UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
DEPARTAMENTO DE INGENIERIA QUIMICA Y BIOTECNOLOGIA

PREDICCIÓN COMPUTACIONAL DE GENES PEQUEÑOS DE ARN NO CODIFICANTE EN GENOMAS BACTERIANOS

TESIS PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN BIOTECNOLOGÍA Y MAGÍSTER EN CIENCIAS DE LA INGENIERÍA MENCIÓN QUÍMICA

ANDRÉS EDUARDO ROGERS AHUMADA

PROFESOR GUÍA:
JUAN A. ASENJO DE L.

MIEMBROS DE LA COMISIÓN:
GONZALO NAVARRO BADINO
MAURICIO GONZALEZ CANALES
IVAN RAPAPORT ZIMERMANN
ORIANA SALAZAR AGUIRRE

SANTIAGO DE CHILE
ABRIL 2012

Resumen

El objetivo de este estudio es desarrollar un método computacional capaz de predecir con alta especificidad y sensibilidad los genes pequeños de ARN no codificante en genomas bacterianos, identificando las variables involucradas de mayor importancia que deben ser consideradas para su correcta clasificación.

El trabajo aquí presentado consistió en investigar y analizar el estado del arte en métodos de predicción computacional de genes pequeños de ARN no codificante en bacterias, recopilar un listado de las variables involucradas y determinar estadísticamente aquellas que diferencian con mayor precisión los genes pequeños de ARN no codificante de secuencias genéticas al azar, comparar distintos métodos de predicción e identificar el que otorgue mejores resultados. Finalmente, se compararon los resultados del método con otros preexistentes y se aplicó el método al genoma completo de *Escherichia coli*.

Los principales resultados obtenidos en este estudio son la identificación de 4 variables que influyen significativamente en la detección correcta de genes pequeños de ARN no codificante. Estas son: $Valor\ z$, $Valor\ de\ partición$, EMP_I y $Porcentaje\ de\ bultos$, las cuales corresponden al subconjunto de variables con mayor capacidad discriminatoria. Por este motivo se recomienda que futuros métodos predictivos consideren la inclusión de estas 4 variables.

Las variables seleccionadas muestran que existe una presión selectiva en la evolución de los genes pequeños de ARN no codificante, la que apunta a aumentar la estabilidad de la molécula al modificar su estructura para disminuir la energía de plegamiento y eliminar subestructuras desestabilizantes no funcionales.

El mejor método de clasificación corresponde al *Perceptrón Multicapa* basado en redes neuronales, con una alta sensibilidad ($88,8\,\%$) y alta especificidad ($85,5\,\%$), teniendo una tasa de falsos positivos relativamente baja ($14,5\,\%$).

Con este subconjunto de variables y el método de clasificación, se realizó una predicción sobre el genoma de la bacteria *Escherichia coli*, generando 1192 predicciones, con un valor de sensibilidad de 30,5 % y un valor predictivo positivo de 1,51 % respecto a los genes pequeños de ARN no codificante conocidos.

Seleccionando las predicciones cercanas a promotores σ^{70} o terminadores intrínsecos (independientes del factor ρ), se obtiene un desempeño predictivo similar al logrado por otros autores en la literatura, con el beneficio adicional de requerir la medición de sólo 4 variables y sin la necesidad de información sobre genes homólogos en organismos cercanos filogenéticamente.

La contribución de este trabajo consiste en profundizar el conocimiento acerca de las características de los genes de ARN no codificante, al haber estudiado las variables utilizadas previamente en la literatura y definir las 4 más relevantes, las cuales se relacionan directamente con la estructura secundaria y su energía mínima de plegamiento.

En segundo lugar se propone un listado de 1192 secuencias del genoma de *Escherichia coli* y un listado más corto con las 5 más probables de ser genes sARN, estas secuencias pueden ser comprobadas experimentalmente.

Estos resultados inciden positivamente en mejorar la calidad de las anotaciones de estos genes en genomas bacterianos, permitiendo mayores avances en estudios de genómica funcional y regulación en redes metabólicas.

Abstract

Small non-coding RNA genes produce a class of newly discovered RNA molecules with catalytic functions within the cells that act as an important part of regulatory networks. Current efforts in the field have created computational methods for genomic prediction, but these methods have encountered many difficulties not found in the prediction of protein-coding genes.

