Workflow para encontrar ORFs en el extremo 5’ y motivos de deslizamiento de la polimerasa en secuencias genómicas de miembros de la Familia Potyviridae

Giannina Stephany Bambarén Capurro
Máster Universitario en Bioinformática y Bioestadística UOC-UB
Bioinformática y Bioestadística Área 3

Nombre Consultor/a: Diego Garrido Martín
Nombre Profesor/a responsable de la asignatura: Ferran Prados Carrasco
Tutor CRAG: Juan José López Moya

Fecha Entrega: 24 de junio de 2020

FICHA DEL TRABAJO FINAL

Título del trabajo:	Workflow para encontrar ORFs en el extremo 5' y motivos de deslizamiento de la polimerasa en secuencias genómicas de miembros de la Familia Potyviridae
Nombre del autor:	Giannina Stephany Bambarén Capurro
Nombre del consultor/a:	Diego Garrido Martín
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega (mm/aaaa):	06/2020
Titulación:	Máster en Bioinformática y Bioestadística
Área del Trabajo Final:	Área 3
Idioma del trabajo:	Español
Palabras clave	ORFs, FIMO, gkm-SVM.

Resumen del Trabajo (máximo 250 palabras): Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.

Los virus de plantas, y en concreto *Potyvirus* e *Ipomovirus*, son patógenos que causan graves pérdidas económicas en cultivos agrícolas. Conocer sus productos génicos puede servir para diseñar estrategias de control de estas patologías.

Existe una necesidad en aumentar el conocimiento sobre patógenos virales de plantas, automatizando la anotación y análisis de sus genomas, en un momento en que la secuenciación de genomas es muy accesible para los investigadores.

Tanto los uORFs como las zonas de deslizamiento de la polimerasa, son eventos que responden a una necesidad evolutiva de los virus sometidos a un cierto tipo de condición. Muchos de estos eventos generan estrategias para que los virus mejoren su adaptabilidad y capacidad infectiva.

El conocimiento de estas zonas nos ayuda a entender conocer su biología y a prevenir sus potenciales riesgos.

En este sentido, este trabajo intenta resolver algunas de estas cuestiones. Se realiza la búsqueda de uORFs en la zonas 5'-UTR de *Ipomovirus*. Además, se prueban FIMO y gkm-SVM como herramientas para la búsqueda de zonas de deslizamiento de la polimerasa.

Finalmente, sobre el uso de estas dos herramientas, se encuentra que gkm-SVM es un algoritmo de clasificación poderoso mucho más preciso que FIMO en la búsqueda de zonas de deslizamiento de la polimerasa, sin embargo, podemos encontrar que los costos de entrenamiento del algoritmo son mucho mayores que la búsqueda de motivos mediante FIMO.

Abstract (in English, 250 words or less):

Plant viruses, in particular *Potyvirus* and *Ipomovirus*, are pathogens that cause serious economic losses in agricultural crops. Knowing their gene products can be useful to design strategies to control these pathologies.

There is a need to increase knowledge about viral plant pathogens, automating the recording and analysis of their genomes, at a time when genome sequencing is very accessible to researchers.

Both the uORFs and the slip zones of the polymerase, are events that respond to an evolutionary need of viruses subject to a certain type of condition. Many of these events generate strategies for viruses to improve their adaptability and infective capacity.

Knowledge of these areas helps us to understand their biology and prevent their potential risks.

In this sense, this work attempts to resolve some of these issues. The uORFs are searched in the 5'-UTR zones of *Ipomovirus.* Additionally, FIMO and gkm-SVM are also tested as tools for the search of polymerase slippage zones.

Finally, on the use of these two tools, it is found that gkm-SVM is a powerful classification algorithm much more accurate than FIMO in the search for polymerase slippage zones, however, the training costs of the algorithm are much higher than the search for motives using FIMO.

Índice

Lista de tablas

Lista de figuras

1. Introducción

1.1. Contexto y justificación del Trabajo

1.1.1. Contexto teórico:

A) Familia *Potyviridae*

Potyviridae es una de las familias de virus de plantas que más especies alberga, llegando a representar cerca del 30% de los virus de plantas conocidos (1). Se han descrito actualmente más de 228 especies para esta familia, agrupadas en 12 géneros (2): *Arepavirus, Bevemovirus, Brambyvirus, Bymovirus, Celavirus, Ipomovirus, Macluravirus, Poacevirus, Potyvirus, Roymovirus, Rymovirus* y *Tritimovirus*, siendo *Potyvirus* el género con más especies (3).

Todos los virus de la familia *Potyviridae* son virus que infectan plantas exclusivamente. Presentan además entre ellos una filogenia muy relacionada, en la cual sus productos génicos tienen generalmente más del 76% de identidad nucleotídica y más del 82% de identidad aminoacídica (4).

Muchas de las especies pertenecientes a esta familia presentan un gran interés socioeconómico ya que están relacionadas con infecciones patógenas en muchas plantas con usos agrícolas (5).

Los virus de esta familia presentan todos genomas de ARN de cadena simple positiva, siendo monopartitos la mayoría, a excepción del género *Bymovirus* que es bipartito. El genoma de los virus monopartitos de la familia se traduce principalmente como una poliproteína, la cual presenta varias secuencias de autocorte en cis y en trans, reconocibles de forma específica por proteasas del virus, dando lugar a los diferentes productos génicos funcionales. En general la poliproteína de los miembros de esta familia se procesa en hasta 10 productos génicos, los cuales presentan una gran conservación de secuencia y organización dentro del genoma (3).

En todos los miembros del género *Potyvirus* se ha descrito la presencia de una pequeña ORF (PIPO) en una fase de lectura diferente a la de la poliproteína viral (6) y que se expresa por un mecanismo peculiar de deslizamiento de la polimerasa (7,8) que puede añadir o quitar una adenina en dominios conservados con al menos seis residuos A consecutivos, generando un producto génico con una parte codificada en la nueva pauta de lectura. Este deslizamiento da lugar a una proteína parcialmente solapante con el

producto P3, localizado en tercera posición desde el extremo N-terminal y denominada P3N-PIPO (6).

Para este trabajo nos hemos enfocado principalmente en los virus de la familia pertenecientes a los géneros *Ipomovirus* y *Potyvirus* de los que a continuación aportamos algunas características de interés.

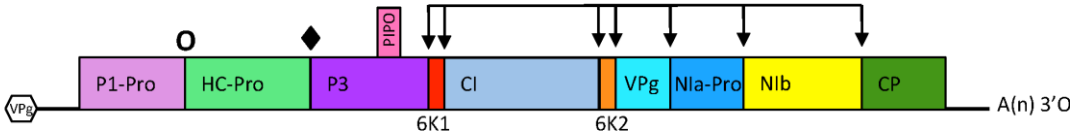

Figura 1: Genoma de Potyvirus.

Genoma de referencia para los virus de esta familia. Imagen obtenida de ICTV: *Potyviridae*. (https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/potyviridae, consultada:02/03/2020)

B) *Ipomovirus*

Los virus del género *Ipomovirus* tienen como vector a la mosca blanca (*Bemisia tabaci*). Su organización y estructura genómica son muy parecidas a la de otros miembros de la familia. Una de las característica de algunos miembros de este género como el virus del amarilleamiento de las venas del pepino (CVYV) y el virus del amarilleamiento de las venas del calabacín (SqVYV) no presentan el producto HC-Pro (Helper Component-Protease) que sí poseen la mayoría de los demás miembros de la familia (9), y en la posición correspondiente del genoma (segundo producto desde el extremo N-terminal de la poliproteína) ha sido reemplazada por una duplicación de la proteína P1 (10,11). A esta P1 duplicada se le denomina "P1b", siendo además la que realiza la función de supresor de silenciamiento, que en otros virus desempeña HCPro (12).

En el grupo de investigación del CRAG recientemente se ha utilizado la técnica de RACE (Rapid Amplification of cDNA Ends) (13) para amplificar la región 5' del genoma del virus CVYV, encontrando que en esta región el virus presentaba 17 nucleótidos más de la referencia de anotación del virus en GeneBank (9). Es interesante resaltar que, en esta pequeña secuencia incorporada al genoma del virus, se observó la presencia de un codón "ATG" que podría dar lugar a una ORF extra, en pauta de lectura diferente de la de la poliproteína. Las ORFs situadas antes del inicio de las proteínas virales se denominan uORF (del inglés "upstream ORF"). En otros virus se ha postulado que las uORF pueden tener un papel en la regulación de la traducción de otros productos génicos, pero en el caso de miembros de la familia *Potyviridae* se ignora si pueden tener alguna función (resultados no publicados).

C) *Potyvirus*

El género *Potyvirus* es el más numeroso y representativo de la familia, e incluye virus que son transmitidos por pulgones de forma no persistente. Este género presenta 192 especies que colectivamente son capaces de infectar a un gran número de plantas huéspedes, aunque individualmente cada uno tiene una gama relativamente restringida a pocas especies. Dentro del género es posible diferenciar algunos grupos con características propias, como por ejemplo los que infectan batatas. Recientemente se ha observado que la mayor parte de los potyvirus que infectan batata presentan en sus genomas dos motivos conservados de deslizamiento de la polimerasa (del tipo G_2A_6), que además de permitir expresar PIPO podrían servir para que se produzca una nueva pauta de lectura en el producto P1, generando un nuevo producto génico conocido como P1N-PISPO. En publicaciones recientes se ha observado que esta proteína puede ser supresor de silenciamiento dotando a los virus que la portan la capacidad de escapar de los mecanismos de defensa inmune del huésped (7,14).

D) Variación génica

Los virus tienen una gran capacidad para generar variación genética que les permita adaptarse a sus huéspedes y sus vectores. En este caso, los virus de la familia *Potyviridae* en su replicación utilizan una RNA polimerasa RNA dependiente (RdRp) propia, y se sabe que este tipo de polimerasas presentan generalmente una tasa de error mayor a la de otras polimerasas. A la posibilidad de cometer errores por sustitución nucleotídica hay que añadir también la recombinación como fuente de variabilidad en sus secuencias (15). Muchos de estos cambios están sometidos a selección positiva y son determinantes para la virulencia (16), adaptación al huéspedes y supresión de la defensa de la planta. Se considera que estos procesos pueden favorecer la aparición de nuevas variantes o especies emergentes, por ejemplo, con nuevas capacidades infectivas en nuevos huéspedes (17,18). La capacidad que tiene esta familia para infectar distintos huéspedes refleja su alto nivel de adaptabilidad.

1.1.2. Justificación del TFM:

Como hemos mencionado, los miembros de la familia *Potyviridae* son virus responsables de graves enfermedades en numerosos cultivos, y que pueden generar grandes daños económicos. Por ello conocer la estructura genómica de los mismos nos permitirá identificar posibles dianas para la lucha antiviral. En este trabajo se pretende realizar la identificación de posibles zonas en el genoma de los virus que presenten la capacidad de producir nuevos productos génicos desconocidos, por ejemplo,

analizando la existencia de fases de lectura en la región 5' no codificante (conocidas como uORF, como ya hemos indicado) y que precede a la poliproteína de los virus. Además, se pretende encontrar regiones de posibles deslizamientos de la polimerasa en motivos similares a G_2A_{6-7}. Partiendo del conocimiento disponible sobre P3N-PIPO y P1N-PISPO, se pretende observar si existen motivos de deslizamiento de polimerasa en otras especies virales que pudieran dar lugar a nuevos productos.

1.2 Objetivos del Trabajo

El objetivo principal de este trabajo engloba el desarrollo de workflows que nos permitan encontrar, y caracterizar en su caso, elementos de interés en el genoma de virus de la familia *Potyviridae* con capacidad de dar lugar a nuevos productos génicos. Para ello se plantean dos objetivos esenciales:

Objetivo 1: Encontrar ORFs en regiones upstream 5' para verificar si su existencia y conservación en miembros de la familia *Potyviridae* y en particular del género *Ipomovirus* pueden asociarse a algún papel funcional.
Objetivo 2: Encontrar motivos de secuencia de deslizamiento de la polimerasa (polymerase slippage) en virus de la familia *Potyviridae*.
Los objetivos 1 y 2 son parte del objetivo principal antes mencionado que englobará todo el trabajo.
Para la realización del primer objetivo se plantea:

- Búsqueda de uORF en todos las miembros de la familia *Potyviridae*.
- Selección de las secuencias que contienen uORFs (candidatas) y búsqueda de uORF en las otras secuencias disponibles de cepas o aislados de la misma especie.
- Cuantificar las secuencias candidatas y realizar un análisis filogénico de estas.

Para la realización del segundo objetivo se plantea:

- Automatizar el análisis de búsqueda de posibles regiones de deslizamiento de la polimerasa que puedan dar lugar a fragmentos génicos desconocidos .
- Generar una herramienta que nos permita la visualización de los resultados para el entendimiento de estos.

Objetivo 3: Como reto adicional para este trabajo se planteó como posible un tercer subobjetivo futuro, no indispensable, que se basaba en el análisis de las posibles ORFs

encontradas en los anteriores atendiendo a la frecuencia de uso de codones “codon usage frequencies” (19). Este análisis podría dar una estimación de la traducibilidad de los productos génicos localizados de forma teórica en los diferentes organismos hospedadores vegetales, y así poder comprobar si hay correlación con la gama de plantas susceptibles a los diferentes virus de la familia. En función del desarrollo de los objetivos principales, se planteará el posible abordaje de este último.

1.3 Enfoque y método seguido

La búsqueda de patrones de secuencia como puedan ser ORFs o frameshifting se puede plantear desde muchas perspectivas, además se disponen de muchas herramientas genómicas para llevar a cabo este tipo de búsquedas. En nuestro trabajo hemos escogido las herramientas que se adaptan mejor al tipo de resultado que estamos buscando y que además tengan la simplicidad para poder llevarlas a cabo.

El primer paso realizado es la instalación el sistema Ubuntu 18.04 basado en Linux 64 bits, en una máquina virtual VirtualBox 6.0 de Oracle VM (Versión 6.0.12) . En esta máquina virtual implementaremos los programas necesarios para el desarrollo de todo el trabajo.

Para el cumplimiento de nuestro primer objetivo implementamos la máquina virtual con las utilidades de NCBI mediante el programa Entrez Utilities “edirect” (disponible: https://www.ncbi.nlm.nih.gov/books/NBK179288/), esta herramienta nos ha permitido la búsqueda y adaptación de todas las secuencia utilizadas a lo largo de este trabajo. Para la búsqueda de ORFs utilizamos el programa ORFfinder de NCBI (disponible: https://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/). Este programa genera una búsqueda de patrones que pueden originar pautas de lectura que inician la traducción de una proteína. Como resultado de estas búsquedas se generan “outputs” que analizaremos mediante R/Bioconductor. Finalmente, analizaremos la filogenia de este género mediante la librería de R/Bioconductor “msa” (20).

