina (C) con la guanina (G), formando las unidades llamadas pares de bases. La unión entre una base, un azúcar y un fosfato, es a lo que se le conoce como nucleótido. Las secuencias de los nucleótidos en las cadenas del ADN, se subdividen en un sentido biológico en secciones, que codifican la información de genes para formar proteínas y moléculas de ARN (Frank-Kamenetskii, 1997).

2.1.2. Ácido ribonucleico (ARN)

El ácido ribonucleico (ARN) es un ácido nucleico similar al ADN, pero es de una sola cadena constituida por un azúcar (ribosa) y grupos de fosfato. A cada azúcar se encuentra unida una de las cuatro bases adenina (A), uracilo (U), citosina (C) o guanina (G). Los ARN codifican información copiada del ADN y hay tres tipos principales, el ARN mensajero (ARNm), ARN ribosomal (ARNr) y ARN de transferencia (ARNt).

2.1.3. Gen

Dentro de las cadenas de las moléculas de ADN, se encuentran regiones (*locus*, del latín, lugar, en plural *loci*) que representan unidades funcionales y físicas hereditarias, llamados genes. Estos locus contienen la información genética específica para codificar proteínas.

2.1.4. Proteína

Las proteínas son moléculas que se componen por una cadena de aminoácidos traducidos de los genes en el ADN y cumplen con diversas funciones en las células tales como estructurales, anticuerpos, mensajeras, entre otras funciones vitales para los organismos.

2.1.5. Estructura secundaria tallo-bucle

La estructura secundaria en tallo-bucle (en inglés *stem-loop* o *hairpin-loop*) se origina mediante el emparejamiento de bases intramoleculares en ácidos nucleicos, principalmente en moléculas de ARN, pero también se presenta en moléculas de ADN monocatenarias. Estas estructuras, exhiben características ventajosas en las interacciones de ADN – ADN y ADN – ARN y cumplen con funciones reguladoras en el metabolismo celular (Broude, 2002).

En la Figura 1, se muestra un ejemplo de una estructura tallo-bucle generada a partir de una matriz CRISPR con la herramienta RNAfold (Gruber et al., 2008). El fragmento de la estructura que representa el tallo, se produce cuando se emparejan las bases complementarias, y el fragmento conocido como bucle, se forma cuando las bases no son complementarias y no existe un emparejamiento de estas.

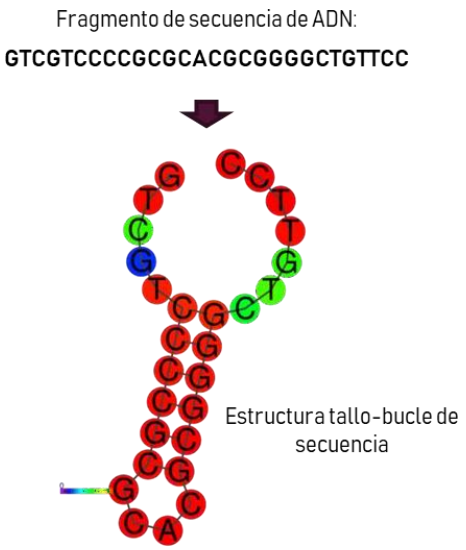

Figura 1. Representación de estructura tallo bucle de un fragmento de secuencia de ADN.

2.1.6. Genoma

Se le denomina genoma al conjunto de material genético hereditario que contiene las instrucciones para el desarrollo y funcionamiento de un organismo. Comúnmente se compone por moléculas de ADN, y algunos virus pueden tener moléculas de ARN.

2.1.7. Metagenómica

La metagenómica es el estudio e identificación del conjunto de genomas microbianos presentes en muestras ambientales o de algún hospedero con el uso de tecnologías de secuenciación masiva o secuenciación de nueva generación (en inglés *Next generation sequencing*, NGS), sin necesidad de aislarlos y cultivarlos en laboratorios. Es decir, permite extraer todos los ácidos nucleicos directamente de las muestras. Esta tecnología fue crucial en la identificación del virus de síndrome respiratorio agudo severo coronavirus 2 (SARS-CoV2) (Schmidt et al., 2021).

A través de los análisis metagenómicos, se obtiene información útil para una mejor compresión de la ecología y evolución de las comunidades de microorganismos de un ecosistema de interés. Esto es posible mediante el uso de programas bioinformáticos que permiten ensamblar los diferentes genomas, caracterizar la abundancia y diversidad de especies, e identificar las funciones que cumplen los microorganismos.

2.1.8. Sistema CRISPR-Cas

El locus CRISPR, por su acrónimo en inglés de "repeticiones palíndromas agrupadas y regularmente espaciadas" (en inglés *Clustered Regularly Interspaced Short Palindromic Repeats*), y las proteínas Cas (en inglés *CRISPR-associated*) son considerados un sistema inmune adaptativo y hereditable encontrado en aproximadamente el 40% de las bacterias y 90% de las arqueas el cual, les permite escindir ácidos nucleicos exógenos comúnmente de bacteriófagos y plásmidos (Hille et al., 2018).

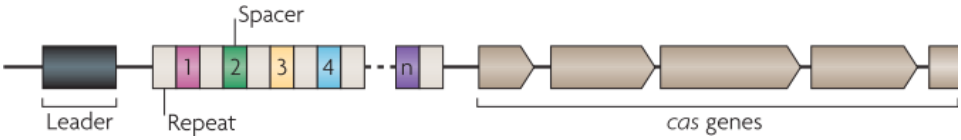

Figura 2. Representación del locus CRISPR-Cas (Marraffini & Sontheimer, 2010).

La estructura del locus CRISPR (Figura 2) consiste en una matriz con subsecuencias de nucleótidos desde dos a cientos de repeticiones directas (*direct repeats - DR*) parcialmente palíndromas que varían de tamaño entre 23 a 47 pb, conservándose entre especies bacterianas y arqueas (Nasko et al., 2019). Estos repetidores se encuentran separados por subsecuencias de nucleótidos únicas a las que se les conoce como espaciadores (*spacers*), los cuales tienen un tamaño de entre 25 a 75 pb, y que se ha demostrado que coinciden con genomas de fagos y plásmidos. Por lo tanto, este sistema representa un registro de infecciones cronológicas donde cada espaciador especifica el ADN invasor como un objetivo de interferencia.

En el extremo inicial izquierdo (5') del locus CRISPR se encuentra una secuencia de nucleótidos no codificantes llamada "líder" la cual es una región rica en adenina (A) y timina (T), y puede ser de 47 – 140 pb. Los genes asociados a CRISPR (genes Cas) que codifican las proteínas que participan en el sistema de defensa, se encuentran de manera adyacente a la matriz CRISPR. Dependiendo de la

clase del sistema, se pueden encontrar en el extremo inicial (5'), o en el extremo final (3') de la matriz CRISPR (Nasko et al., 2019; Shah et al., 2013).

De acuerdo con la naturaleza del sistema CRISPR-Cas, se pueden clasificar en dos clases y seis tipos. Estos tipos, a su vez, se dividen en subtipos. En la clase 1 se encuentran los sistemas de tipo I con subtipos A, B, C, D, E, F, F_T y G, los del tipo III con subtipos A, B, C, D, E y F y del tipo IV con los subtipos A1, A2, A3, B, C, D y E. En la clase 2 se encuentran los sistemas de tipo II con subtipos A, B y C, los del tipo V con los subtipos A, B1, B2, C, D, E, F, F1, F2, F3, G, H, I, J y K y para los sistemas de tipo VI, los subtipos A, B1, B2, C y D. Los subtipos se distinguen por su organización estructural y por las proteínas Cas, ya que algunas, son específicas de subtipos. La principal diferencia entre los sistemas de clase 1 y la clase 2 es la manera en que llevan a cabo el proceso de interferencia.

Las principales etapas de inmunidad de los sistemas CRISPR-Cas, representadas en la Figura 3, son (Hille et al., 2018; Koonin et al., 2019):

i) Adaptación. Un complejo de proteínas Cas se dirige al ADN invasor reconociendo una secuencia de entre 2 – 4 pb conocida como Motivo Adyacente Protospacer (PAM) y seleccionando una parte del ADN, que se le conoce como protoespaciador objetivo (en inglés *protospacer*), lo corta para posteriormente insertar ese segmento entre las repeticiones directas parcialmente palíndromas en la ubicación proximal a la secuencia líder de la matriz. En esta etapa, en la mayoría de los sistemas CRISPR-Cas son las proteínas Cas1 y Cas2 quienes realizan el proceso de adopción.

ii) Expresión. La matriz CRISPR se transcribe en los denominados precursores-ARN CRISPR (pre-crARN), posteriormente estos son procesados por proteínas Cas para crear ARN CRISPR (crARN) maduros. En cierto tipo de sistemas la proteína Cas6 o ARNasas externas son las encargadas de procesar los pre-crARN.

iii) Interferencia. En esta etapa los crARN maduros guían a una o a más nucleasas Cas, hacia el ADN objetivo invasor (protoespaciador) tras detectar la secuencia PAM adyacente; se une a esta con sus bases complementarias para escindirlo. En los sistemas de clase 1 el proceso de interferencia puede ser llevado a cabo por varias proteínas Cas adicionales a las Cas1 o Cas2 como las Cas3, Cas4, Cas9 o una Transcriptasa Inversa (RT), en el caso de los sistemas clase 2 utilizan una única proteína como la Cas9 para escindir el ADN objetivo.

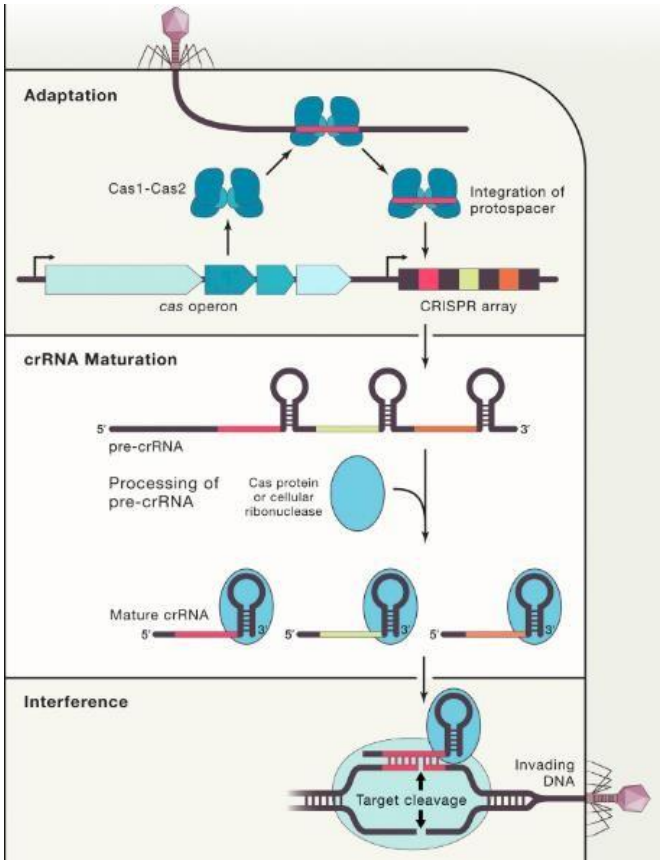

Figura 3. Principales etapas del sistema inmune adaptativo CRISPR-Cas (Hille et al., 2018).

La secuencia líder es la que garantiza la correcta transcripción de pre-crARN y de la integración de nuevos espaciadores entre las repetidores parcialmente palíndromos (Karimi et al., 2018).

Se estipula que en promedio las bacterias contienen tres matrices CRISPR y el número promedio de repetidores por matriz es de trece y la longitud promedio de estas subsecuencias es de 30 nucleótidos. Por otra parte, el promedio de matrices CRISPR en arqueas es de cinco y el número promedio de repetidores por matriz es de dieciocho; la longitud promedio es de 29 nucleótidos (Alkhnbashi et al., 2016; Karimi et al., 2018).

El sistema CRISPR fue reconocido en el año 2000 y desde entonces su área de investigación ha ido creciendo exponencialmente, beneficiando a aplicaciones como la edición genética de genomas y alternativas a antibióticos como la fago-terapia (Azam et al., 2019; Jinek et al., 2012).

2.2. Conceptos computacionales

2.2.1. Inteligencia artificial

La inteligencia artificial (IA) básicamente se puede definir como un campo dentro de las ciencias computacionales que tiene como objetivo el estudio y desarrollo de algoritmos, técnicas, métodos y teorías que simulen el comportamiento inteligente humano para la automatización de tareas, generar nuevo conocimiento en base al aprendizaje y, el razonamiento lógico para la resolución de problemas y la toma de decisiones.

Entre los principales campos de la IA se encuentra el aprendizaje automático, procesamiento de lenguaje natural, visión por computador, procesamiento de voz, minería de datos, robótica, entre otros.

2.2.2. Aprendizaje automático

El aprendizaje automático o máquina (en inglés *machine learning*, ML) es el área que desarrolla máquinas o programas informáticos que, mediante algoritmos y modelos estadísticos, obtengan la capacidad de aprender con base a la experiencia sin intervención humana. Este tipo de modelos son principalmente implementados para la extracción de datos relevantes en grandes y complejos conjuntos de datos.

Las cuatro principales categorías de los enfoques de ML son el aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semi-supervisado y el aprendizaje reforzado. Los algoritmos dentro del aprendizaje supervisado hacen uso de conjuntos de datos asignados a una clase, donde el objetivo es encontrar una función que asigne las variables de entrada con una clase (etiqueta) de salida objetivo. En cambio, en los algoritmos de aprendizaje no supervisado, el conjunto de datos no se encuentra etiquetado y el objetivo es encontrar patrones en los datos que separen los datos en diferentes grupos con algún tipo de significado.

Los algoritmos de aprendizaje semi-supervisado combinan datos etiquetados y no etiquetados para generar predicciones y, finalmente, los de aprendizaje reforzado se entrenan con instrucciones de las acciones a realizar en determinados escenarios y recibe como retroalimentación la información obtenida en respuesta a sus acciones.

Entre los algoritmos predictivos más utilizados en el ML se encuentran los de Regresión lineal, Regresión logística, Árboles de decisiones, Máquinas de soporte vectorial, Bosques aleatorios, las Redes neuronales artificiales, entre otros.

2.2.3. Redes neuronales artificiales

Las redes neuronales artificiales (en inglés *artificial neural network*, ANN) son modelos inspirados en el comportamiento de las redes neuronales biológicas al momento de adquirir conocimiento y generalizar en casos desconocidos. Su arquitectura consiste, de manera general, en un conjunto de

neuronas (nodos), divididas en una o más capas conectadas sucesivamente entre sí; al recibir una señal de entrada, las neuronas la procesan y la transmiten a las neuronas conectadas de la siguiente capa. Cada conexión tiene un peso numérico que modula la fuerza de la señal que se transmite de la capa de entrada hasta la capa de salida.

Las principales aplicaciones de este tipo de modelos son para la selección de características, clasificación, reducción de dimensionalidad o dentro de las arquitecturas de las redes neuronales convolucionales (Koumakis, 2020).

2.2.4. Aprendizaje profundo

El aprendizaje profundo (en inglés *deep learning*, DL) es una rama del ML que surgió como un enfoque específico, utilizando ahora las redes neuronales profundas (por sus siglas en inglés DNN), que son derivadas de las ANN, agregando un mayor número de capas, a lo que se le conoce como profundidad del modelo.

Una de las principales ventajas de los modelos de DL sobre los de ML, es que no se requiere implementar ingeniería de características, es decir, seleccionar las características relevantes para una mejor predicción o clasificación. Esto es debido a que este procedimiento lo realiza el modelo de manera implícita a través de cada capa, donde las primeras capas extraen características sencillas y las últimas capas extraen patrones más significativos de los datos. Además, ofrecen mayor flexibilidad para el manejo de grandes conjuntos de datos con alta dimensionalidad.

Con el aumento masivo de los datos biológicos generados por tecnologías de secuenciación como NGS, ha surgido la necesidad de métodos más eficientes para el procesamiento de grandes conjuntos de datos. En este sentido, los modelos de DL han sido ampliamente utilizados en el área de la genómica, como por ejemplo para predecir e identificar unidades funcionales en las secuencias de ADN, predecir dominios de replicación, el sitio de unión del factor de transcripción (por sus siglas en ingles TFBS), punto de inicio de la transcripción, promotores, potenciador y el sitio de eliminación de genes (Zhang et al., 2019), entre otros. En la transcriptómica, se han implementado modelos de DL para analizar la estructura de secuencias de ARN, predicción de sitios de unión de RBP, sitios de empalme alternativos y tipos de ARN. En la proteómica se ha empleado para identificar estructuras de proteínas, predicción de estructuras secundarias y terciarias, modelos de evaluación de la calidad de proteína, por mencionar algunos (Zhang et al., 2019).