The current work consisted in analyzing the state-of-the-art of the methods of computational prediction of small non-coding RNA genes in bacteria, studying the variables involved and the identification of the ones that are statistically more relevant for differentiating small non-coding RNA genes from random sequences. Finally, comparing the results of the method with previous ones and applying this method on the genome of Escherichia coli.

The main results are the identification of 4 variables with the most significant classifying power: *Z-score*, *Partition value*, MFE_I and *Bulge percentage*.

These variables are related to secondary structure and minimum folding energy of the RNA molecule and show an evolutionary trait of this type of gene, which is to improve the stability of the RNA molecule by minimizing the folding energy and eliminating destabilizing substructures like bulges.

The method with the best classification power was the multi-layer perceptron, showing a high sensibility (88,8 %) and specificity (85,5 %), while maintaining a relatively low false positive rate (14,5 %).

The prediction on the whole genome of Escherichia coli identified 1192 small non-coding RNA genes with sensibility of 30,5 %, according to the known sequences, and a positive predictive value of 1,51 %. Both of these values are similar with the performance of current methods in the field, with the additional benefit of requiring only 4 variables and without the need of knowledge of homologous genes from closely related genomes.

The main contribution of this work is a better the understanding of the most important characteristics of small non-coding RNA genes for their computational prediction, allowing the improvement of the quality of gene annotations and reducing the computing time.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Este trabajo consiste en desarrollar un método computacional que permita predecir eficientemente las secuencias de genes pequeños de ARN no codificante en genomas bacterianos.

La característica de estos genes es que ellos no se traducen en proteínas, sino que se transcriben en moléculas de ARN con actividad regulatoria (Vogel y Sharma, 2005).

Su existencia es conocida desde el año 1981, sin embargo no habían recibido mayor atención hasta que estudios recientes han descubierto su rol en la regulación de una diversa gama de procesos celulares (Eddy, 1999; Argaman et al., 2001; Vogel y Bartels, 2003), motivo por el cual la predicción computacional de genes se ha centrado hasta el momento en la identificación de genes con expresión proteica (Waters y Storz, 2009).

El descubrimiento de este transcriptoma oculto añade un nivel de mayor complejidad a la discriminación entre ARN codificante y no codificante, presentando un desafío al entendimiento de la expresión y regulación de la información genética (Dinger et al., 2008).

Si bien existen métodos computacionales para su predicción, éstos se basan fuertemente en homología de secuencias por lo que su aplicación se limita a genomas de organismos cercanos filogenéticamente y utilizan un algoritmo computacional seleccionado a priori (Vogel y Sharma, 2005; Tran, 2009).

El enfoque de este estudio será recopilar la mayor cantidad posible de variables medidas por métodos computacionales preexistentes y seleccionar estadísticamente aquellas con mayor poder de discriminación para identificar secuencias de genes sARN. Luego, se aplicarán y compararán distintos algoritmos de clasificación, encontrando la combinación óptima entre variables medidas y algoritmo computacional.

La contribución de este trabajo será la determinación de un método computacional predictivo eficiente y aplicable a una mayor variedad de genomas bacterianos, para encontrar secuencias conocidas y proponer genes nóveles. Además se profundizará el conocimiento sobre las características y relevancia de distintas variables que discriminen eficientemente a los genes de ARN no codificante pequeño.

1.1. Dogma central de la biología molecular

La información presente en el código genético está representada por secuencias de los nucleótidos Adenina (A), Citocina (C), Guanina (G) y Tiamina (T) presentes en el ADN de cada organismo. Existen subunidades en el ADN llamadas genes que codifican funciones específicas.

Los procesos básicos de flujo de información consisten en la copia de moléculas de ADN llamada replicación, el traspaso de la información desde los genes a moléculas de ARN llamado transcripción y el traspaso de la información contenida en las moléculas de ARN mensajero a proteínas llamado traducción. Esto se ilustra en la Figura 1.1.

Sin embargo se han descubierto una serie de excepciones al dogma, como el paso de ARN a ADN mediante la enzima transcriptasa reversa, la replicación del ARN y el paso directo de ADN a proteínas. Pero estos casos se consideran aislados ya que ocurren sólo en condiciones particulares, ya sea en virus o en laboratorio.