Para nuestro segundo objetivo hemos utilizado el programa MEME para la búsqueda de motivos de deslizamiento de la polimerasa en un grupo de secuencias del género *Potyvirus* y hemos buscado “matches” de este motivo en todas las secuencias del mismo género, para ello hemos utilizado una herramienta de meme llamada FIMO. De los productos obtenidos en FIMO hemos comprobado si los “matches” obtenidos son realmente significativos mediante la implementación de un algoritmo de clasificación

SVM (Support vector machine), en concreto hemos entrenado y clasificado nuestras secuencias utilizando el paquete gkmSVM de R/Bioconductor.

Finalmente, la incorporación de este workflow la haremos mediante un informe dinámico en R-Markdown.

1.4 Planificación del Trabajo

Este trabajo se ha planteado con una duración de tres meses y medio, el principio se iba a llevar a cabo de forma presencial en los laboratorios del Consorcio CRAG (Centre for research in Agricultural Genomics) que comprende grupos de varias instituciones, CSIC-IRTA-UAB-UB. Debido a las circunstancias actuales motivadas por la declaración del estado de alarma por la emergencia sanitaria causada por la pandemia de COVID-19, y teniendo en cuenta el estado de desarrollo del trabajo al declararse la alarma, hemos realizado modificaciones de la planificación original del trabajo, que ha tenido lugar de forma remota.

Finalmente, el trabajo ha sido supervisado de forma periódica por los tutores del trabajo mediante el foro de la UOC, videollamadas y correo electrónico.

Objetivo 1:

- Actividad 1: será la primera toma de contacto con el laboratorio además se establecerán las pautas del trabajo a realizar. Esta información nos servirá para la redacción de la PAC0.
- Actividad 2: durante este tiempo realizaremos la implementación de los programas necesarios para la preparación de este trabajo.
- Actividad 4: Se redactará la PAC1.
- Actividad 5: Se analizarán los resultados de lo observado en la búsqueda de uORF.
- Actividad 6: se realizará un estudio de la diversidad de las diferentes cepas de las especies que contengan uORFs.
- Actividad 7: se dedicará este tiempo para la redacción de la PAC2.

Objetivo 2:

- Actividad 8: se intentará encontrar los motivos de deslizamiento de la polimerasa en todos los miembros de la familia *Potyviridae*. Se podrán a punto los programas utilizados para la realización de este objetivo MEME, FIMO, gkmSVM.

- Actividad 9: redacción de la PAC3.
- Actividad 10: Automatización e implementación de los resultados obtenidos en un informe dinámico.
- Actividad 11: escritura de la memoria
- Actividad 12: generación de una presentación en PowerPoint para
- Actividad 13: preparación para la presentación oral

El calendario propuesto para realizar estas actividades es el siguiente:

Tabla 1: Calendario propuesto para el desarrollo del trabajo.

	Breve descripción	Inicio	Días	Final
Actividad 1	Preparación del proyecto	02-mar	5	06-mar
Actividad 2	Búsqueda de uORFs	09-mar	15	27-mar
Actividad 3	# Parada por Coronavirus #	13-mar	15	03-abr
Actividad 4	Redacción PAC 1	11-mar	5	16-mar
Actividad 5	Análisis de resultados	29-mar	15	10-abr
Actividad 6	Estudio diversidad uORFs	13-abr	5	17-abr
Actividad 7	Redacción PAC 2	20-abr	3	22-abr
Actividad 8	Búsqueda de motivos	23-abr	32	29-may
Actividad 9	Redacción PAC3	28-may	5	01-jun
Actividad 10	Automatización	01-jun	14	18-jun
Actividad 11	Escritura de la memoria	14-jun	10	24-jun
Actividad 12	Presentación	24-jun	5	28-jun
Actividad 13	Defensa oral	01-jul	-	08-jul

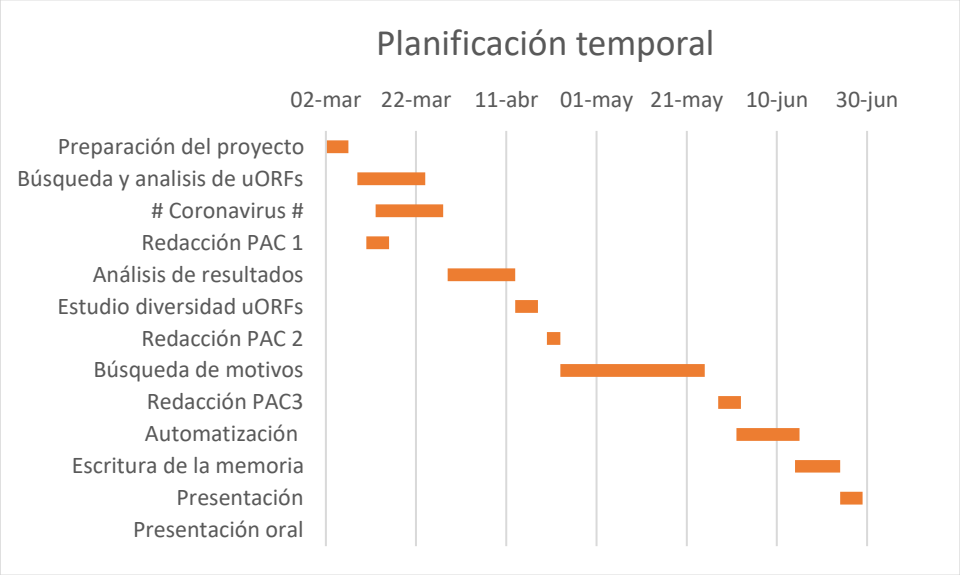

Figura 2: Diagrama de Gantt.

Planificación temporal seguida para el desarrollo de este trabajo.

1.5 Breve sumario de productos obtenidos

Los productos obtenidos de este trabajo son los que han sido entregados periódicamente:

- PAC 1: Plan de Trabajo
- PAC 2 (Avances de proyecto): Análisis de uORF en el extremo 5' de los miembros del género *Ipomovirus* mediante ORFfinder de NCBI
- PAC 3 (Avances del proyecto): Análisis de la búsqueda de motivos de deslizamiento de la polimerasa en secuencias de la familia *Potyvirus* usando MEME, FIMO y el algoritmo gkmSVM con R/Bioconductor.

Finalmente, recopilaremos y uniremos todos estos resultados en una entrega: pipeline, informe de resultados y memoria final del Trabajo.

1.6 Breve descripción de los otros capítulos de la memoria

Hemos diseñado esta memoria en 3 capítulos.
En el **primer capítulo**, expondremos todo lo realizado para la obtención de uORF en las secuencias de *Ipomovirus*, además realizaremos un análisis de estos resultados.
En el **segundo capítulo**, realizaremos la comparación entre la búsqueda de motivos y patrones de deslizamiento de la polimerasa realizada por MEME/FIMO y el algoritmo gkmSVM.
Finalmente, en el **tercer capítulo** intentaremos discutir los resultados obtenidos y extraer conclusiones.

2. Metodología:

2.1 Preparación de la maquinaria de trabajo

Para llevar a cabo el desarrollo de este trabajo, el primer paso que realizamos fue la instalación de una máquina virtual con un sistema de operativo Linux, ya que gran parte del trabajo se ha realizado mediante *command-line*. Para llevo instalamos una máquina virtual VirtualBox 6.0 de Oracle VM (Versión 6.0.12) implementada con Ubuntu 18.04. Dotamos a esta máquina de un disco duro de 25Gb creado dinámicamente.
La segunda parte del análisis y la automatización del mismo se llevó a cabo mediante R (Versión 3.6.1) y su interprete gráfico RStudio (versión 1.2.5001), instalado en Windows 10 x64bits (sistema operativo base). Además, implementamos R con Bioconductor, ya que este software nos permite el análisis de datos moleculares mediante lenguaje de programación R.
Para conectar los dos sistemas operativos utilizamos una carpeta compartida, donde guardamos lo extraído mediante el terminal y lo cargamos mediante las librerías de R específicas para cada archivo.

2.2. Búsqueda de uORF

2.2.1 Obtención de la secuencias:

La secuencias utilizadas en este apartado, pertenecen a la familia *Potyviridae*. Para ello utilizamos cuatro especies del género *Ipomovirus* como son*: Cucumber vein yellowing virus*, *Cassava brown streak virus*, *Squash vein yellowing virus* y *Tomato mild mottle virus*. Como también dos especies del género: *Plum pox virus* y *Tobacco etch virus*.
Nuestro principal objetivo es encontrar uORF en el extremo 5'-UTR de las secuencias del género *Ipomovirus*, de modo que se utilizan las dos especies del género *Potyvirus* para comparar el análisis en otro género de la misma familia.
Para la obtención de secuencias utilizamos las utilidades de programación de Entrez Direct (21) (Entrez Programming Utilities "E-utilities"). Estas utilidades contienen nueve programas "*server-side"* de consultas que generan una interfaz con el sistema de consultas y base de datos de Entrez NCBI (Centro Nacional de Información Biotecnológica).
Instalamos "edirect" siguiendo el protocolo en la terminal (disponible: https://www.ncbi.nlm.nih.gov/books/NBK179288/). Ejecutando el código de descarga obtenemos una carpeta llamada "edirect". La modificación del "PATH" nos permite

ejecutar todos los programas de este directorio. Como dependencias de “edirect” se requiere de “Perl”, disponible en el mismo paquete de descarga de “edirect”.

Para ejecutar las búsquedas de las secuencias utilizamos tres utilidades principalmente:

- ESearch (búsquedas de texto): sirve para responder a consultas de texto en una base de datos determinada, nosotros usamos como base de datos únicamente “Nucleotide” y “Protein”.
- EFetch (descargas de registros de datos): devuelve el resultado de la consulta realizada en un formato especificado. En nuestro caso trabajamos con secuencias en formato “*fasta*”.
- EFilter: restringe los resultados de la consulta.

Como ejemplo a la búsqueda realizada, proporcionamos la búsqueda de “*Cucumber vein yellowing virus*” (CVYV en adelante):

```
$ esearch -db nucleotide -query "Cucumber vein yellowing virus complete" |
  efetch -format fasta > "secuencias/cvyv.fasta"
```

Los parámetros de “*-db*” sirven para especificar la base de datos, “*-query*” donde especificamos la consulta a realizar y con “*-format*” especificamos el formato de salida de la consulta.

Realizamos la misma consulta para todas las especies requeridas y obtuvimos:

Tabla 2: Especies seleccionadas para la búsqueda de uORFs.

Género	*Especie*	Abrv.	Nº secuencias	Archivo
Ipomovirus	*Cassava brown streak virus*	CBSV	12	CBSV.fasta
Ipomovirus	*Cucumber vein yellowing virus*	CVYV	4	cvyv.fasta
Ipomovirus	*Squash vein yellowing virus*	SqVYV	5	svyv.fasta
Potyvirus	*Plum pox virus*	PPV	423	ppv.fasta
Potyvirus	*Tobacco etch virus*	TEV	9	tev.fasta
Ipomovirus	*Tomato mild mottle virus*	UCBSV	1	tmmv.fasta

2.2.2 Adicción de RACE en 5’-UTR CVYV

Como hemos comentado anteriormente, recientemente se ha encontrado que las secuencias anotadas y depositadas en GeneBank de “CVYV” carecen de diecisiete nucleótidos en su extremo 5’-UTR. Para realizar este análisis se añadieron los 17 nucleótidos manualmente a las 4 secuencias. En la **Figura 3**, podemos ver el alineamiento de las secuencias de CVYV antes y después de añadir RACE.

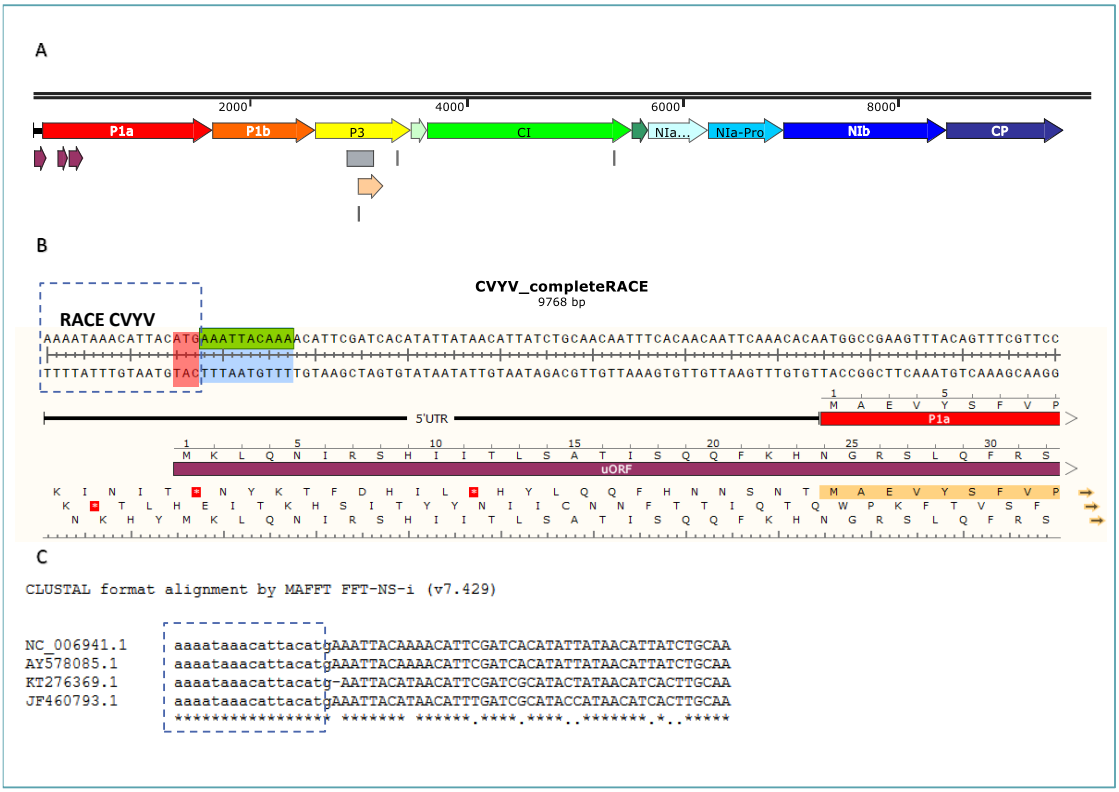

Figura 3: Adicción de la secuencia RACE.

