

Tesis de Máster

BIOINFORMÁTICA APLICADA AL ESTUDIO DE INTERACCIONES sncRNA-ANTITROMBINA

Autor:

Fernando Pérez Sanz



Universidad de Murcia

Máster en Bioinformática

Tutores:

Javier Corral de la Calle
José Luis Fernández Alemán
Ginés Luengo Gil

4 de septiembre de 2014

Máster en Bioinformática

Fernando Pérez Sanz

Índice

1. INTRODUCCIÓN	3
2. OBJETIVOS	6
3. MÉTODOS	7
3.1. <i>Datos</i>	7
3.2. <i>Hardware y software</i>	8
3.3. Bloque I. Análisis funcional y descriptivo	8
3.4. Bloque II. Agrupamiento	9
3.4.1. Alineamiento de secuencias	9
3.4.2. Evaluación de los alineamientos	10
3.4.3. Agrupamiento	10
3.4.4. Secuencias consenso	11
3.5. Bloque III. Predicciones estructurales	11
4. RESULTADOS Y DISCUSIÓN	12
4.1. Bloque I. Resultado del análisis funcional y descriptivo	12
4.2. Bloque II. Resultados del agrupamiento	15
4.3. Bloque III. Resultados de la predicción estructural	19
5. CONCLUSIONES	22
Anexos	24
A. Estructuras 2D y 3D de las secuencias	24
B. Workflow de los <i>scripts</i> Galaxy	33
B.1. Descripción de los <i>scripts</i> que componen el <i>wrapper Retrieve_info</i>	33
B.1.1. Capturas pantalla con resultados del <i>wrapper Retrieve_info</i>	34
B.2. Descripción de los <i>scripts</i> que componen el <i>wrapper RNA_analysis</i>	35
B.2.1. Capturas pantalla con resultados del <i>wrapper RNA_analysis</i>	36
Bibliografía	40

RESUMEN

A pesar de que el genoma humano es activamente transcrit (alrededor del 90 %) apenas un 2 % de éste codifica proteínas. Gran parte de este RNA puede tomar partido en la regulación de procesos celulares. Dentro de este grupo de RNA no codificante encontramos la familia de los RNA pequeños (sncRNA) entre los que se destacan siRNA, miRNA, piRNA o stRNA, que actúan como reguladores de la expresión génica a nivel post-transcripcional, generalmente mediante la formación de complejos ribonucleoproteicos. Recientemente, el grupo de Hematología y Oncología Clínico-Experimental de la Universidad de Murcia ha demostrado la existencia de este tipo de complejos entre sncRNA y ciertas proteínas que intervienen en procesos hemostáticos como la antitrombina (AT), y cuya formación puede afectar directamente a la actividad de dicha proteína.

En este trabajo se aborda el estudio de RNA de pequeño tamaño, a través de un flujo de trabajo metodológico que pasa por el alineamiento de secuencias, análisis estadístico de las mismas con el objetivo de formar grupos estructuralmente similares y seleccionar secuencias representantes de cada grupo. A continuación se realiza tanto la predicción de la estructura secundaria y terciaria, como acoplamiento (*docking*) entre la AT y dichas estructuras, con el objetivo final de encontrar secuencias con potencial capacidad de unión a la AT en el sitio de unión de la heparina. Para ello se han empleado herramientas bioinformáticas y estadísticas existentes, además de desarrollar dos específicas integradas en el entorno Galaxy.

Los resultados del presente estudio, muestran cómo los diferentes algoritmos de alineamiento arrojan resultados muy diferentes con consecuencias en las siguientes fases de estudio. Nuestro estudio se ha seleccionado aquel algoritmo que minimiza la función de entropía de Shannon. Dicho algoritmo ha sido MAFFT-NSi con penalización por gaps máxima (3,0). Así mismo, del agrupamiento de secuencias con $k=9$ se han obtenido grupos estructuralmente muy homogéneos, excepto 2 grupos a los que se les ha aplicado nuevamente el agrupamiento creando subgrupos. Por último se han determinado estructuras secundarias y terciarias de 26 secuencias seleccionadas (originales y consenso) y se ha realizado *docking* con todas ellas generando 130 posibles soluciones sobre la potencialidad de interacción entre los sncRNA y la AT.

Estos estudios muestran sncRNA que se unen en el sitio de unión de la heparina o sus proximidades, y que podrían interferir en la unión de este cofactor de AT, lo que podría provocar un cambio en la actividad inhibitoria de este potente anticoagulante.

Keywords: sncRNA, Antitrombina, alineamiento múltiple de secuencias, clasificación, predicción estructural, *docking*.

1. INTRODUCCIÓN

Una de las grandes sorpresas de la Biología moderna fue descubrir que, el material genético que codifica las 20.000 proteínas de la especie humana apenas si representa el 2 % del genoma total. La vasta mayoría del genoma se consideraba DNA basura. Sin embargo, con la llegada de las nuevas tecnologías de secuenciación masiva, se determinó que al menos el 90 % del genoma es activamente transcrit, encontrándose una gran variedad de RNA no codificante. Inicialmente se argumentó que podría ser ruido transcripcional, aunque recientes evidencias demuestran que ese “RNA basura” del genoma en realidad desempeña un papel clave en diversos procesos celulares[1].

A pesar del enorme progreso en el conocimiento de la regulación génica mediada por RNA, hay todavía todo un “submundo” inexplorado de procesos regulados por RNA en los que está aún por descubrir el papel que juegan como reguladores de los mismos [2]. Formando parte de los RNA no codificantes se encuentran los RNA pequeños (sncRNA), éstos han irrumpido con fuerza como activos represores de la expresión génica en plantas, animales y muchos hongos. Toda una familia de sncRNA entre los que se incluyen por su conocida función reguladora, RNA pequeño de interferencia (siRNA), micro RNA (miRNA), RNA asociados a Piwi (piRNA), RNA pequeños temporales (stRNA) [3], que actúan como reguladores de la expresión génica, mediante procesos que activan nucleasas específicas que cortan el RNA diana, deadenilación y decaimiento o de inhibición traduccional por desestabilización del RNA mensajero (mRNA). Se ha estimado que podrían regular la expresión de hasta un tercio de los genes de mamíferos [2].

Esta regulación es llevada a cabo específicamente por complejos ribonucleoproteicos (RNP), donde la proteína contiene uno, o más frecuentemente, múltiples dominios de unión al RNA [4]. Este tipo de complejos aunque están ampliamente caracterizados, ha sido mucho menos estudiada su función reguladora que en el caso de las interacciones proteína-proteína (PPI) o DNA-proteína. Además, sumado a su papel en el control de la expresión génica, las interacciones RNA-proteína regulan otros procesos biológicos fundamentales que van desde la replicación y transcripción del DNA, a la resistencia a patógenos o la replicación viral [5].

Actualmente, continúan descubriéndose nuevos genes regulados por sncRNA que sugieren que estas moléculas están implicadas en el control de gran parte de los procesos fisiológicos. Algunos estudios demuestran que muchos de ellos están específicamente expresados en ciertos órganos, tipos celulares o etapas del desarrollo [6]. Concretamente se ha constatado la importancia del papel que desempeñan los sncRNA en ciertas fases de los procesos hemostáticos como la regulación de la expresión del factor tisular (TF) [7].

Pero el papel de los sncRNA en hemostasia no se restringe a su capacidad de regulación transcripcional, también pueden jugar un papel relevante en la regulación de la función de proteínas hemostáticas. En este sentido, el grupo de Hematología y Oncología Clínico-Experimental de la Universidad de Murcia ha demostrado la existencia de complejos RNP entre sncRNA y la antitrombina (AT). Esta glicoproteína, perteneciente a la super familia de inhibidores de la serín proteasa (serpina), es el principal anticoagulante endógeno ya que inhibe de forma rápida y eficaz diferentes factores implicados en la coagulación sanguínea (FXa, FIXa, FXIa, FXIIa, y sobre todo, de ahí su nombre, la trombina o FIIa).

Por ello, su deficiencia incluso moderada incrementa significativamente el riesgo trombótico [8].

Empleando diferente metodología, especialmente resonancia de plasmones de superficie (BiacoreTM), se comprobó que el RNA puede unirse a la AT en el sitio de unión de la heparina compitiendo con la unión de este cofactor. Este efecto tendría consecuencias protrombóticas ya que impediría la activación de la AT por la heparina. No obstante, si esta unión es posible, especulamos que otros sncRNA podrían unirse a la AT con las mismas consecuencias que la heparina, activar este anticoagulante (y por tanto actuar de forma similar a la heparina (*heparin-like*)). [7].

Este mayor conocimiento sobre los sncRNA y sus funciones se ha debido en cierta medida, al desarrollo de herramientas y tecnología experimentales (secuenciación masiva, transfección, experimentos con luciferasa o con proteína verde fluorescente -GFP-, etc.) pero también al desarrollo de herramientas bioinformáticas de predicción *in silico*. Actualmente, el campo de la Bioinformática está desarrollándose muy rápido, aportando herramientas fundamentales para alcanzar el conocimiento que se tiene sobre los sncRNA. Alineadores de secuencias, implementaciones de algoritmos estadísticos multivariantes, herramientas predictoras de estructuras espaciales (2D, 3D), simuladores de interacciones moleculares, son algunas de las herramientas que facilitan a los investigadores el entendimiento del papel que juegan los sncRNA como reguladores en multitud de procesos moleculares.

El alineamiento múltiple de secuencias (*MSA*) es el núcleo principal en el análisis de secuencias biológicas. La reconstrucción de árboles filogenéticos, predicción de estructuras o descubrimiento de motivos (*motifs*) a través de modelos ocultos de Markov (*HMM*) son algunas de las tareas que requieren del *MSA* para inferir identidades estructurales o funcionales. Esa omnipresente necesidad de computar alineamientos ha hecho de este campo uno de los más activos con más de 100 métodos diferentes publicados en los últimos 30 años, y numerosas publicaciones relacionadas con su caracterización, evaluación y comparación [9, 10, 11].

Tanto los resultados arrojados por un *MSA* determinado, como los obtenidos por diferentes métodos de alineamiento, tienen un efecto importante sobre los posteriores análisis que se realicen sobre los datos. Tanto es así, que la principal característica a tener en cuenta en un nuevo método de alineamiento es la precisión más que la rapidez [9].

El alineamiento óptimo sólo es factible para unas pocas secuencias, por el alto coste computacional. Sin embargo para un conjunto de decenas o cientos de secuencias se emplean algoritmos heurísticos. Los más relevantes son los progresivos, en los que el alineamiento múltiple se construye a base de alineamientos por pares a los que de forma progresiva se le van incorporando secuencias [12].

Habitualmente, el siguiente paso al alineamiento de secuencias, suele ser un tratamiento estadístico, que permita establecer una relación estructural y/o funcional entre las secuencias, al objeto de determinar posibles relaciones filogenéticas entre homólogas de diferentes especies, o bien establecer estructuras similares en secuencias con diferente funcionalidad. Por tanto se hace necesario el uso de algún tipo de técnica estadística que facilite la clasificación del conjunto de secuencias por similitudes.

En este sentido, la clasificación consiste en realizar una partición del conjunto original de datos en subconjuntos homogéneos, siguiendo un determinado criterio de clasificación, donde cada elemento pertenece a un único subconjunto [13]. Los métodos de clasificación pueden ser jerárquicos o itera-

tivos (no jerárquicos), en el primer caso no se define a priori el número de grupos y en el segundo es necesario hacerlo [13, 14]. Ambas técnicas son empleadas en Bioinformática dentro de la búsqueda de agrupamientos por homologías [15, 16, 17, 18].

Existen otra serie de técnicas estadísticas multivariantes como el análisis de componentes principales (PCA) o escalado multidimensional (MDS) en las que partiendo de un número determinado de variables, mediante combinaciones lineales se obtienen unas pocas variables incorrelacionadas que explican la mayor parte de la variabilidad [13]. A pesar de no ser técnicas de clasificación, dentro del campo del análisis de secuencias se han empleado para “clasificar” dichas secuencias mediante la asignación de cada una de ellas a una componente [15, 19, 20].

Las secuencias tanto peptídicas como nucleotídicas, en estado natural en disolución, no son estables como conformaciones extendidas, sino que, debido principalmente a la capacidad de formar puentes de hidrógeno, tienden a plegarse constituyendo las estructuras secundarias, cuya disposición tridimensional se denomina estructura terciaria [21].

Tener la capacidad de predecir estas conformaciones y más aún poder predecir las interacciones entre diferentes estructuras (RNA-proteína), es un tema crucial en la investigación biomédica por la importancia que está adquiriendo el conocimiento en la regulación de proteínas por RNA. Por tanto, llegar a comprender los mecanismos moleculares de la formación de estos complejos es uno de los mayores retos de la Biología estructural.

Aunque menos precisos que los métodos experimentales (cristalografía, resonancia magnética nuclear, microscopía electrónica), las predicciones computacionales pueden ser suficientemente precisas como para dar “pistas” sobre estas interacciones que sirvan para guiar de forma más concreta posteriores experimentos [22].

Siguiendo esta línea, en los últimos años, se han desarrollado muchas aproximaciones computacionales para predecir las interacciones RNA-Proteína [5]. Herramientas tales como Patchdock [23], Haddock [24], Rosetta [25], 3dRNA [26], CatRapid [27], abordan el problema empleando diferentes metodologías tales como el uso de Machine Learning (SVM, Random Forest), incluyendo o no información sobre estructuras secundarias del RNA, incorporando información sobre fuerzas electrostáticas, aproximaciones exclusivamente estructurales, etc. [28, 29, 30].

En el presente trabajo se aborda el estudio de RNA de pequeño tamaño, con el objetivo final de encontrar secuencias con potencial capacidad de unión a la AT en el sitio de unión de la heparina, a través de un flujo metodológico que pasa por el análisis, alineamiento, clasificación, predicción de estructuras, y finalmente simulación de las interacciones RNA-Proteína haciendo uso de algunas de las herramientas existentes así como desarrollando dos específicas. El descubrimiento de secuencias concretas de RNA que se unan a la AT compitiendo con la heparina o actuando como *heparin-like* podría tener importantes consecuencias terapéuticas. Poder inferir el efecto funcional que la unión de un RNA tiene en la AT podría permitir el desarrollo de nuevos fármacos potencialmente útiles como antídotos de la heparina o como nuevos anticoagulantes.