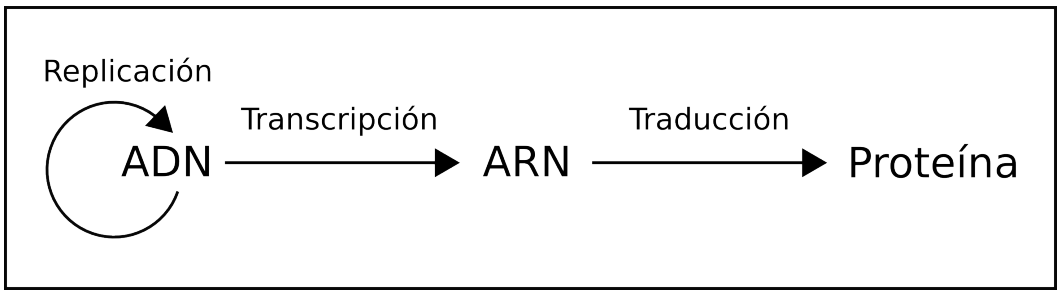


Figura 1.1: El Dogma Central de la Biología Molecular

1.1.1. Expandiendo el dogma

Si bien el dogma central continúa siendo válido en cuanto a la descripción del proceso general de formación de proteínas desde el material genético, en años recientes ha sido evidente que existe una mayor cantidad de información dentro del genoma que la contenida en los genes que codifican proteínas; esta información es la que controla el momento adecuado y la rapidez en el proceso de síntesis de proteínas (Perkins et al., 2005).

En organismos complejos se observa que sólo una minoría (2-3 %) de los transcritos genéticos codifica proteínas, el resto había sido llamado "ADN basura" y se explicaba su existencia como vestigios de genes codificantes perdidos durante la evolución (Perkins et al., 2005).

Sin embargo, estudios recientes demuestran que los transcritos no codificantes juegan un papel crítico en la regulación de la expresión génica a través de mecanismos de control que alteran la estabilidad del ARN mensajero, el inicio de la transcripción de genes o la traducción de proteínas (Perkins et al., 2005).

1.2. Revisión de los genes pequeños de ARN no codificante

1.2.1. Definición

Los genes de ARN no codificante son un tipo de secuencias genómicas, las cuales transcriben moléculas que funcionan directamente como ARN, en vez de traducirse en proteínas (Eddy, 2002), cumpliendo importantes funciones estructurales, catalíticas y regulatorias dentro de la célula (Chen et al., 2002; Mimouni et al., 2009; Rivas et al., 2001).

Ejemplos clásicos son los genes de ARN ribosomal (rARN) y ARN de transferencia (tARN), los cuales son bien conocidos y están presentes en todo tipo de vida celular (Nawrocki y Eddy, 2009). Sin embargo, se conoce la existencia de otros tipos de genes de ARN no codificante con roles estructurales y regulatorios cuyo número e importancia se creía marginal hasta hace poco tiempo (Eddy, 2002).

Estudios recientes (Eddy, 1999; Argaman et al., 2001; Vogel y Bartels, 2003) han encontrado que esta nueva clase de genes de ARN no codificante son abundantes en todos los dominios de la vida, indicando que tienen una antigua historia evolutiva (Gottesman, 2005).

Debido a estos hallazgos, los genes sARN han atraído una atención considerable como una nueva clase de reguladores de la expresión génica en procesos celulares críticos (Nawrocki y Eddy, 2009; Gottesman, 2005; Wassarman et al., 2001; Szymanski et al., 2007; Argaman et al., 2001; Vogel y Sharma, 2005; Kulkarni y Kulkarni, 2007; Dinger et al., 2008).

En los genomas bacterianos, los genes que codifican moléculas de ARN con actividad regulatoria son conocidos como genes pequeños de ARN no codificante, comúnmente llamados sRNAs en la literatura por su nombre en inglés *small non-coding RNAs*, ya que son secuencias cortas con un largo que varía entre 50 y 400 nucleótidos (Vogel y Sharma, 2005).

Para seguir la convención, en adelante estos genes seran referidos como genes sARN, aludiendo específicamente a los genes pequeños de ARN no codificante de origen bacteriano.

En la Figura 1.2 se muestra un esquema simplificado de las clases de ARN en genomas bacterianos.