(A) Visión completa del mapa de CVYV. Imagen obtenida mediante SnapGene. **(B)** Secuencia de 17 nucleótidos RACE (**---**) la cual presenta un ATG en 5' UTR (señalado en rojo). Imagen obtenida mediante SnapGene. **(C)** Salida del alineamiento de todos los aislados de CVYV con RACE. Alineamiento realizado con MAFFT (disponible: https://www.ebi.ac.uk/Tools/msa/mafft/) con formato de salida ClustalW.

2.2.3 ORFfinder

Un marco de lectura abierto ORF (Open Reading Frame) es una sección de la cadena de nucleótidos que no presenta codones de parada y puede dar lugar a una proteína candidata. Los uORFs son pequeñas secuencias en la región upstream no codificante, que presentan un papel en la modulación del inicio de la traducción de los CDS en eucariotas (22). En el caso de la familia *Potyviridae* muchos de los virus presentan estrategias para competir contra los ARNm celulares de la planta por factores de traducción limitados (23) , de modo que algunos de ellos han desarrollado elementos de ARN en sus regiones no traducibles (UTR) tanto en 5' como en 3' (24) como estrategia para competir eficientemente por los recursos de la planta.

Open Reading Frame Finder (ORFfinder) de NCBI es una herramienta de análisis gráfico que busca marcos de lectura abiertos en secuencias de ADN. Esta herramienta encuentra los sitios de ORF en todos las direcciones y marcos de lectura, dando como

resultado la cadena de aminoácidos producida por este ORF y su posición dentro de la cadena de nucleótidos.

Hemos descargado ORFfinder en la terminal siguiendo los siguientes pasos:

```
$ mkdir orffinder
$ cd orffinder
$ wget ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/ORFfinder/linux-i64/ORFfinder.gz
$ gunzip ORFfinder.gz
$ chmod 755 ORFfinder
$ export PATH=$PATH:$PWD
```

Para poder ejecutar el programa necesitamos ir hacia la carpeta y exportar el PATH.

Los patrones de búsqueda que hemos utilizado para todas las secuencias analizadas han sido:

- **-s** = 0 : solo inicios "ATG"
- **-ml** =30: tamaño mínimo del ORF de 30 nucleótidos
- **-strand** = plus : solo en la cadena positiva
- **-outfmt: f**ormato de salida del resultado :
  - "0" lista de ORF en formato fasta (péptidos),
  - "1" CDS en formato fasta,
  - "2" archivo en formato "ASN1", con las características del ORF,
  - "3" coordenadas del ORF en formato ".txt"
- **-in:** destino del archivo de entrada
- -**out:** destino del archivo de salida

Ejemplo de búsqueda de ORF en CVYV:

```
$ ORFfinder -in /home/uoc/secuencias/seqorf/cvyv.fasta -s 0 -ml 30 -strand plus -outfmt 3 -out /home/uoc/secuencias/seqorf/cvyv_orf/cvyv_orf.txt
```

Hemos aplicado el mismo tipo de búsqueda para todas las especies y hemos guardado el resultado en carpetas individuales para cada especie, finalmente en la **Tabla 3** podemos ver el resultado obtenido para cada especie.

Tabla 3: Número de secuencias obtenidas por especie.

Género	Especie	Abr.	N.º secuencias	Carpeta seqorf
Ipomovirus	*Cassava brown streak virus*	CBSV	12	cbsv_orf.asn cbsv_orf_list.fasta
Ipomovirus	*Cucumber vein yellowing virus*	CVYV	4	cvyv_orf.asn cvyv_orf_list.fasta
Ipomovirus	*Squash vein yellowing virus*	SqVYV	5	svyv_orf.asn svyv_orf_list.fasta
Potyvirus	*Plum pox virus*	PPV	423	ppv_orf.asn ppv_orf_list.fasta
Potyvirus	*Tobacco etch virus*	TEV	9	tev_orf.asn tev_orf_list.fasta
Ipomovirus	*Tomato mild mottle virus*	UCBSV	1	tev_orf.asn tev_orf_list.fasta

2.1.3 Análisis en R:

Para poder analizar los resultados obtenidos mediante ORFfinder en R, primero cargamos los archivos y las convertimos en un “dataframe”. Para ello utilizamos las librerías de Bioconductor “Biostring” (25) y “Seqinr”(26).

Para seleccionar las especies que presentan uORFs, realizamos dos consultas mediante la librería de R “sqldf”, donde preguntamos sobre las secuencias que presentan un inicio anterior al inicio del ORF más largo. Este ORF más largo será la poliproteína del virus y cualquier inicio anterior a este dará lugar a un uORF.

Ejemplo de consulta realizada para CVYV: La primera pregunta nos devuelve el inicio de los ORF que presentan productos mayores a 8000 nucleótidos. En la segunda pregunta usamos el inicio de poliproteína más tardío para obtener el nombre de la secuencia o secuencias que presentan inicios alternativos a la poliproteína.

```
> sqldf('SELECT start, Name FROM df_cvyv WHERE length > 8000 ORDER BY start ASC')

> (result_cvyv<-sqldf('SELECT ID, Name, ORF,start, stop,sequence FROM df_cvyv WHERE (start < 127 AND length < 8000) ORDER BY start ASC'))
```

Finalmente, hemos guardado la secuencia de aminoácidos de las especies que presentan uORF. Con dichas secuencias realizaremos un alineamiento mediante la librería de Bioconductor “msa” (20).

2.3. Búsqueda de zonas de deslizamiento de la polimerasa

2.3.1 Obtención de secuencias

De la misma forma que en el apartado anterior, hemos obtenido las secuencias mediante “edirect”.

Vamos a crear dos motivos de deslizamiento de la polimerasa. El primer motivo estará basando en el zona de slippage que genera el producto en trans P1N-PISPO. Para ello utilizaremos secuencias de doscientos nucleótidos de todas las cepas de la especie “*Sweet potato feathery mottle virus*” (SPFMV en adelante). Utilizamos las secuencias de las posiciones 1300-1500, ya que es en esta zona donde frecuentemente ocurre el slippage.

```
$ esearch -db nucleotide -query "Sweet potato feathery mottle virus [ORGN] complete genome" | efetch -format fasta -seq_start 1300 -seq_stop 1500 > feath_200t.fasta
$ grep ">" feath_200t.fasta
```

De esta búsqueda hemos obtenido 55 secuencias de 200 nucleótidos de SPFMV.

El segundo motivo de slippage se encuentra en la zona del producto de la poliproteína P3, este deslizamiento produce una proteína solapante P3N-PIPO, esto se produce en todos los miembros de la familia *Potyviridae*. Para la búsqueda vamos a utilizar solo secuencias de *Potyvirus*, ya que este es el género más numeroso. Para ello, realizamos una búsqueda de secuencias con una mayor extensión que en el caso de P1, ya la ubicación del producto P3 es diferente en cada especie. De forma que utilizamos secuencias de 2400 nucleótidos, en las posiciones 2400-4800.

```
$ esearch -db nucleotide -query ‘Potyvirus [ORGN]’ | efetch -format fasta -seq_start 2400 -seq_stop 4800 > potyvirus_p3_meme.fasta
```

De esta búsqueda obtuvimos 2361 secuencias.

En la **Tabla 4** mostramos los comandos de búsqueda utilizados, así como también la cantidad de secuencias obtenidas en cada búsqueda.

Tabla 4: Comandos para la búsqueda en edirect.

Búsqueda edirect (NCBI)	**Descripción**	**Nº**
esearch -db nucleotide -query "Sweet potato feathery mottle virus [ORGN] complete genome" | efetch -format fasta > swp.fasta	Todas las cepas de SPFMV	55
esearch -db nucleotide -query "Sweet potato virus complete genome NOT (partial cds)" | efilter -query 'Potyvirus [ORGN]'| efetch -format fasta > sweetpotato_fimo.fasta	Todos los *Potyvirus* huéspedes de batata	112
esearch -db nucleotide -query "virus complete genome NOT (partial cds)" | efilter -query "Potyvirus[ORGN]"| efetch -format fasta > potyvirus_fimo.fasta	Todas las cepas del g. *Potyvirus*	2351
esearch -db nucleotide -query "virus complete genome NOT (partial cds)" | efilter -query 'Potyviridae[ORGN]'| efetch -format fasta > potyviridae_fimo.fasta	Todas las cepas de la fam. *Potyviridae*	2739

La utilización de estas secuencias las describiremos en el siguiente apartado.

2.3.2 Descubrimiento de motivos: MEME 5.1.1

Para descubrir motivos de secuencia hemos utilizado el programa MEME (Multiple EM for Motif Elicitation), mediante "*command-line*" descargando el programa en el terminal UNIX y siguiendo los pasos especificados en la guía de instalación (disponible: http://meme-suite.org/doc/install.html?man_type=web). Para su instalación también tuvimos que instalar las dependencias requeridas para su desarrollo como: "zlib1g-dev", "autoconf", "automake", "libtool" y "default-jdk". La instalación de estas las realizamos mediante el comando "apt-get install"

MEME permite la búsqueda de motivos en secuencias nucleotídicas (ADN o ARN) o aminoacídicas no alineadas y permite realizar un amplio análisis de estas. Para la búsqueda de motivos, MEME utiliza un algoritmo que maximiza las expectativas probabilísticas de una secuencia y estima cuantas veces ocurre está en cada secuencia con métodos probabilísticos y discretos (27). MEME es también una colección de herramientas (Motif Discovery, Motif Enrichment Analysis, Motif Search, Motif Comparison, Gene Regulation, etc) que permiten una búsqueda de motivos y análisis más especializada.

En nuestro caso los parámetros que hemos tenido en cuenta para la búsqueda de motivos en MEME han sido principalmente:

- nmotifs: donde añadimos el número de motivos que deseamos
- minw: tamaño mínimo del motivo
- maxw: tamaño máximo del motivo

- objfun= classic : nos permite puntuar los motivos utilizando el estadístico E-value.

Las salidas que se obtienen en MEME es una carpeta que contiene:

- Imágenes de cada motivo en formato ".eps"
- Imágenes de cada motivo en formato ".png"
- Resumen *html* de los motivos encontrados y su localización en las secuencias
- Documento de texto con información detallada de cada motivo, como "scoring matrix", "regular expression", "position-specific probability matrix", etc.
- Resumen XML de los motivos encontrados

Para la búsqueda de P1 hemos utilizado el siguiente comando:

```
$ meme /home/uoc/feath_200t.fasta -dna -oc /home/uoc/meme_out/meme2 -nostatus -time 3000 -mod zoops -nmotifs 20 -minw 6 -maxw 16 -objfun classic -revcomp -markov_order 0
```

En este caso hemos obtenido 20 motivos y el cuarto (**Figura 4**) de ellos es el que presenta el patrón especifico buscado.

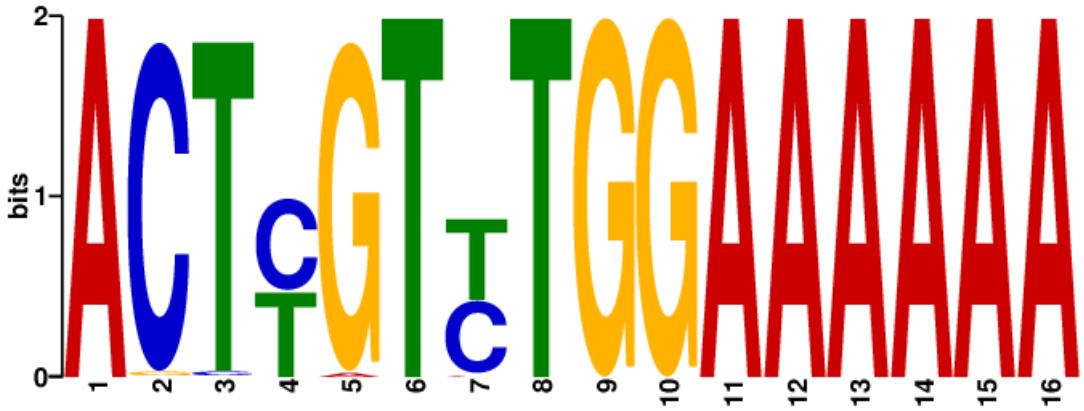

Figura 4: Logo del motivo para zona de "slippage" del producto P1.
Logo obtenido a partir del descubrimiento de motivos con **MEME 5** en las secuencias de nucleótidos de las cepas de la especie SPFMV. En la posición **9-16** del logo encontramos el motivo característico de la zona de deslizamiento de la polimerasa en el producto P1 para la formación de P1N-PISPO

Para el descubrimiento del motivo en P3 se utilizó:

```
$ meme /home/uoc/potyvirus_p3_meme.fasta -dna -oc /home/uoc/meme_out/meme3_p3 -mod zoops -nmotifs 30 -minw 20 -maxw 30 -objfun classic
```

Para este motivo elevamos mucho la cantidad de motivos requeridos ya que al tener secuencias de orígenes diferentes nos costó mucho más encontrar el motivo de

deslizamiento buscado. Finalmente, obtuvimos 30 motivos, siendo el número 20 (**Figura 5**) el cual presentaba el motivo esperado.

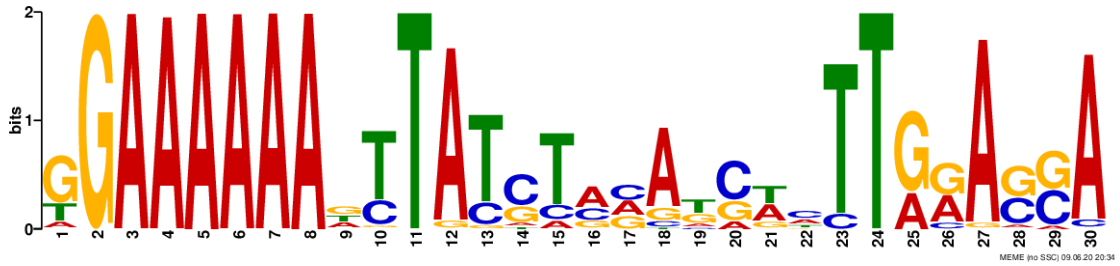

Figura 5: Logo del motivo para la zona del "slippage" del producto P3.

Logo obtenido a partir del descubrimiento de motivos con **MEME 5** en las secuencias de nucleótidos de las cepas de las especies de Potyvirus. En la posición **2-8** del logo encontramos el motivo característico de la zona de deslizamiento de la polimerasa en el producto P3 que origina el producto P3N-PIPO.