2.2.5. Redes neuronales convolucionales

Las redes neuronales convolucionales (en inglés *convolutional neural network*, CNN) son un tipo de algoritmos de DL que inicialmente fueron propuestas para el procesamiento de imágenes. Debido a su gran capacidad en la extracción, selección y reducción de características de los datos mediante sus capas convolucionales, han sido implementadas para el procesamiento de datos como texto, voz y video en tareas como el procesamiento de lenguaje natural, reconocimiento de voz y reconocimiento de objetos. Dentro del área biológica, este tipo de modelos han sido utilizados para la predicción de la expresión génica, clasificación de proteínas y la predicción de estructuras genéticas (Zhang et al.,

2019). La arquitectura básica de una CNN se compone por una o varias capas convolucional, capas de agrupación o submuestreo y la capa de red neuronal totalmente conectada, como se observa en la Figura 4.

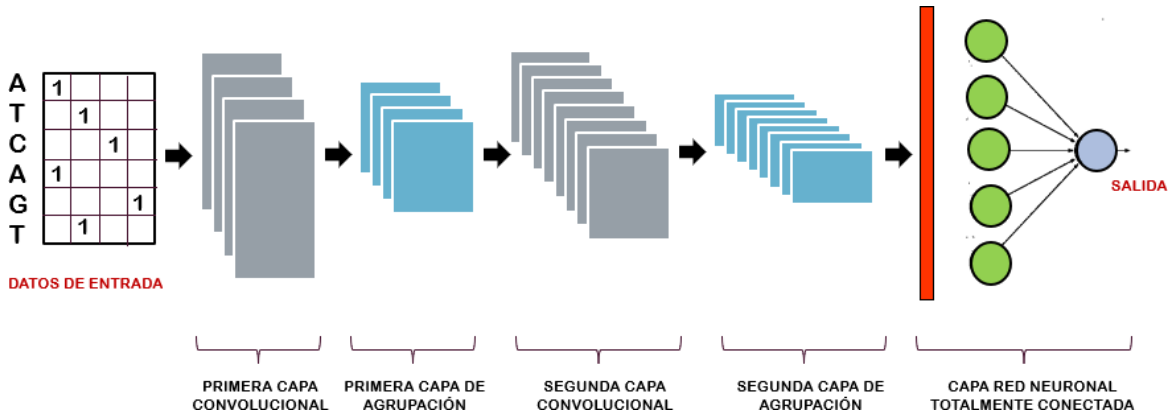

Figura 4. Representación de la arquitectura de una red neuronal convolucional.

Las CNN reciben los datos de entrada en matrices de un tamaño fijo para ser analizadas en la capa de convolución. En esta capa se realizan operaciones de convolución que consisten en el cálculo del producto escalar de la matriz con los datos de entrada contra los vectores de peso, también conocidos como filtros o núcleos (*kernels*), con la finalidad de extraer y generar mapas de características o activación. Para realizar el proceso de convolución, como se representa en la Figura 5, el filtro se desliza de izquierda a derecha a través de la matriz de entrada, este proceso se repite de acuerdo con el número de filtros definidos en la arquitectura. Los hiperparámetros principales que definen el volumen de salida de las capas convolucionales son el tamaño del filtro, el salto (*stride*), que es el número de posiciones en que se deslizará el filtro, cuan mayor es el salto, más pequeños resultan los mapas de características, y el relleno de ceros útil para preservar la dimensión de la entrada en los mapas (Emmert-Streib et al., 2020).

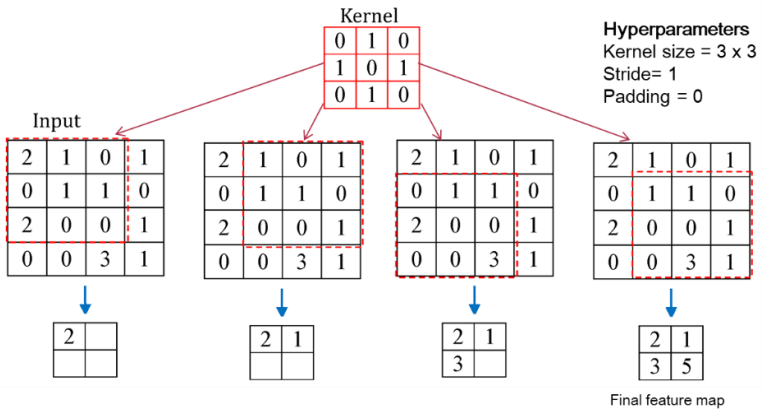

Figura 5. Ejemplo de la operación de convolución de una matriz de 4x4 y un filtro de 3x3. Imagen modificada de Naranjo-Torres et al., 2020.

A los resultados de las operaciones de los filtros, se les aplica una función de activación no lineal, para finalmente, generar los mapas de características que pasarán a la siguiente capa de la red. Existen diferentes tipos de funciones de activación (ReLU, sigmoide, tanh, etc.), una de las más utilizadas es la función de unidad lineal rectificada (en inglés *rectified linear unit*, ReLU), matemáticamente se define como la ecuación 1 (Naranjo-Torres et al., 2020):

$$f(x) = \max(0, x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases} \quad (1)$$

La siguiente capa es la de agrupación, donde las operaciones realizadas en esta capa tienen como objetivo reducir el número de parámetros a entrenar. De manera similar a los filtros, se pueden definir hiperparámetros para los núcleos de agrupación como: el tamaño, el salto y relleno con ceros. Las principales funciones implementadas son la de agrupación máxima y agrupación por promedio, donde la primera, es la que ha demostrado mejor eficiencia en el procesamiento de imágenes. Los núcleos de agrupación máxima (en inglés *max pooling*) extraen el valor máximo dentro de cada sub espacio analizado en el mapa de características. En la Figura 6 se muestra un ejemplo de los resultados de las operaciones (Emmert-Streib et al., 2020).

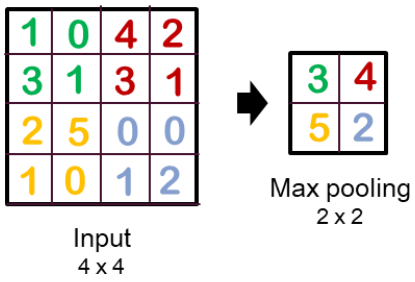

Figura 6. Ejemplo de operación del kernel de agrupación máxima de 2x2.

De manera opcional, después de la capa de agrupación, se implementa una capa que incluye la regularización para evitar el problema de sobreajuste y mejorar la generalización del modelo CNN final. Una de las técnicas mayormente empleadas es la denominada abandono (en inglés *dropout*), que consiste en deshabilitar aleatoriamente un determinado porcentaje de neuronas ocultas de la red y de esta manera se evita que el modelo memorice los pesos durante el entrenamiento.

Finalmente, las salidas de la capa de agrupación/capa de regularización, las unidades neuronales se transforman a un vector de una dimensión para ser introducido como entrada a la última capa totalmente conectada (en inglés *fully connected*, FC). Esta última capa profundiza la búsqueda de patrones en las características de alto nivel recibidas como entrada para mapearlas a una etiqueta (clase) de salida mediante una función de activación que calcula la probabilidad para cada clase. En los modelos de clasificación multiclase, se utiliza la función Softmax, que calcula las probabilidades de cada clase usando una función exponencial normalizada que determina la clase con la probabilidad más alta obtenida. Las probabilidades de clasificación se encuentra en el rango entre 0 y 1, y la suma total de las mismas es igual a 1 (Sony et al., 2021).

CAPÍTULO 3. METODOLOGÍA

En este capítulo se presenta la metodología empleada y como parte fundamental, el desarrollo del modelo de aprendizaje profundo propuesto. Este tiene como objetivo, realizar la caracterización o detección de matrices CRISPR en secuencias de nucleótidos bacterianas y la extracción de las subsecuencias espaciadoras para que, posteriormente, puedan ser utilizadas para realizar predicciones de interacción fago - bacteria hospedera.

Básicamente, el proceso utilizado consiste en las siguientes etapas: (i) Extracción y depuración de los datos: en esta etapa se exportaron los datos anotados del sistema CRISPR-Cas de la base de datos (BD) CRISPRCasdb (Pourcel et al., 2020) y se realizó la depuración para obtener los datos correspondientes a las secuencias de nucleótidos bacterianas; (ii) Definición del conjunto de datos de entrenamiento y prueba: en esta etapa se realizó un análisis exploratorio de las subsecuencias de las repeticiones directas y espaciadores que integran las matrices CRISPR en las secuencias de la etapa i. Como resultado, se generó el conjunto de datos de entrenamiento y prueba; (iii**)** Modelo computacional: en esta etapa se desarrolló el modelo de red neuronal convolucional (CNN) basado en aprendizaje profundo, para detectar las regiones de matrices CRISPR en secuencias bacterianas y la aplicación de las métricas de evaluación del rendimiento del modelo; (iv) Posprocesamiento de los resultados obtenidos del modelo: en esta etapa, se filtraron los resultados de la red mediante sus probabilidades de predicción para definir umbrales de confianza; asimismo, se realizó un proceso adicional para las estructuras CRISPR con longitudes inferiores o superiores a la estándar; (v) Aplicación del modelo computacional en un genoma bacteriano completo: se procesó una secuencia de nucleótidos de una bacteria real (*Actinoalloteichus sp. AHMU CJ021*), a partir de los resultados predictivos, se extrajeron las subsecuencias espaciadoras y se realizó su alineamiento con la base de datos de bacteriófagos con la finalidad de identificar al virus anfitrión de la bacteria analizada. A continuación, se describe en forma detallada cada una de estas etapas de la metodología de solución.

3.1. Exploración y depuración del conjunto de datos

CRISPRCasdb (Pourcel et al., 2020), es una base de datos (BD) relacional que contiene datos predichos y validados del sistema CRISPR-Cas de arqueas y bacterias, obtenidos mediante el análisis de secuencias genómicas y de cromosomas disponibles en la BD de nucleótidos GenBank de NCBI, con la herramienta CRISPRCasFinder.

La BD CRISPRCasdb fue usada en este trabajo dado a su constante actualización. Los datos extraídos corresponden a la fecha del 03 de julio del 2020. De manera general, se descargó un total de 20,189

matrices CRISPR que fueron identificadas en 9,174 genomas bacterianos, que pertenecen a 4,153 especies diferentes. La información detallada se encuentra en la Tabla 5.

Tabla 5. Resumen de datos exportados de CRISPRCasdb al 3 de julio del 2020.

Datos exportados	Cantidad
Repeticiones directas	387,242
Espaciadores	367,053
ID matrices CRISPR	20,189
ID secuencias bacterianas	9,174
Taxones bacterianos	4,153

3.1.1. Número de matrices CRISPR por secuencia bacteriana

Un genoma bacteriano puede contener más de un sistema CRISPR-Cas funcional a lo largo de su secuencia, así que el primer análisis fue contabilizar el número de matrices CRISPR contenidas en las 9,174 secuencias bacterianas exportadas de CRISPRCasdb.

La distribución y densidad de la frecuencia del número de matrices CRISPR por secuencia bacteriana son presentados en un diagrama de dispersión en la Figura 7a; y en un diagrama de violín en la Figura 7b. La cantidad mínima de matrices CRISPR es de 1 y la máxima es de 37. En las medidas estadísticas de tendencia central, tanto la mediana como la moda indican 2 matrices por genoma; el rango intercuartílico indica que el 25% de las secuencias contiene 1 matriz, el 50%-75% se concentran en 2 matrices y el 25% de las secuencias restantes llegan a un máximo de 3, mientras las cantidades superiores de matrices CRISPR, hasta 37, reflejan casos atípicos.

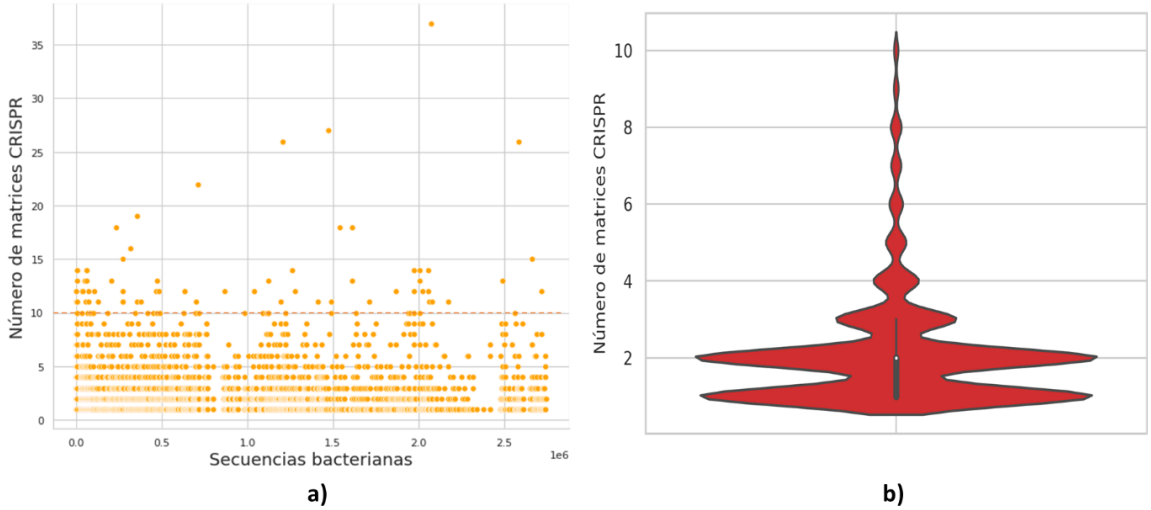

*Figura 7. Número de matrices CRISPR en los 9,174 genomas de bacterias. **a)** Dispersión del número de matrices CRISPR por genoma. **b)** Diagrama de densidad. En este gráfico se omitieron las frecuencias atípicas superiores a 10 matrices CRISPR.*

En la Tabla 6 se muestra la información de los 10 genomas bacterianos con el mayor número de matrices CRISPR. El genoma con mayor cantidad de matrices CRISPR pertenece a la bacteria *Actinoalloteichus sp. AHMU CJ021* con 37 y, en segundo lugar, se encuentra la secuencia de la bacteria *Actinoalloteichus hoggarensis* con 27 matrices CRISPR; interesantemente la mayoría de los genomas bacterianos son de la clase *Actinomycetia* y del filo *Cyanobacteria*.

Tabla 6. Las 10 secuencias bacterianas con mayor número de matrices CRISPR.

Clase	Nombre científico	No. de matrices
Actinomycetia	*Actinoalloteichus sp. AHMU CJ021*	37
Actinomycetia	*Actinoalloteichus hoggarensis*	27
Actinomycetia	*Nocardiopsis alba ATCC BAA-2165*	26
Actinomycetia	*Streptomyces sp. WAC08241*	26
Actinomycetia	*Actinomyces sp. oral taxón 414*	22
Actinomycetia	*Streptomyces sp. Tu6071*	19
Cyanobacteria (filo)	*Tolypothrix tenuis PCC 7101*	18
Cyanobacteria (filo)	*Aulosira laxa NIES-50*	18
Actinomycetia	*Actinoalloteichus sp. ADI127-7*	18
Cyanobacteria (filo)	*Nostoc sp. PCC 7107*	16

Las bacterias podrían registrar un mayor número de matrices CRISPR, dependiendo del ecosistema en el que se encuentran. Si se encuentran con mayor exposición a los virus que las infectan podrían verse obligadas a desarrollar más sistemas CRISPR-Cas donde almacenen diferentes fragmentos de genomas invasores e incluso de genomas repetidos de diferentes regiones de la secuencia, que les permitan identificar y defenderse fácilmente de los fagos o plásmidos en una nueva reinfección.

3.1.2. Número de repeticiones directas y espaciadores por cada una de las matrices CRISPR

Asimismo, se analizó la estructura de las 20,189 matrices CRISPR contenidas en los 9,174 genomas y se contabilizó la cantidad de repeticiones directas y de los espaciadores por matriz con la finalidad de conocer la cantidad de subsecuencias que conforman estas regiones.

En la Figura 8a se muestra el diagrama de dispersión de la cantidad de repeticiones directas, mientras en la Figura 8b muestra el diagrama de violín con la distribución y densidad de frecuencia. En promedio cada matriz tiene 12 repeticiones directas y existiendo una matriz con 588.