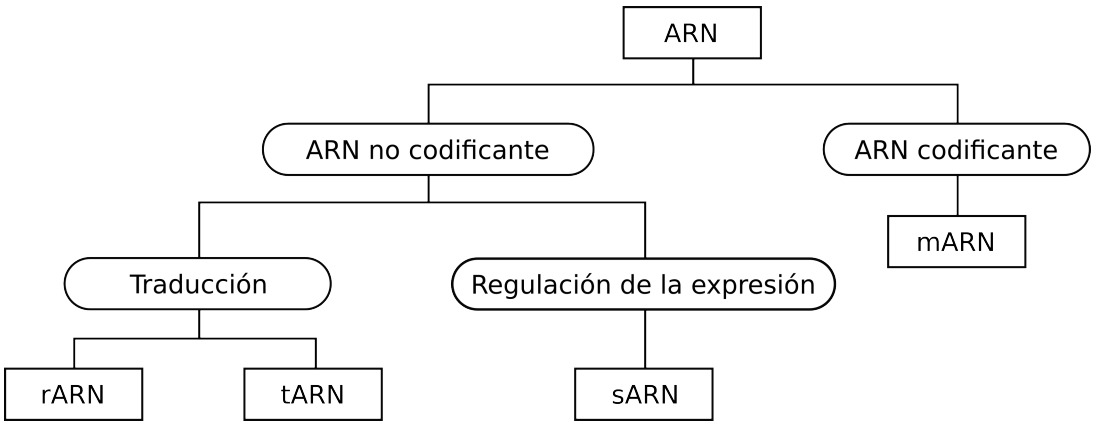


Figura 1.2: Clasificación funcional del ARN en genomas bacterianos

1.2.2. Descubrimiento

Moléculas de ARN con función regulatoria fueron descubiertas por primera vez en 1981, cuando se encontró que un nucleótido de 108 pares de bases nombrado *RNA I* bloqueaba la replicación del plasmidio *ColE1* mediante unión por complementariedad de bases con el partidor de ARN que inicia su replicación (Waters y Storz, 2009).

En 1984 se identificó el gen *MicF* ubicado dentro del cromosoma de *Escherichia coli*, el cual transcribe una molécula de ARN de 174 pares de bases que inhibe la traducción del ARN mensajero que codifica la proteína de membrana *OmpF*. Estos primeros sARNs fueron identificados experimentalmente mediante geles debido a su abundancia, pero su importancia y real contribución no fue totalmente apreciada durante décadas (Waters y Storz, 2009).

1.2.3. Reconocimiento de su importancia

En los años 2001 y 2002, cuatro grupos (Argaman et al., 2001; Chen et al., 2002; Rivas et al., 2001; Wassarman et al., 2001) publicaron la identificación de un gran número de genes sARN encontrados mediante un análisis computacional de conservación de secuencias en regiones intergénicas del genoma de la bacteria *Escherichia coli* (Livny y Waldor, 2007).

Desde entonces la apreciación de la importancia de los genes sARN ha crecido enormemente, debido a que múltiples estudios genómicos han mostrado su diversidad y amplia distribución en una gran variedad de organismos (Hershberg et al., 2003; Wassarman, 2007; Vogel y Sharma, 2005; Gottesman, 2005; Matera et al., 2007).

Actualmente el ARN está emergiendo como el factor clave en la regulación celular y ocupando un área central de investigación en biología molecular, ya que toma roles activos en las múltiples capas de regulación de la expresión génica desde transcripción, maduración de ARN, modificación de ARN y regulación de la traducción (Flamm et al., 2005; Backofen et al., 2007).

Las diversas funciones que desempeña el sARN en el proceso de regulación de la expresión génica lo convierte en un objetivo importante de estudio para el desarrollo de intervenciones terapéuticas (Kaikkonen et al., 2011).

Resulta sorprendente que muchas de las clases de sARN fueran descubiertas recientemente o sus números incrementados en análisis recientes, lo que apunta a la relativa dificultad de descubrir las secuencias de sARN comparado al descubrimiento de genes codificantes de proteínas. Al mismo tiempo esto abre la opción de que existan muchas más clases de genes sARN todavía por descubrir (Nawrocki y Eddy, 2009; Eddy, 2002; Machado-Lima et al., 2008).

Resulta entonces un problema de gran interés el descubrir genes sARN nóveles mediante el análisis computacional de secuencias (Nawrocki y Eddy, 2009; Backofen et al., 2007; Machado-Lima et al., 2008; Pichon y Felden, 2008; Vogel y Sharma, 2005), siendo la identificación computacional de genes sARN es uno de

los problemas más importantes y desafiantes de la biología computacional y bioinformática en la actualidad (Machado-Lima et al., 2008; Tran et al., 2009).