Como podemos ver, el logo de P3 a diferencia del de P1, la primera “G” que da lugar a la zona de slippage no está tan conservada dentro del género *Potyvirus*, de forma que presenta una mejor frecuencia, pero las 6 adeninas seguidas si están bien representadas.

2.3.3 Búsqueda de ocurrencias: FIMO

Además del descubrimiento de motivos, MEME presenta herramientas que nos permiten la búsqueda de dicho motivos en otras bases de datos de secuencias, una de estas herramientas es FIMO (28), la cual escanea las secuencias dadas buscando la ocurrencia de un motivo concreto.

La herramienta FIMO “Find Individual Motif Occurrences” trata cada motivo de forma individual, de modo que para poder realizar la búsqueda debemos tener el motivo deseado en formato MEME. Para convertir los motivos a analizar en formato MEME seguimos la guía de formatos de MEME “MEME Motif Format” (disponible: http://meme-suite.org/doc/meme-format.html). Para convertir los motivos 4 de P1 y 20 de P3 en formato MEME lo hicimos de forma manual utilizando la información de la salida de texto (“meme.txt”) de los motivos.

FIMO utiliza el estadístico p-value utilizando un algoritmo de programación dinámica que convierte las puntuaciones de las secuencias en probabilidades. Además también realiza la corrección del p-value mediante el método de Benjamini y Hochberg (29) (obteniendo las q-value de cada secuencia).

Las salidas obtenidas en FIMO son un directorio donde tenemos 5 archivos “fimo” que contienen un resumen y los valores estadístico de cada ocurrencia:

- fimo.html: archivo en formato *html*
- fimo.tsv: un archivo TSV (valores separados por tabulaciones)
- fimo.gff: un archivo de formato GFF3
- cisml.xml: que proporciona los resultados en el esquema CisML
- fimo.xml: que describe las entradas a FIMO y hace referencia al archivo CISML

El comando utilizado para las búsqueda en FIMO es:

$ fimo –oc <directorio de salida> <motivo.meme> <secuencias.fasta>

Para obtener los valores de FIMO realizamos 3 búsquedas para cada motivo, en la **Tabla 5** describimos estos comandos utilizados, como también el tipo de secuencia utilizada en cada motivo.

Tabla 5: Comandos utilizados en FIMO.

Búsqueda en FIMO	Motivo	Secuencias utilizadas
fimo --oc /home/uoc/meme_out/fimo2 /home/uoc/meme4_motif.meme /home/uoc/swp.fasta	P1	Cepas SPFMV (55)
fimo --oc /home/uoc/meme_out/fimo3 /home/uoc/meme4_motif.meme /home/uoc/sweetpotato_fimo.fasta	P1	Todos los *Potyvirus* huéspedes de batata (112)
fimo --oc /home/uoc/meme_out/fimo4 /home/uoc/meme4_motif.meme /home/uoc/potyvirus_fimo.fasta	P1	Todas las cepas del g. *Potyvirus* (2351)
fimo --oc /home/uoc/meme_out/fimo_p3_1 /home/uoc/memep3.meme /home/uoc/sweetpotato_fimo.fasta	P3	Todos los *Potyvirus* huéspedes de batata (112)
fimo --oc /home/uoc/meme_out/fimo_p3_2 /home/uoc/memep3.meme /home/uoc/potyvirus_fimo.fasta	P3	Todas las cepas del g. *Potyvirus* (2351)
fimo --oc /home/uoc/meme_out/fimo_p3_3 /home/uoc/memep3.meme /home/uoc/potyviridae_fimo.fasta	P3	Todas las cepas de la fam. *Potyviridae* (2739)

2.3.4 Algoritmo gkm-SVM: gapped *k*-mer SVM (Support Vector Machine)

Un k-mer se refiere a una conjunto de letras de longitud “k”. Los oligómeros de longitud k son características ampliamente utilizadas para modelar propiedad y funciones de secuencias de ADN o proteínas. Los motivos de secuencia se diferencian de los k-mers, ya que permiten cierto grado de tolerancia a errores (30). Por otro lado, cuando los k-mers son más largos tiene como limitación que la probabilidad de observarlo se vuelve muy pequeña, ya que son patrones más restrictivos (31). Esta restricción se ha solucionado mediante el uso de gapped k-mers, que son k-mers que pueden presentar huecos o espacios, flexibilizando así la estimación de las frecuencias del k-mer.

Las “máquinas vectores de soporte” SVM, son un conjunto de algoritmos de aprendizaje supervisado enfocados en la resolución de problemas de clasificación binaria y regresión(32). Se basa en el cálculo de hiperplanos en el espacio multidimensional. A modo que el hiperplano de separación quede definido por las observaciones de cada clase. Los SVM son ampliamente utilizados en “machine learning” para la clasificación de textos (32). En el caso que los datos no se puedan separar de forma lineal, se utilizan “kernels” o funciones de similitud, donde se busca obtener un parámetro “C” óptimo que minimice el coste de función del algoritmo.

El clasificador gkm-SVM (33) (máquinas de vectores de soporte con núcleos de k-mer con espacios) es una herramienta para la estimación robusta de k-mers(31), que entre sus parámetros de implementación nos permite incluir los valores de “k”, tamaño del k-mer sin hueco, “L” longitud de k-mer con hueco, de modo que valores más altos de “L” permiten que el aprendizaje sea más flexible. El algoritmo gkm-SVM se han utilizado y se ha diseñado para aprender modelos predictivos de secuencias de ADN reguladoras humanas, con lo cual ha desarrollado una estructura eficiente para el cálculo de la matrix string-kernel(31).

En nuestro trabajo hemos implementado el algoritmo gkm-SVM en R para la clasificación de nuestras secuencias obtenidas mediante FIMO.

2.3.5. Preparación de las secuencias FIMO en R

Hemos utilizado el archivo de salida “fimo.gff” para convertir la información obtenida en un dataframe. Para ello hemos utilizado la librería de Bioconductor “rtracklayer” (34), útil para la lectura de archivos en formato “.gff”. Una vez obtenido el dataframe hemos realizado la partición del mismo en datos de entrenamiento y datos para la evaluación. Hemos asignado de forma aleatoria el 70% de los datos para entrenamiento del algoritmo y 30% para la evaluación del mismo.

Para la creación del kernel y para el algoritmo de entrenamiento, necesitamos dos tipos de conjuntos de datos: secuencias positivas y secuencias negativas.

Para la obtención de las secuencias positivas utilizamos el dataframe creado anteriormente y lo convertimos en un archivo en formato “fasta”, donde como identificador tendrá el ID de la secuencia, así como su posición en el genoma del virus y la secuencia será el “match” encontrado por FIMO.

Para la creación de las secuencias negativas utilizamos el identificador de las secuencias del dataframe y generamos una secuencia de 16 nucleótidos de extensión para motivo de P1 y 30 para el motivo de P3, con una posición al azar en el genoma de cada secuencia.

Lo que se busca es tener un set de secuencias que no estén relacionadas con las secuencias positivas, de este modo poder entrenar el algoritmo teniendo como clase positiva el set de secuencias positivas y como clase negativa el set de secuencias negativas. De la misma forma que las secuencias positivas, recuperaremos las secuencias negativas en formato “fasta”.

Generamos el set de secuencias positivas y negativas tanto para el dataframe de prueba como para el de entrenamiento.

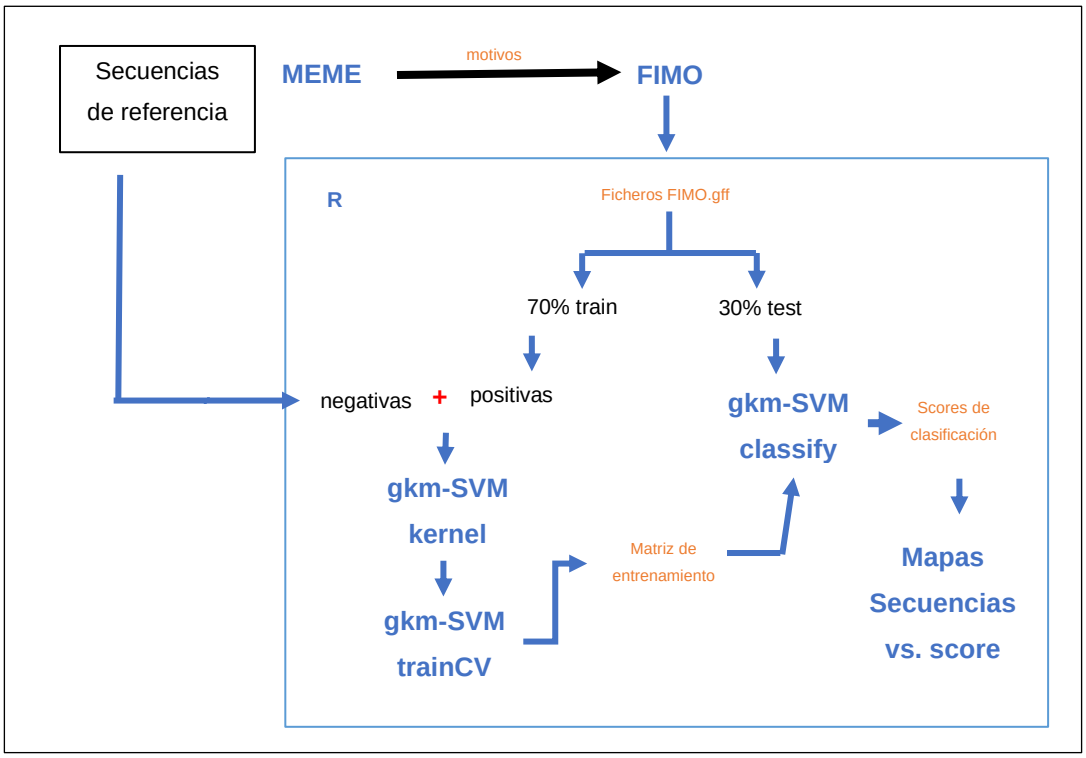

Figura 6: Diagrama del pipeline seguido para la búsqueda de zonas de slippage.

2.3.6. Aplicación de gkm-SVM

Utilizamos la función "**gkmsvm_kernel**" de la librería de R gkm-SVM (33) para generar nuestro kernel de entrenamiento con las secuencias positivas y negativas correspondiente.

Un vez obtenido el kernel utilizamos la función "**gkmsvm_trainCV**" para realizar el entrenamiento de las secuencias positivas y negativas con 5-fold Cross Validation. En el caso de la evaluación de P3 teníamos kernels mucho más pesados de modo que no utilizamos 5-fold, sino que dejamos la CV default (2-fold CV) del algoritmo.

Del entrenamiento del algoritmo hemos obtenido la matriz de clasificación que utilizaremos para testar nuestro set de evaluación, para ello utilizamos la función "gkmsvm_classify". Este algoritmo utiliza como parámetros "default" un valor de k=6 y L=10. De este algoritmo obtenemos el "score de clasificación" para cada secuencia.

Además, creamos secuencias en mosaico mediante sliding windows (35) de 10-mers (de igual tamaño que el parámetro L) de las secuencias mejor y peor calificadas, con el fin de tener un trazado visual de las secuencias *versus* su score obtenido. Los mapas de estas secuencias los generamos mediante la librería de R "ggplot2".

En la **Figura 6** nos podemos hacer una idea gráfica del trabajo descrito en este apartado.

3. Resultados

3.1. Búsqueda de uORF

En la búsqueda mediante ORFfinder se obtuvo que todas las especies presentan un gran número de ORF en sus secuencias. Sin embargo, mediante las consultas realizadas en R obtenemos que solo presentan uORFs tres especies, dos de ellas los *Ipomovirus* CVYV y SqVYV y el *Potyvirus* PPV. En la **Tabla 6** podemos ver un resumen de los datos obtenidos en las búsquedas con ORFfinder.

Tabla 6: Tabla resumen de la búsqueda de uORFs.

Género	Especies	*Nº secuencias*	Nº ORF	Nº uORF
Ipomovirus	*Cassava brown streak virus* (CBSV)	12	624	0
Ipomovirus	*Cucumber vein yellowing virus* (CVYV)	4	274	3
Ipomovirus	*Squash vein yellowing virus* (SqVYV)	5	269	5
Potyvirus	*Plum pox virus* (PPV)	423	21227	22
Potyvirus	*Tobacco etch virus* (TEV)	9	456	0
Ipomovirus	*Tomato mild mottle virus* (TMMV)	1	60	0

Para CVYV encontramos que presenta 2 uORF (**Tabla 7**) en la posición 14 del genoma de sus 4 cepas, ya que **NC_006941 y AY578085** pertenecen a la misma cepa (36).

Tabla 7: uORFs para CVYV.

ID	Name	ORF	start	stop	sequence
NC_006941	ORF56_NC_006941	ORF56	14	124	MKNRSHTSATSKHNGRSRSNR
KT276369	ORF64_KT276369	ORF64	14	46	MNYTDR
AY578085	ORF56_AY578085	ORF56	14	124	MKNRSHTSATSKHNGRSRSNR

En el caso de PPV, partíamos de un gran número de cepas, de modo que hemos encontrado muchos ORF en total. Sin embargo, solo 22 de ellos son uORFs. En la **Tabla 8** mostraremos solo 5 de ellos.

Tabla 8: uORFs para PPV.

ID	Name	ORF	start	stop	sequence
AY953263	ORF10_AY953263	ORF10	10	114	MRSNSAKSRSKNKNSN
LC375132	ORF56_LC375132	ORF56	11	61	MRTNSAKS
LC375129	ORF53_LC375129	ORF53	11	115	MRSNSAKSRSKNKNSSN
LC374992	ORF9_LC374992	ORF9	11	91	MRSNSAKSRSKSTNK
AB576055	ORF50_AB576055	ORF50	11	115	MRSNSAKSRSKNKNSN

Para SqVYV encontramos 4 uORF (**Tabla 9**) en la posición 14, en este caso tenemos que NC_010521 y EU259611 pertenecen a la misma cepa (11).

Tabla 9: uORFs para SqVYV.

ID	Name	ORF	start	stop	sequence
NC_010521	ORF15_NC_010521	ORF15	14	82	MNHNDKSNHCHSY
KT721735	ORF16_KT721735	ORF16	14	106	MKHNDKSANCYVHTTS
KU953950	ORF54_KU953950	ORF54	14	70	MNHNDSHYTKS
JF897996	ORF15_JF897996	ORF15	14	82	MNHNDKSNHCHSY
EU259611	ORF15_EU259611	ORF15	14	82	MNHNDKSNHCHSY

Como hemos visto en las tablas anteriores, en el caso de los *Ipomovirus* que presentan uORFs, estos son pequeños péptidos de unos 15-20 aminoácidos que no superponen la poliproteína viral (**Figura 7**), en el caso de PVV encontramos tanto uORFs pequeños como de considerable tamaño.