2. OBJETIVOS

Los objetivos planteados en el presente trabajo fueron:

Bloque I:

- Analizar de forma descriptiva y funcional las secuencias objeto de estudio.

Bloque II:

- Realizar agrupamiento de secuencias basada en similitud estructural.

Bloque III:

- Predecir la estructura secundaria y terciaria de una serie de secuencias seleccionadas.
- Predecir la potencial capacidad de unión entre las estructuras calculadas anteriormente y la AT

A su vez, se han planteado unos objetivos operativos:

- Manejar diferentes herramientas de alineamiento múltiple de secuencias, conocer sus características y ámbito de aplicación.
- Programar un algoritmo de análisis de la calidad de los alineamientos.
- Manejar un entorno de computación estadística (R).
- Desarrollar *scripts* en R y en Shell Bash, tanto para manipulación de ficheros como para automatizar ciertos análisis.
- Programar herramientas en el entorno Galaxy que integren parte de los trabajos realizados sobre las secuencias objeto de estudio.
- Desarrollar un flujo de trabajo metodológico específico, orientado al estudio fundamentalmente estructural de RNA de pequeño tamaño.

En consecuencia, el presente documento ha quedado organizado como sigue: en las secciones 3.1 y 3.2 se describe brevemente la procedencia de los datos, el equipo informático y el software empleado. En las secciones 3.3 y 4.1 correspondientes al bloque I, se realiza el análisis descriptivo y funcional de las secuencias objeto de estudio. En el bloque II (secciones 3.4 y 4.2) se describe la metodología seguida y los resultados obtenidos tras alinear y agrupar las secuencias. En las secciones 3.5 y 4.3, correspondientes al bloque III, se describe las herramientas empleadas para las predicciones estructurales, así como las estructuras obtenidas. Por último, en la sección 5 se presentan las conclusiones del presente trabajo, así como algunas ideas futuras.

A continuación se muestra un diagrama con el flujo de trabajo seguido (Figura 1). Además se han incluido una serie de anexos al documento, con algunas representaciones gráficas y los *scripts* empleados.

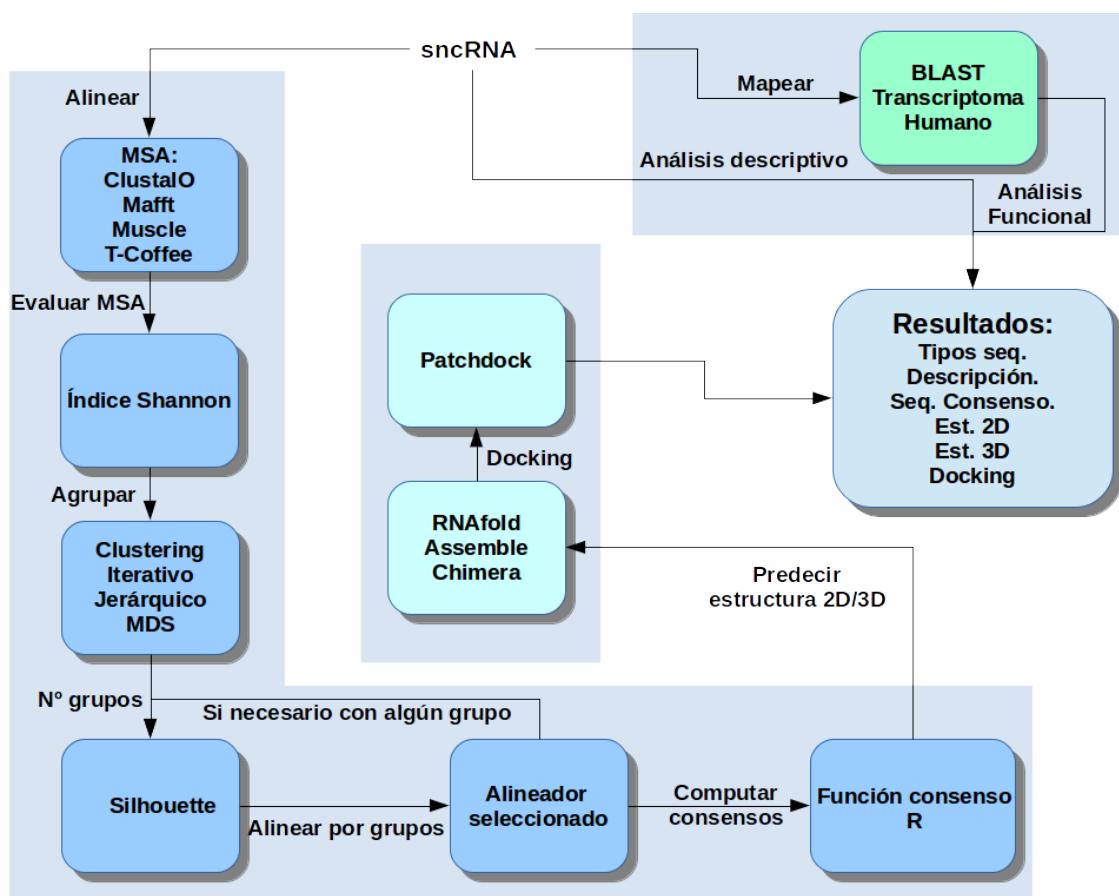


Figura 1: Flujo de trabajo seguido en el presente estudio.

3. MÉTODOS

3.1. *Datos*

Las secuencias de nucleótidos que se unen a la AT y se han empleado en el presente trabajo, proceden de estudios experimentales realizados por el grupo de investigación de Hematología y Oncología Clínica Experimental, Hospital General Universitario Morales Meseguer, Universidad de Murcia-IMIB. Estos datos, propiedad intelectual de mencionado grupo, están pendientes de publicación y podrían ser objeto de patente por lo que no pueden ser mostradas en este trabajo.

Brevemente, se describe el procedimiento empleado para conocer qué sncRNA se unen a la AT. Para la obtención del sncRNA se cultivaron 50 millones de células endoteliales (EA.hy926, hibridoma formado en laboratorio a partir de células endoteliales fusionadas con adenocarcinoma de pulmón) y empleando un kit específico para purificar RNA de menos de 200 bases se obtuvieron un total de 70 μg . de sncRNA. El sncRNA se pasó por una columna FPLC con AT unida a colas de histidina, se lavó extensivamente con tampón basal y con 500 mM NaCl y finalmente los sncRNA unidos a la AT se eluyeron con 3M de NaCl. Al final del proceso se obtuvieron 50 ng. de sncRNA que fue sometido a secuenciación masiva de RNA pequeño empleando la plataforma de LC-Sciences.

El resultado de la secuenciación se procesó eliminando la información no necesaria procedente

del secuenciador resultando un fichero en formato FASTA¹ que contenía 2.256.096 secuencias, correspondientes a 898 secuencias únicas de longitudes inferiores a 40 nucleótidos, con el que se ha realizado este trabajo.

3.2. Hardware y software

Todos los cálculos y cómputos realizados en local han sido ejecutados en un equipo informático portátil, constituido por un procesador Intel® Core™ i7 CPU Q 720 @ 1,60GHz x 8, con 8GB de memoria RAM y tarjeta gráfica ATI HD 5730 con 1 GB de memoria dedicada.

El software empleado en la realización del presente trabajo que se menciona a continuación, ha sido ejecutado en su totalidad sobre una plataforma linux.

- R-RStudio (análisis estadístico, agrupamientos, Shannon, gráficas estadísticas, secuencias consenso), ClustalO, T-Coffee, MUSCLE² y MAFFT (alineamiento de secuencias), BLAST³ (mapeo de secuencias), RNAFold (estructuras secundarias), Assemble2-Chimera (estructuras terciarias), PatchDock (docking), Galaxy (plataforma bioinformática).

A continuación, se menciona otro software probado en el transcurso del trabajo y que finalmente no ha sido empleado por diferentes motivos.

- 3dRPC, SwissDock, Jalview, 3dRNA, Rosetta, Mirtools, Force Field Explorer, Ugene.

3.3. Bloque I. Análisis funcional y descriptivo

En primer lugar, y con el objetivo de determinar a qué tipo de RNA pertenece cada secuencia, se ha realizado un mapeo de las mismas contra la base de datos del transcriptoma humano descargada del NCBI (<http://www.ncbi.nlm.nih.gov/>) a través del programa BLAST [31]. Esta herramienta es ampliamente usada para la búsqueda de similitudes en estructuras proteicas y nucleotídicas mediante alineamiento y comparación con las secuencias presentes en las bases de datos.

En este caso, para la ejecución del BLAST, se ha construido una herramienta específica (Anexo B.1) integrada en el entorno web Galaxy [32, 33], que se ejecuta en local contra la mencionada base de datos. Dentro de los parámetros de configuración de BLAST, se ha seleccionado el algoritmo *blastn-short*, específico para secuencias de nucleótidos de pequeña longitud, o el ajuste del *e-value* a $1 \cdot 10^{-5}$, pues debido a la escasa longitud de las secuencias ha sido necesario ampliar el margen de error para obtener mapeo positivo.

```
blastn -db ~/bioapps/ncbi-blast-2.2.29+/db/rna -query $input
-task blastn-short -max_target_seqs 10 -word_size 7
-evalue 100 -outfmt 7 -out $temp
```

Al fichero resultante, se le ha aplicado un *script* programado específicamente para este propósito (Anexo B.1), en el que mediante la combinación de instrucciones de Shell Bash (entorno Linux) y

¹FASTA: FAST-All

²MUSCLE: MUltiple Sequence Comparison by Log- Expectation

³BLAST: Basic Local Alignment Search Tool

script en el lenguaje y entorno de programación para análisis estadístico R [34] se obtiene una estadística sobre la composición funcional de las secuencias.

Al igual que en el caso del análisis funcional, para cuantificar la estructura de datos objeto de estudio, se ha realizado una herramienta específica integrada en Galaxy, que ejecuta un *script* con una serie de instrucciones dirigidas al tratamiento de los datos (Anexo B.1), con la finalidad de obtener parámetros tales como la distribución de abundancia de secuencias y de lecturas, según su longitud, o proporción de nucleótidos.

3.4. Bloque II. Agrupamiento

3.4.1. Alineamiento de secuencias

Un paso previo al agrupamiento de secuencias propiamente dicho, consiste en el alineamiento de las mismas. Mediante este proceso se consigue un ajuste de unas secuencias a otras, con la finalidad de inferir similitudes estructurales, resultando un conjunto de datos de igual longitud debido a la inserción de *gaps* o saltos. Este paso es necesario para la construcción de la matriz de disimilitud basada en las diferencias estructurales de las secuencias.

Existe una gran variedad de herramientas de alineamiento, cada cual con unas características propias tanto en velocidad del proceso como en calidad de alineamiento entre otras, que la hacen idónea para un conjunto de datos determinados. En este caso se han utilizado cuatro de las más conocidas herramientas [10] de alineamiento múltiple de secuencias: *Muscle*[35], *T-Coffee*[12, 36], *ClustalO*[37], *MAFFT*[38, 39]. Todas ellas han sido ejecutadas en local con parámetros establecidos por defecto. En el caso de MAFFT, que desde las primeras pruebas realizadas mostró mayor calidad de alineamiento para estos datos, se han probado varios de los algoritmos que implementa, así como distintos valores de penalización por *gap*. Con ClustalO y MAFFT se ha probado también la versión web server aunque en ambos casos la ejecución local era más rápida. El resto de alineadores aunque también se ha intentado probar la versión web server, tienen limitación en el número de secuencias, con lo que sólo se ha podido ejecutar en local. Con cada uno de ellos se ha obtenido un fichero FASTA de alineamientos que posteriormente ha sido evaluado.

```
t_coffee -in=rnas.fa Mpcma Mmafft_msa Mclustalw_msa Mmuscle_msa Mt_coffee_msa
          -output=score_html clustalw_aln fasta_aln score_ascii phylip -tree
          -outorder=on -outorder=input -run_name=result -multi_core=6 -quiet=stdout

clustalo -i RNA-trim5-3-sinN.fa -t RNA --cluster-size=300
          --clustering-out=cluster.out -o clustalo.aln --outfmt=fa --threads=8

clustalo --infile clustalo-I20140707-170912-0542-83571961-pg.upfile --threads 8
          --MAC-RAM 8000 --verbose --outfmt fa --outfile clustalo.fasta
          --output-order tree-order --seqtype dna

muscle -in rnas.fa -verbose -quiet -fasta -out muscle.aln -group

mafft --thread 2 --thredit 0 --reorder --auto input >output
# Todos los parametros por defecto
```

```

mafft --thread 2 --threaddit 0 --reorder --op 3.0 --maxiterate 1000 --6merpair input >output
# FFTINS gaps penalty

mafft --thread 2 --threaddit 0 --reorder --maxiterate 1000 --6merpair >output
# FFTINS

mafft --thread 2 --threaddit 0 --reorder --maxiterate 1000 --retree 1 --globalpair input >output
# GINSi globalpair

mafft --thread 2 --threaddit 0 --reorder --op 3.0 --maxiterate 1000 --retree 1
--globalpair input > output # GINSi con gaps penalty

mafft --reorder --op 3.0 --maxiterate 1000 --retree 1 --localpair input >output # LINS localpair

```

3.4.2. Evaluación de los alineamientos

Con el objetivo de poder decidir cuál de las herramientas de alineamiento empleadas es la que ofrece mejor resultado y por tanto la que se selecciona para el resto del trabajo, se hace necesario aplicar a todos los alineamientos un mismo criterio de puntuación, una misma función con la que poder compararlos, de forma que ayude a la toma de decisión. En este caso se ha programado una función sobre el entorno de R basado en la mínima entropía de Shannon [40, 41], que calcula para cada posición del alineamiento la puntuación en esa posición.