1.2.4. Funciones

Hasta hace poco se pensaba que la expresión génica ocurría principalmente mediante el control de la transcripción mediante proteínas represoras o activadoras. Sin embargo estudios recientes revelan que existen otros mecanismos de regulación basados en moléculas de ARN (Johansson, 2003).

La función general de los genes sARN es la de regular la expresión génica en distintos niveles, ya sea controlando el inicio de la transcripción, modificando la estabilidad del mARN, modulando la actividad de proteínas y regulando la traducción (Mimouni et al., 2009; Wassarman, 2007; Gottesman, 2005; Repoila y Darfeuille, 2009).

Los genes sARN regulan una gran cantidad de procesos biológicos (Livny y Waldor, 2007) permitiendo la adaptación celular en respuesta a cambios en el medio ambiente (Brouns et al., 2008; Pichon y Felden, 2008; Repoila y Darfeuille, 2009).

Algunas de las funciones celulares específicas que se sabe son reguladas por genes sARN son: respuesta a estrés celular, respuesta a cambios de nutrientes en el medio, motilidad celular, homeostasis del carbono, homeostasis de la membrana celular, respuesta *SOS*, producción de toxinas y antitoxinas, coordinación de la patogenicidad (virulencia), detección de quórum (*Quorum Sensing*), control de replicación plasmidial y viral, resistencia adquirida a bacteriófagos (Johansson, 2003; Vogel y Sharma, 2005; Vogel y Papenfort, 2006; Szymanski et al., 2007; Brouns et al., 2008; Pichon y Felden, 2008; Perez et al., 2009; Repoila y Darfeuille, 2009; Voss et al., 2009; Toledo-Arana et al., 2007; Bejerano-Sagie y Xavier, 2007).

1.2.5. Mecanismos de acción

Los genes sARN pueden modular el proceso de transcripción, traducción y modificar la estabilidad del mARN mediante diversos mecanismos (Waters y Storz, 2009).

El principal mecanismo es la unión por complementariedad de bases con un mARN específico, con lo cual se logra bloquear el acceso al sitio de unión a ribosoma o el codón de inicio, impidiendo así la traducción del mARN objetivo (Gottesman, 2005; Perez et al., 2009; Waters y Storz, 2009).

La molécula de ARN de doble hebra resultante de la unión es posteriormente degradada por ribonucleasas (Gottesman, 2005; Perez et al., 2009; Waters y Storz, 2009).

Se ha encontrado además que la unión complementaria de bases entre estos sARN regulatorios y sus objetivos requiere de la presencia de la proteína *Hfq* que actúa como chaperona (Storz et al., 2004).

Otros mecanismos incluyen el silenciamiento mediante interacción con ADN y la unión a proteínas, con lo cual se confiere una especificidad por mRNA objetivos, modulando así su actividad (Eddy, 2002; Storz et al., 2004; Masse, 2003; Repoila y Darfeuille, 2009).

1.2.6. Estructura secundaria

Las moléculas de ARN transcritas por los genes sARN tienen una fuerte tendencia a adoptar una conformación plegada mediante el apareamiento de bases complementarias en la misma hebra, resultando en una estructura secundaria relacionada directamente con su función (Lee y Kim, 2008).

Una molécula de ARN puede adoptar múltiples estructuras distintas al haber diversas formas posibles de apareamiento entre sus bases. Al conjunto de todas estas estructuras posibles se le llama ensamble de estructuras secundarias de una secuencia.

Sin embargo, se acepta que la estructura que cuenta con mayor probabilidad de ser adoptada es aquella que posee la energía mínima de plegamiento (EMP). Esta estructura puede ser calculada mediante el programa *RNAfold* (Hofacker, 2004).

Las estructuras secundarias pueden ser representadas de diversas formas, siendo las dos más comunes la notación de "puntos y paréntesis" y la representación bidimensional. Para ejemplificar se toma la siguiente secuencia de ARN:

UGGGAUGAUGAGCGUUUACCGGUCUGAGCUGUGAUGACAUCGGCACUGUCUGACU

La notación de "puntos y paréntesis" representa los nucleótidos desapareados como puntos y los nucleótidos apareados como paréntesis complementarios. A continuación se muestra la estructura secundaria con la energía mínima de plegamiento (EMP) para la secuencia anterior en esta notación:

..((((...((.((.....)).))...((((((....)).)))).)...))))....