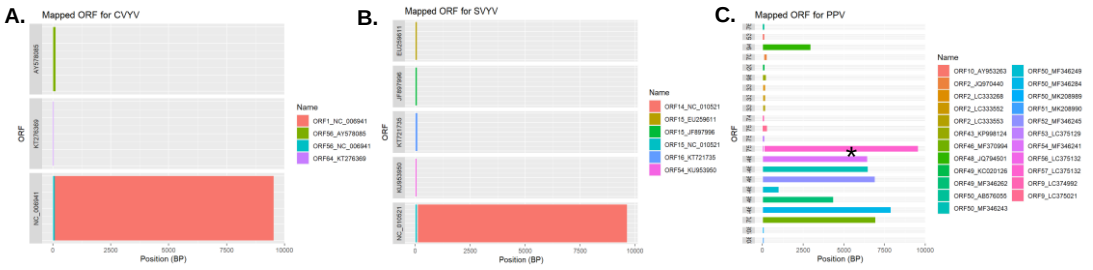

Figura 7: Mapa con la ubicación de los uORFs para CVYV, SqVYV y PPV.

Los mapas se han realizado en R usando "ggplot2" donde se señala la posición de los uORFS respecto al tamaño de la poliproteína: **(A)** Mapa de CVYV, la poliproteína está representada en color rojo y los uORF se sitúan al inicio de la secuencia, sin superponerse. **(B)** Mapa para SqVYV, la poliproteína está representada en rojo y los uORFs se encuentran al inicio de las secuencias, de la misma manera que para CVYV. **(C)** Mapa para PPV, en este caso la poliproteína (*) es la que presenta el mayor tamaño, pero podemos apreciar que esta virus presenta uORFs con tamaños muy variables.

Hemos realizado un alineamiento de estos uORFs (**Figura 8**) para observar si sus secuencias se encuentran conservadas. Hemos descartado el alineamiento para PPV, ya que como pudimos ver en la gráfica anterior, presenta uORFs muy diversos. En el caso de los *Ipomovirus* vemos que no hay una conservación clara entre sus secuencias. Tenemos además muy pocas especies en estudio, como es el caso de CVYV que solo podemos comparar dos uORFs de 4 secuencias en total, sin embargo, estos dos uORFs son completamente diferentes. En el caso de SqVYV sus uORFs parecen estar mejor conservados al inicio de estos, sin embargo, no tenemos suficiente cantidad de secuencias a estudio como para poder sacar conclusiones.

A.

```
MUSCLE 3.8.31

Call:
   msa(alg_cvyv, "Muscle")

MsaAAMultipleAlignment with 3 rows and 21
    aln                   names
[1] MKNRSHTSATSKHNGRSRSNR ORF56_NC_006941
[2] MKNRSHTSATSKHNGRSRSNR ORF56_AY578085
[3] ---------------MNYTDR ORF64_KT276369
Con MKNRSHTSATSKHNGRSRSNR Consensus
```

B.

```
ClustalOmega 1.2.0

Call:
   msa(alg_svyv, "ClustalOmega")

MsaAAMultipleAlignment with 5 rows and 16
    aln               names
[1] MNHNDSHYTKS----- ORF54_KU953950
[2] MNHNDKSNHCHSY--- ORF15_NC_010521
[3] MNHNDKSNHCHSY--- ORF15_JF897996
[4] MNHNDKSNHCHSY--- ORF15_EU259611
[5] MKHNDKSANCYVHTTS ORF16_KT721735
Con MNHNDKSNHCHSY--- Consensus
```

Figura 8: Salida de R de alineamientos realizados con "msa".

(A) Alineamiento para CVYV, alineamiento realizado mediante “Muscle”, se observa que no existe conservación entre sus dos uORFs. **(B)** Alineamiento para SqVYV, alineamiento realizado mediante “ClustalOmega”, en el caso vemos que existe una cierta conservación entre sus uORFs.

Finalmente, hemos realizado un breve estudio de filogenia sobre las proteínas de la cápside de todas las especies de *Ipomovirus* (*Cassava brown streak virus, Cumcumber vein yellowing virus, Squash vein yellowing virus, Sweet potato mild mottler virus, Tomato mild mottle virus, Uganda cassava Brown streak virus*) y además hemos añadido las especies de *Potyvirus* con las que hemos realizado el análisis (*Plum pox virus* y *Tobacco etch virus*). Hemos escogido esta proteína ya que es la que más conservada entre la familia *Potyviridae* (37) tratando de que nos sirva de sesgo para comprobar si existen similitudes previas entre las especies que nos ayuden a entender la aparición de uORFs en algunas de las especies y en otras no.

El árbol filogenético nos muestra que los *Ipomovirus* se han agrupados en dos clusters diferentes **(Figura 9).** Un grupo lo forman las especies CBSV, CVYV y UCVSV y el otro clúster está dividido en tres ramas donde una de ellas las forma los *Potyvirus* que se han situado contiguos formando un rama junto a TMMV y SPFMV. Finalmente, en la

última rama de este clúster tenemos a SqVYV, de este modo no observamos un patrón de agrupación relacionado entre la cápside vírica y la presencia/ausencia de uORFs.

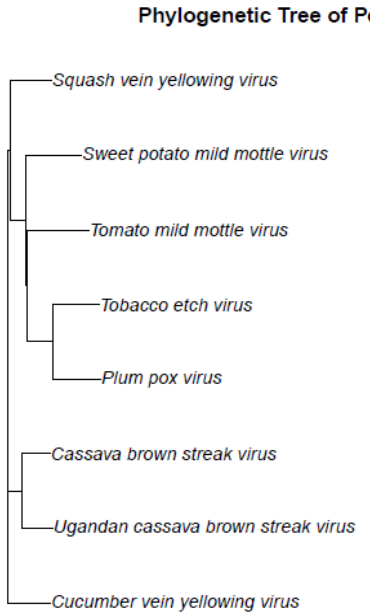

Figura 9: Árbol filogenético de las especies de Ipomovirus y dos de Potyvirus.

Árbol realizado mediante un alineamiento de las secuencias con ClustalOmega (librería Biconductor "msa") y resuelto mediante *Neighbor-joining (librería de R "ape"*(38)*).*

3.2 Zonas de deslizamiento de la polimerasa

Hemos realizados dos tipos de búsquedas de zonas de deslizamiento de la polimerasa. La primera está basada en el motivo de deslizamiento de la polimerasa en P1 en las secuencias de SPFMV. A partir de este motivo hemos realizado búsquedas del mismo con FIMO. FIMO ha encontrado al menos dos ocurrencias en cada secuencia consultada. Lo esperado en esta búsqueda, es que al menos hubiera una ocurrencia en cada secuencia, ya que el motivo está basado en secuencias del producto P1 únicamente.

Adicionalmente, hemos realizado una segunda búsqueda de patrón de deslizamiento G_2A_6 en las ocurrencias encontradas por FIMO. De esta segunda búsqueda de patrón encontramos que de las ocurrencias encontradas en FIMO para SPFMV solo 55 de ellas presentan el patrón de deslizamiento real. Este resultado es congruente ya que el motivo para P1 realizado con MEME lo obtuvimos con 55 secuencias de SPFMV.

En el caso de la búsqueda con FIMO del motivo de P1 para todos los virus huéspedes de batata, hemos obtenido de la segunda búsqueda de patrón que el 23% de las ocurrencias encontradas por FIMO realmente presentan el patrón buscado. En el caso

de secuencias de miembros del género *Potyvirus* solo el 6% de las ocurrencias encontradas por FIMO presenta el patrón real de deslizamiento.

Tabla 10: Resumen de FIMO para el motivo en P1.

Motivo	Secuencias utilizadas	Ocurrencias FIMO	Patrón real G2A6(P1)
P1	Cepas SPFMV (55)	130	55
P1	Todos los *Potyvirus* huéspedes de batata (112)	285	81
P1	Todas las cepas del g. *Potyvirus* (2351)	5951	356

El segundo caso de búsqueda de patrón de deslizamiento de la polimerasa, lo realizamos con un motivo ubicado en P3. Para la búsqueda de este motivo utilizamos todas las secuencias encontradas para miembros del género *Potyvirus*. El motivo encontrado no presenta la primera "G" del motivo G_2A_6 en todos sus miembros, de modo que el motivo es menos restrictivo que el anterior ya que permite variaciones a AGA_6 o TGA_6.

En la búsqueda realizada con FIMO encontramos que existen más de 4 ocurrencias del motivo en cada secuencia. Este resultado se nos hace complicado de aceptar, ya que esperamos que haya al menos una ocurrencia en cada secuencia, en la zona de P3 y quizás otra en P1.

Tabla 11: Resumen de FIMO para el motivo en P3.

Motivo	Secuencias utilizadas	Ocurrencias FIMO	Patrón real GA6(P3)
P3	Todos los *Potyvirus* huéspedes de batata (112)	624	238
P3	Todas las cepas del g. *Potyvirus* (2351)	11220	2452
P3	Todas las cepas de la fam. *Potyviridae* (2739)	12504	2634

En el caso del motivo P3 (**Figura 5**), este es mucho más largo que el motivo de P1 de modo que admite variaciones más amplias en la búsqueda de "matches" ya que presenta pocos nucleótidos con un 100% de frecuencias en una posición determinada . Realizando una segunda búsqueda del patrón de deslizamiento en los ocurrencias obtenidas por FIMO, vemos que para el caso de las secuencias de virus huéspedes de batata, encontramos que el 38% de las ocurrencias presentan el motivo real. En el caso de los miembros del género *Potyvirus* solo el 21,8% de las ocurrencias de FIMO presentan el patrón real y para los miembros de la familia *Potyviridae* este porcentaje es de muy parecido, concretamente es del 21%.

3.2.1 Resultados de FIMO para P1 en secuencias de SPFMV.

Hemos comprobado la efectividad de clasificación de FIMO comparándola con el algoritmo gkm-SVM. En la **Figura 10-A** vemos que la distribución de q-values de las ocurrencias obtenidas por FIMO, distinguimos en esta distribución presenta dos tendencias, ya que existen gran número de secuencias que presentan valores q-value cercanos a cero, por tanto, significativos, como también vemos que existe un número considerable de falsos positivos (altos q-value).

Sin embargo, si miramos los puntajes de clasificación obtenidos mediante el algoritmo gkm-SVM a partir de las secuencias positivas de evaluación, notamos que la mayoría de estas secuencias presenta "scores" muy altos (**Fig. 10-B**), es decir, en la mayoría de ellas el clasificador ha discriminado correctamente, esto se debe a que el algoritmo de entrenamiento presenta un valor de AUC máximo (**Fig. 10-C**). Este valor nos dice que el clasificador es capaz de discriminar completamente las secuencias negativas de las positivas.

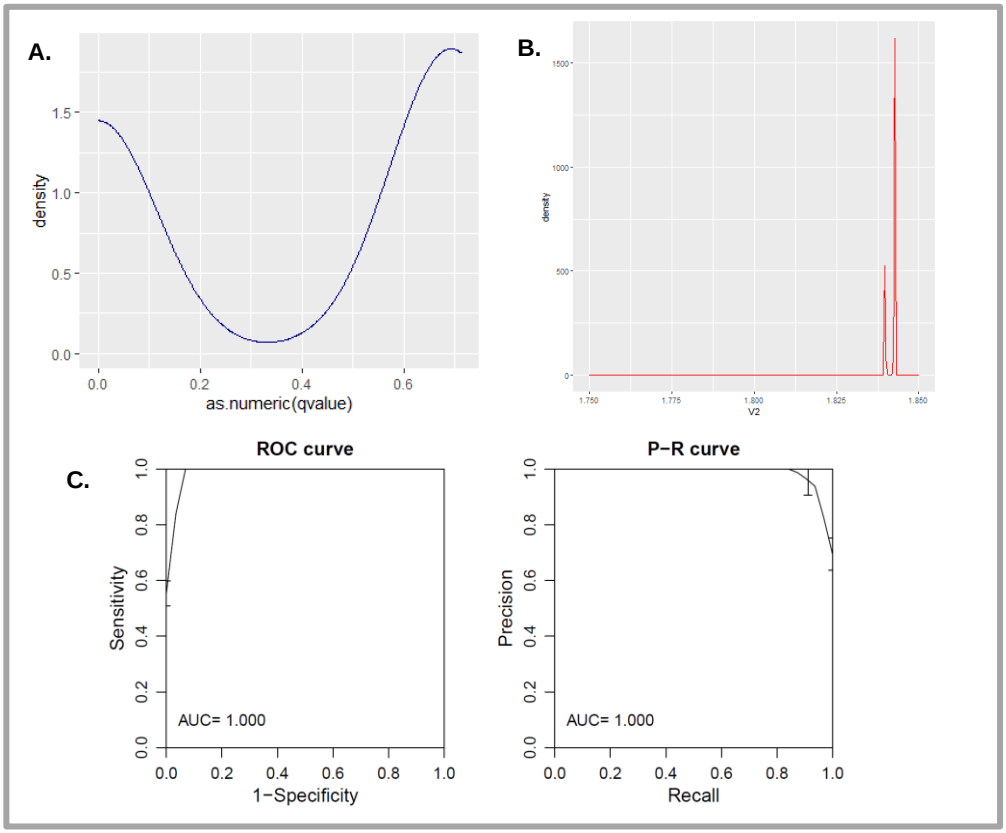

Figura 10: Análisis del motivo P1 con FIMO y gkm-SVM.

(A) Distribución de los q-values para las ocurrencias encontradas por FIMO para el motivo de P1 en las secuencias de SPFMV, gráfica realizada con "ggplot2". **(B)** Gráfica de la distribución de los puntajes de las secuencias positivas evaluadas mediante el algoritmo de clasificación "gkm-SVM", realizada con "ggplot2". **(C)** Curvas ROC y P-R obtenidas del algoritmo de entrenamiento "gkm-SVM" 5-fold CV.

Hemos realizado un mapeo de cada cambio de nucleótidos mediante el algoritmo sliding Windows, en secuencias de 10-mers para una cepa del virus SPFMV primero en la región de 1-2000 bp y luego de 7000-9000 bp, y hemos aplicado el algoritmo de clasificación en las dos secuencias, como resultado hemos obtenido un mapa (**Figura 11**) de del recorrido de las secuencias versus el puntaje obtenido.