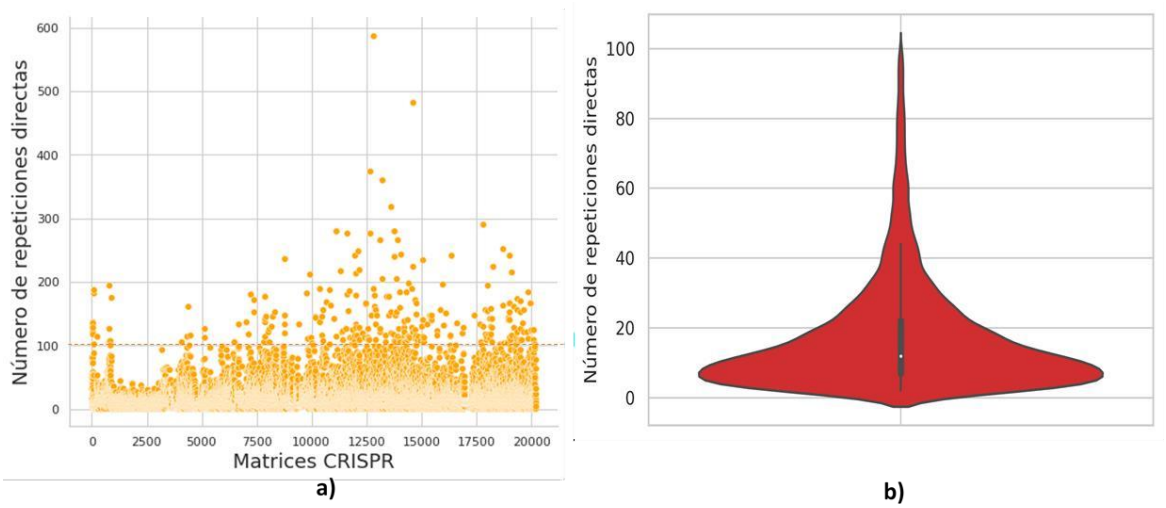

Figura 8. Número de repeticiones directas en las matrices CRISPR: ***a) N****úmero de repeticiones por matriz CRISPR.* ***b)*** *Diagrama de densidad. En este gráfico se omitieron las frecuencias atípicas superiores a 100 repeticiones directas por matriz CRISPR.*

El número de repeticiones directas siempre es mayor por uno que el número de espaciadores (Tabla 7), debido a que las repeticiones directas flanquean a los espaciadores.

Tabla 7. Comparación del resumen estadístico del análisis de la cantidad de repeticiones directas y espaciadores por matriz CRISPR.

	Repeticiones directas	**Espaciadores**
Matrices CRISPR analizadas	20,189	20,189
Cantidad mínima	2	1
Cantidad máxima	588	587
Media aritmética	19	18
Desviación estándar	23	23
Mediana	12	11
Moda	4	3
Primer cuartil-Q1 (25%)	7	6
Segundo cuartil-Q2 (50%)	12	11
Tercer cuartil-Q3 (75%)	23	22

3.1.3. Longitud en nucleótidos de las repeticiones directas y espaciadores

El análisis de las longitudes en nucleótidos de las repeticiones directas y espaciadores permitió conocer las medidas estadísticas básicas de las matrices CRISPR anotadas, lo cual fue útil en el procesamiento del conjunto de datos y la formulación del modelo computacional propuesto.

La Figura 9a presenta la longitud en nucleótidos (nt) de cada una de las 387,242 repeticiones directas analizadas; mientras que la Figura 9b muestra su correspondiente diagrama de violín con la distribución y la densidad de las frecuencias de estas subsecuencias. Se puede observar que la longitud mínima es de 23, la mediana de 29 y la máxima de 51 nucleótidos.

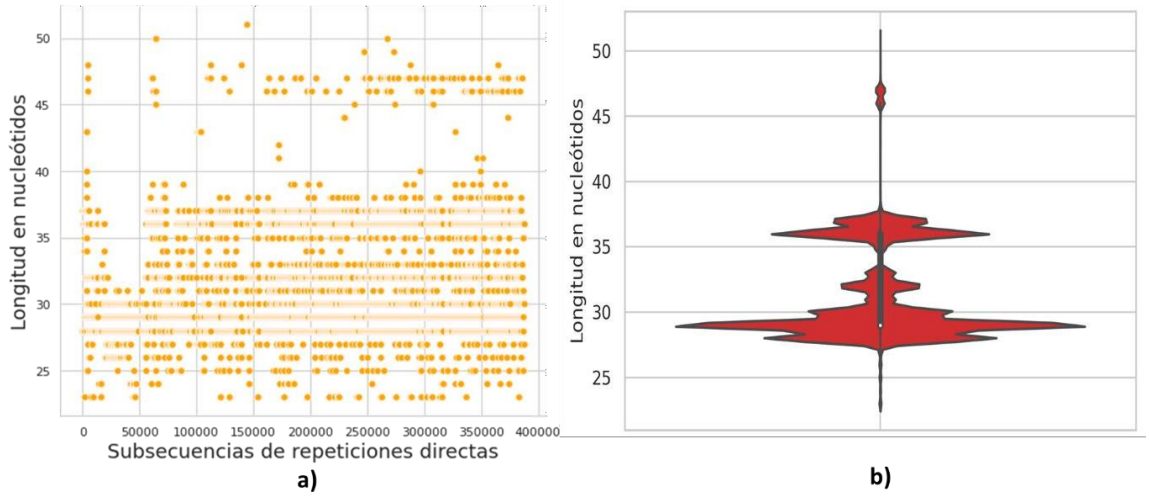

Figura 9. Análisis de las longitudes en nucleótidos de las 387,242 repeticiones directas del conjunto de datos: ***a)*** *Diagrama de dispersión de la longitud en nt de las repeticiones directas.* ***b)*** *Distribución y densidad de la longitud en nt de las repeticiones directas.*

De la misma manera, la Figura 10a presenta la longitud en nucleótidos (nt) de los 367,053 espaciadores analizados; mientras que la Figura 10b muestra el diagrama de violín de su distribución y densidad. Se obtuvo que la longitud mínima es de 15, la mediana es de 33 y la máxima de hasta 108 nucleótidos, lo que indica que los espaciadores tienen un mayor rango de longitudes que los repetidores.

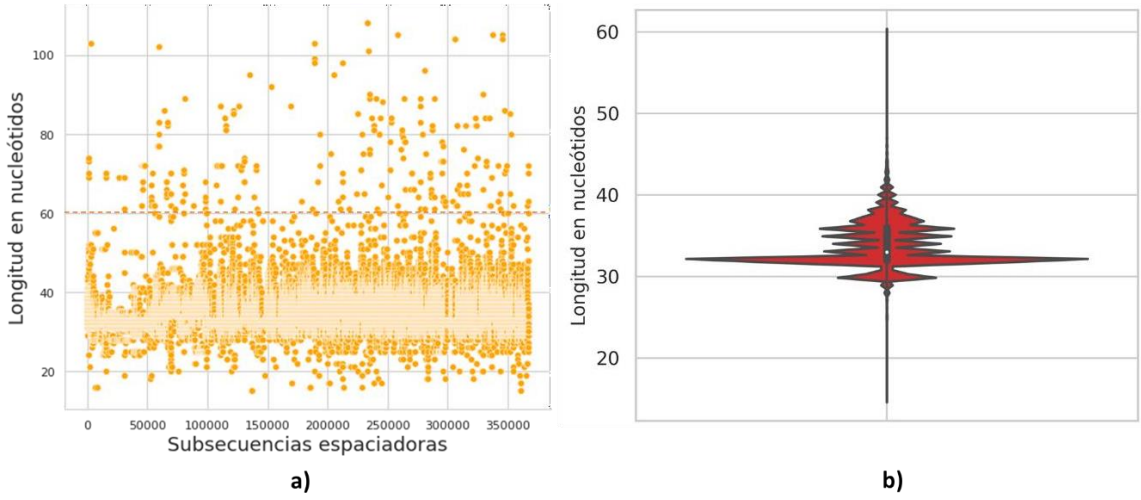

Figura 10. Longitudes de los 367,053 espaciadores del conjunto de datos: ***a)*** *Dispersión de las longitudes de los espaciadores.* ***b)*** *Diagrama de densidad. En este gráfico se omitieron las frecuencias atípicas superiores a 60 nucleótidos de longitud, por lo que sólo se graficaron las frecuencias correspondientes a 366,739 espaciadores.*

Para examinar de manera más detallada las similitudes y diferencias de las longitudes en nucleótidos de las repeticiones directas y espaciadores, se realizó un análisis de los parámetros de tendencia central y dispersión de ambos conjuntos (ver Tabla 8). A pesar de que existen espaciadores más grandes que los repetidores (108 vs 51), el promedio de longitudes tiende a ser similar, siendo 33 y 31 nucleótidos, respectivamente.

Tabla 8. Comparación de las longitudes en nucleótidos de las repeticiones directas y espaciadores.

	Repeticiones directas	**Espaciadores**
Subsecuencias analizadas	387,242	367,053
Longitud mínima	23	15
Longitud máxima	51	108
Media aritmética	31	33
Desviación estándar	3	3
Mediana	29	33
Moda	29	32
Primer cuartil-Q1 (25%)	29	32
Segundo cuartil-Q2 (50%)	29	33
Tercer cuartil-Q3 (75%)	36	36

3.1.4. Análisis del agrupamiento de repetidores y espaciadores

Las repeticiones directas tienen la particularidad de ser de tamaños poco variables, encontrarse de manera repetida, sólo siendo separadas por una subsecuencia espaciadora de tamaño similar, y presentar patrones parcialmente palíndromos. Se espera que el modelo computacional propuesto en este trabajo sea capaz de reconocer estos patrones. Por tal motivo, se realizó el análisis de la conservación de las repeticiones directas y espaciadores.

Para analizar el nivel de conservación de las subsecuencias exportadas de CRISPRCasdb, se usó la herramienta bioinformática denominada CD-HIT-EST (Huang et al., 2010), la cual, mediante su algoritmo heurístico, realiza agrupaciones (*clusters*) de grandes cantidades de secuencias biológicas en función con un porcentaje de similitud definido, con la finalidad de reducir el tamaño o eliminar la redundancia en conjuntos de datos.

Con el objetivo de eliminar la redundancia de subsecuencias idénticas y evitar un sesgo en el conjunto de entrenamiento, se procesó las 387,198 repeticiones directas y los 367,053 espaciadores con el programa CD-HIT-EST a un porcentaje de identidad del 100%. En la Tabla 9 se presentan los resultados obtenidos. En las repeticiones directas, se obtuvo un total de 20,032 grupos de los cuales

11,693 grupos, que representan el 58%, contienen una sola repetición directa que no son idénticas a otras subsecuencias. El grupo con la mayor cantidad subsecuencias contiene 49,122 repeticiones directas 100% similares y la subsecuencia representante de este grupo pertenece a una secuencia de la especie *Escherichia coli*.

En cuanto al agrupamiento de los espaciadores, se obtuvo un total de 227,417 grupos de los cuales 192,404 grupos, que representan el 84%, contienen un solo espaciador (Tabla 9). El grupo con la mayor cantidad de subsecuencias contiene 475 espaciadores idénticos con 100% de similitud y la subsecuencia representante de este grupo pertenece a una secuencia de la especie *Mycobacterium tuberculosis PanR0317*.

Tabla 9. Agrupamiento al 100% de similitud de repeticiones directas y espaciadores.

	Repeticiones directas	**Espaciadores**
Cantidad de subsecuencias analizadas	387,242	367,053
Número de grupos obtenidos	20,032	227,417
Número de grupos con una subsecuencia	11,693	192,404
Cantidad mayor de subsecuencias en un grupo	49,122	475
Taxón de la subsecuencia representante	562	1346776
Especie de la subsecuencia representante	*Escherichia coli*	*Mycobacterium tuberculosis PanR0317*

Las repeticiones directas tuvieron un mayor nivel de similitud o conservación, debido a que se formó un menor número (20,032) de grupos que contienen subsecuencias 100% idénticas en comparación con los espaciadores (227,417 grupos obtenidos); además, en las repeticiones directas sólo el 58% de los grupos contienen una sola subsecuencia. Estos resultados indican que las repeticiones directas tienen una tendencia a conservarse en las diferentes matrices CRISPR e incluso entre diferentes genomas de bacterias que pueden o no estar relacionadas taxonómicamente, es decir, que pueden o no derivar de un mismo ancestro común.

En cambio, los espaciadores mostraron tener menor similitud entre ellos dado que el 84% de grupos contienen una sola subsecuencia, lo que indica que los espaciadores no son regiones conservadas entre las bacterias. Esto se debe a que estas subsecuencias representan el material exógeno que integran las bacterias en sus matrices CRISPR para utilizarlos como defensa en nuevas reinfecciones. Por lo tanto, los fragmentos que integran las bacterias en su sistema de autoinmunidad dependen de los virus a los que se encuentren expuestas y a la región específica que seleccionen para escindir el ADN invasor.

3.1.5. Análisis de diversidad de especies bacterianas en grupos de las repeticiones directas

Para conocer a detalle la conservación estructural de las repeticiones directas, se analizaron los 20,032 grupos obtenidos en el agrupamiento al 100% de identidad (similitud). En la Tabla 10 se muestra la

información de los cuatro grupos más diversos, es decir, que agruparon repeticiones directas que fueron encontradas en diferentes especies bacterianas.

El grupo de repeticiones directas con la mayor diversidad contabilizó un total de 688 especies bacterianas diferentes que comparten repetidores con un 100% de similitud y el repetidor representante de este grupo pertenece a un genoma de la bacteria *Escherichia coli*. A su vez, se encontró que un total de 16,804 grupos (83% del total de grupos) contienen repeticiones directas encontradas en la misma especie bacteriana y no contienen diversidad.

El segundo grupo registró un total de 388 especies diferentes donde la repetición directa representante del grupo pertenece a la especie *Salmonella entérica*. El tercer grupo con mayor diversidad registró 264 especies diferentes y la repetición representante pertenece a la especie *Salmonella enterica subsp. enterica serovar Pullorum* y finamente, el cuarto grupo registró un total de 236 especies diferentes y la repetición directa representante pertenece a la especie *Klebsiella aerogenes*.

Tabla 10. Grupos de repeticiones directas al 100% de similitud con mayor diversidad de especies diferentes.

	Primer grupo	**Segundo grupo**	**Tercer grupo**	**Cuarto grupo**
ID de grupo	7141	8509	8460	11340
Número de especies diferentes	688	388	279	236
Taxón de la especie representante	562	28901	605	548
Nombre científico de la especie representante	*Escherichia coli*	*Salmonella enterica*	*Salmonella enterica subsp. serovar Pullorum*	*Klebsiella aerogenes*

3.2. Modelo computacional

Mediante el análisis de las subsecuencias de nucleótidos que integran las matrices CRISPR en las secuencias bacterianas exportadas de CRISPRCasdb, fue posible generar información de utilidad para el desarrollo del modelo computacional que caracterice las matrices CRISPR en las secuencias de nucleótidos de bacterias.

Se concluyó que las repeticiones directas tienen un alto porcentaje de conservación entre diferentes secuencias bacterianas e incluso entre bacterias de diferentes especies o cepas. Por consiguiente, se consideran viables para ser reconocidas como rasgos característicos de las matrices CRISPR dentro de las secuencias de nucleótidos, con el objetivo de identificar y extraer estas regiones mediante el modelo con técnicas de aprendizaje profundo. Posteriormente, las subsecuencias espaciadoras extraídas, pueden ser utilizadas para la definición de interacciones de infección bacteriófago – hospedero.

3.2.1. Definición del proceso para la identificación de matrices CRISPR en genomas de bacterias

En el análisis de los datos exportados de CRISPRCasdb se encontró que la estructura mínima de las matrices CRISPR se conforma por una sola subsecuencia espaciadora, la cual representa el fragmento de fago integrado, con sus dos subsecuencias de repeticiones directas flanqueantes, la primera en su extremo 5' que corresponde al lado izquierdo y la segunda al extremo 3' que corresponde al lado derecho, como se muestra en el ejemplo de la Figura 11.

Figura 11. Ejemplo de la estructura mínima de una matriz CRISPR en secuencias bacterianas.

Como se presentó en la sección 3.2.2, se calculó que la mediana del número de espaciadores por matriz CRISPR es de 11 y que la matriz con mayor cantidad de espaciadores contiene un total de 587, pero, se encontró que la mayoría eran matrices con la estructura mínima. Lo que indica que el tamaño de la estructura de estas regiones es muy variable ya que depende de la cantidad de fragmentos integrados o heredados por las bacterias.