Cabe notar que en el ejemplo existen varios pares *G-U*, los que se forman además de los pares típicos de *Crick* como *C-G* y *A-U*. Este par ocurre de forma tan frecuente como los pares de *Crick*, también se produce por puentes de hidrógeno y tiene una estabilidad similar. El par *G-U* tiene una gran importancia en la formación de la estructura secundaria en moléculas de ARN y en su interacción con otras moléculas (Varani y McClain, 2000).

La representación bidimensional muestra un boceto de la estructura secundaria que adopta la molécula de ARN, la misma estructura representada anteriormente en notación de "puntos y paréntesis" se muestra en esta notación en la Figura 1.3. Esta figura cuenta además con la ventaja de incorporar todos los elementos estructurales típicos que pueden estar presentes en una estructura secundaria de ARN.

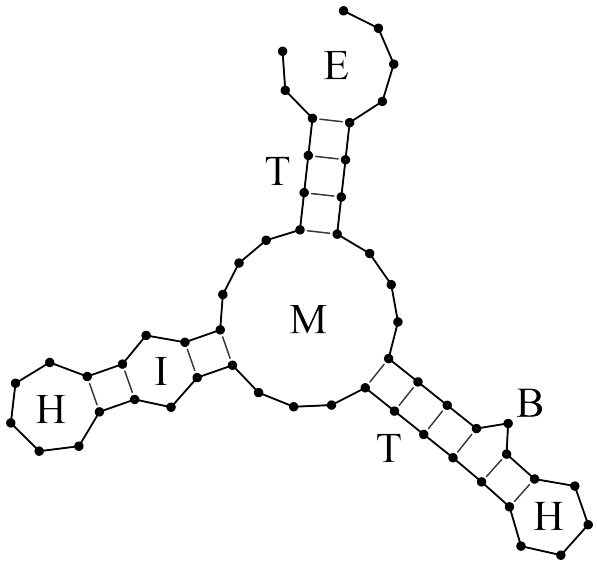


Figura 1.3: Representación de la estructura secundaria de una molécula de ARN. Cada punto corresponde a un nucleótido y las líneas internas señalan las bases apareadas. Además se muestran los 6 tipos básicos de subestructuras existentes: (E) muestra una subestructura de bases externas, (T) muestra una estructura de tallo, (M) muestra un bucle múltiple, (I) muestra un bucle interno, (H) muestra un bucle en forma de horquilla, (B) muestra una estructura de bulto.

También resulta útil tener una forma de medir la similitud entre estructuras secundarias del mismo largo. Para ello se utiliza la noción de "distancia en pares de bases", que consiste en comparar ambas estructuras en su notación de "puntos y paréntesis" y contar el número de caracteres distintos.

Por ejemplo, la distancia en pares de bases entre la estructura ..(..).. y la estructura .((..)). es igual a 2.

1.2.7. Características comunes

La mayoría de los genes sARN encontrados se localizan en regiones intergénicas del genoma; es decir, regiones del genoma flanqueadas por genes que codifican proteínas, muchas veces en marcos de lectura en dirección opuesta a los genes circundantes (Vogel y Sharma, 2005; Altuvia, 2007).

En general, los genes cuentan con un promotor y una señal terminadora fuerte independiente del factor ρ, tienen un mayor contenido de bases G y C que una región promedio y el largo del transcrito varía típicamente de 50 a 400 nucleótidos (Vogel y Sharma, 2005; Altuvia, 2007).

Los genes sARN muestran un proceso evolutivo único, ya que son indiferentes a substituciones de bases mientras se conserve la estructura secundaria de la molécula de ARN, por lo cual no existe una fuerte conservación de la secuencia primaria salvo en organismos cercanos filogenéticamente (Vogel y Sharma, 2005; Tabei y Asai, 2009).

También se ha encontrado que los transcritos de genes sARN adoptan estructuras secundarias con energías mínimas de plegamiento (EMP) menores que las de secuencias del mismo largo generadas aleatoriamente, siempre y cuando éstas conserven las mismas frecuencias dinucleotídicas (Clote et al., 2005).

1.3. Métodos de identificación y predicción

Los genes sARN han sido en su mayoría descubiertos con una preselección de candidatos mediante un método computacional predictivo, seguido de la validación experimental de estos candidatos (Kulkarni y Kulkarni, 2007; Altuvia, 2007; Livny y Waldor, 2007).