En el plot obtenido vemos que obtenemos un puntaje máximo en una zona próxima a los 1300bps, esta zona coincide con el lugar donde se produce el deslizamiento en este virus, en el gráfico de debajo vemos que también hay un pico importante en la zona de 8600 bps, pero el puntaje que obtiene este pico es menor que el de clasificación (<1.8), con lo cual hablaríamos una zona neutral donde pueden a ver nucleótidos parecidos al motivo.

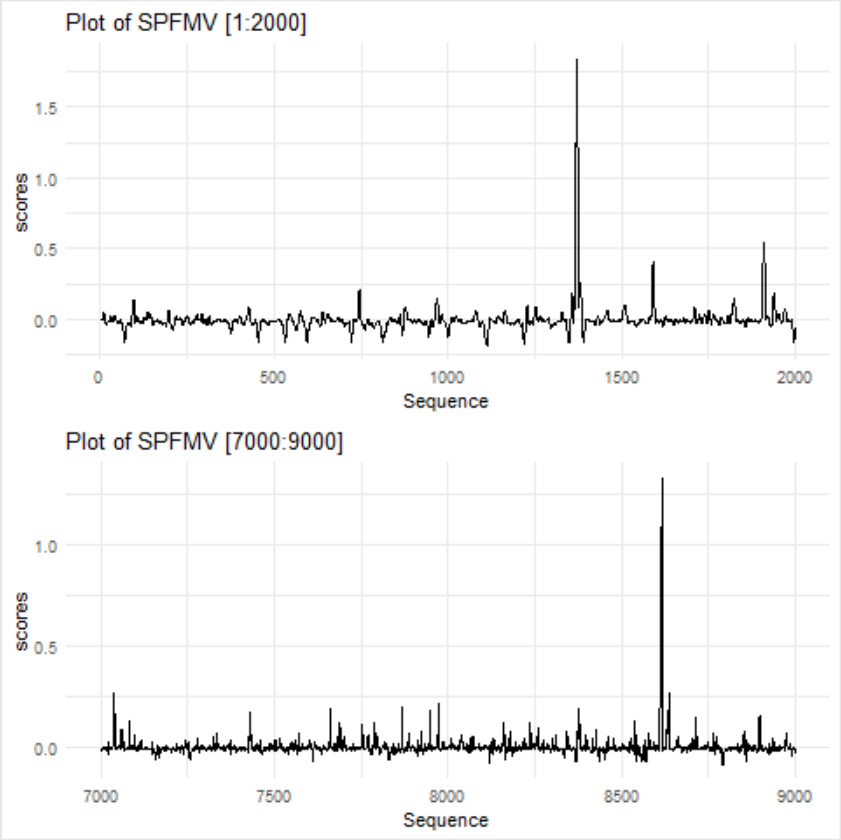

Figura 11: Secuencia de SPFMV en mosaico vs. score obtenido con gkm-SVM.

Se han obtenido secuencia en mosaico de 10-mers que muestran el cambio de un nucleótido. El pico que presenta un mayor puntaje el que se encuentra en la coordenadas de frecuente ubicación del motivo de P1.

3.2.1 Resultados de FIMO para P1 y P3 en secuencias de virus huéspedes de batata.

Los resultado de la búsqueda del motivo en estas secuencias también nos marcan un tendencia bimodal en la distribución de q-values (**Figura 12-A**) de las ocurrencia de FIMO, de modo que podemos esperar que en este caso también tengamos gran cantidad de falsos positivos. Mediante el algoritmo gkm-svm podemos observar un resultado similar, ya que pocas secuencias presentan de evaluación presentan realmente puntajes altos (**Figura 12-B**), por tanto, significativos, para la clasificación, de modo que tenemos muchas secuencias con ocurrencias parecidas al motivos, pero idénticas.

Para el motivo de P3 en estas secuencias observamos que la gráfica de distribución de los q-values de fimo presenta q-values menores que en el caso de P1, pero al igual que en P1 son menos las ocurrencias que presentan q-values cercanos a cero. Con el puntaje de las secuencias de evaluación obtenidas con el algoritmo de clasificación gkm-SVM nos ocurre lo mismo, tenemos muchas secuencias con puntajes neutrales y muy pocas con puntajes altos similares al motivo.

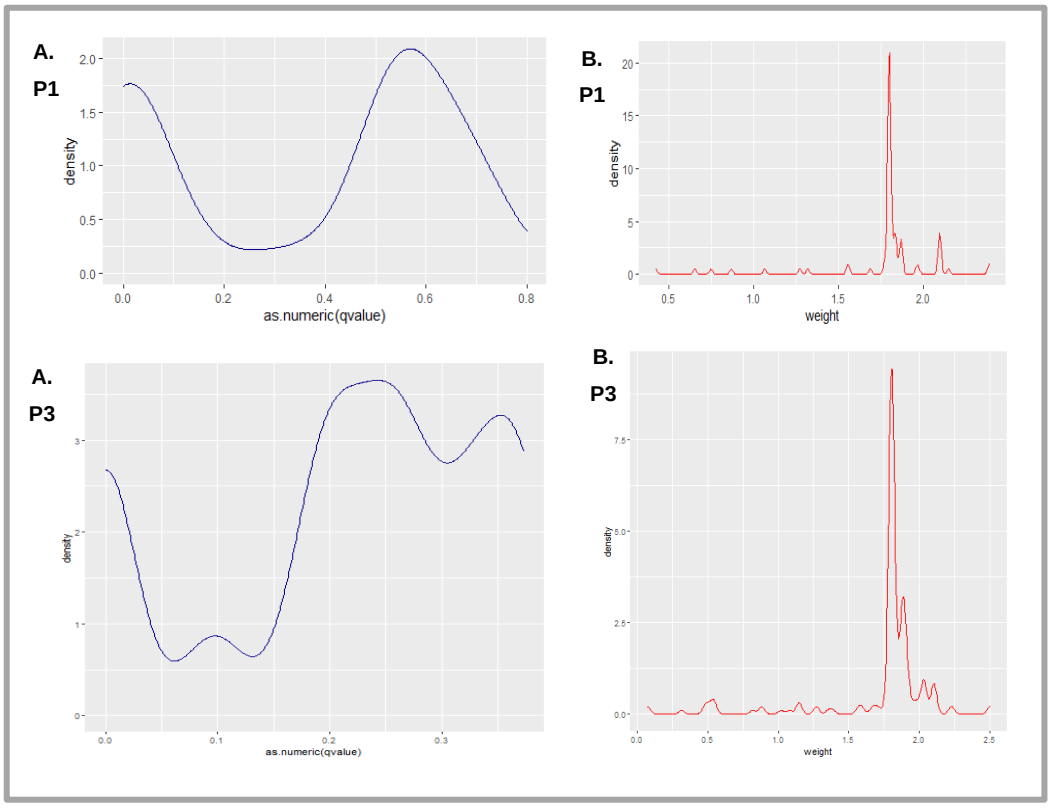

Figura 12: Distribución de q-values y puntajes de clasificación para ocurrencias de FIMO en el motivo P1 y P3 en secuencias de virus huéspedes de batata.

(A) Diagrama de distribución de q-value de las ocurrencias obtenidas por FIMO de los motivos en secuencia de virus huéspedes de batata. **(B)** Diagrama de distribución de los puntajes obtenidos del algoritmo de clasificación gkm-SVM para la ocurrencias encontradas por FIMO en secuencias de virus huéspedes de batata. Ambos gráficos se realizaron en “ggplot”.

En ambos casos, el algoritmo de entrenamiento para P1 y P3 nos da un AUC alta (**Figura 13**). En el caso de P1 AUC=0.994 y en P3 AUC=0.973, de modo que el algoritmo es un buen clasificador para nuestras secuencias.

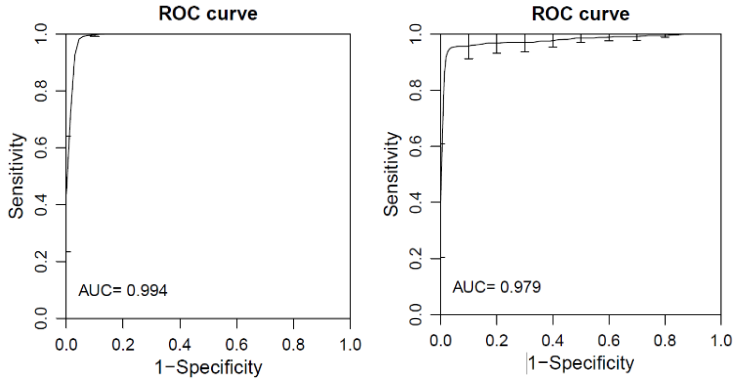

Figura 13: Curvas ROC de gkm-SVM en virus huéspedes de batata.

Derecha: curva ROC para el motivo P1, valor AUC= 0.994. **Izquierda**: Curva ROC para el motivo P3, valor AUC= 0.979. Ambos algoritmos entrenados con 5-fold CV.

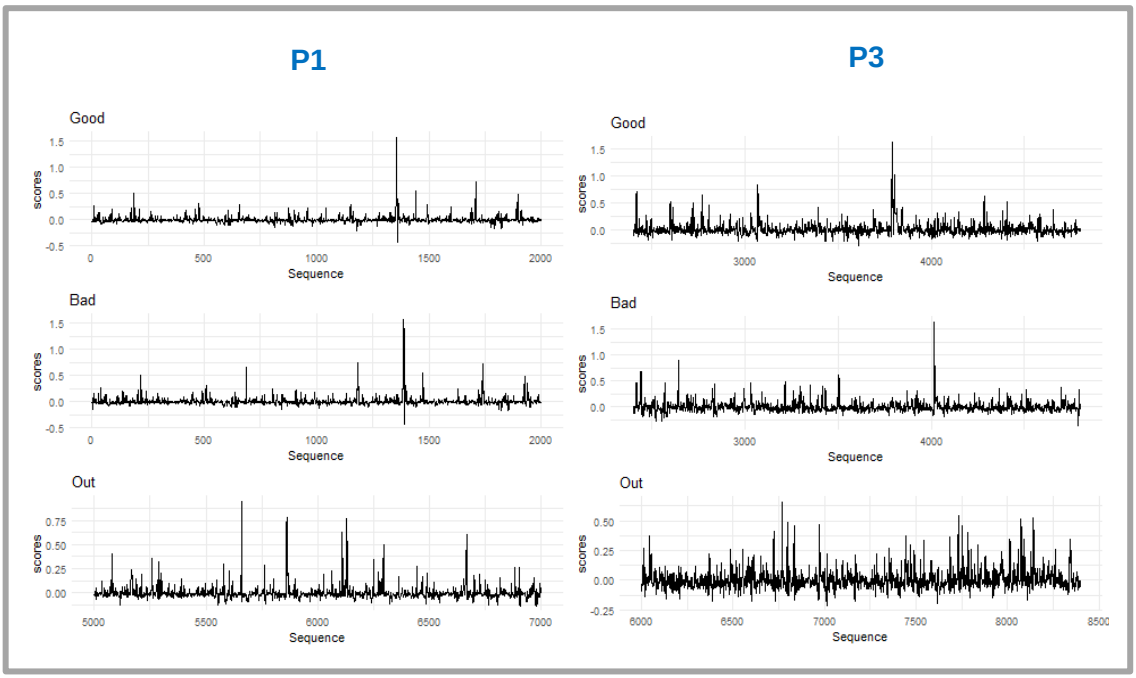

Figura 14: Secuencias en mosaico calificadas con gkm-SVM entrenado con secuencias de virus huéspedes de batata.

El mapa muestra un cambio de nucleótidos a partir de secuencias de 10-mers, los picos de máximo puntaje corresponden a zonas donde se encuentra el motivo.

Hemos realizado la predicción de secuencias en mosaico de 10-mers. De la misma forma que el ejemplo anterior, mediante el algoritmo sliding windows, para los motivos P1 y P3. Para la realización de estas secuencias en mosaico extrajimos la secuencia

que presentó mayor y el peor puntaje en predicción de las secuencias de evaluación y también generamos una secuencia el mosaico en una zona fuera del lugar donde se encuentran estos motivos de deslizamiento. El resultado obtenido es para ambos motivos un pico de máximo puntaje en regiones relacionadas con la ubicación del motivo (**Figura 14**), tanto en las secuencias con buen y mal puntaje. En el caso de los secuencias que se encuentran fuera de la zona de localización de los motivos, vemos que los picos obtenidos presentan puntajes poco representativos para el algoritmo, de modo que son zonas que no presentan el motivo.

3.2.1 Resultados de FIMO para P1 y P3 en secuencias de miembros del género *Potyvirus*.

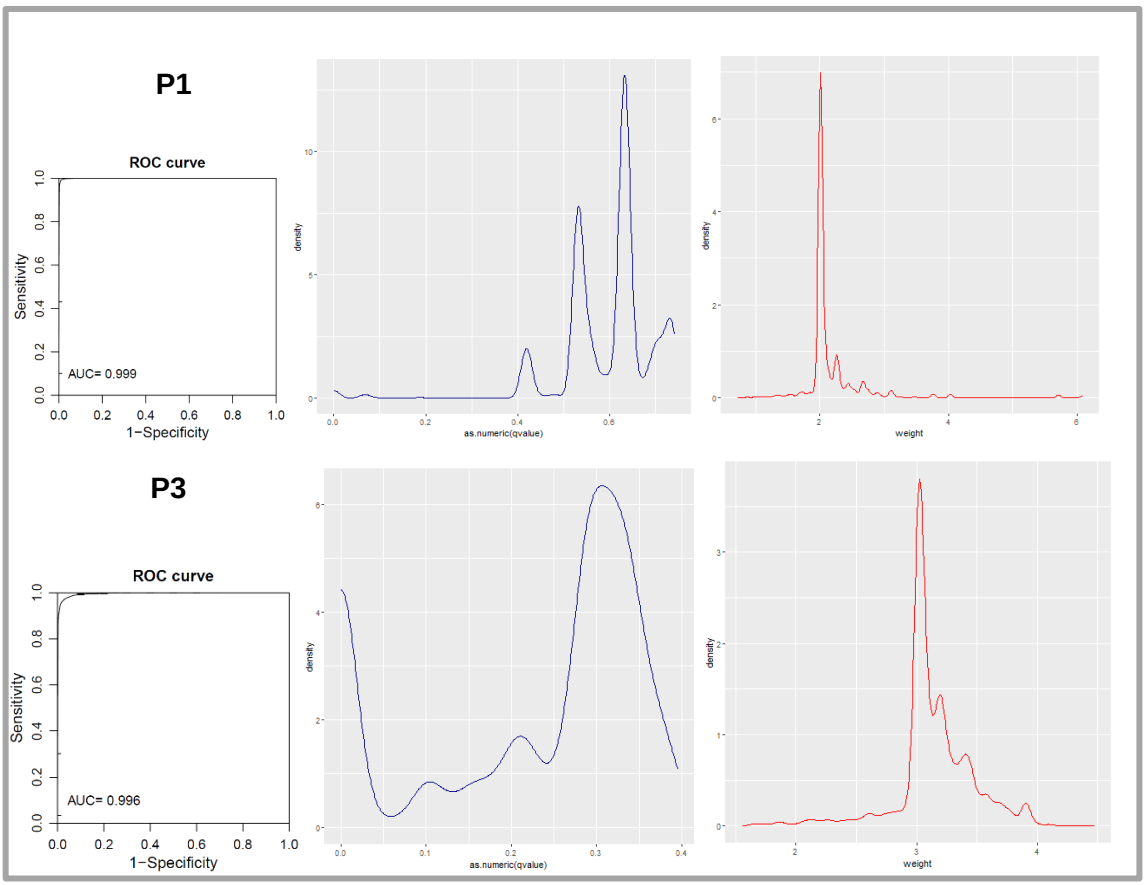

Figura 15: Análisis de ocurrencias de FIMO en secuencias del género Potyvirus.