Por tal motivo, para la detección de las matrices CRISPR en las secuencias de nucleótidos bacterianas, se decidió basarse en el patrón de las matrices con la estructura mínima. De modo que, cuando se enfrente a matrices CRISPR que contienen dos o más espaciadores, se identifiquen por fragmentos que sigan el patrón estructural de las estructuras mínimas, a los cuales llamamos *estructura CRISPR*. Es decir, las matrices con más de un espaciador, se identificarán dividiéndolas en subsecuencias que cumplen con la estructura mínima de una matriz (un espaciador y sus dos repeticiones directas flanqueantes), por lo tanto, la cantidad de estructuras CRISPR a extraer de una matriz CRISPR, será la misma de la cantidad de espaciadores que contenga.

En la Figura 12, se representa el ejemplo de una matriz CRISPR que contiene dos espaciadores, por lo que se fragmentará en dos estructuras CRISPR con un traslape de nucleótidos entre las mismas de la longitud de la repetición directa. Por lo tanto, si la matriz CRISPR contiene ***n*** subsecuencias espaciadoras, se extraerán ***n*** estructuras CRISPR.

Figura 12. Ejemplo de fragmentación de una matriz CRISPR a estructuras CRISPR.

Considerando lo anterior y que los genomas bacterianos pueden llegar a tener longitudes de entre 1 a 14.5 millones de nucleótidos, es necesario la aplicación de un preprocesamiento a la secuencia de nucleótidos de entrada al modelo computacional. El preprocesamiento de los datos de entrada consiste en la fragmentación de la secuencia del genoma bacteriano a analizar en *k*-meros del tamaño estándar de las matrices CRISPR con la estructura mínima. Los *k*-meros son subsecuencias de una longitud *k* extraídos de una secuencia de nucleótidos; el proceso para obtener todos los *k*-meros posibles inicia tomando los primeros *k* caracteres, luego se aplica un *salto* de caracteres para la extracción del siguiente *k*-mero y así sucesivamente se va iterando sobre la *longitud* total de la cadena. La cantidad de *k*-meros a extraer de una secuencia está dada por la ecuación (2).

$$\boldsymbol{k-\text{meros} = \frac{(Longitud - k)}{salto} + 1} \quad \textbf{(2)}$$

En la Figura 13 se muestra un ejemplo de la fragmentación en *k*-meros de una secuencia de 33 nucleótidos de longitud. De acuerdo con la ecuación (2), para calcular la cantidad de *k*-meros, si consideramos la secuencia de 33 nt de longitud (*Longitud* = 33), *k*-meros de 8 nt (*k* = 8) y saltos de 5 nt (*salto* = 5), como resultado de la fragmentación se obtendrían 6 subsecuencias.

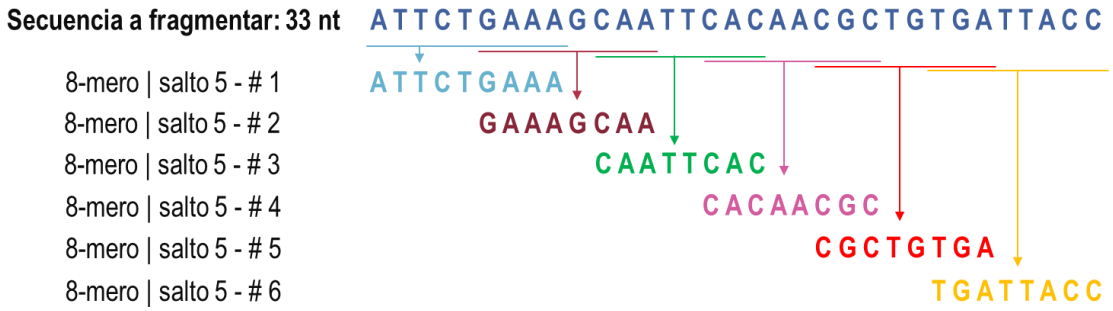

Figura 13. Ejemplo de fragmentación de secuencia en k-meros de longitud 8.

A partir del análisis de las matrices CRISPR (sección 3.2.2), se definió el tamaño estándar de las estructuras como de 100 nucleótidos, considerando 32 nucleótidos para las repeticiones directas (32 en cada extremo), lo cual es cercano a la longitud de la media aritmética y 36 nucleótidos para los espaciadores tomando en cuenta que es la longitud marcada por el tercer cuartil que representa el 75% de las subsecuencias.

Dado que el modelo computacional solo puede recibir entradas de subsecuencias de nucleótidos de una longitud fija, sin importar la *longitud* del genoma bacteriano completo, las subsecuencias de entrada a analizar son en *k*-meros de 100 nucleótidos (*k* = 100) con saltos de 5 nucleótidos (*salto* = 5).

De modo que el modelo computacional recibe las subsecuencias y extrae los 32 nt en ambos extremos de las subsecuencias de entrada y evalúa si contienen los patrones característicos de las repeticiones

directas para finalmente clasificarlas como estructuras CRISPR verdaderas (positivas) o falsas (negativas). La Figura 14 muestra la representación del proceso de la extracción y evaluación de los extremos de las subsecuencias extraídas de la secuencia bacteriana a analizar.

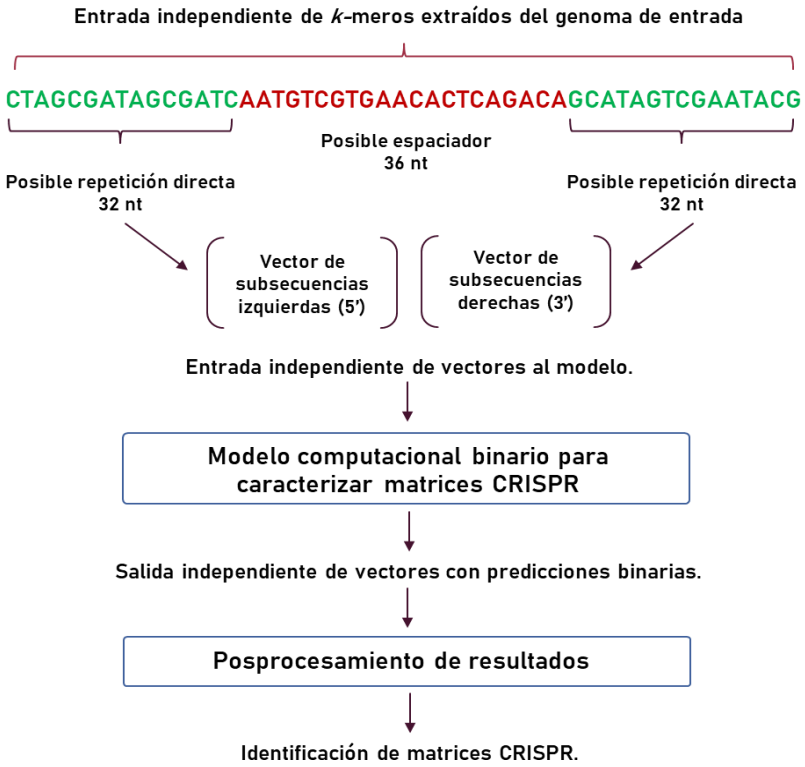

Figura 14. Representación del proceso de extracción y evaluación de las subsecuencias de entrada al modelo binario.

3.2.2. Generación del conjunto de datos

Con el objetivo de identificar las potenciales regiones que representan matrices CRISPR, el modelo computacional binario debe evaluar los 32 nucleótidos extraídos de ambos extremos de las subsecuencias de 100 nt obtenidas a partir de la fragmentación de la secuencia bacteriana recibida como entrada. Para ello, la evaluación de los extremos izquierdos (5') y derechos (3') de cada subsecuencia, se realiza de manera independiente.

El modelo computacional binario para reconocerá los patrones estructurales de las subsecuencias de repeticiones directas a partir del conjunto de repeticiones directas positivas y un segundo conjunto de repeticiones directas negativas.

En la sección 3.2.4. se analizó la conservación estructural de las 387,242 repeticiones directas exportadas de CRISPRCasdb generando 20,053 grupos con similitudes de secuencias del 100%. A partir de estos grupos, se extrajo la subsecuencia representativa construyendo el conjunto de datos de la clase positiva, la cual no tiene redundancia.

Como ya se mencionó, se estableció 32 nucleótidos de longitud para las subsecuencias de entrada del modelo, donde cada nucleótido es representado por su letra correspondiente. Sin embargo, muchas repeticiones directas son de una longitud menor, siendo de 23 nucleótidos la longitud mínima encontrada. Por lo tanto, se procesó las 20,032 repeticiones directas de la siguiente manera:

- Si la longitud era menor a 32 nt, se concatenó el número de nucleótidos faltantes en el extremo izquierdo con los de los espaciadores reales. Por otra parte, si la longitud era mayor a 32 nt, se truncó la subsecuencia en su extremo izquierdo, a 32.

- De manera similar, se aplicó el mismo proceso al conjunto inicial de ejemplos positivos en el extremo derecho.

- Las 40,064 repeticiones directas generadas fueron sometidas nuevamente a CD-HIT-EST con un porcentaje de similitud al 100% para eliminar nuevamente la redundancia en dicho conjunto, ya que las subsecuencias que ya cumplían con la longitud de 32 nucleótidos, en los dos procesos anteriores pasaron sin modificación.

De esta manera, se obtuvo un total de 37,374 repeticiones directas sin redundancia que conformaron el conjunto de datos de la clase positiva, donde cada subsecuencia tuvo una longitud de 32 nucleótidos y con su etiqueta de clase positiva o verdadera representada por el número 1.

Por otra parte, para generar el conjunto de datos para la clase negativa, se tomó el conjunto de 12,000 repeticiones directas falsas propuestas por Wang & Liang (2017), que utilizaron para el desarrollo de su modelo CRISPR *Random Forest* (CRF). Estas subsecuencias fueron generadas aleatoriamente siguiendo la organización estructural de repeticiones directas reales, con la implementación de un modelo de primer orden de Markov.

En primer lugar, se verificó que este conjunto de datos negativo, no tuviera similitud a un porcentaje mayor a 40% con las subsecuencias del conjunto de datos positivo generado. También se verificó que no existiera redundancia de subsecuencias entre las 12,000 repeticiones directas falsas. Ambas verificaciones se realizaron con el uso de CD-HIT-EST. Adicionalmente, se agregaron 187 repeticiones falsas o negativas que sólo contienen combinaciones de uno o dos nucleótidos, o la letra **N** de acuerdo con la nomenclatura de la Unión Internacional de Química Pura y Aplicada (IUPAC), representa a nucleótidos no definidos.

Posteriormente, se realizó el análisis de las longitudes de las subsecuencias, debido a que se detectaron varias repeticiones directas con longitudes cortas. En la Figura 15 se muestra el diagrama de dispersión de las longitudes en nucleótidos del conjunto negativo completo y el diagrama de densidad. Asimismo, en la Tabla 11 se presenta el resumen estadístico básico. Se encontró que las subsecuencias tienen longitudes de 28 hasta 60 nucleótidos, siendo muchas de longitud inferior a 32 nucleótidos, que es la longitud definida para las subsecuencias de entrada al modelo.

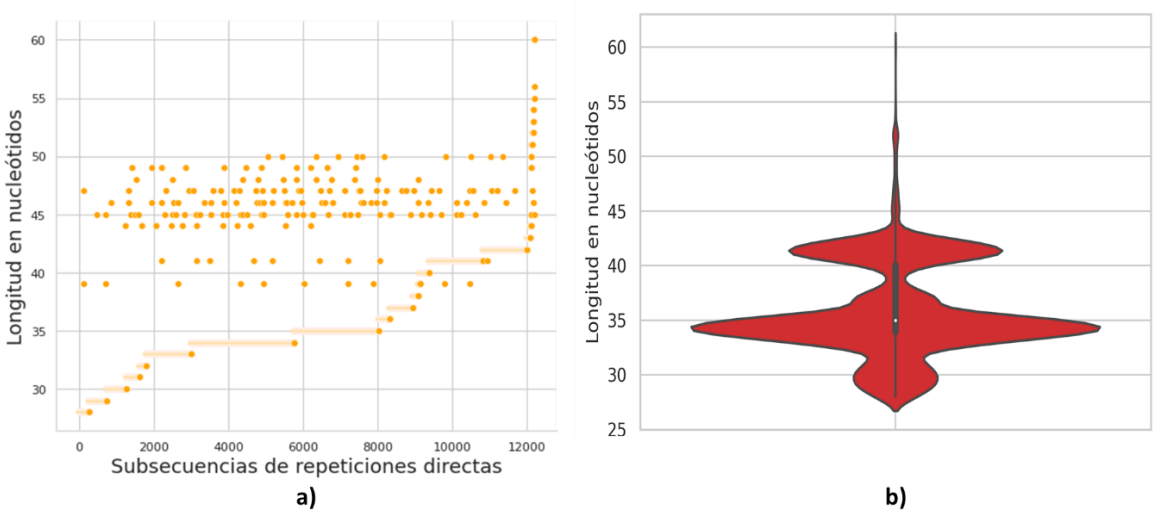

Figura 15. Longitudes las 12,187 repeticiones directas de la clase negativa: a) Dispersión de las longitudes. b) Diagrama de densidad de las longitudes de las repeticiones directas.

Tabla 11. Resumen estadístico de las longitudes en nucleótidos de las 12,187 repeticiones directas negativas.

	Repeticiones directas negativas
Subsecuencias analizadas	12,187
Longitud mínima	28
Longitud máxima	60
Media aritmética	35
Desviación estándar	4
Mediana	35
Moda	34
Primer cuartil-Q1 (25%)	34
Segundo cuartil-Q2 (50%)	35
Tercer cuartil-Q3 (75%)	40

En vista de lo anterior, se aplicó el mismo proceso a las 12,187 repeticiones directas negativas que a las repeticiones del conjunto positivo.

- Si la longitud era menor a 32 nt, se concatenó el número de nucleótidos faltantes en el extremo izquierdo, con espaciadores reales; si la longitud era mayor a 32 nt, se truncó la subsecuencia en su extremo izquierdo.

- De manera similar, se aplicó el mismo proceso al conjunto inicial de ejemplos negativos en el extremo derecho.

- Las 21,878 repeticiones directas negativas generadas fueron sometidos a CD-HIT-EST con un porcentaje de similitud al 100% para eliminar la redundancia en el conjunto procesado.

Finalmente, se obtuvo un total de 21,366 repeticiones directas falsas como conjunto de datos para la clase negativa, cada subsecuencia con una longitud de 32 nucleótidos y con su etiqueta de la clase negativa o falsa representada por el número 0.

En resumen, el conjunto de datos del modelo binario propuesto se constituyó por un total de 58,740 repeticiones directas, el cual fue formado por 37,374 subsecuencias de repeticiones directas de la clase positiva y 21,366 subsecuencias de repeticiones directas de la clase negativa.

3.2.3. Preprocesamiento del conjunto de datos

Se tomó un total de 46,992 repeticiones entre positivas y negativas que representan el 80% de las subsecuencias totales, como subconjunto de datos para el entrenamiento del modelo y las restantes 11,748 repeticiones directas que representan el 20% de las subsecuencias, como subconjunto de datos de prueba y validación para el modelo. Se implementó el balanceo de la clase positiva y negativa en ambos subconjuntos extraídos.

Una vez teniendo el conjunto de datos particionado en entrenamiento y prueba, se procedió a codificar en forma categórica las subsecuencias de nucleótidos, conformadas principalmente por las letras A, C, G, T y N. Para ello, cada letra en las subsecuencias de nucleótidos se transformó a un valor numérico. Como se muestra en la Figura 16, la letra N o cualquier otra se transformó a 0, Adenina a 1, Citosina a 2, Guanina a 3 y Timina a 4.

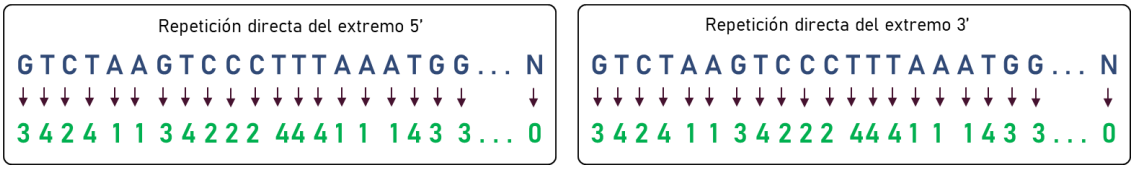

Figura 16. Representación de la transformación de subsecuencias a datos numéricos categóricos.