Existe una gran cantidad de métodos predictivos enfocados a genes sARN. Sin embargo, no existe un método que tenga una gran eficiencia y especificidad, lo cual hace difícil la comprobación experimental de las secuencias debido a la gran cantidad de falsos positivos en las predicciones, por lo que se considera éste un problema abierto en bioinformática (Machado-Lima et al., 2008).

1.3.1. Métodos de identificación en laboratorio

Se utilizan diversos métodos experimentales para identificar moléculas sARN y sus genes en procariontes, como la detección global de transcritos no codificantes usando microarreglos, mediante clonación *shotgun*, co-purificación con proteínas de unión a RNA como *Hfq*, secuenciamiento de alta capacidad de cADN (*RNA-Seq*), clonamiento de ARNs pequeños abundantes luego de fraccionamiento por tamaño en geles de polyacrilamida y confirmación en *Northern blots*. (Vogel y Sharma, 2005; Sharma y Vogel, 2009)

Si bien existen estos métodos, resulta extremadamente difícil detectar genes sARN nóveles en experimentos bioquímicos (Uzilov et al., 2006), ya que muchos sARN sólo se expresan ante condiciones celulares particulares (Eddy, 2002), dependiendo de la fase de crecimiento o se relacionan a estrés celular, por lo que probablemente existen muchos genes sARN que no han sido descubiertos experimentalmente por ser transcritos bajo circunstancias específicas (Altuvia, 2007).

Desafortunadamente, la validación experimental de los transcritos es muy costosa, por lo cual resulta importante filtrar computacionalmente los candidatos para reducir los transcritos a validar (Kin et al., 2007).

Resulta entonces necesaria la generación de métodos computacionales que disminuyan la cantidad de experimentos e integren la creciente cantidad de información genómica disponible (Kin et al., 2007; Tran, 2009).

1.3.2. Métodos computacionales predictivos

Se han creado diferentes métodos. Estos se basan en las características comunes de los sARNs encontrados a la fecha y generalmente se separan en dos clases: los métodos basados en homología de la secuencia y métodos *"de novo"*.

Sin embargo, todos ellos generan una gran cantidad de falsos positivos, lo que dificulta el análisis a gran escala de un genoma completo (Tran, 2009).

Métodos basados en homología

Los métodos basados en homología buscan secuencias nucleotídicas altamente conservadas entre regiones intergénicas de genomas bacterianos cercanos filogenéticamente, como por ejemplo entre *Escherichia coli*

y enterobacterias cercanas como *Salmonella typhimurium* y *Yersinia pestis*. Además, toman en cuenta la conservación de la estructura secundaria y la existencia de un promotor y señal de término a una distancia de 50-400 nucleótidos (Vogel y Sharma, 2005).

Este tipo de métodos ha sido ampliamente utilizado para la predicción de genes sARN con una secuencia homóloga conocida. Mientras resulta bastante eficiente para identificar genes sARN en diversos genomas, no sirven para identificar genes sARN nóveles (Tran, 2009).

Algunos de estos métodos son:

- QRNA (Rivas y Eddy, 2001)
- RNAz (Washietl, 2005)
- Dynalign (Uzilov et al., 2006)

Métodos *"de novo"*

Los métodos *"de novo"* utilizan algoritmos sofisticados de aprendizaje automático y no dependen de conocimientos previos sobre el microorganismo a estudiar (Vogel y Sharma, 2005).

Este tipo de método consiste en el entrenamiento de un algoritmo clasificador que recibe las variables de un grupo de datos clasificados como casos positivos y negativos, para luego realizar predicciones sobre nuevos grupos de datos sin clasificar.

La efectividad del método depende de la capacidad de identificar las características comunes de los genes sARN para ser medidas como variables, que los distingan de otros tipos de secuencias como genes codificantes y secuencias al azar.

Algunos de estos métodos son:

- CONC (Liu et al., 2006)
- RSSVM (Xu et al., 2009)

Dificultades de la predicción computacional

El análisis computacional de secuencias genómicas ha sido eficaz para la predicción de genes que codifican proteínas. Sin embargo, los genes sARN presentan un nuevo conjunto de desafíos para la genómica computacional (Eddy, 2002; Livny y Waldor, 2007).

Los genes sARN muestran un proceso evolutivo único, en el cual las sustituciones de bases distantes se correlacionan con el fin de conservar la estructura secundaria de la molécula de ARN. Por ende, se debe tomar en cuenta no sólo la conservación de la secuencia primaria sino también la estructura secundaria (Eddy, 2002; Tabei y Asai, 2009).