En **azul,** distribución de los q-values para las ocurrencias encontradas de los motivos P1 y P3 con FIMO de secuencias del género *Potyvirus*. En **rojo**, los puntajes de clasificación de las secuencias de evaluación del algoritmo entrenado con gkm-SVM. Los valores AUC del algoritmo de entrenamiento para ambos motivos son mayores a 0.99.

Como en los casos anteriores, en la **Figura 15** vemos que las ocurrencias obtenidas para los miembros del género *Potyvirus* de los motivos P1 y P3 presentan un gran número de falsos negativos. Por otro lado, vemos como a medida que el número de

secuencias va aumentando, el algoritmo gkm-SVM presentan más sensibilidad de discriminación de forma que obtenemos más número de secuencias positivas de evaluación con puntajes neutros y escasas secuencias con puntajes altos y significativos.

Entre ambos motivos, podemos ver que P3 presenta mejores resultados que P1 ya que presenta q-value menores y además parece tener ligeramente más secuencias con puntajes altos. Este hecho puede deberse a que el motivo de P3 es menos restrictivo que el de P1. Además, el motivo de P1 está basado en la zona de deslizamiento de la polimerasa ocurrida en SPFMV para la generación en trans del producto P1N-PISPO, este deslizamiento ocurre como estrategia de algunos virus huéspedes de batata. Con lo cual la búsqueda del motivo de P1 en todos los *Potyvirus* es menos eficaz que la búsqueda del motivo de P3, ya que el deslizamiento de la polimerasa el P3 ocurre la mayoría de las especies de familia *Potyviridae*.

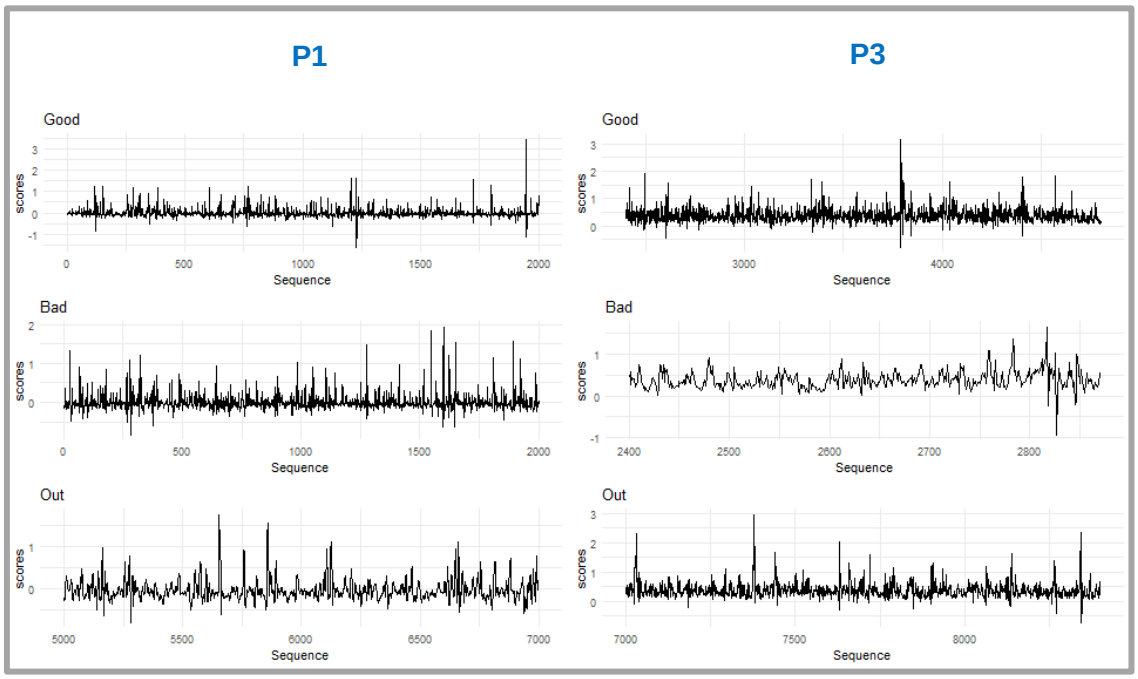

Figura 16: Secuencias en mosaico calificadas con gkm-SVM entrenado con secuencias de Potyvirus.

Secuencias en mosaico realizadas con 10-mers que nos muestran el cambio de un nucleótido. Vemos que para los dos motivos existe un pico de puntaje máximo en las regiones frecuentes del motivo. Para el motivo de P3 encontramos que existe un pico con un puntaje aceptable en una zona fuera de la región del motivo, puede deberse a lo poco restrictivo del motivo.

En las gráficas de las secuencias en mosaico **(Figura 16)** de las cepas con mejor y peor puntaje, como la zonas fuera del lugar de localización del algoritmo, vemos que para ambos motivos se observan picos de alto puntaje en lugares relacionados con la

ubicación del motivo, en el caso de la secuencia peor calificada vemos que los picos que presentan son de bajos puntajes en los dos motivos.

Finalmente, en el caso de los puntajes en zonas fuera del motivo vemos para el motivo de P1 existen picos de bajos puntajes, mientras que para P3 observamos picos con puntajes neutrales.

3.2.1 Resultados de FIMO para P3 en secuencias de miembros de la familia *Potyviridae*.

Como en los casos anteriores, en esta búsqueda también encontramos altos valores de falsos negativos para las búsqueda de FIMO, lo que también se reafirma en los puntajes obtenidos con gkm-SVM ya que muy pocas secuencias de evaluación presentaron puntajes altos y muchas de ellas presentan puntajes neutrales (**Figura 14**). Como en los demás casos hemos obtenido un valor de AUC muy alto para el algoritmo de entrenamiento.

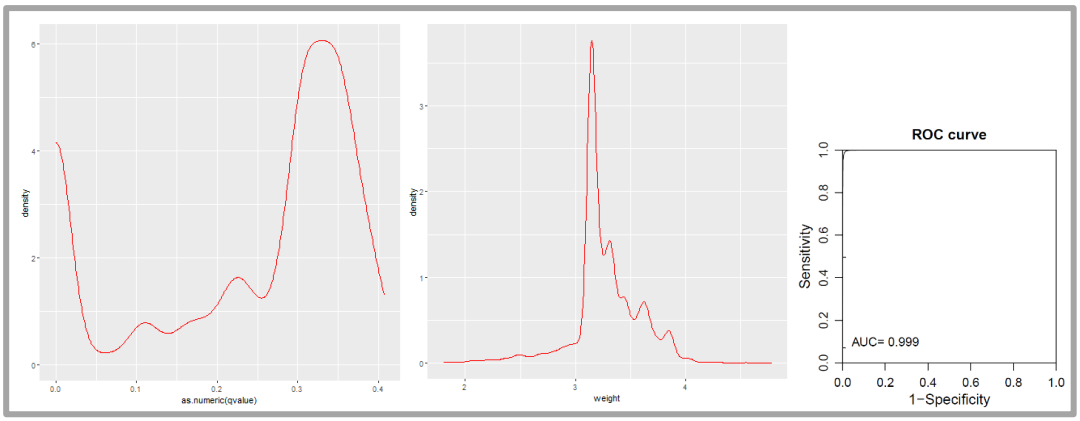

Figura 17: Análisis FIMO/gkm-SVM en secuencias de especies de la Familia Potyviridae.

A la **derecha** vemos la distribución de las q-values y a la **izquierda** la distribución de los puntajes de las secuencias de la Familia *Potyviridae* generadas por FIMO y analizadas mediante el algoritmo gkm-SVM, el algoritmo de entrenamiento obtenido mediante gkm-SVM nos da un valor de AUC=0.999.

Finalmente, del análisis de las secuencias en mosaico hemos obtenido que para la secuencia mejor calificada presenta un pico de alto puntaje en una zona relacionada con el lugar de deslizamiento que representa el motivo P3. Tanto la secuencia peor calificada como la que se encuentra fuera de la zona del motivo presenta puntajes bajos poco significativos (**Figura 18**) .

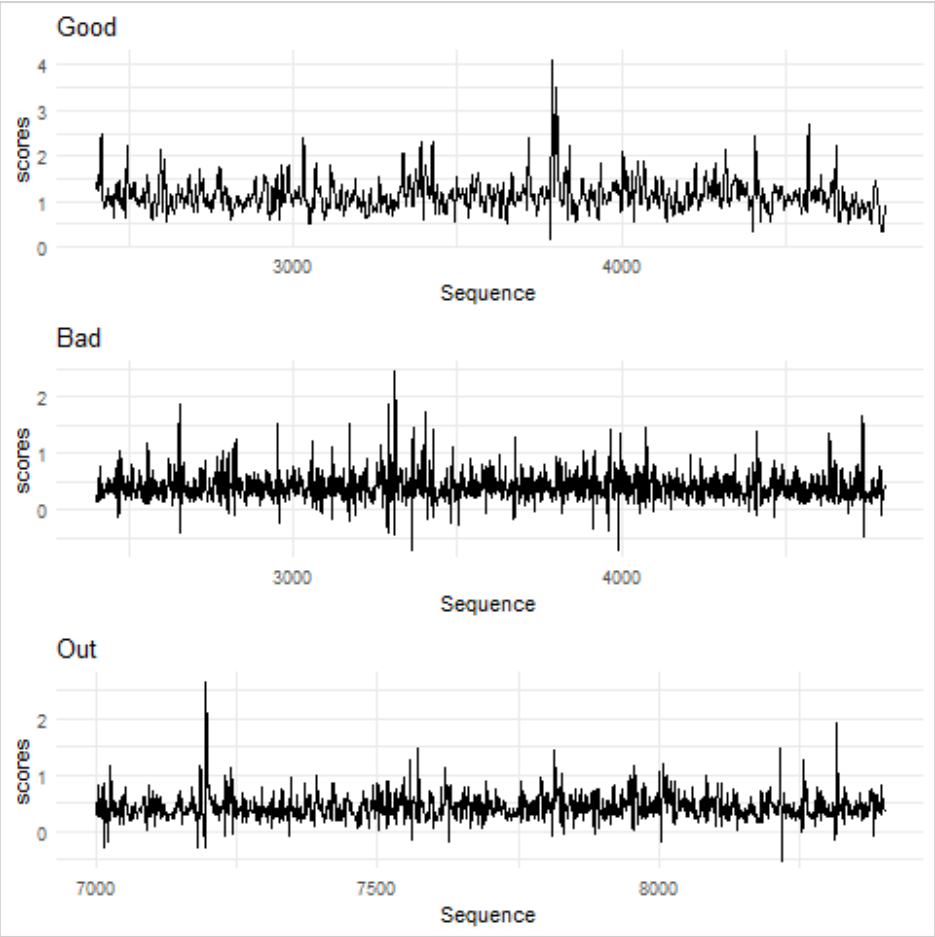

Figura 18: Secuencia en mosaico vs. Scores del algoritmo de clasificación entrenado con secuencias de especies de la Familia Potyviridae.

Se observan un mapa de secuencias de 10-mers que presentan un cambio de nucleótido. Podemos ver que las secuencias mejores calificados por el algoritmo presentan un pico de puntaje mayor a las secuencias mal calificadas o las que están fuera de la zona del motivo.

4. Discusión

Para el trabajo realizado en la búsqueda de ORFs hemos encontrado ciertas limitaciones entre la realización de un protocolo que puede ser compatible con nuestro conocimiento y con el tiempo que disponíamos para realizar el trabajo. De forma que decidimos abordar la exploración de uORFs con herramientas sencillas como ORFfinder de NCBI.

Lo que se pretendía es conocer la existencia de uORFs en las secuencias del género *Ipomovirus*, a causa de la introducción de 17 nucleótidos en la región 5'-UTR de las secuencias del virus CVYV, ya que en esta mini secuencia faltante en 5' se encuentra un ATG en la posición 14, que podría iniciar una nueva pauta de lectura.

Una de las principales limitaciones que encontramos mientras realizábamos el trabajo, fue que existen muy pocas secuencias de este virus anotadas, de modo que disponíamos de muy pocas fuentes para realizar una comparación clara entre los uORFs encontrados para este virus.

Con relación a los resultados encontrados en CVYV, nos llamó la atención que a pesar de que todas las secuencias presentan el "ATG", los uORFs encontrados no fuesen homogéneos ni todas sus cepas lo presentasen. Esto nos da una pista de la gran variabilidad que presenta la zona 5-UTR en estos virus. De momento, en el laboratorio de Virología del CRAG, quedan pendientes cuestiones por resolver acerca de la presencia del "ATG" en la posición 14. Una de las conjeturas que se pretende abordar es que responda a una necesidad estructural del virus.

Hemos encontrado también, que *Squash vein yellowing virus* presenta uORFs con una conservación media entre sus péptidos. En el caso de *Plum pox virus*, presentan uORFs muy largos en algunas de sus cepas, lo que podría generar una confusión en la polimerasa a la hora de la traducción del virus. Esto quizás puede tener un explicación evolutiva que aún no conocemos y puede llevar a crear una nueva diana de investigación.