La entropía total es la suma de las entropías de todas las posiciones. Así, el alineamiento con el menor valor del cálculo, es el que, teóricamente, ha resultado como el mejor.

$$S = - \sum p_a^i \cdot \log_2 p_a^i$$

donde p_a^i es la probabilidad del elemento a en la posición (columna) i .

3.4.3. Agrupamiento

Mediante el análisis de conglomerados [14], se pretende agrupar los objetos de interés en grupos lo más homogéneos posibles según el criterio establecido, en este caso la similitud estructural. Para ello se ha seguido los siguientes pasos:

- Agrupamiento (*clustering*) iterativo [42], con grupos de 2 a 20 sobre el alineamiento seleccionado previamente. Se ha considerado más apropiado el uso de K-medoides y no K-means como algoritmo de *clustering* por dos motivos fundamentales: para un número elevado de datos K-medoides resulta menos afectado por posibles valores extremos (el cálculo del centro de cada grupo se hace en función de la mediana y no de la media) y es más apropiado cuando se trabaja con una matriz de disimilitudes como es el caso [43].

- Aplicar la función *Silhouette*[44] (S_i) sobre los grupos. Los agrupamientos con valores de S más elevados indican una mejor asignación de los elementos a los grupos de ese agrupamiento.

$$S_i = (b_i - a_i) / \max(a_i, b_i)$$

donde a_i es la disimilaridad media entre el elemento i y el resto de elementos de su grupo y b_i es la disimilaridad media entre el elemento i y los elementos del grupo más cercano.

- *Clustering jerárquico* [45] sobre las secuencias. El *clustering* iterativo ayuda a definir el número de grupos, pero el agrupamiento se ha realizado mediante el jerárquico, cortando el dendrograma a nivel que se obtenga el número de grupos determinados mediante K-medoides y la función *Silhouette*.

- Con el fin de visualizar y validar en cierta medida que el *clustering* realizado ha generado grupos relativamente homogéneos se ha aplicado un MDS. Esta técnica, más apropiada para matrices de distancias que el PCA [46, 47, 20, 48], permite obtener una configuración de puntos que los represente en un espacio de reducidas dimensiones (2-3) con la menor pérdida de información.

3.4.4. Secuencias consenso

Con el objetivo de realizar una predicción de estructuras secundarias y terciarias de cada grupo de secuencias obtenido, se han computado tantos alineamientos múltiples como grupos generados en el paso anterior. Cada alineamiento se ha hecho esta vez con las secuencias pertenecientes a ese grupo. De cada uno de ellos, se ha calculado la secuencia consenso que se empleará para las predicciones posteriores, a través de una modificación de la función *consensus* específica del paquete *Sginr* de R, de tal forma que se calcula para cada posición el nucleótido más abundante sin tener en cuenta los *gaps*.

Todos los pasos que componen este bloque han sido integrados en una herramienta específicamente programada para ser ejecutada desde el entorno Galaxy, cuyo funcionamiento está recogido en el esquema del anexo B.2.

3.5. Bloque III. Predicciones estructurales

Las predicciones estructurales se han realizado con 26 secuencias pertenecientes a 12 secuencias consenso de otros tantos grupos, 12 secuencias originales de cada grupo (la que más lecturas tenía en cada uno), y 2 secuencias pertenecientes a 2 grupos de 1 sola secuencia.

La predicción de estructuras secundarias se ha llevado a cabo mediante RNAFold [49]. Este programa está integrado dentro de los servicios web proporcionados por el servidor ViennaRNA (<http://rna.tbi.univie.ac.at/>). RNAFold hace las predicciones a partir cadenas sencillas, buscando siempre la estructura de mínima energía libre⁴.

La reconstrucción de las estructuras terciarias se ha realizado *ab initio*, mediante un proceso semiautomático en el que se seleccionan de manera sucesiva fragmentos (hélices o hebras) de la estructura secundaria generándose automáticamente el fragmento terciario correspondiente. Para ello se ha empleado Assemble2 [50] en local, con un plugin instalado del visualizador y editor de estructuras terciarias UCSF Chimera [51], de forma que se ha podido supervisar la recreación del modelo en tiempo real.

⁴Potencial termodinámico de un sistema. Función de estado que da la condición de equilibrio y espontaneidad para una reacción química.

Una vez construidas las 26 estructuras terciarias, se ha procedido a realizar un *docking* estructural, entre cada una de las estructuras creadas, y la cadena I de la AT (PDB: 1AZX), a la que se le ha eliminado el pentasacárido (elemento básico de la heparina que se une a la AT). Para llevar a cabo este proceso se ha empleado PatchDock [23], que utiliza un algoritmo con criterios puramente morfológicos, creando una lista de potenciales complejos proteína-ligando ordenados por complejidad. Puesto que cada proceso de *docking* suponía un tiempo medio de 20 minutos. Se ha realizado tanto en local como contra el servidor de PatchDock (<http://bioinfo3d.cs.tau.ac.il/PatchDock/>) con el fin de poder lanzar varios procesos simultáneamente.

4. RESULTADOS Y DISCUSIÓN

Esta sección presenta los resultados obtenidos de los tres bloques. En el primer bloque se muestran los datos resultantes del análisis estadístico descriptivo y funcional. El segundo bloque recoge los resultados obtenidos de la clasificación en grupos, así como de las secuencias consenso. Hay que indicar aquí, que debido a la existencia de propiedad intelectual sobre las secuencias, tampoco se pueden mostrar la relación de nucleótidos que constituyen las secuencias consenso. Por último, en el tercer bloque se muestran las predicciones sobre estructuras secundarias y terciarias, así como el resultado obtenido del *docking* estructural.

4.1. Bloque I. Resultado del análisis funcional y descriptivo

Análisis funcional

El resultado del mapeo de las secuencias contra el transcriptoma humano mediante la herramienta BLAST se ha obtenido seleccionando el emparejamiento de más alta puntuación. Destacar que estos resultados hay que tomarlos como algo meramente descriptivo, pues debido a la escasa longitud de las secuencias (similares a miRNA), la probabilidad de emparejamiento completo con cualquier otra secuencia de la base de datos es muy alta, lo que no garantiza que la identificación sea correcta.

Por otro lado, dado que los RNA objeto del presente estudio, son un conglomerado de RNA pequeños y posibles fragmentos de RNA de diferentes tipologías ha sido imposible mapear contra una base de datos específica de miRNA. A pesar de lo mencionado se ha considerado conveniente mostrar la distribución de las distintas tipologías obtenidas del mapeo (Figura 2).

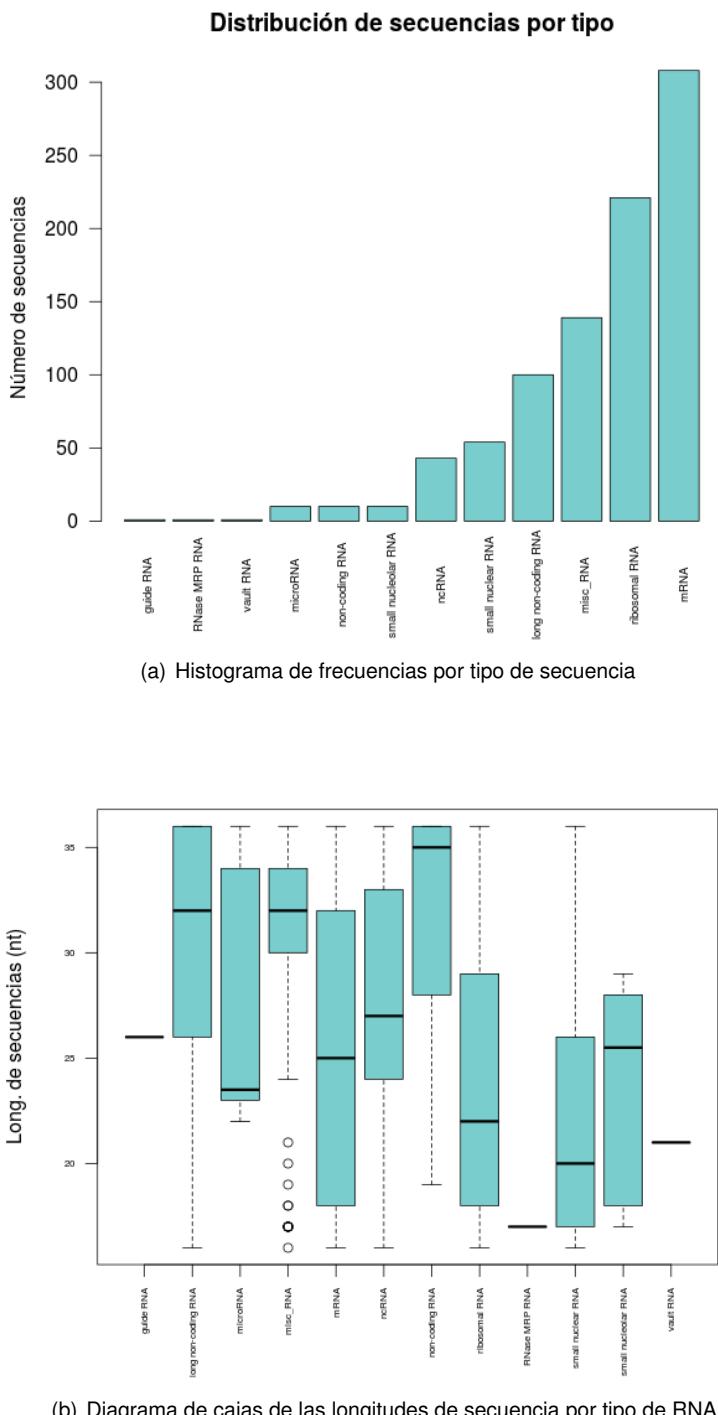


Figura 2: Distribución de la abundancia de las distintas tipologías de RNA (a). Distribución de las longitudes de secuencias para cada tipología (b).

Análisis descriptivo

El fichero de trabajo (Tabla 1) consta de 898 secuencias diferentes, con una longitud mínima de 15 bases, máxima de 35 y longitud media de 25,2 bases. El número de lecturas totales (copias de las secuencias) es de 2.256.096 secuencias distribuidas entre las 898 secuencias únicas. Respecto a la distribución de la abundancia de nucleótidos, la adenina aparece en un 17,6 %, la citosina en un 22,5 %, la guanina en un 33,7 % y el uracilo en un 26,2 %.

Número de secuencias	898
Número de lecturas	2.256.096
Longitud mínima	15 nt
Longitud máxima	35 nt
Longitud media	25,2 nt
Porcentaje A	17,6 %
Porcentaje G	33,7 %
Porcentaje C	22,5 %
Porcentaje U	26,2 %

Tabla 1: Resumen de la cuantificación de las secuencias

Tal como se observa en el histograma de frecuencias (Figura 3), entre las secuencias de menor longitud (15-16 nt) y las de mayor longitud (33-35 nt) se concentra el 39,64 % de todas las secuencias. Estas 4 tipologías aglutinan 356 secuencias diferentes.

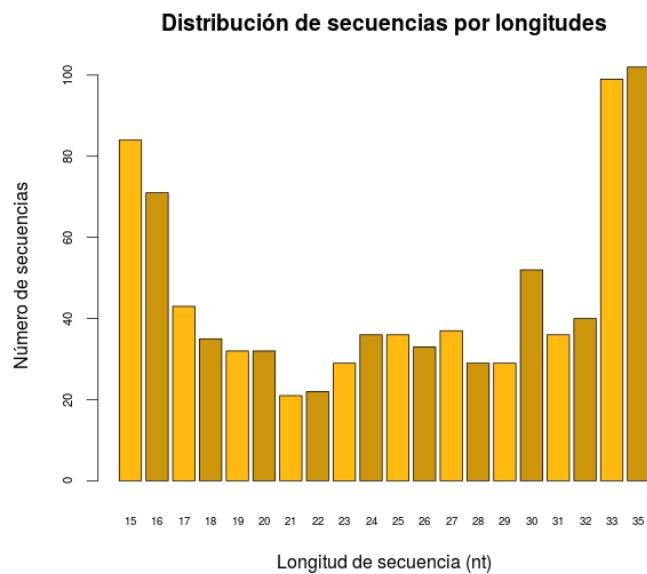


Figura 3: Distribución de las secuencias según el número de nucleótidos que contienen. Las secuencias más cortas, así como las más largas son las que muestran más abundancia respecto al resto.

Representando la longitud de las secuencias frente al número de lecturas (Figura 4), se obtiene una distribución diferente. En este caso, en los extremos de la gráfica siguen apareciendo altos valores de frecuencia, algo esperado considerando el alto número de secuencias que hay en esos grupos.

Sin embargo también se observa que hay varios grupos de secuencias (longitudes 20, 26, 30, 31 y 32) con altos niveles de lecturas donde no hay un alto número de secuencias, lo que hace pensar que hay unas pocas secuencias con gran número de copias. Estos 9 grupos de secuencias (longitudes 15, 16, 20, 26, 30-35) aportan el 75,7 % de las lecturas.

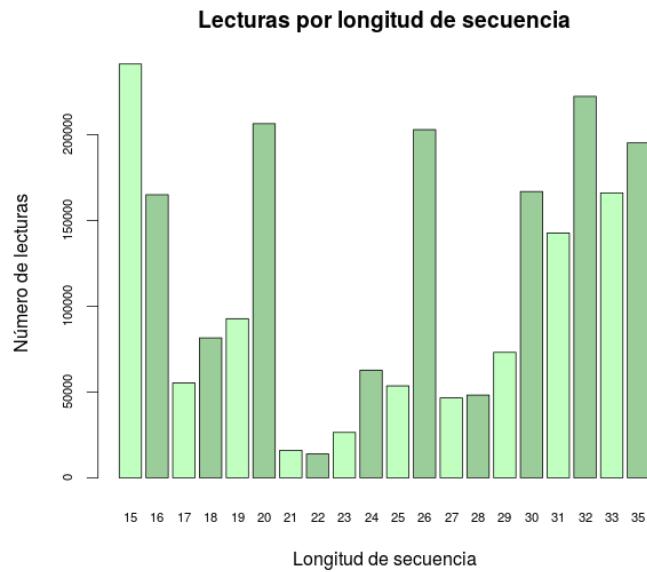


Figura 4: Distribución de las secuencias según el número de lecturas.