Los tensores o también conocidos como vectores multidimensionales, son la estructura de datos básica que utilizan los modelos de aprendizaje profundo para procesar los datos. Tanto como los datos de entrada como los objetivos (etiquetas) de las redes neuronales, deben de transformarse a tensores de datos de tipo punto flotante (float32), a este proceso se le denomina vectorización de datos.

Las subsecuencias de nucleótidos en valores numéricos, fueron vectorizadas con la técnica conocida como *one-hot encoding*, que es ampliamente utilizada para la codificación de datos categóricos. Dicha técnica consiste en codificar las variables categóricas en vectores binarios con una dimensión igual al número de categorías y las variables binarias toman el valor de 0 o 1 para indicar la inclusión o exclusión de cada categoría en los datos originales. En la Figura 17 se muestra el ejemplo de la aplicación de *one-hot encoding* a las repeticiones directas. Cada nucleótido en su valor numérico, se codificó en un vector binario con longitud de los cinco nucleótidos N, A, C, G y T, generando así, un tensor binario de 2-dimensiones, donde el número de filas del tensor es igual a 32 (longitud de

nucleótidos definidos para las repeticiones directas) y el número de columnas es igual a 6 (cinco nucleótidos posibles a presentarse dentro de las repeticiones y adicionalmente, se añadió una columna más de relleno (*padding*) con ceros).

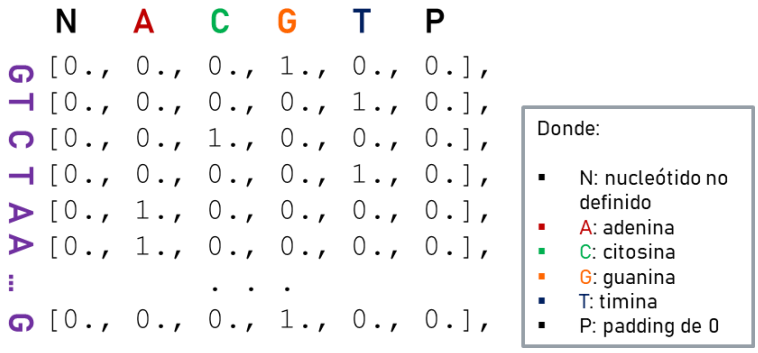

Figura 17. Ejemplo del tensor binario de 2-dimensiones obtenido mediante one-hot encoding de las repeticiones directas.

De esta manera, se obtuvo un tensor binario 2D para cada repetición directa dentro del conjunto de entrenamiento y prueba. Posteriormente, se realizó un reajuste para transformarlos a tensores binarios de 3-dimensiones y así puedan ser procesados correctamente por el modelo propuesto. La dimensión resultante de los tensores fue de [1 x 32 x 6], donde el 1 indica en la CNN el canal, 32 filas donde cada una corresponde a un nucleótido en las repeticiones directas de entrada y las 6 columnas que representan los cinco nucleótidos y la sexta columna el *padding* con ceros.

3.2.4. Arquitectura del modelo binario para la caracterización de matrices CRISPR

En el desarrollo del modelo binario se utilizó la interfaz funcional de Keras, la cual es un entorno de trabajo de alto nivel para Python ampliamente utilizado en el campo del aprendizaje profundo, que puede ser implementado con Tensorflow, Theano o CNTK (en inglés *Microsoft Cognitive Toolkit*) como backend. Keras proporciona simplicidad y gran capacidad para la construcción de modelos secuenciales complejos con parámetros ajustables.

En este trabajo de maestría se implementó una red neuronal convolucional (CNN) binaria (Figura 18) para la caracterización de matrices CRISPR en secuencias de nucleótidos bacterianas. El diseño y la arquitectura final de la CNN se obtuvo mediante la prueba de diferentes hiperparámetros comúnmente recomendados en el área de aprendizaje profundo.

De manera general, la arquitectura es la siguiente: inicialmente se encuentra la capa de entrada, luego contiene dos capas convolucionales, donde cada una es seguida por la capa de agrupación MaxPooling, una capa de abandono (*dropout*), y finalmente por la capa totalmente conectada (*fully-connected*) con una salida binaria.

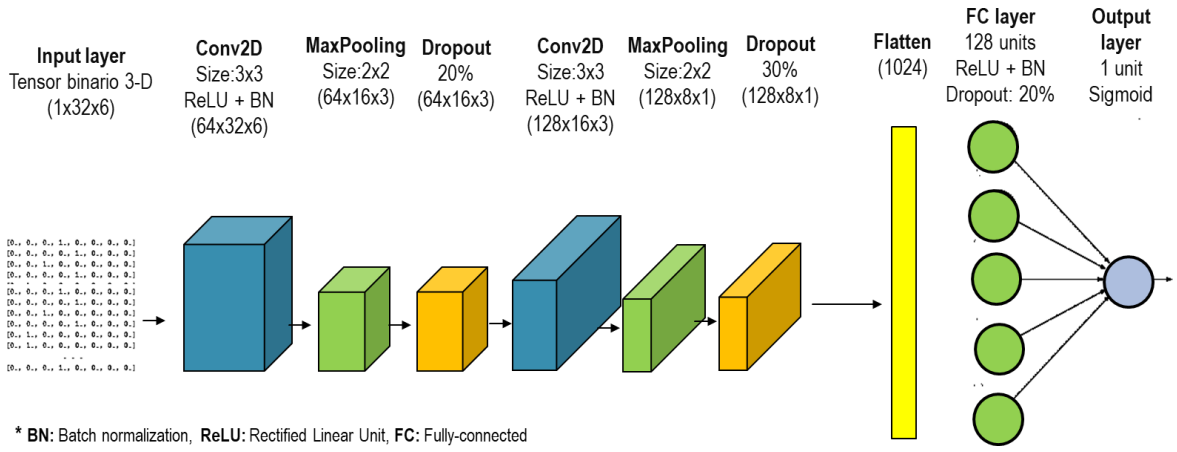

Figura 18. Arquitectura de la red neuronal convolucional binaria propuesta para caracterizar matrices CRISPR.

A continuación, se describe brevemente la arquitectura del modelo propuesto:

- **Capa de entrada (*input layer*).** La capa de entrada contiene el tensor binario de 3-dimensiones (1x32x6) sin procesar.

- **Primera capa convolucional (*Conv2D*).** En la primera capa convolucional se aplicaron 64 filtros con un tamaño de 3x3. En cada filtro se calcula un producto punto entre sus pesos y una pequeña región bidimensional del tensor de entrada para extraer características significativas de los datos; a esta operación entre los filtros y el tensor de entrada se le llama convolución. Se aplicó la función de activación ReLU y la técnica normalización por lotes (*batch normalization*) con el objetivo de normalizar las activaciones de salida. La salida obtenida de esta capa fue un tensor de 64x32x6, el cual recibe el nombre de mapa de características.

- **Primera capa de agrupación (*MaxPooling2D*).** En la primera capa de agrupación se recibió como entrada el mapa de características de 64x32x6. Esta capa tiene como objetivo la reducción del tamaño del mapa de características en sus dimensiones espaciales (ancho y alto) con la operación MaxPooling. Los núcleos de agrupación implementados fueron de tamaño 2x2 y como salida final se obtuvo un tensor de 64x16x3.

- **Primera capa de abandono (*Dropout*).** Primera capa donde se implementó la técnica de regularización dropout para evitar el sobre aprendizaje (*overfitting*) de la CNN recibiendo el tensor de 64x16x3 y aplicándole un porcentaje de abandono del 20% en el mapa de características; la dimensión del tensor de salida fue la misma de 64x16x3.

- **Segunda capa convolucional (*Conv2D*).** La segunda capa convolucional recibió como entrada el mapa de características de 64x16x3 y se le aplicaron 128 filtro con un tamaño de

3x3. Se aplicó la función de activación ReLU y la técnica *batch normalization* para la normalización de las activaciones de salida. La salida de esta capa fue un tensor de 128x16x3.

- **Segunda capa de agrupamiento (*MaxPooling2D*).** La segunda capa de agrupación recibió el mapa de características de 128x16x3, tras la implementación de la operación MaxPooling con los núcleos de agrupación de 2x2, se redujo la dimensión del tensor a 128x8x1.

- **Segunda capa de abandono (*Dropout*).** En la segunda capa de abandono, se recibió el tensor de 128x8x1 y se le aplicó un porcentaje de abandono del 30% en el mapa de características, la dimensión del tensor de salida fue la misma de 128x8x1.

- **Capa totalmente conectada (*fully-connected layer*).** Inicialmente, el tensor de la capa anterior con una dimensión de 128x8x1 fue aplanado (*flatten*) en un tensor de 1-dimensión dando como resultado un vector de dimensión 1024 como entrada. Este vector contiene las características extraídas de los datos en las capas convolucionales. Se implementó un total de 128 neuronas, la función de activación ReLU, la técnica *batch normalization* y la técnica de regularización *dropout* con un porcentaje del 20%.

- **Capa de salida (*output layer*).** Es la encargada de indicar la clasificación final de los datos de entrada de acuerdo con el cálculo de puntuación para cada clase. Contiene 1 neurona y se implementó la función de activación sigmoidal (*sigmoid*) para proporcionar una salida de clasificación binaria. Recibió como entrada un tensor con una dimensión de 128 y genera como salida un tensor un 1 cuando la clasificación final es positiva o un 0 cuando la clasificación es negativa.

3.2.5. Métricas de evaluación

El modelo computacional propuesto, fue entrenado para minimizar el error al realizar las clasificaciones binarias a datos no vistos en el entrenamiento. Este error es evaluado con las métricas de evaluación de rendimiento del modelo como la exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*), puntuación F1 (*F1-score*) y el coeficiente de correlación de Matthews (MCC).

Estas métricas se obtienen a partir de los resultados de la matriz de confusión probabilística sobre el conjunto de prueba, que compara las etiquetas reales con las etiquetas obtenidas tras la clasificación de los ejemplos de la siguiente manera:

- Ejemplos verdaderos clasificados correctamente como positivos (*True positive*, TP).
- Ejemplos falsos clasificados correctamente como negativos (*True negative*, TN).
- Ejemplos falsos clasificados incorrectamente como positivos (*False positive*, FP).
- Ejemplos verdaderos clasificados incorrectamente como negativos (*False negative*, FN).

La métrica de exactitud (*accuracy*), indica el porcentaje de ejemplos correctamente clasificados y se denota con la ecuación (3):

$$\boldsymbol{accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN}} \quad \textbf{(3)}$$

La precisión (*precision*) representa el porcentaje de ejemplos positivos clasificados correctamente y se denota con la ecuación (4):

$$\boldsymbol{precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \quad \textbf{(4)}$$

La exhaustividad (*recall*) representa el porcentaje de ejemplos positivos reales que se clasificaron correctamente y se denota con la ecuación (5):

$$\boldsymbol{recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad \textbf{(5)}$$

La puntuación F1 (*F1-score*) es una media armónica entre la métrica de la precisión y la exhaustividad, se denota con la ecuación (6):

$$\boldsymbol{F1} = \boldsymbol{2}.\frac{\boldsymbol{precision}\,.\boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}} \quad \textbf{(6)}$$

La métrica del coeficiente de correlación de Matthews (MCC) es utilizada para medir la calidad de modelos de clasificación, debido a que se considera una medida equilibrada y es recomendada cuando se cuenta con clases desbalanceadas, lo cual es nuestro caso. Es el coeficiente de correlación que va de -1 a +1 donde +1 es una clasificación perfecta y se denota con la ecuación (7):

$$\boldsymbol{MCC} = \frac{\boldsymbol{TP}\,.\boldsymbol{TN} - \boldsymbol{FP}\,.\boldsymbol{FN}}{\sqrt{(\boldsymbol{TP} + \boldsymbol{FP})(\boldsymbol{TP} + \boldsymbol{FN})(\boldsymbol{TN} + \boldsymbol{FP})(\boldsymbol{TN} + \boldsymbol{FN})}} \quad \textbf{(7)}$$

3.3. Posprocesamiento de los resultados obtenidos del modelo

La CNN binaria asigna una probabilidad de predicción a cada extremo de las subsecuencias evaluadas la cual indica si contienen patrones característicos de repeticiones directas. Estas probabilidades van de 0 a 1 (donde 0 es negativo y 1 es positivo). Para clasificar si los *k*-meros evaluados por el modelo propuesto son estructuras CRISPR que pueden representar o ser parte de una matriz CRISPR, fue necesaria la aplicación de un posprocesamiento a los resultados obtenidos.

Inicialmente, se definieron dos umbrales de confianza; en el primer umbral de "estructuras CRISPR confiables" se extrajeron todos los *k*-meros con una probabilidad de predicción ≥0.95 en ambos extremos de la subsecuencia y en el segundo umbral de "estructuras CRISPR potenciales" se extrajeron los *k*-meros con una probabilidad de predicción ≥0.91 en al menos uno de sus extremos y que la diferencia de probabilidades entre sus extremos sea ≤0.04.

Dado a que, en el preprocesamiento de la secuencia de entrada, los *k*-meros fueron extraídos con traslape, las subsecuencias que fueron extraídas de manera contigua contienen fragmentos de la misma región de la secuencia principal. Por lo tanto, a partir de una determinada cantidad de *saltos* en la extracción de *k*-meros, las subsecuencias extraídas contienen diferentes regiones de la secuencia de la matriz real.

Considerando lo anterior, se aplicó un procedimiento para evitar que dos o más *k*-meros contiguos identifiquen a una misma estructura CRISPR real. Al descartar los *k*-meros contiguos, se verificó la identificación y recuperación de las estructuras CRISPR putativas de la longitud estándar definida en este trabajo y de longitudes superiores e inferiores con las mayores probabilidades de predicción. Para esto, si entre dos *k*-meros existe un traslape que sea ≤40 nucleótidos (5 *saltos*), son considerados contiguos y se evalúan de la siguiente manera:

- Estructuras CRISPR de longitud estándar: se identificó como estructura CRISPR de longitud estándar al *k*-mero que contenía las mayores probabilidades de predicción en ambos extremos en un conjunto de *k*-meros contiguos, los demás fueron descartados (Figura 19a).

- Estructuras CRISPR de longitudes largas: se identificó como estructura CRISPR de longitud superior a la estándar cuando un primer *k*-mero (en un conjunto de contiguos) tenía la mayor probabilidad sólo en el extremo izquierdo y un siguiente *k*-mero contiguo tenía la mayor probabilidad sólo en el extremo derecho (Figura 19b). Al primer *k*-mero se le concatenó el fragmento faltante del segundo y todos los demás *k*-meros contiguos fueron descartados.

- Estructuras CRISPR de longitudes cortas: de forma contraria a las estructuras de longitudes largas, se identificó como estructura CRISPR de longitud inferior a la estándar cuando un primer *k*-mero contiguo tenía la mayor probabilidad sólo en el extremo derecho y un siguiente *k*-mero contiguo tenía la mayor probabilidad sólo en el extremo izquierdo (Figura 19c). Al primer *k*-mero se le hizo un corte en el extremo izquierdo correspondiente a los nucleótidos de los *saltos* hacia el segundo *k*-mero y todos los demás k-meros contiguos fueron descartados.

Una vez posprocesados, se formaron las matrices CRISPR con base a la separación de nucleótidos entre las estructuras. Para ello, si existía una estructura con una separación ≥205 nucleótidos con la anterior, fue considerada como otra o parte de otra matriz CRISPR y las menores como parte de una misma matriz CRISPR.

Finalmente, en las estructuras CRISPR confiables sólo se conservaron las matrices CRISPR que estaban conformadas por ≥2 estructuras y en el umbral de estructuras CRISPR potenciales sólo se conservaron las matrices CRISPR conformadas por ≥5 estructuras. Este filtro se aplicó ya que se sabe que la confianza de interacción aumenta cuando una mayor cantidad de espaciadores CRISPR en la bacteria son homólogos a un mismo bacteriófago (Edwards et al., 2016; Ruohan et al., 2022).