En relación con el trabajo realizado sobre las zonas de deslizamiento de la polimerasa, encontramos que existen muchas herramientas predictoras de zonas de variación génica, pero muchas de ellas están enfocadas en organismos con genomas grandes. Normalmente muchas herramientas están diseñadas para el análisis del genoma humano. Muchas otras han sido adaptadas por los investigadores según sus necesidades, pero la información acerca de estas herramientas (manuales, tutoriales, etc.) no está disponible. En otros casos la limitación fue que no disponemos de los conocimientos necesarios para llevar a cabo alguno de los protocolos.

Los *Potyviridae*, son virus pequeños, de 10Kb aproximadamente, además, géneros como *Potyvirus* abarcan la totalidad de especies de esta familia. Muchos de estos virus se encuentran bien anotados y cuentan con muchas especies depositadas en las bases de datos, sin embargo, otras especies carecen de anotación.

Hasta ahora la anotación de genomas virales se hace de forma manual por parte de investigadores especializados, y en ocasiones esto retrasa la disponibilidad de la información por saturación (demasiadas nuevas secuencias obtenidas y depositadas en las bases de datos). Un ejemplo de ello es que pocas especies abarcan la totalidad de secuencias anotadas como es el caso de *Potato virus Y* o *Plum pox virus*. Que son especies ampliamente estudiadas.

Sobre la búsqueda de zonas de deslizamiento de la polimerasa, lo que se pretendía es encontrar motivos que nos ayuden a caracterizar bien estas zonas o verificar si existen zonas potenciales de frameshifting en las secuencias de Potyvirids. Con relación a ello encontramos que el descubrimiento, búsqueda y análisis de motivos de secuencia es un protocolo óptimo para abordar la búsqueda de zonas de deslizamiento en virus *Potyviridae*. De este modo establecimos un protocolo sencillo de llevar a cabo mediante la utilización de MEME/FIMO. Además, probamos la capacidad de discriminación de FIMO analizando sus resultados mediante un algoritmo de clasificación bajo supervisión como es gkm-SVM.

Otra estrategia que podríamos haber abordado en relación con estas dos herramientas hubiese sido la utilización del algoritmo para encontrar zonas de posible deslizamiento, entrenando el algoritmo con fragmentos de tamaño determinado de posibles zonas de slippage, a modo que las zonas que presenten mayor puntaje nos sirvan para encontrar un motivo de secuencia mejor definido. El problema de esta estrategia es que el costo computacional para la creación del kernel para entrenar las secuencias es muy pesado y largo, de modo que nos hace más complicado la búsqueda de patrones por ensayo y error.

Encontramos que los resultados de FIMO nos dan secuencias aproximadas con el motivo buscado, pero no exactas, con lo cual tenemos gran cantidad de secuencias que el algoritmo gkm-SVM clasifica como “neutrales” para el motivo y por consiguiente una gran cantidad de falsos negativos. Por otro lado, los resultados de FIMO son compatibles con motivos reconocidos como zonas de deslizamiento de la polimerasa, de modo que los resultados calificarían como “posibles zonas de deslizamiento”.

Hay una amplia literatura acerca de la variabilidad de los virus de ARN de cadena sencilla (15,37). En el caso de los *Potyvirus* se describen gran cantidad de especies, debido a la capacidad de especiación y adaptabilidad que presenta esta especie a la

hora de hospedar una nueva planta. De este modo que no podemos despreciar el conocimiento acerca de estas “posibles zonas”. Como vimos en los resultados del motivo de P3, al ser menos restrictivo que P1, se encuentra más representado en las especies, lo que nos podría hacernos pensar que estos virus las presenta como “almacén” de lugares de adaptación para ser usados en el momento oportuno.

Finalmente, encontramos necesario disponer de herramientas de análisis y visualización sencillas de usar, que permitirán encontrar nuevos productos génicos en el genoma de virus de una forma rápida y automatizada.

5. Conclusiones

Primero, la búsqueda de uORFs en miembros del género *Ipomovirus* no ha mostrado resultados significativos que nos hagan sospechar la necesidad de estos uORFs por parte de los virus. En concreto, no hemos encontrado que la especie *Cumcumber vein yellowing virus*, presenten uORFs con una necesidad funcional o como estrategia evolutiva del virus, ya que los péptidos obtenidos a partir de estos uORFs no presentan relación entre ellos. Sin embargo, no disponemos de la cantidad suficiente de secuencias analizadas como para establecer una conclusión clara. En este sentido, en la actualidad el laboratorio explora otro tipo de hipótesis relacionadas ahora con un necesidad estructural del virus.

Segundo, la utilización de FIMO como herramienta para la búsqueda de zonas de deslizamiento de la polimerasa tiene repercusiones, ya que si tenemos motivos muy laxos como en el caso de P3 tendremos muchos falsos positivos en la búsqueda. En este sentido, la utilización de un algoritmo clasificador bajo supervisión ha sido de gran utilidad para generar una discriminación de los resultados obtenidos con FIMO. Hemos encontrado que el algoritmo gkm-SVM es un herramienta muy potente para la clasificación de secuencias de elementos reguladores del ADN (31). Además, la utilización del potente algoritmo de clasificación “Suport Vector Machine” preparado para el análisis de secuencia de patrón con agujeros también nos fue útil para el análisis de motivos de secuencia.

Hemos encontrado que los resultados obtenidos con FIMO pueden representar la adaptabilidad evolutiva “potencial” que presentan estas especies.

Finalmente, encontramos una necesidad en desarrollar o poner a punto más herramientas relacionadas con el análisis de virus, ya que muchos pueden presentar una letalidad potencial para sus huéspedes. Con lo cual, la necesidad en el conocimiento de patógenos puede encontrar como solución la automatización en la anotación y análisis de sus genomas, ya que en actualidad la secuenciación de genomas es muy accesible para los investigadores.

4. Glosario

AUC: área debajo de la curva

Bioconductor: proyecto de código abierto para el análisis de datos en Genómica basado en lenguaje R.

Command-line: serie de líneas de texto que permiten dar instrucciones a un programa informático.

Cross-Validation CV: técnica para evaluar los resultados de un análisis estadístico y garantizar que los resultados son validos independientes de la partición de los datos en entrenamiento y prueba.

Dataframe: hoja de datos estructurada

Fasta: formato de texto para secuencias, que presentan “>” para identificar la cabecera y luego incorpora la secuencia de datos.

Frameshifting: evento que ocurra durante la traducción donde se producen múltiples proteína a partir de un solo ARN mensajero.

kernel: matriz positiva definida, utilizado normalmente para funciones de machine learning.

k-mers: Subsecuencia de longitud “k”.

Motivo: secuencia patrón de nucleótidos o aminoácidos que por su conservación puede estar relacionado con un significado biológico.

Ocurrencia: frecuencia en la que ocurre un evento.

PATH: variable del entorno del sistema operativo Linux.

q-value: parámetro estadístico que controla la cantidad de Falsos Positivos.

Secuencia en mosaico:

Slippage: zonas de deslizamiento de la polimerasa.

Terminal UNIX: Intérprete de comandos para el sistema Linux.

5. Bibliografía

1. Fauquet C, Mayo MA, Maniloff J, Desselberger U, Ball LA. Virus taxonomy - eighth report of the International Committee on the taxonomy of viruses. Vol. 83, The Viruses. 2005. 988–992 p.
2. Stephen J Wylie, Alice Kazuko Inoue-Nagata, Jan Kreuze, Juan José López-Moya, Kristiina Mäkinen KO and AW. Potyviridae [Internet]. Virus Taxonomy: 2019 Release. 2019 [cited 2020 Jun 15]. Available from: https://talk.ictvonline.org/ictv-reports/ictv_online_report/positive-sense-rna-viruses/w/potyviridae
3. Wylie SJ, Adams M, Chalam C, Kreuze J, López-Moya JJ, Ohshima K, et al. ICTV virus taxonomy profile: Potyviridae. J Gen Virol. 2017;98(3):352–4.
4. Adams MJ, Antoniw JF, Fauquet CM. Molecular criteria for genus and species discrimination within the family Potyviridae. Arch Virol [Internet]. 2005;150(3):459–79. Available from: https://doi.org/10.1007/s00705-004-0440-6
5. Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, et al. Top 10 plant viruses in molecular plant pathology. Mol Plant Pathol. 2011;12(9):938–54.
6. Chung BY-W, Miller WA, Atkins JF, Firth AE. An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci. 2008 Apr 15;105(15):5897–902.
7. Mingot A, Valli A, Rodamilans B, San León D, Baulcombe DC, García JA, et al. The P1N-PISPO trans -Frame Gene of Sweet Potato Feathery Mottle Potyvirus Is Produced during Virus Infection and Functions as an RNA Silencing Suppressor . J Virol. 2016;90(7):3543–57.
8. Olspert A, Chung BY, Atkins JF, Carr JP, Firth AE. Transcriptional slippage in the positive-sense RNA virus family Potyviridae . EMBO Rep. 2015;16(8):995–1004.
9. Janssen D, Martín G, Velasco L, Gómez P, Segundo E, Ruiz L, et al. Absence of a coding region for the helper component-proteinase in the genome of cucumber vein yellowing virus, a whitefly-transmitted member of the Potyviridae. Arch Virol. 2005;150(7):1439–47.
10. Valli A, López-Moya JJ, García JA. Recombination and gene duplication in the evolutionary diversification of P1 proteins in the family Potyviridae. J Gen Virol. 2007;88(3):1016–28.
11. Li W, Hilf ME, Webb SE, Baker CA, Adkins S. Presence of P1b and absence of HC-Pro in Squash vein yellowing virus suggests a general feature of the genus

Ipomovirus in the family Potyviridae. Virus Res. 2008 Aug;135(2):213–9.

12. Valli A, Martin-Hernandez AM, Lopez-Moya JJ, Garcia JA. RNA Silencing Suppression by a Second Copy of the P1 Serine Protease of Cucumber Vein Yellowing Ipomovirus, a Member of the Family Potyviridae That Lacks the Cysteine Protease HCPro. J Virol. 2006;80(20):10055–63.
13. Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci U S A. 1988 Dec;85(23):8998–9002.
14. Untiveros M, Olspert A, Artola K, Firth AE, Kreuze JF, Valkonen JPT. A novel sweet potato potyvirus open reading frame (ORF) is expressed via polymerase slippage and suppresses RNA silencing. Mol Plant Pathol. 2016;17(7):1111–23.
15. Nigam D, LaTourrette K, Souza PFN, Garcia-Ruiz H. Genome-Wide Variation in Potyviruses. Front Plant Sci. 2019;10(November):1–28.
16. Bedhomme S, Lafforgue G, Elena SF. Multihost experimental evolution of a plant RNA virus reveals local adaptation and host-specific mutations. Mol Biol Evol. 2012;29(5):1481–92.
17. García-Arenal F, Fraile A, Malpica JM. V Ariability and G Enetic S Tructure of P Lant V Irus P Opulations . Annu Rev Phytopathol. 2001;39(1):157–86.
18. Huang L, Li Z, Wu J, Xu Y, Yang X, Fan L, et al. Analysis of genetic variation and diversity of rice stripe virus populations through high-throughput sequencing. Front Plant Sci. 2015;6(MAR):1–7.
19. Athey J, Alexaki A, Osipova E, Rostovtsev A, Santana-Quintero L V., Katneni U, et al. A new and updated resource for codon usage tables. BMC Bioinformatics. 2017;18(1):1–10.
20. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics [Internet]. 2015 Aug 26;31(24):3997–9. Available from: https://doi.org/10.1093/bioinformatics/btv494
21. Kans J. Entrez Direct: E-utilities on the UNIX Command Line. In: Entrez Programming Utilities Help. Entrez Program Util Help [Internet]. 2013;(Md):1–86. Available from: http://www.ncbi.nlm.nih.gov/books/NBK179288/
22. Zhang H, Wang Y, Lu J. Function and Evolution of Upstream ORFs in Eukaryotes. Trends Biochem Sci [Internet]. 2019;44(9):782–94. Available from: http://www.sciencedirect.com/science/article/pii/S0968000419300556
23. Carrington JC, Freed DD. Cap-independent enhancement of translation by a plant potyvirus 5’ nontranslated region. J Virol. 1990 Apr;64(4):1590–7.
24. Roberts R, Zhang J, Mayberry LK, Tatineni S, Browning KS, Rakotondrafara AM. A Unique 5’ Translation Element Discovered in Triticum Mosaic Virus. J Virol.

2015 Dec;89(24):12427–40.

25. Pagès H, Aboyoun P GR. Biostrings: Efficient manipulation of biological strings. R package version 2.56.0. DebRoy S. 2020.
26. D. Charif and J.R. Lobry. Seqin R 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. Springer Verlag. 2007.
27. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings Int Conf Intell Syst Mol Biol. 1994;2:28–36.
28. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics [Internet]. 2011 Feb 16;27(7):1017–8. Available from: https://doi.org/10.1093/bioinformatics/btr064
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995;57(1):289–300.
30. Yousef M, Khalifa W, Acar İE, Allmer J. MicroRNA categorization using sequence motifs and k-mers. BMC Bioinformatics [Internet]. 2017;18(1):170. Available from: https://doi.org/10.1186/s12859-017-1584-1
31. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. PLOS Comput Biol [Internet]. 2014 Jul 17;10(7):e1003711. Available from: https://doi.org/10.1371/journal.pcbi.1003711
32. Lantz B. Machine Learning with R [Internet]. Packt Publishing; 2013. (Community experience distilled). Available from: https://books.google.es/books?id=ZQu8AQAAQBAJ
33. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics [Internet]. 2016/04/19. 2016 Jul 15;32(14):2205–7. Available from: https://pubmed.ncbi.nlm.nih.gov/27153639
34. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics [Internet]. 2009 May 25;25(14):1841–2. Available from: https://doi.org/10.1093/bioinformatics/btp328
35. Proutski V, Holmes E. SWAN: sliding window analysis of nucleotide sequence variability. Bioinformatics. 1998 Jun;14(5):467–8.
36. Janssen D, Martín G, Velasco L, Gómez P, Segundo E, Ruiz L, et al. Absence of a coding region for the helper component-proteinase in the genome of cucumber vein yellowing virus, a whitefly-transmitted member of the Potyviridae. Arch Virol. 2005 Jul;150(7):1439–47.

37. Gibbs AJ, Hajizadeh M, Ohshima K, Jones RAC. The Potyviruses: An Evolutionary Synthesis Is Emerging. Viruses. 2020 Jan;12(2).
38. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics [Internet]. 2004 Jan 22;20(2):289–90. Available from: https://doi.org/10.1093/bioinformatics/btg412