4.2. Bloque II. Resultados del agrupamiento

Alineamientos

Como resultado de los alineamientos, se han obtenido 10 ficheros, cada uno de los cuales ha sido evaluado mediante el índice de mínima entropía de Shannon [52]. El resultado obtenido (Tabla 2) muestra que el mejor alineador para las secuencias objeto de estudio y con la función de evaluación aplicada ha sido MAFFT con la estrategia FFT-NS-i y con penalización por *gap* de 3,0 [38, 39], con un valor de entropía de 15,21. Esta estrategia de refinamiento iterativa a pesar de ser algo más lento que las estrategias progresivas que implementa MAFFT [39], tardó sólo 24 segundos en realizar el alineamiento en el equipo donde se ejecutó el programa.

Fichero alineamiento	Entropía
Rcoffee.fa	72,24
clustalo.fa	87,22
clustaloweb.fa	75,09
muscle.fa	75,98
mafftdefault.aln	20,65
mafftFFTINS.aln	15,21
mafftFFTINSGapdef.aln	21,89
mafftGINS.aln	20,39
mafftGINSGAP3.aln	22,51
mafftLINS.aln	17,40

Tabla 2: Entropías. Los ficheros siguen el mismo orden en el que están las líneas de ejecución del apartado 3.4.1

Agrupamiento

A partir del fichero de secuencias alineadas procedentes de MAFFT-NS-i, se ha procedido a realizar el agrupamiento. En primer lugar, con el fin de valorar el número de grupos en los que clasificar el conjunto de datos se ha realizado un *clustering* iterativo k-medoides para un número de grupos (K) de 2 a 10, pues se buscaba un número reducido de éstos con sus elementos lo mejor clasificados posible.

El resultado (Figura 5) muestra como aparecen máximos de la función Silhouette en $K=2,4,9$. En pruebas posteriores con escalado multidimensional (MDS) una $K=9$ ha mostrado generar grupos más homogéneos que con 2 o 4 grupos, por lo que se ha escogido este.

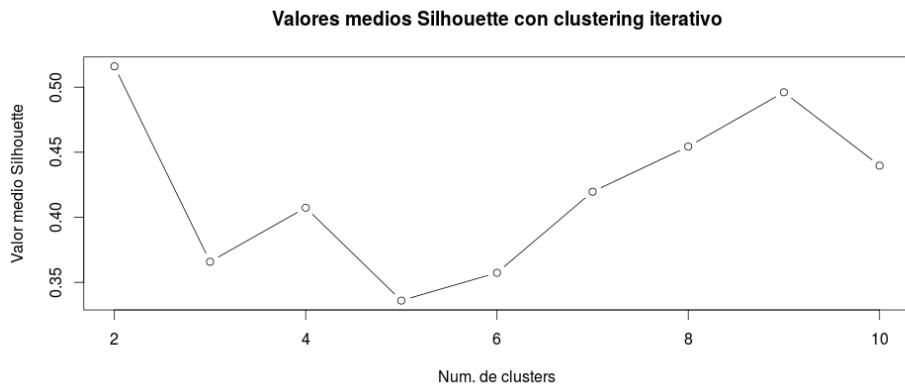


Figura 5: Valores *Silhouette* para distintos valores de K

En la siguiente gráfica (Figura 6) se muestra un dendrograma con los grupos formados al cortar para un valor de $K=9$.

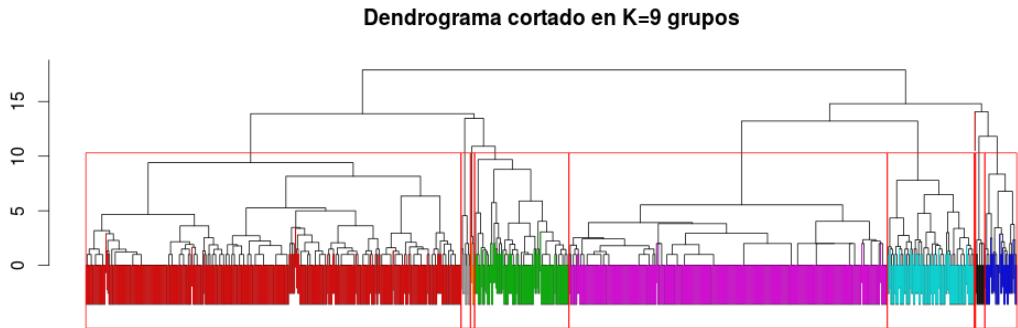


Figura 6: Dendrograma cortado en K=9 grupos

Para poder representar el conjunto de datos en un espacio de reducidas dimensiones (2-3), y valorar visualmente el agrupamiento [46, 47, 20, 48], tal como se menciona en el apartado 3.4.3, se ha realizado un escalado multidimensional (MDS) o coordenadas principales. Del resultado del análisis se obtiene que las dos primeras componentes supone una bondad de representación del 95,56% y las tres primeras un 97,92%.

En la Figura 7 (a) la distribución espacial de los datos muestra que se han creado grupos claramente diferenciados, excepto algunos puntos en el extremo inferior izquierdo, donde parece estar mezclados con los de otro grupo. Para comprobar si esos puntos “anómalos” están realmente mal agrupados, o por el contrario están en otro plano, se ha realizado también una representación de las tres primeras componentes (Figura 7 (b)), donde puede observarse que esos datos están en un plano posterior (puntos de color rojo) y no mezclados con otro grupo.

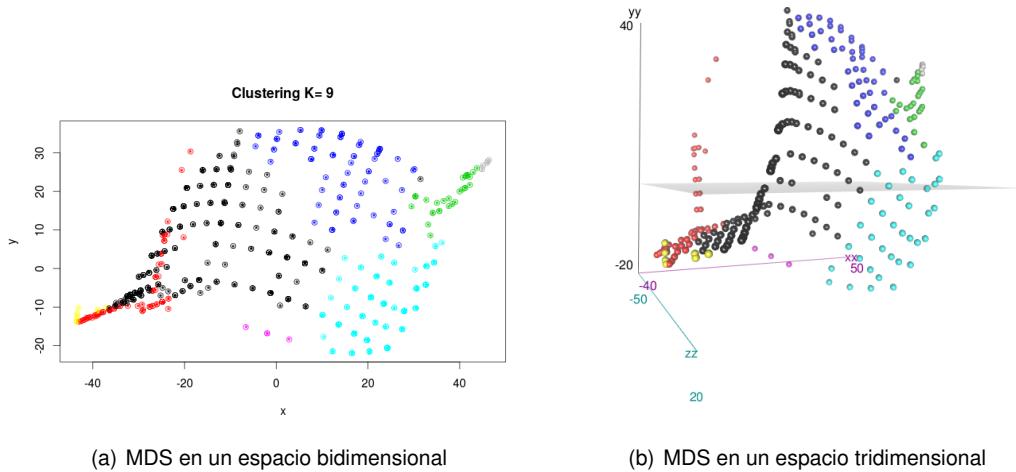


Figura 7: Escalado multidimensional. En (a) se observan algunos puntos superpuestos entre dos grupos. La representación en tres dimensiones muestra que no hay solapamiento, sino que están en un plano posterior.

Una vez establecidos los grupos, se extraen las secuencias originales (sin alinear) pertenecientes a cada grupo en tantos ficheros como grupos para, a continuación, alinear los elementos de cada uno

de ellos. Con esto se pretende obtener alineamientos de mejor calidad, al emplear menos secuencias y más similares que en el primer alineamiento realizado, ejecutando para cada caso el alineador que mejor resultado había dado en el proceso anterior. El objetivo final de este proceso es obtener secuencias consenso de cada grupo, lo más representativas posible de las secuencias originales.

Los nuevos alineamientos han sido evaluados con la función de mínima entropía encontrándose que los grupos 2 y 3 han arrojado valores similares o superiores (15,05 y 19,80) que el alineamiento original, por lo que estos dos ficheros se han sometido nuevamente a agrupamiento, dando como resultado cuatro subgrupos para el grupo 2 y tres para el grupo 3.

Secuencias consenso

Del alineamiento obtenido de cada grupo, se ha calculado la secuencia consenso, donde el nucleótido en la posición i se determina como el más abundante en toda la columna i del alineamiento.

En la siguiente figura (8), se muestran los histogramas de frecuencia de los nucleótidos para las secuencias consenso de todos los grupos que tienen más de una secuencia.

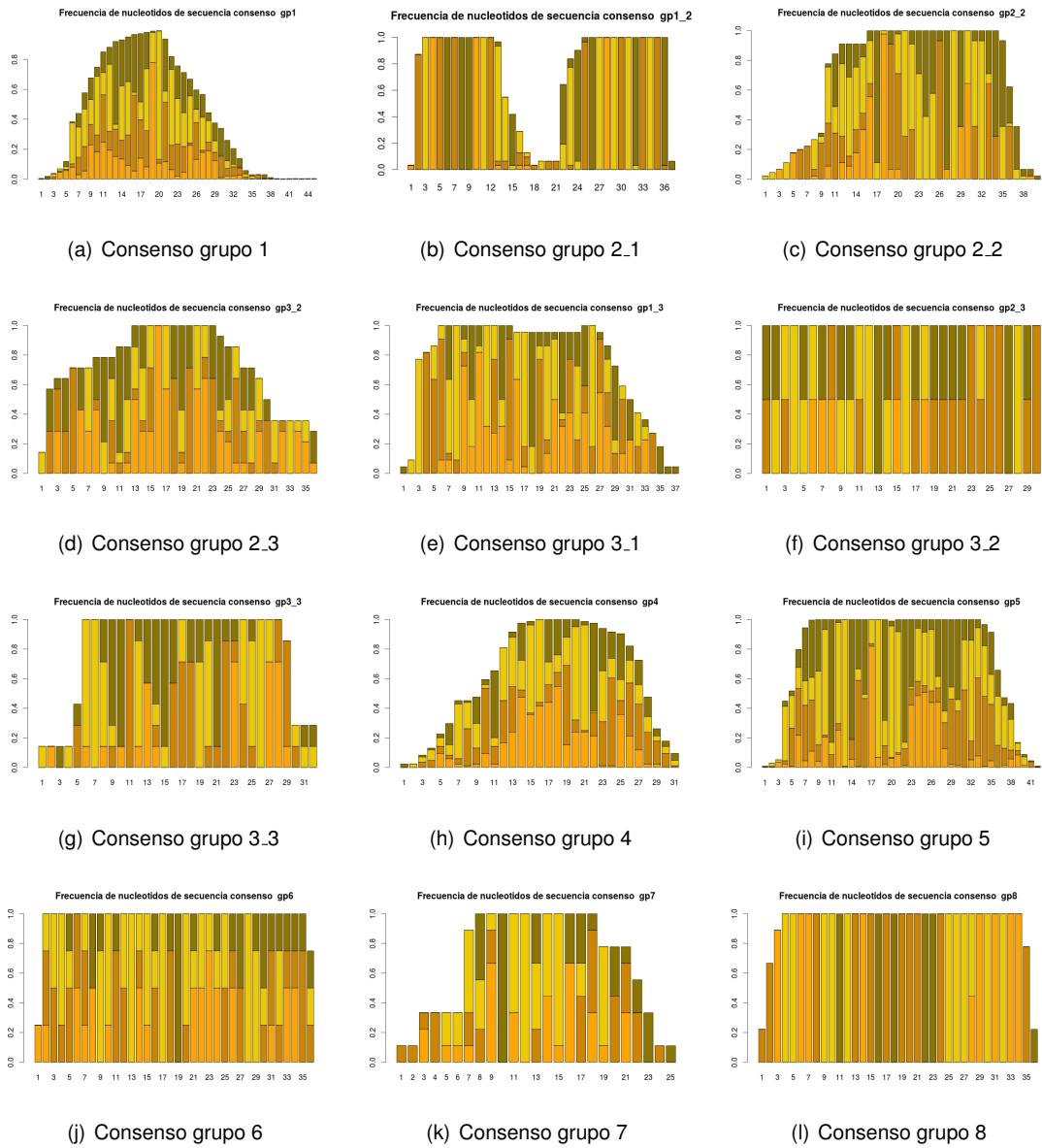


Figura 8: Histogramas de frecuencia de nucleótidos para los distintos agrupamientos. Por los mencionados motivos de propiedad intelectual, se han codificado los nucleótidos por colores, cambiando la asignación de color de unas gráficas a otras.

4.3. Bloque III. Resultados de la predicción estructural

El anexo A recoge las estructuras secundarias obtenidas en función del mínimo de energía libre. A modo de ejemplo en la figura 9.a) se muestra el resultado de predicción de estructura secundaria correspondiente a la secuencia con mayor número de lecturas de las obtenidas experimentalmente. La tabla 3 recoge los valores de las energías libres de cada estructura creada. Valores más bajos (gp5, gp8, gp2_1, gp13, sec669) indican conformaciones más estables, lo que habría que tener en cuenta en el momento de trasladar estas pruebas al laboratorio, en caso de que alguna de estas estructuras arrojase un resultado positivo al realizar el *docking*. En la figura 9(b) se muestra la estructura terciaria correspondiente a la mencionada secuencia con mayor número de lecturas, obtenida

con Assemble2. Todas las estructuras terciarias están recogidas en el anexo A.