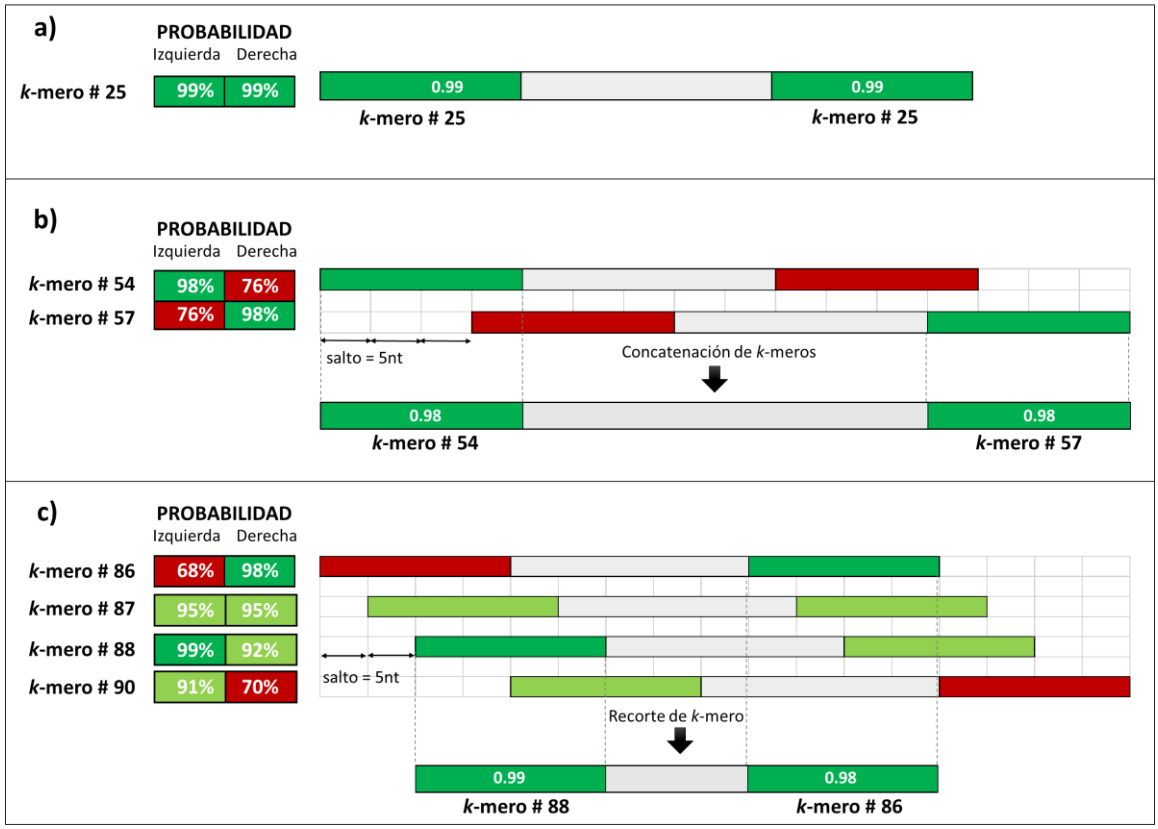

Figura 19. Procedimiento para la identificación de estructuras CRISPR en k-meros contiguos. a) Identificación de estructura de longitud estándar con las mayores probabilidades en sus extremos; b) Concatenación de k-meros contiguos con la mayor probabilidad en sus extremos complementarios para la identificación de estructuras largas; c) Recorte del k-mero en el extremo izquierdo correspondiente al número de nucleótidos de los saltos al siguiente k-mero con la mayor probabilidad complementaria al primero para la identificación de estructuras cortas.

3.4. Aplicación del modelo computacional en un genoma bacteriano completo

En la etapa final de la metodología seguida, se llevó a cabo la aplicación del modelo computacional propuesto en una secuencia de nucleótidos real. Para su evaluación se seleccionó el genoma bacteriano de la especie *Actinoalloteichus sp. AHMU CJ021* con taxón *2072503*, el cual contiene un total de 37 matrices CRISPR validadas, a partir de las cuales se extrajeron 275 estructuras CRISPR. Esta secuencia era la que contenía la mayor cantidad de matrices del conjunto de datos utilizado y cuenta con una longitud de su genoma de 6,822,750 de nucleótidos.

Por último, una vez caracterizadas las matrices CRISPR en la secuencia de nucleótidos evaluada, se realizó la extracción de las subsecuencias espaciadoras del umbral de estructuras CRISPR confiables para someterlas a alineamiento contra las secuencias de la base de datos de genomas de bacteriófagos de NCBI, con la herramienta bioinformática BLAST (en inglés *Basic Local Alignment Search Tool*).

Esta herramienta utiliza un algoritmo heurístico que compara secuencias biológicas para encontrar homólogos entre las secuencias recibidas como entrada con las secuencias de una base de datos previamente definida (Altschul et al., 1990). De esta manera, fue posible realizar un análisis de la taxonomía de los espaciadores con la herramienta MEGAN6 y establecer potenciales interacciones infecciosas entre la bacteria analizada y los bacteriófagos homólogos de los espaciadores en las matrices CRISPR caracterizadas mediante el modelo propuesto.

CAPÍTULO 4. RESULTADOS Y DISCUSIÓN

En este capítulo se detallan los resultados obtenidos del entrenamiento, evaluación y aplicación del modelo computacional propuesto. Se presentan los resultados de las métricas de evaluación de la Red Neuronal Convolucional (CNN) binaria sobre el conjunto de datos de prueba, los resultados de la caracterización de matrices CRISPR de la secuencia de nucleótidos bacteriana real analizada, el análisis del alineamiento de las potenciales subsecuencias espaciadoras contra la base de datos de virus y el análisis taxonómico de los resultados para la definición de interacciones bacteriófago – bacteria hospedera.

4.1. Evaluación del rendimiento del modelo computacional propuesto

Como ya se mencionó en la sección 3.3.4., la CNN propuesta fue desarrollada usando la biblioteca de Keras del lenguaje de programación Python y utilizando Tensorflow como *backend*. El entrenamiento y las pruebas del modelo se realizaron en la plataforma Google Colaboratory, el cual es un servicio web que permite el acceso gratuito por tiempo limitado a un servidor GPU de Google para el desarrollo de algoritmos en Python. Se realizaron diferentes entrenamientos para ajustar manualmente los hiperparámetros y así seleccionar aquellos que proporcionaron el mayor rendimiento sobre el conjunto de datos de prueba.

En la Tabla 12 se muestra el conjunto de hiperparámetros seleccionados para la arquitectura y el entrenamiento del modelo final. Como resultado, se obtuvo un 96.29% en la métrica de exactitud (*accuracy*), 96.00% de precisión, recuerdo (*recall*) y puntuación F1 (*score-F1*) y, 92.04% de coeficiente de correlación de Mathews (MCC).

Tabla 12. Configuración e hiperparámetros del modelo final.

Arquitectura de la CNN		
Capa	Configuración de Capa	Función de activación
Convolution2D	64 x (3,3)	ReLU
Batch normalization	N.A.	N.A.

Max Pooling2D	(2,2)	N.A.
Dropout	$p = 0.2$	N.A.
Convolution2D	128 x (3,3)	ReLU
Batch normalization	N.A.	N.A.
Max Pooling2D	(2,2)	N.A.
Dropout	$p = 0.4$	N.A.
Fully connected	64 neuronas	ReLU
Batch normalization	N.A.	N.A.
Dropout	$p = 0.2$	N.A.
Output layer	1 neurona	Sigmoid
Hiperparámetros de entrenamiento de la CNN		
Tamaño de lote	32	
Número de épocas	20	
Tasa de aprendizaje	0.0006	
Optimizador	Adam	
Función de pérdida	Binary cross-entropy	

Mientras se realizaba el entrenamiento de la CNN, se generaron las curvas de aprendizaje del modelo para observar su rendimiento tanto el conjunto de entrenamiento como en el de prueba (Figura 20), y detectar si se presenta algún problema de bajo aprendizaje (*underfit*) o de sobre aprendizaje (o*verfit*). En la Figura 20a se muestran las curvas de la exactitud. Como se puede observar, a partir de la época 7, ambos conjuntos se mantuvieron convergiendo con valores similares hasta la época 15 donde se presentó una ligera separación entre las curvas indicando que en el modelo tuvo menor exactitud en el conjunto de prueba que en el de entrenamiento. Finalmente, cambiaron nuevamente su dirección a unirse en la última época por lo que se garantiza que la CNN no presentará *underfit* ni *overfit*.

En la Figura 20b se muestran las curvas de la función de pérdida, también llamada función de costos u objetivo. Esta función es la encargada de calcular las predicciones correctas e incorrectas, y es minimizada durante el entrenamiento. El comportamiento en el conjunto de entrenamiento y de prueba, se mantuvieron convergiendo de manera similar; se unieron aproximadamente en la época 10 y también se presentó una ligera separación, hasta cambiar nuevamente su dirección a unirse en la última época.

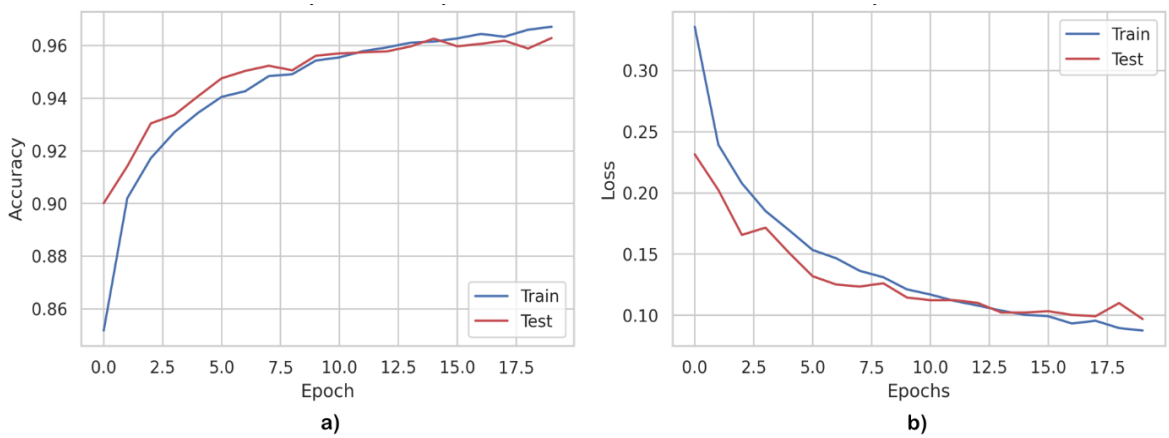

Figura 20. Curvas de aprendizaje en el conjunto de entrenamiento y prueba del modelo CNN. a) Rendimiento de exactitud. b) Rendimiento de la función de pérdida.

En la Figura 21 se muestra el gráfico de la matriz de confusión obtenida de la evaluación del conjunto de prueba. Se obtuvo un total de 276 ejemplos de falsos negativos (FN), lo cual representa el 4% de estructuras CRISPR perdidas, ya que son clasificados erróneamente como no-CRISPR (negativas). Asimismo, 160 ejemplos fueron falsos positivos (FP), lo cual, también representa el 4% de estructuras falsas clasificadas erróneamente como verdaderos CRISPR. Los principales ejemplos donde el modelo tuvo errores de clasificación fue en las estructuras CRISPR cortas, es decir, con longitud inferior a la estándar definida de 100 nucleótidos.

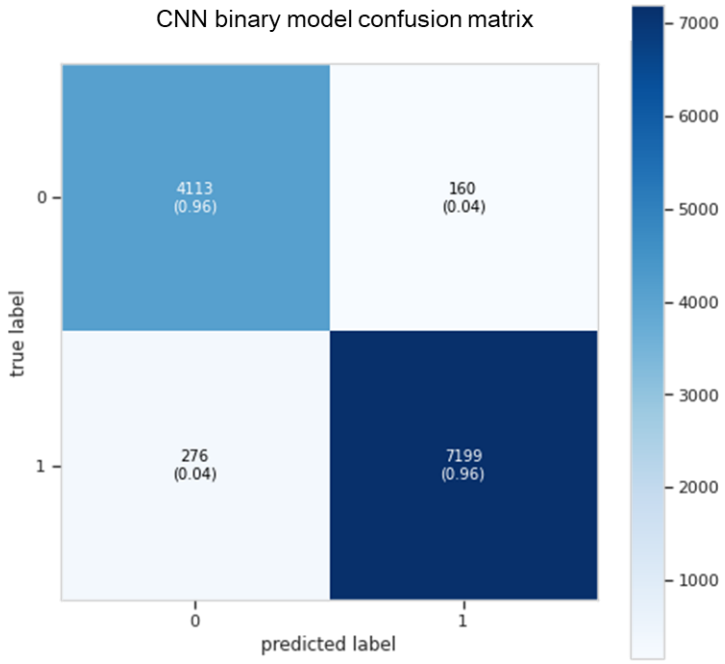

Figura 21. Matriz de confusión del conjunto de prueba con la CNN binaria.

4.2. Aplicación del modelo computacional para la caracterización de matrices CRISPR

Una vez que se generó un buen modelo CNN, este se aplicó sobre la secuencia del genoma completo de la bacteria *Actinoalloteichus sp. AHMU CJ021* con taxón *2072503*, que fue la bacteria con la mayor cantidad de matrices CRISPR validadas en la BD utilizada CRISPRCasdb. Esta bacteria fue reportada en sedimentos marinos (*NCBI BioSample: SAMN08364581*) y, para el presente análisis, se descargó su genoma de NCBI. Como primer paso, se realizó el preprocesamiento a la secuencia bacteriana de 6,822,750 nucleótidos de *longitud*, como se describe en la sección 3.3.1 de la metodología. A partir de la fragmentación de la secuencia, se obtuvo un total de 1,364,531 *k*-meros con longitud de 100 nucleótidos, extraídos con *saltos* de 5 nucleótidos. Este conjunto de *k*-meros fue introducido como datos de entrada a la CNN binaria. El modelo asignó a cada *k*-mero una probabilidad de predicción, en ambos extremos, de contener subsecuencias de repeticiones directas. Posteriormente, se aplicó el posprocesamiento de resultados a la salida del modelo, que consistió en el análisis de los *k*-meros evaluados con base a los umbrales de confianza (de estructuras CRISPR de longitud estándar, de longitudes de mayor longitud y de menor longitud), y como paso final, el filtro de las matrices CRISPR por la cantidad de estructuras que las integran (sección 3.4.).

En la Tabla 13 se presentan los principales resultados obtenidos de la evaluación de los *k*-meros de la secuencia bacteriana *Actinoalloteichus sp. AHMU CJ021*. Tras el posprocesamiento realizado con el umbral 0.95 definido para "estructuras CRISPR confiables" se obtuvo un total de 6,437 *k*-meros, mientras que con el umbral 0.91 definido para "estructuras CRISPR potenciales" se obtuvo un total de 25,010 *k*-meros. La eliminación de las matrices CRISPR integradas por menos de 2 estructuras CRISPR en el primer umbral generó 2,872 estructuras confiables que integran 1,135 matrices CRISPR; y con la eliminación de las matrices integradas por menos de 5 estructuras CRISPR en el segundo umbral, se obtuvo como resultado final 7,963 estructuras potenciales que integran 1,198 matrices CRISPR.

Tabla 13. Resultados de la caracterización de matrices CRISPR en la bacteria Actinoalloteichus sp. AHMU CJ021.

	Posprocesamiento de resultados				**Resultado final**	
No. *k*-meros evaluados	Umbral de confianza	No. *k*-meros posprocesados	No. Matrices posprocesadas	Filtro de matrices por estructuras	No. *k*-meros finales	No. Matrices finales
1,364,531	≥ 0.95	6,437	4,700	≥2	2,872	1,135
	≥ 0.91	25,010	10,722	≥5	7,963	1,198

4.2.1. Validación de estructuras CRISPR conocidas en los resultados finales

La secuencia bacteriana analizada mediante el modelo propuesto tiene un total de 37 matrices CRISPR conocidas y anotadas en CRISPRCasdb, a partir de las cuales se extrajeron 275 estructuras CRISPR con sus longitudes reales. Se realizó el análisis para verificar que el modelo computacional haya clasificado correctamente dichas estructuras CRISPR. Para ello, se realizó una búsqueda de las estructuras conocidas en los resultados finales obtenidos en los dos umbrales de confianza de estructuras CRISPR clasificadas por el modelo. Se obtuvo la correcta clasificación de 180 estructuras conocidas en el umbral de "estructuras CRISPR confiables" y 41 en el umbral de "estructuras CRISPR potenciales", lo que da como resultado final un total de 221 estructuras (80.36%). A su vez, estas estructuras clasificadas correctamente corresponden a 31 (93.93%) de las 37 matrices CRISPR conocidas.

En la Figura 22 se presenta la comparación de la cantidad de estructuras que integran a las 37 matrices conocidas y anotadas de la bacteria de prueba, con la cantidad de estructuras que logró identificar y clasificar correctamente el modelo dentro de los dos umbrales de confianza. Interesantemente las 6 matrices CRISPR conocidas que la CNN propuesta no logró identificar, ya que no se identificó ninguna estructura de las mismas, fueron aquellas que están integradas por menos de 5 estructuras CRISPR.

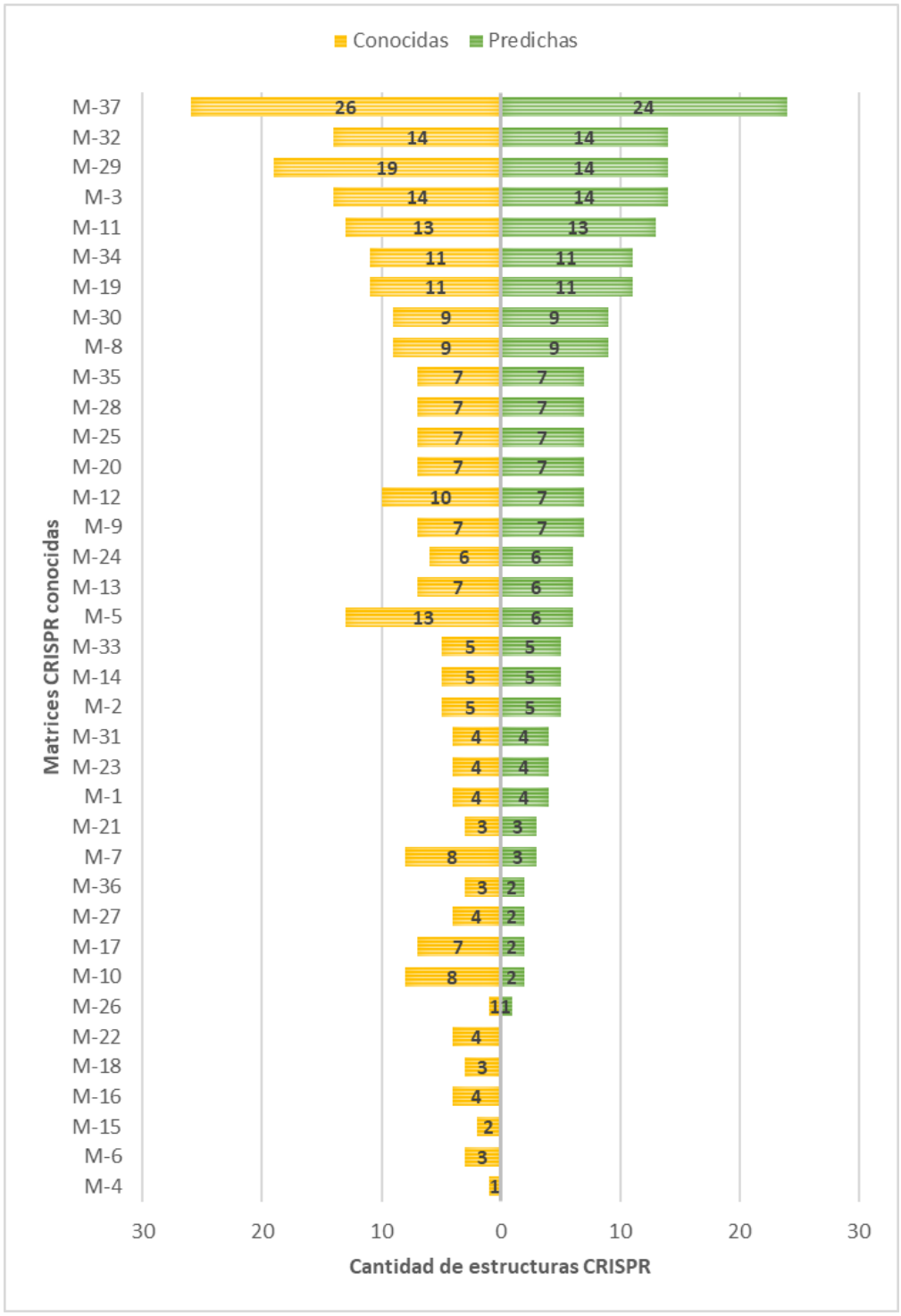

Figura 22. Comparación del número de estructuras conocidas con las predichas por el modelo, de las 37 matrices CRISPR conocidas y anotadas. En el eje X, se representa la cantidad de estructuras conocidas (barras izquierdas) y predichas (barras derechas) de las matrices CRISPR; y en el eje Y se representa el identificador de cada una de las matrices CRISPR anotadas y conocidas en CRISPRCasdb.

4.2.2. Búsqueda de similitud de los espaciadores predichos y conocidos con genomas de bacteriófagos y su análisis taxonómico

De las 1,135 matrices CRISPR predichas en el primer umbral de confianza, se extrajeron los 2,872 espaciadores y de las 1,198 matrices en el segundo umbral, se extrajeron los 7,963 espaciadores con longitudes de 22 – 51 nucleótidos. De manera independiente, se sometieron a una búsqueda (alineamiento) contra las secuencias de bacteriófagos no redundantes de la BD de Virus de NCBI creada el 8 de febrero de 2022; esto permitió encontrar la similitud de los espaciadores predichos mediante el modelo propuesto con los genomas de fagos anotados y conocidos para identificar interacciones fago – bacteria hospedera.

Para el alineamiento independiente de los espaciadores predichos y los conocidos, se utilizó BLASTn v2.12.0., que es específico para la comparación de secuencias de nucleótidos. Se utilizaron los parámetros predeterminados a excepción de los siguientes tres: e-value de 1 (probabilidad de obtener una coincidencia de manera aleatoria con respecto al tamaño de la BD de referencia), porcentaje de identidad de 80% (porcentaje mínimo de similitud de la secuencia de consulta con la secuencia objetivo de la BD de referencia) y porcentaje de cobertura de 80% (porcentaje mínimo de la secuencia de consulta que debe estar cubierta por la secuencia objetivo).

Como resultado, se obtuvo que 1,365 (47.52%) de los 2,872 espaciadores predichos en el primer umbral encontraron secuencias similares (*hits*), 2,408 (30.23%) de los 7,963 espaciadores predichos en el segundo umbral encontraron *hits* y 92 de los 275 espaciadores conocidos encontraron *hits* con los protoespaciadores en los genomas de los bacteriófagos de la BD. Estos resultados son consistentes con otros estudios reportados (Dion et al., 2021; Edwards et al., 2016; Shmakov et al., 2017), en donde los espaciadores predichos encontraron pocos homólogos, en el rango de 1%-69% en las BD de fagos conocidos.

A partir de los *hits* obtenidos del alineamiento con BLASTn, se realizó el análisis taxonómico de los espaciadores predichos en el primer umbral de confianza y de los espaciadores conocidos. Para ello se utilizó MEGAN6, que es una herramienta bioinformática que utiliza el algoritmo Ancestro Común más Bajo (en inglés *Lowest Common Ancestor*, *LCA*) para analizar grandes conjuntos de secuencias biológicas y asignarles un nivel taxonómico utilizando la taxonomía de NCBI. Cada nivel refleja el nivel de conservación de la secuencia. Además, permite generar gráficos y estadísticos para comparar la diversidad taxonómica de diferentes conjuntos de secuencias (Huson et al., 2007).

Los resultados mostraron 1,507 espaciadores sin *hits*, 63 espaciadores sin asignación taxonómica y 1,302 espaciadores asignados a una taxonomía en el dominio de Virus de la siguiente manera:

- 17 espaciadores asignados al nivel dominio de *Virus*.
- 1 espaciador asignado al nivel reino de *Monodnaviria;* dentro de este clado, es decir, en los descendientes de este ancestro común, se encontraron 5 espaciadores asignados al nivel familia de *Inoviridae* y 3 espaciadores asignados al nivel familia de *Microviridae*.

- 8 espaciadores asignados al nivel reino en *Virus de bacterias no clasificados*.

- 227 espaciadores asignados al nivel orden de *Caudovirales*. Dentro de este clado se encontraron también otros 1,041 espaciadores siguientes niveles taxonómicos. A nivel familia: *Siphoviridae* con 838, *Myoviridae* con 126, *Podoviridae* con 32, *Caudovirales de muestras ambientales* con 21, *Caudovirales no clasificados* con 16 y *Autographiviridae* con 8 espaciadores asignados.

Por otra parte, en el análisis taxonómico de los 275 espaciadores conocidos, se detectaron los 183 espaciadores sin *hits*, 3 espaciadores sin asignación taxonómica y 89 espaciadores asignados a una taxonomía en el dominio de Virus de la siguiente manera:

- 1 espaciador asignado al nivel dominio de *Virus*.

- 13 espaciadores asignados al nivel orden de *Caudovirales*, dentro de este mismo clado se encontraron 75 espaciadores en los siguientes niveles taxonómicos. A nivel familia: *Siphoviridae* con 61, *Myoviridae* con 8, *Podoviridae* con 3 y *Caudovirales de muestras ambientales* con 3 espaciadores asignados.

A partir de los resultados obtenidos en los análisis taxonómicos, se confirmó que existe una semejanza en la asignación taxonómica de los espaciadores predichos y los conocidos a partir de su similitud con las secuencias de los fagos, encontrada mediante el alineamiento.

En la Figura 23 se muestra la comparación del árbol filogenético de los espaciadores predichos y los conocidos des-colapsado hasta nivel especie. Este gráfico nos permite observar la relación por descendencia en común de la taxonomía de los espaciadores. Como se puede observar, en ambos casos, la mayoría de los espaciadores encontraron homología con secuencias de bacteriófagos al nivel orden de *Caudovirales*. Así mismo, dentro de este mismo clado, las familias más prevalentes y que se identificaron para los espaciadores predichos y conocidos, fueron *Siphoviridae*, *Myoviridae*, *Podoviridae* y *Caudovirales de muestras ambientales*.

Por lo tanto, a partir de la homología encontrada entre los espaciadores predichos, con sus potenciales protoespaciadores objetivos en el genoma del bacteriófago, se infiere que la bacteria *Actinoalloteichus sp. AHMU CJ021* tuvo interacción con fagos a nivel familia de *Siphoviridae*, *Myoviridae*, *Podoviridae*, *Caudovirales de muestras ambientales* y de manera adicional a lo encontrado para los espaciadores conocidos, con las familias *Caudovirales no clasificados* y *Autographiviridae*, todos estos dentro del nivel orden de *Caudovirales*.

Con menor prevalencia, pero también de manera adicional a los espaciadores conocidos, se encontraron potenciales interacciones con bacteriófagos derivados de un ancestro en común del nivel familia de *Inoviridae* y *Microviridae* del nivel reino de *Monodnaviria*; y con *Virus de bacterias no clasificados* a nivel reino.

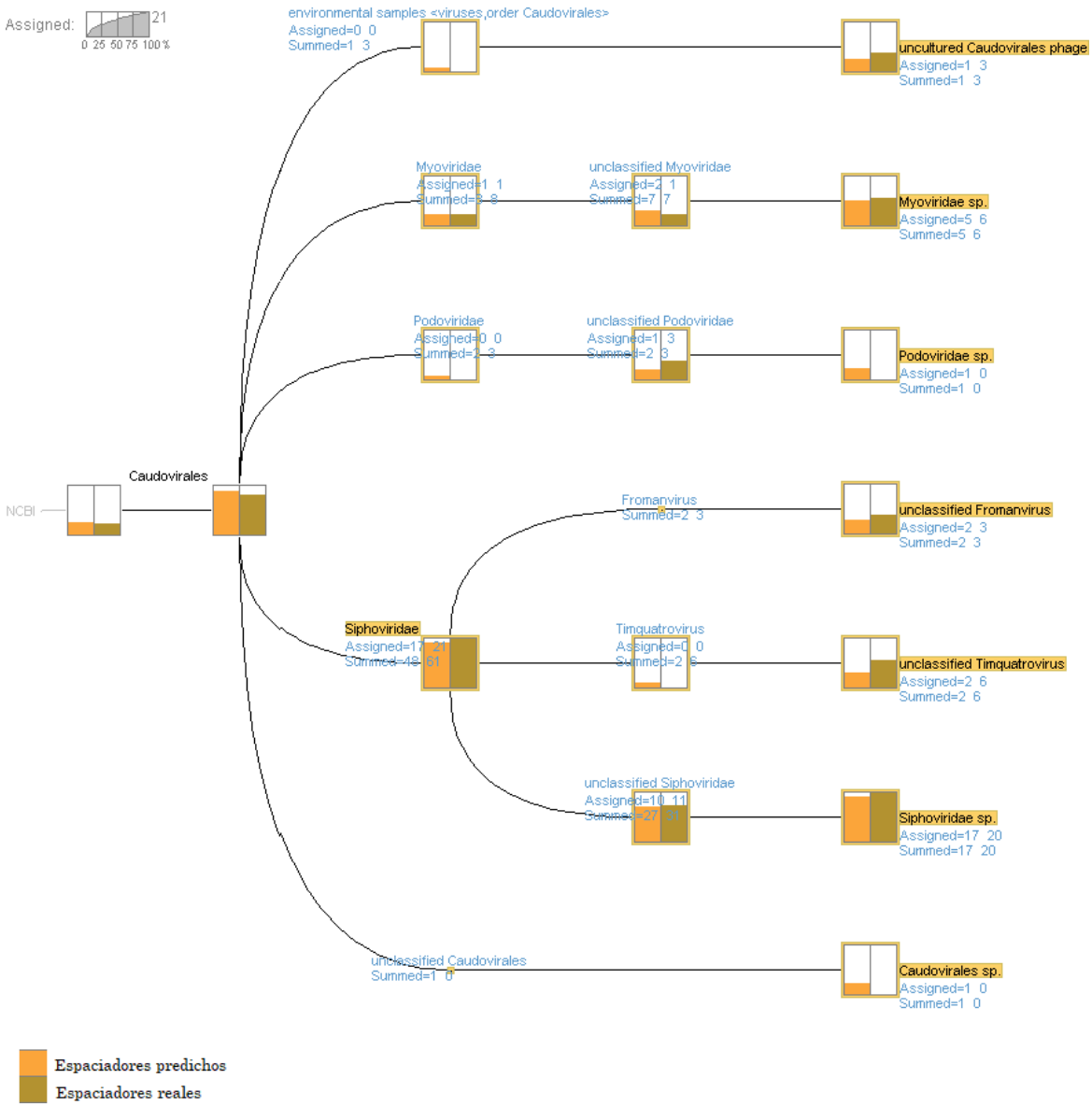

Figura 23. Comparación del árbol filogenético de los espaciadores predichos y los conocidos en porcentajes normalizados en el clado con mayor semejanza, des-colapsado hasta nivel especie.

CAPÍTULO 5. CONCLUSIONES Y TRABAJO FUTURO

Los bacteriófagos (fagos) tienen la capacidad de modificar la dinámica de las comunidades bacterianas tras la infección de las mismas. Sin embargo, nos enfrentamos a que, a pesar de que estos son reconocidos como una de las entidades más abundantes en la Tierra, la base de datos donde se encuentran caracterizados con su determinada clasificación taxonómica, es hasta 20 veces más reducida que la de bacterias (Dion et al., 2021). Por lo tanto, conocer las interacciones entre los bacteriófagos – bacterias hospederas, es una vía para comprender el papel ecológico que estos cumplen y cómo influyen en los cambios de composición y función del nicho donde radican. Así como también, estas interacciones proporcionan información que contribuye a la caracterización de nuevos fagos.

Con el objetivo de aportar a la definición de interacciones bacteriófago – bacteria huésped, en el presente trabajo de investigación, se desarrolló un modelo computacional basado en una red neuronal convolucional (CNN), la cual es una técnica de aprendizaje profundo, para caracterizar matrices CRISPR en genomas bacterianos. A diferencia de las herramientas reportadas a la fecha (Mitrofanov et al., 2021; Wang at al, 2017; R. Zhang et al., 2021), no es necesario el proceso manual de extracción de características para alimentar al modelo, sino que lo hace a través del análisis directo de la secuencia de nucleótidos.

En las métricas para la evaluación del rendimiento en la identificación de matrices, se obtuvo un 96.29% de exactitud, 96.00% de Precisión, Exhaustividad y Puntuación-F1 y 92.04% de Coeficiente de correlación de Mathews, sobre el conjunto de prueba.