Estructura	E. libre (Kcal/mol)	Estructura	E. libre (Kcal/mol)
gp1	-7,70	gp3.3	-8,90
gp4	-6,50	sec1	-5,50
gp5	-11,70	sec20	-0,60
gp6	-7,00	sec3	-5,50
gp7	-6,00	sec308	-8,30
gp8	-14,00	sec32	-1,90
gp9sec	-0,70	sec35	-13,40
gp1_2	-13,70	sec2	-9,20
gp2_2	-8,90	sec41	-3,70
gp3_2	-6,80	sec179	-4,40
gp4_2sec	-5,80	sec93	-7,30
gp1_3	-5,50	sec669	-14,80
gp2_3	-6,50	sec84	-1,20

Tabla 3: Valores de energía libre resultante de las predicciones de estructuras 2D

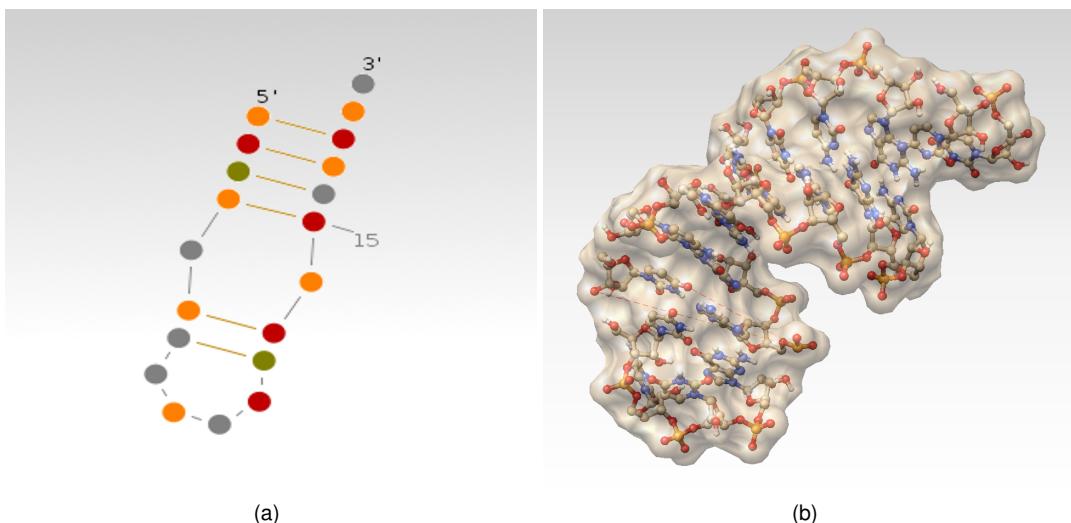


Figura 9: Estructura secundaria (a) (predicción RNAFold) y estructura terciaria (b) (Predicción Assemble2).

Como resultado de la simulación del complejo sncRNA-AT, se han seleccionado las cinco conformaciones con mejor puntuación calculadas por PatchDock para cada secuencia, obteniendo un total de 130 estructuras posibles. Todas ellas han sido examinadas con Chimera añadiendo la heparina al complejo para comprobar si efectivamente el sncRNA se ha unido a la AT en el sitio de unión de la heparina o con la suficiente proximidad a éste como para interferir en la unión del pentasacárido. En la figura 10, se muestran distintos ejemplos de *docking* con resultado claramente positivo, muy próximo al sitio o en zona totalmente opuesta al sitio de unión de la heparina, así como una muestra de la AT-heparina.

De las 26 secuencias, en 15 de ellas hubo una interferencia clara en el sitio de unión de la heparina, 6 se unieron con la suficiente cercanía como para poder interferir en la unión de la heparina, y 5 se unieron en la parte totalmente opuesta al sitio de unión objetivo.

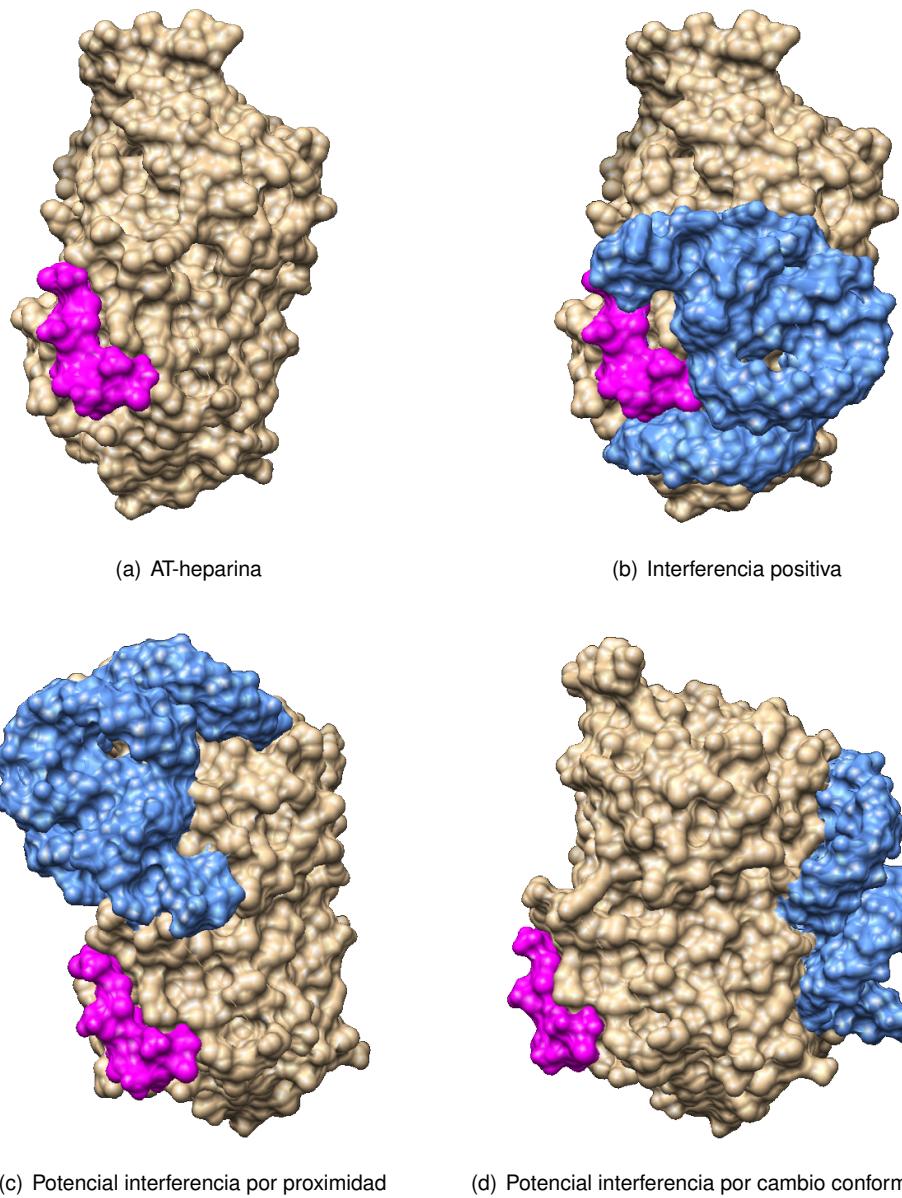


Figura 10: Representación tridimensional de *docking* en diferentes situaciones. La heparina es representada con color magenta y el RNA con color azul.

5. CONCLUSIONES

En este trabajo se han alcanzado los objetivos propuestos.

En primer lugar se ha logrado realizar un análisis cuantitativo y funcional de las secuencias objeto de estudio, desarrollando para ello una serie de *scripts* en distintos entornos y lenguajes.

En segundo lugar, se ha logrado obtener grupos de secuencias estructuralmente homogéneas empleando para ello herramientas y técnicas bioinformáticas y estadísticas, que han servido posteriormente para obtener un reducido número de secuencias con las que poder predecir la posible interacción con la AT.

Por último, se evidencia de forma *in silico* a través de este trabajo, la potencial capacidad de unión entre los sncRNA de pequeño tamaño con la AT en el sitio de unión de la heparina o en su proximidad, lo que sin duda, debe afectar a la capacidad de unión del cofactor en mayor o menor grado. De esta forma podrían actuar como antídotos de heparinas de bajo peso molecular y pentasacáridos, antídotos no disponibles en la actualidad. No obstante, algunas de estas moléculas de RNA también podrían causar la activación de la AT, como lo hace la heparina, pudiendo actuar como nuevos anti-coagulantes sin los efectos adversos de las heparinas.

A nivel operativo, los objetivos que habían sido establecidos, han sido también alcanzados.

- Desarrollo de herramientas en el entorno Galaxy, específicas para este trabajo.
- Manejo de diferentes herramientas bioinformáticas (R, Galaxy, T-Coffee, MUSCLE, MAFFT, ClustalO, BLAST, RNAFold, Assemble2, Chimera, PatchDock).
- Programación de *scripts* para automatizar o desempeñar determinadas tareas.
- Implementar en el lenguaje R el algoritmo de mínima entropía de Shannon específico para evaluación de alineamientos.

Se ha establecido, por tanto, una metodología, diseñada específicamente para el análisis de secuencias de sncRNA y predicción de sus posibles interacciones con proteínas, en este caso la AT.

Finalmente, en base a los resultados obtenidos en el presente estudio, se plantean algunas ideas de mejora y trabajos futuros.

- La principal limitación del estudio, es la derivada de todos los estudio *in silico*; se basa en simulaciones establecidas en condiciones muy específicas donde no conocemos todas las variables biológicas, y por tanto, los resultados obtenidos deben ser validados experimentalmente.
- El estudio se sustenta en la selección de todos los sncRNA que se unen a la AT, sin diferenciar por afinidades. La elución se realiza en un solo paso con NaCl 3M, sin gradiente, por razones tanto metodológicas (para obtener suficiente sncRNA que secuenciar) como económicas (el coste de secuenciación masiva todavía es significativo). Por tanto, hay una importante heterogeneidad en la muestra de partida que podría tener importantes consecuencias funcionales.

- Para alcanzar una mayor precisión en la predicción de las interacciones sncRNA-AT, sería conveniente emplear alguna herramienta que base sus cálculos no sólo en algoritmos estructurales o de complementariedad de formas, sino que también tenga en cuenta interacciones electrostáticas. En este sentido las herramientas examinadas en el transcurso de este trabajo o empleaban un algoritmo estructural o uno electrostático pero no ambos. Así mismo, de las herramientas que se han probado, la mayoría no ha dado resultados satisfactorios, bien por un tiempo de ejecución excesivo (alrededor de 15 h para predecir una estructura terciaria en el caso de Rosetta) o bien por errores de ejecución.
- Desarrollar una aplicación que integre todo el proceso de análisis y predicción.

Anexos

A. Estructuras 2D y 3D de las secuencias

A continuación se muestran las estructuras secundarias y terciarias descritas en la sección 4.3

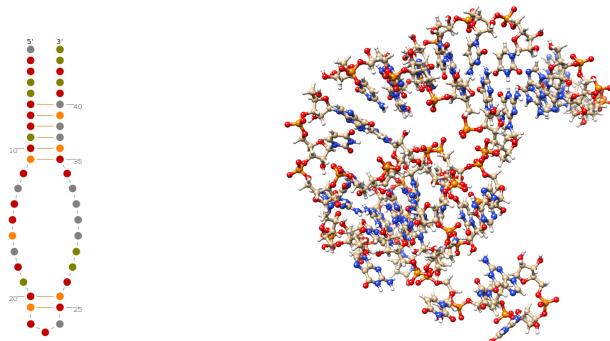


Figura 11: Estructura secundaria (a) y estructura terciaria (b) gp1.

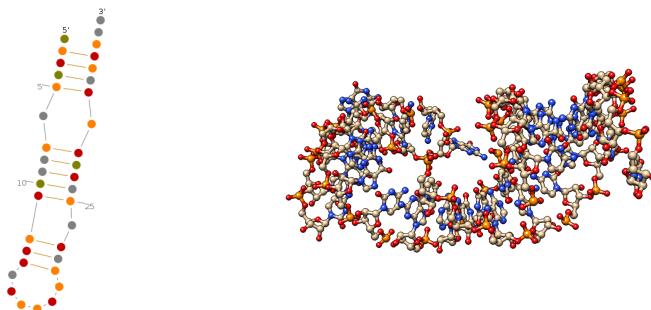


Figura 12: Estructura secundaria (a) y estructura terciaria (b) gp1_2.

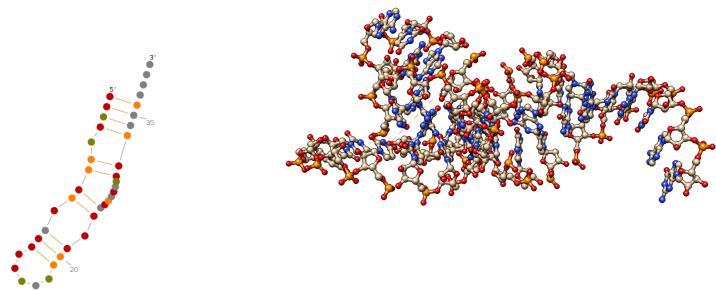


Figura 13: Estructura secundaria (a) y estructura terciaria (b) gp2_2.

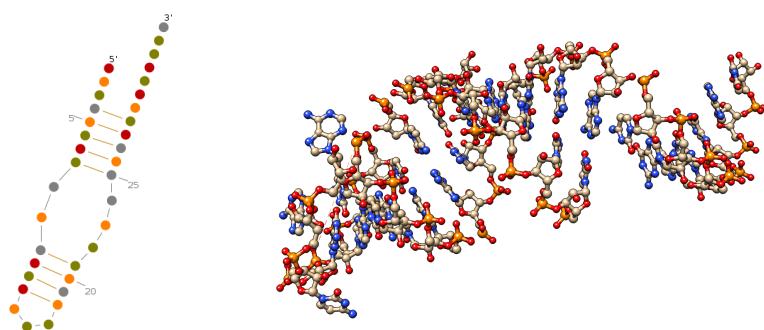


Figura 14: Estructura secundaria (a) y estructura terciaria (b) gp3_2.

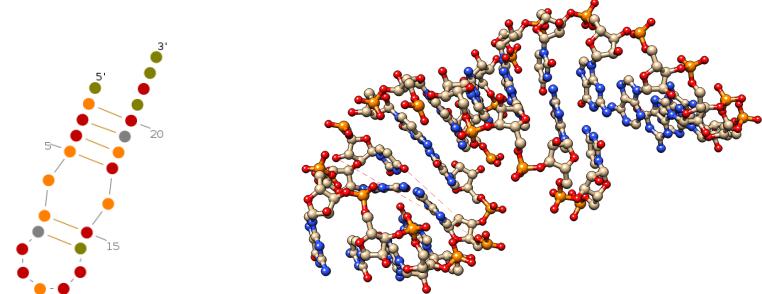


Figura 15: Estructura secundaria (a) y estructura terciaria (b) gp4_2.

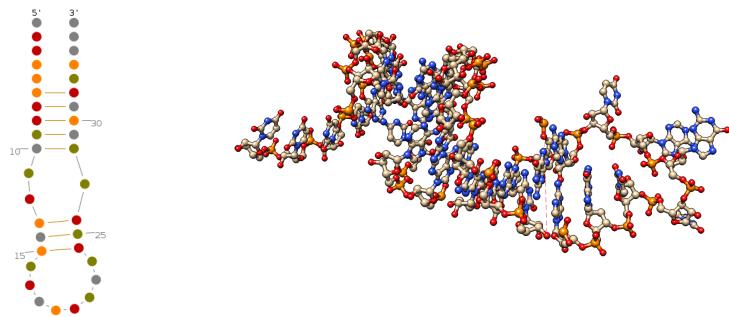


Figura 16: Estructura secundaria (a) y estructura terciaria (b) gp1_3.

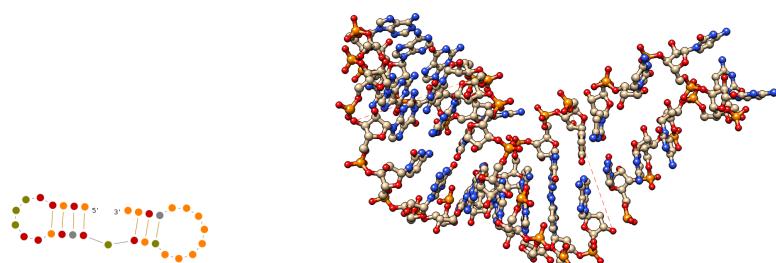


Figura 17: Estructura secundaria (a) y estructura terciaria (b) gp2_3.

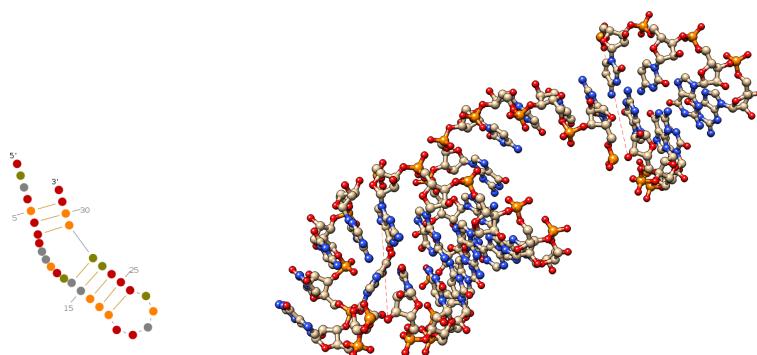


Figura 18: Estructura secundaria (a) y estructura terciaria (b) gp3_3.

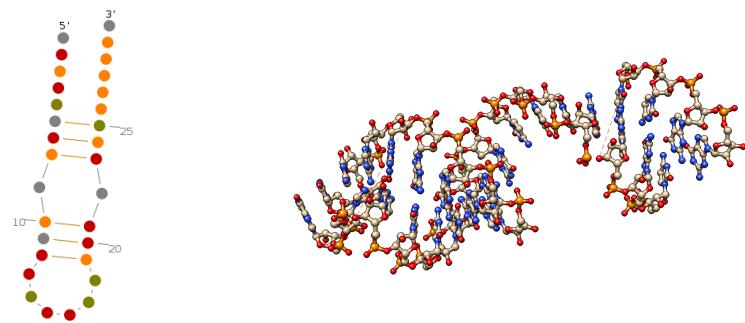


Figura 19: Estructura secundaria (a) y estructura terciaria (b) gp4.

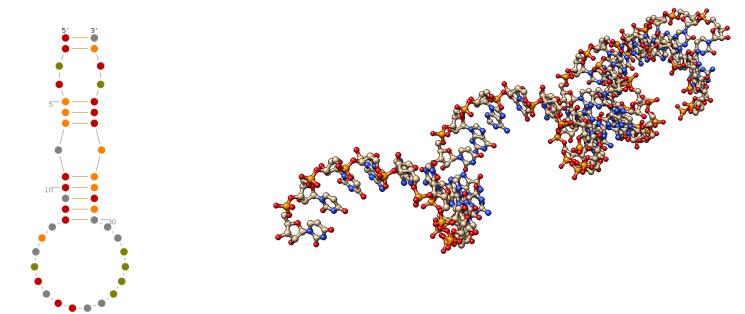


Figura 20: Estructura secundaria (a) y estructura terciaria (b) gp5.

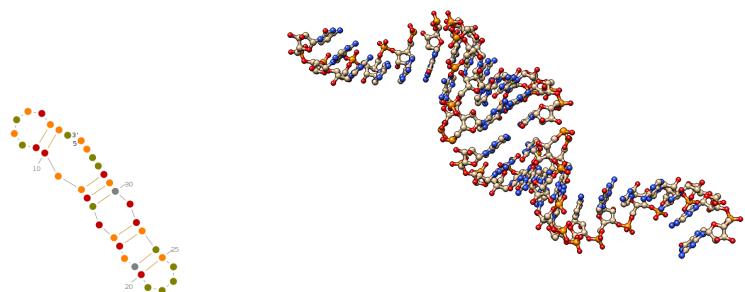


Figura 21: Estructura secundaria (a) y estructura terciaria (b) gp6.

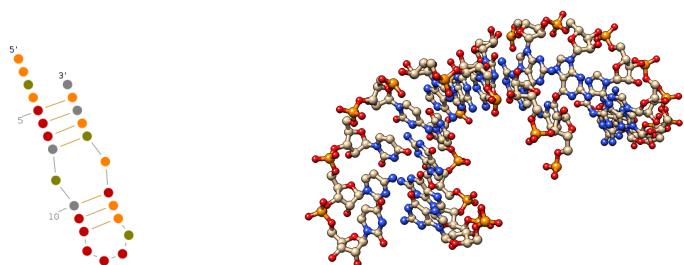


Figura 22: Estructura secundaria (a) y estructura terciaria (b) gp7.

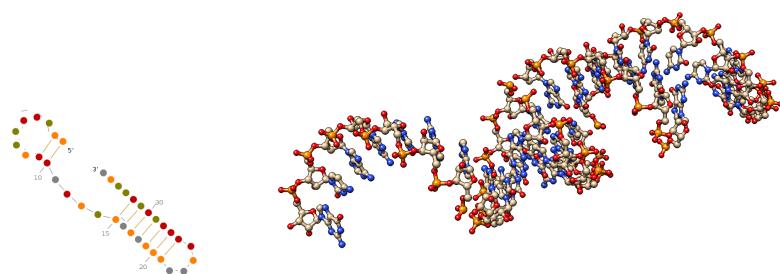


Figura 23: Estructura secundaria (a) y estructura terciaria (b) gp8.

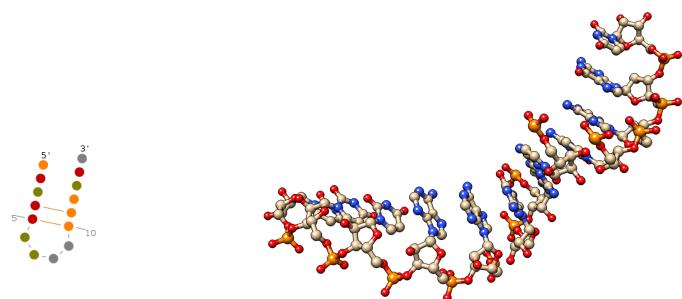


Figura 24: Estructura secundaria (a) y estructura terciaria (b) gp9sec.

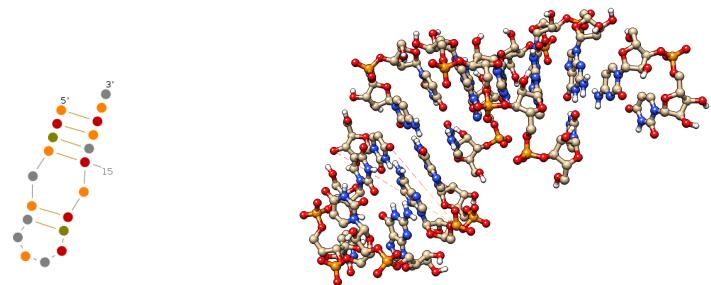


Figura 25: Estructura secundaria (a) y estructura terciaria (b) sec1.

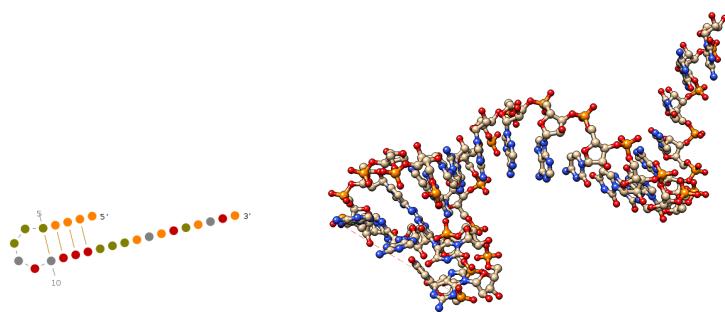


Figura 26: Estructura secundaria (a) y estructura terciaria (b) sec179.

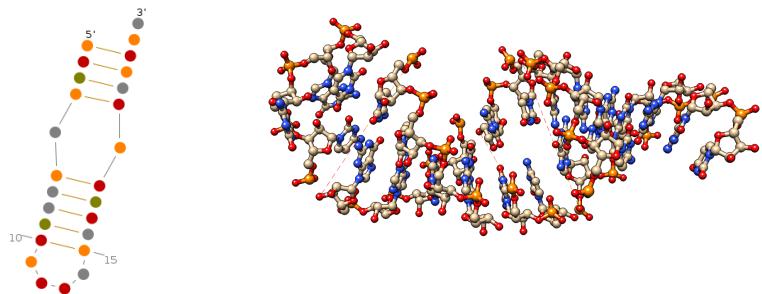


Figura 27: Estructura secundaria (a) y estructura terciaria (b) sec2.

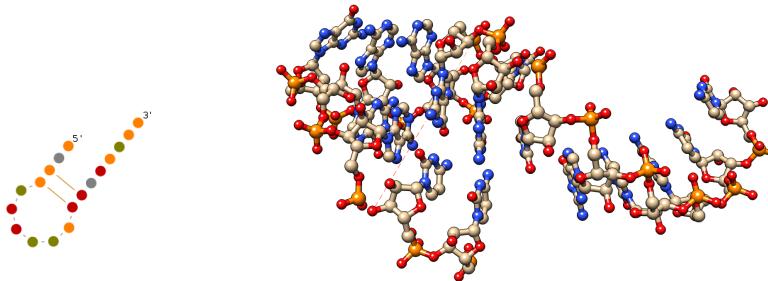


Figura 28: Estructura secundaria (a) y estructura terciaria (b) sec20.

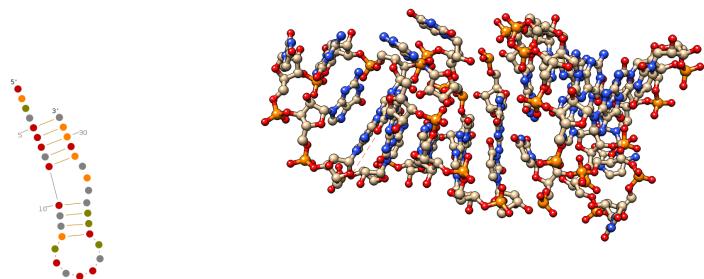


Figura 29: Estructura secundaria (a) y estructura terciaria (b) sec3.

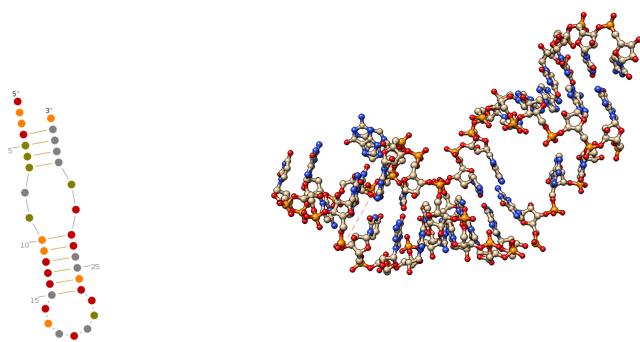


Figura 30: Estructura secundaria (a) y estructura terciaria (b) sec309.

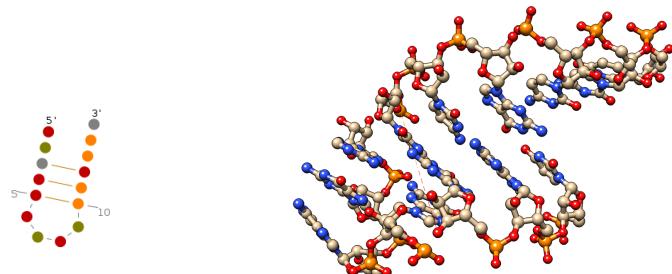


Figura 31: Estructura secundaria (a) y estructura terciaria (b) sec32.

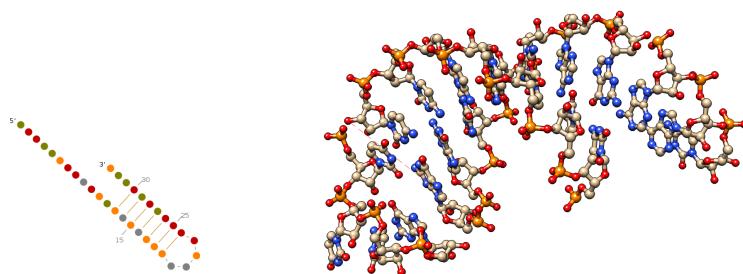


Figura 32: Estructura secundaria (a) y estructura terciaria (b) sec35.