Se caracterizaron las matrices CRISPR de la bacteria *Actinoalloteichus sp. AHMU CJ021* y los resultados fueron extraídos a partir de dos umbrales de confianza definidos. En el primero, se encontraron 1,135 y en el segundo 1,198 matrices CRISPR. Finalmente, se realizó la extracción de las subsecuencias espaciadoras de las matrices predichas en ambos umbrales de “estructuras CRISPR confiables” y de “estructuras CRISPR potenciales” para su alineamiento con los genomas de bacteriófagos anotados en la base de datos de NCBI Virus. El 47.52% y el 30.23% de los espaciadores predichos en el primer y segundo umbral respectivamente, encontraron su homólogo con su potencial protoespaciador en la base de datos de bacteriófagos.

Este resultado es congruente con la idea de que la poca homología encontrada para los espaciadores CRISPR en los genomas de bacteriófagos, se debe a que los protoespaciadores objetivo se encuentren en fagos que aún no han sido caracterizados y no se encuentran anotados en las BD de referencia, o bien, por su constante mutación en su evolución para evitar ser reconocidos por la bacteria huésped.

Aunado a que la bacteria analizada fue extraída de una muestra de sedimento marino, y se sabe que en el océano, se ha encontrado la mayor cantidad de virus no identificados, a lo que se le denomina "materia viral oscura" (Edwards et al., 2016; Nasko et al., 2019; Roux et al., 2015; Ruohan et al., 2022).

No obstante, se encontraron nuevas potenciales interacciones con bacteriófagos dentro de las familias *Caudovirales no clasificados* y *Autographiviridae* dentro del nivel orden de *Caudovirales* y con las familias *Inoviridae* y *Microviridae* dentro del nivel reino de *Monodnaviria;* y con *Virus de bacterias no clasificados* a nivel reino.

En vista de que los resultados obtenidos son alentadores, se acepta la hipótesis planteada en el presente trabajo, ya que la red neuronal convolucional implementada, fue capaz de reconocer los patrones estructurales de las matrices CRISPR en la secuencia de nucleótidos analizada y, con los espaciadores extraídos, se encontraron potenciales interacciones con sus bacteriófagos homólogos.

Como trabajo futuro se plantea la automatización del modelo computacional para recibir como entrada de datos diferentes genomas bacterianos y mejorar su rendimiento. La implementación de un modelo multiclase, que sea entrenado para especificar la taxonomía de la bacteria analizada, de la cual, se caractericen matrices CRISPR con el primer modelo. Esto con la finalidad de realizar anotaciones específicas de interacción bacteriófago – bacteria hospedera, incluso si se desconoce la clasificación de la bacteria de entrada.

Así también, la generación de una base de datos de los espaciadores de las matrices CRISPR predichas, ya que al tratarse de potenciales protoespaciadores de fagos conocidos y desconocidos, esta información podría contribuir a estudios biotecnológicos de la tendencia de los sitios objetivo de la maquinaria del sistema CRISPR-Cas. Aunado a ello, podría aportar a la caracterización de fagos desconocidos con la información taxonómica asignada a la bacteria de donde fueron extraídos, mediante el modelo multiclase.

Referencias

Alkhnbashi, O. S., Costa, F., Shah, S. A., Garrett, R. A., Saunders, S. J., & Backofen, R. (2014). CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics*, *30*(17), i489–i496. https://doi.org/10.1093/BIOINFORMATICS/BTU459

Alkhnbashi, O. S., Meier, T., Mitrofanov, A., Backofen, R., & Voß, B. (2019, July 19). CRISPR-Cas bioinformatics. *Methods*. https://doi.org/10.1016/j.ymeth.2019.07.013

Alkhnbashi, O. S., Mitrofanov, A., Bonidia, R., Raden, M., Tran, V. D., Eggenhofer, F., … Backofen, R. (2021). CRISPRloci: comprehensive and accurate annotation of CRISPR–Cas systems. *Nucleic Acids Research*, *49*(W1), W125–W130. https://doi.org/10.1093/NAR/GKAB456

Alkhnbashi, O. S., Shah, S. A., Garrett, R. A., Saunders, S. J., Costa, F., & Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. *Bioinformatics*, *32*(17), i576–i585. https://doi.org/10.1093/bioinformatics/btw454

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. https://doi.org/10.1016/S0022-2836(05)80360-2

Azam, A. H., & Tanji, Y. (2019, March). Bacteriophage-host arm race: an update on the mechanism of phage resistance in bacteria and revenge of the phage with the perspective for phage therapy. *Applied Microbiology and Biotechnology*, Vol. 103, pp. 2121–2131. https://doi.org/10.1007/s00253-019-09629-x

Beller, L., & Matthijnssens, J. (2019). What is (not) known about the dynamics of the human gut virome in health and disease. *Current Opinion in Virology*, *37*, 52–57. https://doi.org/10.1016/J.COVIRO.2019.05.013

Biswas, A., Fineran, P. C., & Brown, C. M. (2014). Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics*, *30*(13), 1805–1813. https://doi.org/10.1093/BIOINFORMATICS/BTU114

Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C., & Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, *17*(1), 1–14. https://doi.org/10.1186/s12864-016-2627-0

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., & Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics 2007 8:1*, *8*(1), 1–8. https://doi.org/10.1186/1471-2105-8-209

Broude, N. E. (2002). Stem-loop oligonucleotides: a robust tool for molecular biology and biotechnology. *Trends in Biotechnology*, *20*(6), 249–256. https://doi.org/10.1016/S0167-7799(02)01942-X

Coclet, C., & Roux, S. (2021). Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology*, *49*, 117–126. https://doi.org/10.1016/J.COVIRO.2021.05.003

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., … Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research*, *46*(W1), W246–

W251. https://doi.org/10.1093/nar/gky425

Crawley, A. B., Henriksen, J. R., & Barrangou, R. (2018). CRISPRdisco: An Automated Pipeline for the Discovery and Analysis of CRISPR-Cas Systems. *The CRISPR Journal*, *1*(2), 171–181. https://doi.org/10.1089/crispr.2017.0022

Dion, M. B., Plante, P. L., Zufferey, E., Shah, S. A., Corbeil, J., & Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Research*, *49*(6), 3127–3138. https://doi.org/10.1093/nar/gkab133

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23*(1), 205–211. https://doi.org/10.1142/9781848165632_0019

Edgar, R. C. (2007). PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics 2007 8:1*, *8*(1), 1–6. https://doi.org/10.1186/1471-2105-8-18

Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, *40*(2), 258–272. https://doi.org/10.1093/femsre/fuv048

Emmert-Streib, F., Yang, Z., Feng, H., Tripathi, S., & Dehmer, M. (2020). An Introductory Review of Deep Learning for Prediction Models With Big Data. *Frontiers in Artificial Intelligence*, *0*, 4. https://doi.org/10.3389/FRAI.2020.00004

Frank-Kamenetskii, M. D. (1997). Biophysics of the DNA molecule. *Physics Reports*, *288*(1–6), 13–60. https://doi.org/10.1016/S0370-1573(97)00020-3

Godde, J. S., & Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*, *62*(6), 718–729. https://doi.org/10.1007/s00239-005-0223-z

Grissa, I., Vergnaud, G., & Pourcel, C. (2007a). CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(SUPPL.2), W52–W57. https://doi.org/10.1093/nar/gkm360

Grissa, I., Vergnaud, G., & Pourcel, C. (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics 2007 8:1*, *8*(1), 1–10. https://doi.org/10.1186/1471-2105-8-172

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA Websuite. *Nucleic Acids Research*, *36*(suppl_2), W70–W74. https://doi.org/10.1093/NAR/GKN188

Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., & Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. *Cell*, *172*(6), 1239–1259. https://doi.org/10.1016/j.cell.2017.11.032

Huang, Y., Niu, B., Gao, Y., Fu, L., & Li, W. (2010). CD-HIT Suite: A web server for clustering and comparing biological sequences. *Bioinformatics*, *26*(5), 680–682. https://doi.org/10.1093/bioinformatics/btq003

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/GR.5969107

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal:

prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics 2010 11:1*, *11*(1), 1–11. https://doi.org/10.1186/1471-2105-11-119

Jansen, R., Van Embden, J. D. A., Gaastra, W., & Schouls, L. M. (2002). Identification of a novel family of sequence repeats among prokaryotes. *OMICS A Journal of Integrative Biology*, *6*(1), 23–33. https://doi.org/10.1089/15362310252780816

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Karimi, Z., Ahmadi, A., Najafi, A., & Ranjbar, R. (2018). Bacterial CRISPR Regions: General Features and their Potential for Epidemiological Molecular Typing Studies. *The Open Microbiology Journal*, *12*(1), 59–70. https://doi.org/10.2174/1874285801812010059

Khawaldeh, S., Pervaiz, U., Elsharnoby, M., Alchalabi, A. E., & Al-Zubi, N. (2017). Taxonomic classification for living organisms using convolutional neural networks. *Genes*, *8*(11). https://doi.org/10.3390/genes8110326

Koonin, E. V., & Makarova, K. S. (2019, May). Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 374. https://doi.org/10.1098/rstb.2018.0087

Koumakis, L. (2020). Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, *18*, 1466–1473. https://doi.org/10.1016/J.CSBJ.2020.06.017

Li, M., & Zhang, W. (2021). PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Briefings in Bioinformatics*. https://doi.org/10.1093/BIB/BBAB348

M, D., N, L., & R, O. (1997). Searching for patterns in genomic data. *Trends in Genetics : TIG*, *13*(12), 497–498. https://doi.org/10.1016/S0168-9525(97)01347-4

Manrique, P., Dills, M., & Young, M. J. (2021). Bacteriophages of the Human Microbiome. In *Encyclopedia of Virology* (pp. 283–290). https://doi.org/10.1016/b978-0-12-809633-8.21226-0

Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., … Kyrpides, N. C. (2006). The integrated microbial genomes (IMG) system. *Nucleic Acids Research*, *34*(suppl_1), D344–D348. https://doi.org/10.1093/NAR/GKJ024

Marraffini, L. A., & Sontheimer, E. J. (2010, March 2). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature Reviews Genetics*, Vol. 11, pp. 181–190. https://doi.org/10.1038/nrg2749

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K. S., Koonin, E. V, & Backofen, R. (2021). CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Research*, *49*(4), e20–e20. https://doi.org/10.1093/nar/gkaa1158

Moreno del Castillo, M. C., Valladares-García, J., Halabe-Cherem, J., Moreno del Castillo, M. C., Valladares-García, J., & Halabe-Cherem, J. (2018). Microbioma humano. *Revista de La Facultad de Medicina (México)*, *61*(6), 7–19. https://doi.org/10.22201.fm.24484865e.2018.61.6.02

Naranjo-Torres, J., Mora, M., Hernández-García, R., Barrientos, R. J., Fredes, C., & Valenzuela, A. (2020). A Review of Convolutional Neural Network Applied to Fruit Image Processing. *Applied Sciences 2020, Vol. 10, Page 3443*, *10*(10), 3443. https://doi.org/10.3390/APP10103443

Nasko, D. J., Ferrell, B. D., Moore, R. M., Bhavsar, J. D., Polson, S. W., & Wommack, K. E. (2019). Crispr spacers indicate preferential matching of specific virioplankton genes. *MBio*, *10*(2). https://doi.org/10.1128/mBio.02651-18

Nguyen, N. G., Tran, V. A., Ngo, D. L., Phan, D., Lumbanraja, F. R., Faisal, M. R., … Satou, K. (2016). DNA Sequence Classification by Convolutional Neural Network. *Journal of Biomedical Science and Engineering*, *09*(05), 280–286. https://doi.org/10.4236/jbise.2016.95021

Nooij, S., Schmitz, D., Vennema, H., Kroneman, A., & Koopmans, M. P. G. (2018, April 23). Overview of virus metagenomic classification methods and their biological applications. *Frontiers in Microbiology*, Vol. 9. https://doi.org/10.3389/fmicb.2018.00749

Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C. P. L. F., & Backofen, R. (2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. *GigaScience*, *9*(6), 1–12. https://doi.org/10.1093/GIGASCIENCE/GIAA062

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J. P., Couvin, D., Toffano-Nioche, C., & Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Research*, *48*(D1), D535–D544. https://doi.org/10.1093/nar/gkz915

Reiman, D., Metwally, A., & Dai, Y. (2017). Using convolutional neural networks to explore the microbiome. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 4269–4272. https://doi.org/10.1109/EMBC.2017.8037799

Rho, M., Wu, Y.-W., Tang, H., Doak, T. G., & Ye, Y. (2012). Diverse CRISPRs Evolving in Human Microbiomes. *PLOS Genetics*, *8*(6), e1002441. https://doi.org/10.1371/JOURNAL.PGEN.1002441

Rousseau, C., Gonnet, M., Le Romancer, M., & Nicolas, J. (2009). CRISPI: A CRISPR interactive database. *Bioinformatics*, *25*(24), 3317–3318. https://doi.org/10.1093/bioinformatics/btp586

Roux, S., Hallam, S. J., Woyke, T., & Sullivan, M. B. (2015). Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *ELife*, *4*(JULY2015). https://doi.org/10.7554/ELIFE.08490.001

Rowan-Nash, A. D., Korry, B. J., Mylonakis, E., & Belenky, P. (2019). Cross-Domain and Viral Interactions in the Microbiome. *Microbiology and Molecular Biology Reviews*, *83*(1). https://doi.org/10.1128/mmbr.00044-18

Ruohan, W., Xianglilan, Z., Jianping, W., & Shuai Cheng, L. I. (2022). DeepHost: phage host prediction with convolutional neural network. *Briefings in Bioinformatics*, *23*(1). https://doi.org/10.1093/BIB/BBAB385

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *Https://Home.Liebertpub.Com/Crispr*, *3*(6), 462–469. https://doi.org/10.1089/CRISPR.2020.0059

Schmidt, B., & Hildebrandt, A. (2021). Deep learning in next-generation sequencing. *Drug Discovery Today*, *26*(1), 173–180. https://doi.org/10.1016/J.DRUDIS.2020.10.002

Shah, S. A., Erdmann, S., Mojica, F. J. M., & Garrett, R. A. (2013). Protospacer recognition motifs. *RNA Biology*, *10*(5), 891–899. https://doi.org/10.4161/rna.23764

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2017). The CRISPR spacer space is dominated by sequences from species-specific mobilomes. *MBio*, *8*(5). https://doi.org/10.1128/MBIO.01397-17/SUPPL_FILE/MBO004173485SF4.PDF

Sony, S., Dunphy, K., Sadhu, A., & Capretz, M. (2021). A systematic review of convolutional neural network-based structural condition assessment techniques. *Engineering Structures*, *226*, 111347. https://doi.org/10.1016/J.ENGSTRUCT.2020.111347

Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, *22*(3). https://doi.org/10.1093/BIB/BBAA177

Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M., & Moore, J. H. (2018). Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics*, *85*, 168–188. https://doi.org/10.1016/J.JBI.2018.07.015

Wang, K., & Liang, C. (2017). CRF: Detection of CRISPR arrays using random forest. *PeerJ*, *2017*(4), e3219. https://doi.org/10.7717/peerj.3219

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., … Ahlgren, N. A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/nargab/lqaa044

Zhang, F., Zhao, S., Ren, C., Zhu, Y., Zhou, H., Lai, Y., … Huang, Z. (2018). CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions. *Communications Biology*, *1*(1), 180. https://doi.org/10.1038/s42003-018-0184-6

Zhang, F., Zhou, F., Gan, R., Ren, C., Jia, Y., Yu, L., & Huang, Z. (2019). PHISDetector: a web tool to detect diverse in silico phage-host interaction signals. *BioRxiv*, 661074. https://doi.org/10.1101/661074

Zhang, Q., & Ye, Y. (2017). Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. *BMC Bioinformatics 2017 18:1*, *18*(1), 1–12. https://doi.org/10.1186/S12859-017-1512-4

Zhang, R., Mirdita, M., Karin, E. L., Norroy, C., Galiez, C., & Sö Ding, J. (2021). SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics*, *37*(19), 3364–3366. https://doi.org/10.1093/BIOINFORMATICS/BTAB222

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., & Peng, S. (2019). Deep learning in omics: a survey and guideline. *Briefings in Functional Genomics*, *18*(1), 41–57. https://doi.org/10.1093/BFGP/ELY030

Zielezinski, A., Deorowicz, S., & Gudyś, A. (2022). PHIST: fast and accurate prediction of prokaryotic hosts from metagenomic viral sequences. *Bioinformatics*, *38*(5), 1447–1449. https://doi.org/10.1093/BIOINFORMATICS/BTAB837

Zuallaert, J., Godin, F., Kim, M., Soete, A., Saeys, Y., & De Neve, W. (2018). SpliceRover: interpretable convolutional neural networks for improved splice site prediction. *Bioinformatics*, *34*(24), 4180–4188. https://doi.org/10.1093/bioinformatics/bty497