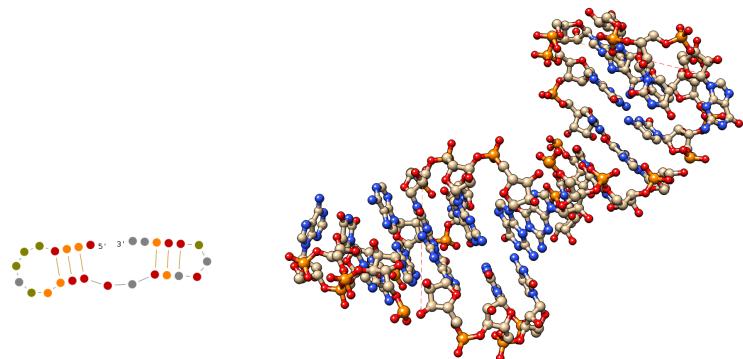


Figura 33: Estructura secundaria (a) y estructura terciaria (b) sec41.

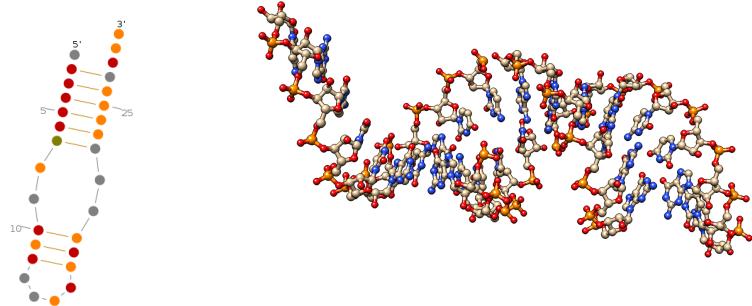


Figura 34: Estructura secundaria (a) y estructura terciaria (b) sec669.

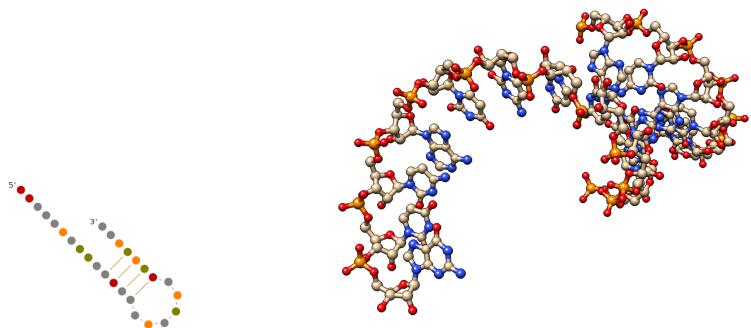


Figura 35: Estructura secundaria (a) y estructura terciaria (b) sec84.

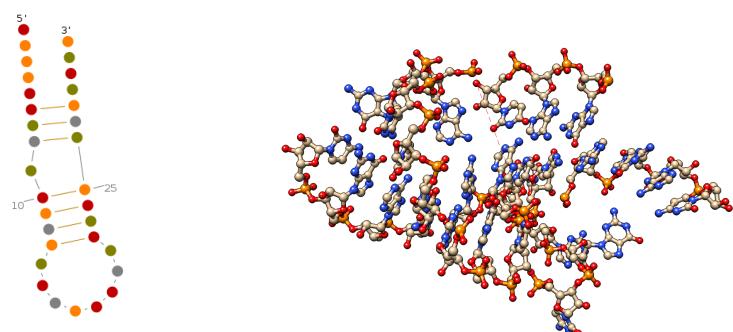


Figura 36: Estructura secundaria (a) y estructura terciaria (b) sec93.

B. Workflow de los *scripts* Galaxy

B.1. Descripción de los *scripts* que componen el *wrapper* Retrieve_info

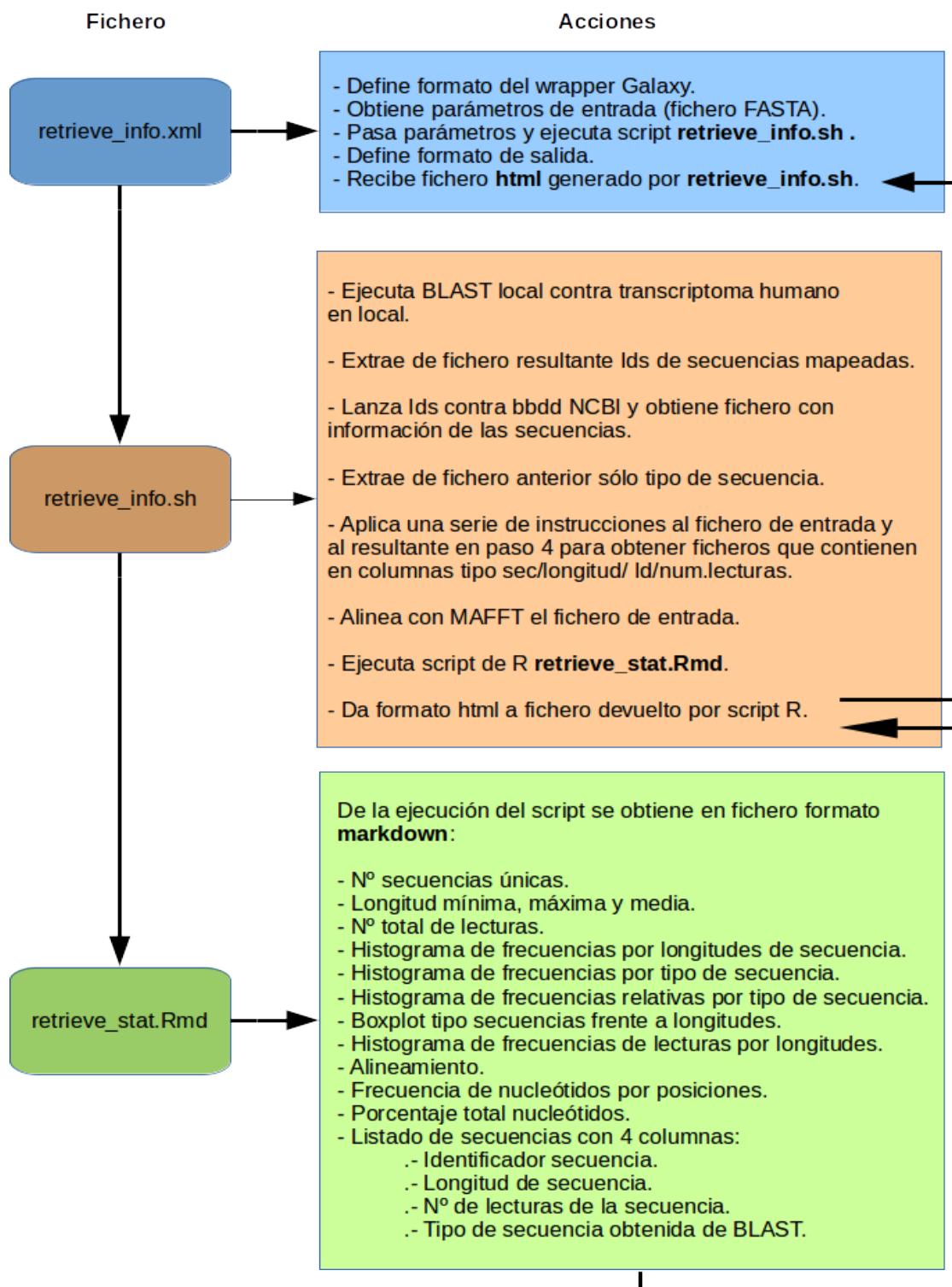


Figura 37: Workflow *scripts* retrieve_info

B.1.1. Capturas pantalla con resultados del wrapper Retrieve_info

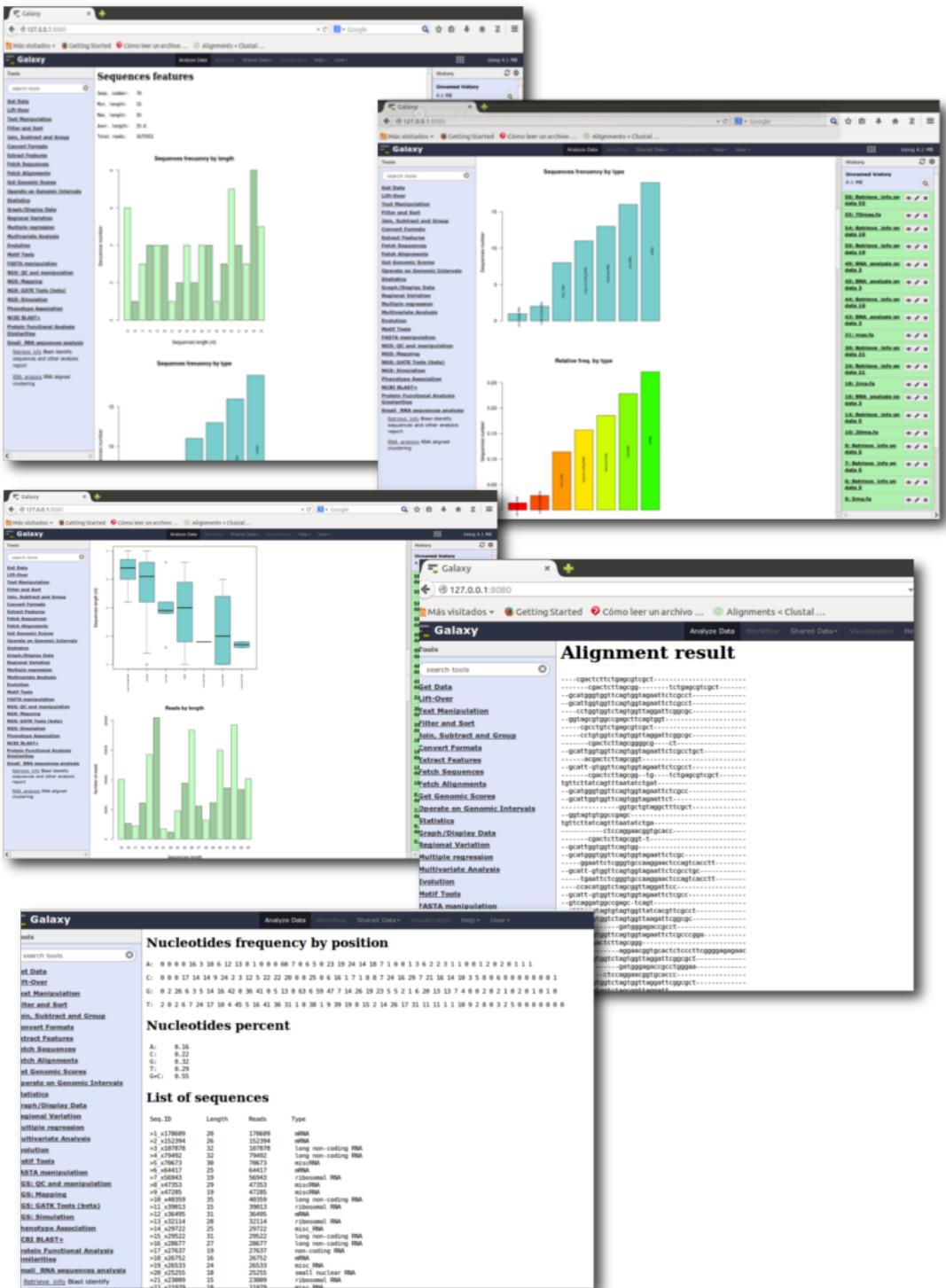


Figura 38: Resultados de retrieve.info

B.2. Descripción de los *scripts* que componen el *wrapper RNA_analysis*

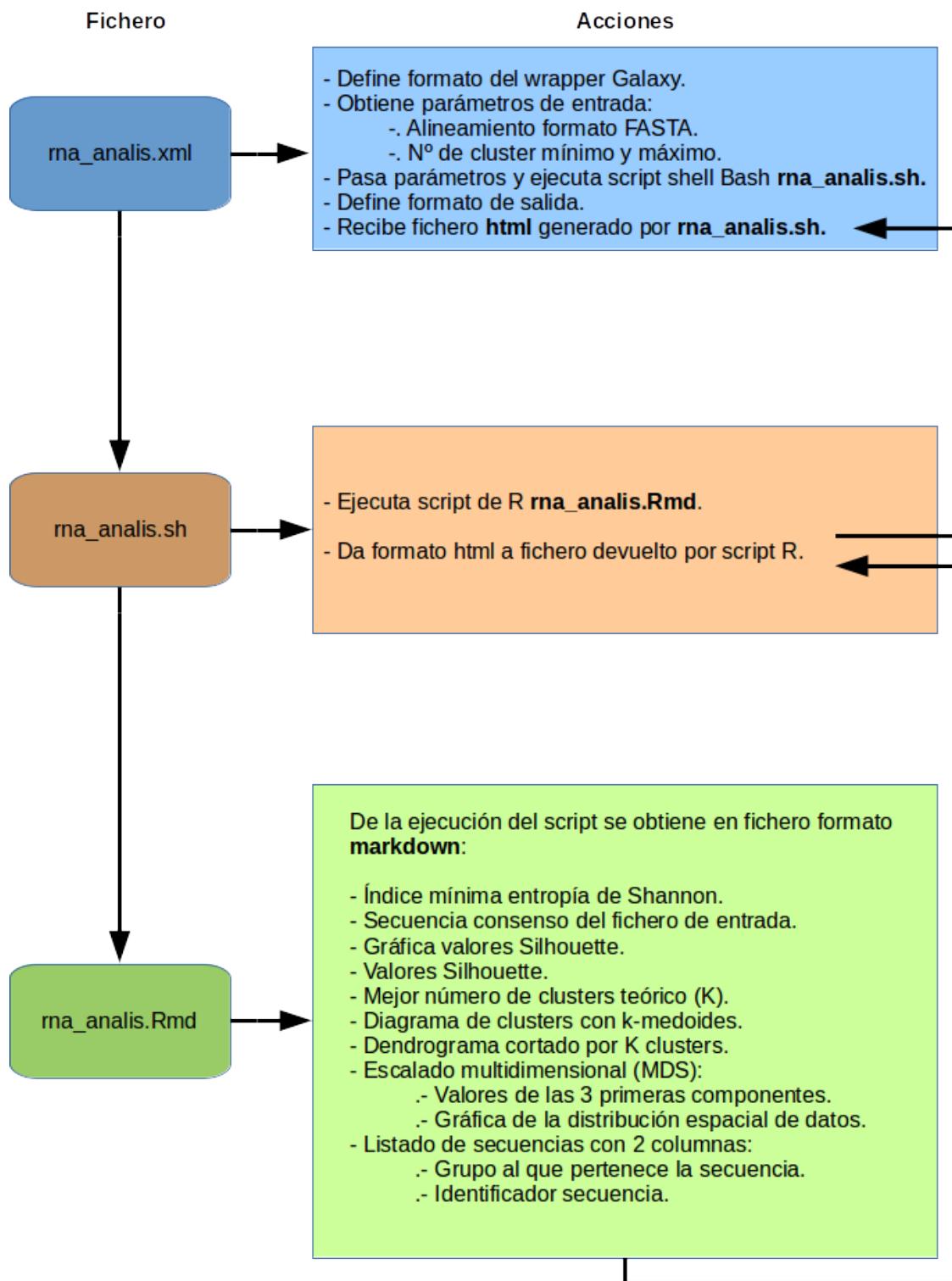


Figura 39: Workflow *scripts rna_analysis*

B.2.1. Capturas pantalla con resultados del wrapper RNA_analysis

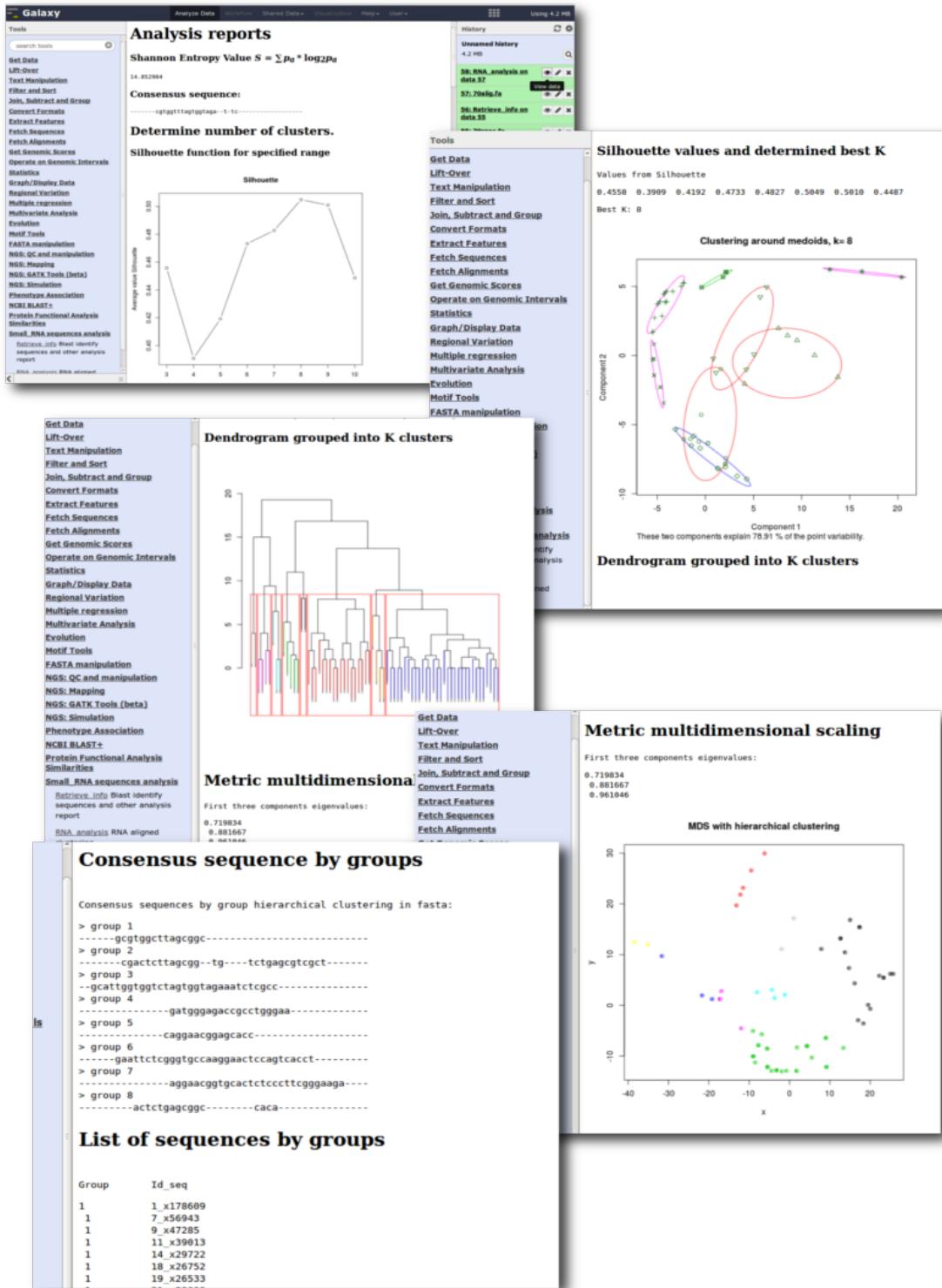


Figura 40: Resultados de RNA_analysis

Bibliografía

- [1] Ewan A. Gibb, Carolyn J. Brown, and Wan L. Lam. The functional role of long non-coding RNA in human carcinomas. *Molecular Cancer*, 10(1):38, 2011.
- [2] Willemijn M. Gommans and Eugene Berezikov. Controlling miRNA regulation in disease. In Jian-Bing Fan, editor, *Next-Generation MicroRNA Expression Profiling Technology*, number 822 in Methods in Molecular Biology, pages 1–18. Humana Press, 2012.
- [3] E. Jean Finnegan and Marjori A. Matzke. The small RNA world. *Journal of Cell Science*, 116(23):4689–4693, January 2003.
- [4] Tina Glisovic, Jennifer L. Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–1986, June 2008.
- [5] Carmen M. Livi and Enrico Blanzieri. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. *BMC Bioinformatics*, 15(1):123, 2014.
- [6] Raúl Teruel, Irene Martínez-Martínez, José A. Guerrero, Rocío González-Conejero, María E. de la Morena-Barrio, Salam Salloum-Asfar, Ana B. Arroyo, Sonia Águila, Nuria García-Barberá, Antonia Miñano, Vicente Vicente, Javier Corral, and Constantino Martínez. Control of post-translational modifications in antithrombin during murine post-natal development by miR-200a. *Journal of Biomedical Science*, 20(1):29, 2013.
- [7] Raúl Teruel, Javier Corral, Virginia Pérez-Andreu, Irene Martínez-Martínez, Vicente Vicente, and Constantino Martínez. Potential role of miRNAs in developmental haemostasis. *PLoS ONE*, 6(3):e17648, 2011.
- [8] Irene Martínez-Martínez, José Navarro-Fernández, Alice Østergaard, Ricardo Gutiérrez-Gallego, José Padilla, Nataliya Bohdan, Antonia Miñano, Cristina Pascual, Constantino Martínez, María Eugenia de la Morena-Barrio, Sonia Aguila, Shona Pedersen, Søren Risom Kristensen, Vicente Vicente, and Javier Corral. Amelioration of the severity of heparin-binding antithrombin mutations by posttranslational mosaicism. *Blood*, 120(4):900–904, July 2012.
- [9] Elmar Pruesse, Jörg Peplies, and Frank Oliver Glöckner. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *BMC Bioinformatics*, 28(14):1823–1829, 2012.
- [10] Julie D. Thompson, Frédéric Plewniak, and Olivier Poch. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research*, 27(13):2682–2690, 1999.
- [11] Nadia Essoussi, Khaddouja Boujenfa, and Mohamed Limam. A comparison of MSA tools. *Bio-information*, 2(10):452–455, 2008.
- [12] C. Notredame, D. G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217, 2000.
- [13] Cuadras, C.M. *Nuevos métodos de análisis multivariante*. CMC Editions, 2014.
- [14] Peña, D. *Análisis de datos multivariantes*. MCGRAW-HILL, 2002.

- [15] Wei Chen, Clarence K. Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. A comparison of methods for clustering 16s rRNA sequences into OTUs. *PLoS ONE*, 8(8):e70837, 2013.
- [16] Thomas J. Sharpton, Guillaume Jospin, Dongying Wu, Morgan GI Langille, Katherine S. Pollard, and Jonathan A. Eisen. Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. *BMC Bioinformatics*, 13(1):264, 2012.
- [17] Linxia Wan, Jiandong Ding, Ting Jin, Jihong Guan, and Shuigeng Zhou. Automatically clustering large-scale miRNA sequences: methods and experiments. *BMC Genomics*, 13:S15, 2012.
- [18] Geetha, T. et al. Distance-based k-medoids clustering for gene expression data. *IUP Journal of Information Technology*, 6(2):7, 2010.
- [19] Bo Wang and Michael A. Kennedy. Principal components analysis of protein sequence clusters. *Journal of Structural and Functional Genomics*, 15(1):1–11, 2014.
- [20] Raffaella Piccarreta and Orna Lior. Exploring sequences: a graphical tool based on multi-dimensional scaling. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(1):165–184, 2010.
- [21] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4th edition, 2002.
- [22] Tomasz Puton, Lukasz Kozlowski, Irina Tuszynska, Kristian Rother, and Janusz M. Bujnicki. Computational methods for prediction of protein–RNA interactions. *Journal of Structural Biology*, 179(3):261–268, September 2012.
- [23] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J. Wolfson. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research*, 33(Web Server issue):W363–367, July 2005.
- [24] Sjoerd J. de Vries, Aalt D. J. van Dijk, Mickaël Krzeminski, Mark van Dijk, Aurelien Thureau, Victor Hsu, Tsjerk Wassenaar, and Alexandre M. J. J. Bonvin. HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins: Structure, Function, and Bioinformatics*, 69(4):726–733, December 2007.
- [25] Steven A. Combs, Samuel L. Deluca, Stephanie H. Deluca, Gordon H. Lemmon, David P. Nannemann, Elizabeth D. Nguyen, Jordan R. Willis, Jonathan H. Sheehan, and Jens Meiler. Small-molecule ligand docking into comparative models with rosetta. *Nature Protocols*, 8(7):1277–1298, 2013.
- [26] Yunjie Zhao, Yangyu Huang, Zhou Gong, Yanjie Wang, Jianfen Man, and Yi Xiao. Automated and fast building of three-dimensional RNA structures. *Scientific Reports*, 2, October 2012.
- [27] Federico Agostini, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. catRAPID omics: a web server for large-scale prediction of protein–RNA interactions. *Bioinformatics (Oxford, England)*, 29(22):2928–2930, November 2013.
- [28] Usha K. Muppirala, Vasant G. Honavar, and Drena Dobbs. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, 12(1):489, December 2011.
- [29] K Usha. Computational tools for investigating RNA-protein interaction partners. *Journal of Computer Science & Systems Biology*, 06(04), 2013.

- [30] Lian Yi Han, Cong Zhong Cai, Siew Lin Lo, Maxey C.M. Chung, and Yu Zong Chen. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, 10(3):355–368, March 2004.
- [31] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
- [32] Daniel Blankenberg, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. Galaxy: A web-based genome analysis tool for experimentalists. In *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc., 2001.
- [33] Jeremy Goecks, Anton Nekrutenko, James Taylor, and \\$author.firstName \\$author.lastName. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, 2010.
- [34] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
- [35] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.
- [36] Andreas Wilm, Desmond G. Higgins, and Cédric Notredame. R-coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Research*, 36(9):e52, 2004.
- [37] Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1), 2011.
- [38] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 2002.
- [39] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 2013.
- [40] Claude Elwood Shannon and Warren Weaver. *The mathematical theory of communication*. University of Illinois Press, 1963.
- [41] William S.J. Valdar. Scoring residue conservation. *Proteins*, 48(2):227–241, 2002.
- [42] Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, SIGMOD '98, page 73–84. ACM, 1998.
- [43] Santhanam T. Velmurugan T. A survey of partition based clustering algorithms in data mining: An experimental approach. *Information Technology Journal*, 2011.
- [44] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

- [45] Dan Wei, Qingshan Jiang, Yanjie Wei, and Shengrui Wang. A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics*, 13:174, 2012.
- [46] J. A. Costa, A. Tenreiro Machado and Maria Dulce Quelhas. Multidimensional scaling applied to histogram-based DNA analysis. *International Journal of Genomics*, 2012:e289694, 2012.
- [47] Adam Hughes, Yang Ruan, Saliya Ekanayake, Seung-Hee Bae, Qunfeng Dong, Mina Rho, Judy Qiu, and Geoffrey Fox. Interpolative multidimensional scaling techniques for the identification of clusters in very large sequence sets. *BMC Bioinformatics*, 13:S9, 2012.
- [48] Jiwoong Kim, Yongju Ahn, Kichan Lee, Sung Hee Park, and Sangsoo Kim. A classification approach for genotyping viral sequences based on multidimensional scaling and linear discriminant analysis. *BMC Bioinformatics*, 11:434, 2010.
- [49] Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck, and Ivo L. Hofacker. The vienna RNA websuite. *Nucleic Acids Research*, 36:W70–W74, 2008.
- [50] Fabrice Jossinet, Thomas E. Ludwig, and Eric Westhof. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2d and 3d levels. *BMC Bioinformatics*, 26(16):2057–2059, 2010.
- [51] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.
- [52] Virpi Ahola, Tero Aittokallio, Mauno Vihinen, and Esa Uusipaikka. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics*, 7(1):484, 2006.
- [53] D. Charif and J.R. Lobry. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U. Bastolla, M. Porto, H.E. Roman, and M. Vendruscolo, editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York, 2007. ISBN : 978-3-540-35305-8.
- [54] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011.
- [55] Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.
- [56] E. Paradis, J. Claude, and K. Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.
- [57] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. ISBN 978-0-387-75968-5.
- [58] K.P. Schliep. Phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.
- [59] Hadley Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011.
- [60] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.