Por este motivo, los métodos computacionales desarrollados para identificar genes que codifican proteínas fallan en identificar a los genes sARN, lo que hace que el descubrimiento de estos genes requiera de un método diferente al de los genes que codifican proteínas (Machado-Lima et al., 2008).

Aunque se ha dedicado una cantidad significativa de investigación a la predicción computacional de los genes sARN a partir de secuencias genómicas, los resultados no han sido del todo positivos, ya que a diferencia de los genes que codifican proteínas, los genes sARN no cuentan con atributos genómicos aparentes (Rivas y Eddy, 2001; Kavanaugh y Dietrich, 2009).

Una de las características que los hacen especialmente difíciles de identificar es su pequeño tamaño, el cual varía entre 50 a 400 pares de bases, a diferencia de genes de proteínas que pueden medir varios miles de pares de bases (Argaman et al., 2001; Machado-Lima et al., 2008; Vogel y Papenfort, 2006; Tran et al., 2009).

Otra característica es la ausencia de un marco de lectura abierto como el de los genes que codifican proteínas, por lo que no existen señales estadísticas fuertes en la composición de sus secuencias, como las

desigualdades en el uso de codones en el caso de genes de proteínas (Nawrocki y Eddy, 2009; Vogel y Papenfort, 2006; Tran et al., 2009).

Además, la comparación de secuencias es poco sensible ya que usan un alfabeto reducido y menos informativo de sólo 4 letras (Nawrocki y Eddy, 2009; Vogel y Papenfort, 2006; Tran et al., 2009).

Estos genes también tienen la característica de evolucionar rápidamente a nivel de secuencia primaria, conservando sólo la estructura secundaria relacionada a su función, por lo que existe menor conservación de secuencia entre especies lejanas filogenéticamente (Backofen et al., 2007; Machado-Lima et al., 2008; Vogel y Papenfort, 2006; Tran et al., 2009).

Por último, aunque algunos genes sARN tienen promotores y terminadores conocidos, la identificación de estas señales regulatorias resulta actualmente un difícil desafío (Nawrocki y Eddy, 2009; Vogel y Papenfort, 2006; Tran et al., 2009).

Si bien los métodos predictivos basados en homología son fiables al estar respaldados por bases de datos curadas de familias conocidas de genes sARN (Nawrocki y Eddy, 2009), tienen el inconveniente de estar limitados en su aplicación sólo para predecir genes sARN con homólogos conocidos, por lo que no permiten el descubrimiento de genes nóveles (Tran et al., 2009).

Por otro lado, los métodos predictivos *"de novo"* actualmente tienen una alta tasa de falsos positivos, por lo que son útiles como un filtro seguido por comprobación experimental, pero esto los hace poco adecuados para el análisis automatizado de genomas a gran escala (Tran et al., 2009).

Necesidad de nuevos métodos predictivos

Con la reciente expansión en la disponibilidad de de material genómico secuenciado, existe la necesidad de mejorar la capacidad de anotación de genes y elementos estructurales de una manera rápida, eficiente y precisa (Chen et al., 2002; Vogel y Wagner, 2007; Tran et al., 2009).

Aunque grupos han publicado métodos predictivos para anotar genes sARN en genomas bacterianos, mucho queda por hacer en este campo (Chen et al., 2002; Kavanaugh y Dietrich, 2009; Tran et al., 2009).

Para avanzar en el conocimiento biológico, se requieren métodos computacionales que puedan detectar con precisión los genes sARN en secuencias genómicas (Uzilov et al., 2006), permitiendo así la identificación a gran escala de genes sARN nóveles y facilitando la exploración de los mecanismos de regulación génica (Xu et al., 2009).

1.4. Estadística

Se utilizan dos herramientas estadísticas que vale la pena mencionar, ellas son:

1.4.1. Coeficiente de Correlación de Pearson

Este coeficiente mide el grado de correlación lineal que existe entre dos variables aleatorias cuantitativas, tomando valores entre -1 y 1. Si el valor es cercano a 0 significa que no existe una correlación lineal, si es cercano a -1 quiere decir que existe correlación lineal inversa y si es cercano a 1 existe una correlación lineal directa.

En la Ecuación 1.1 se muestra la forma de calcularla para las variables x e y, siendo σ_x la varianza de la variable x y σ_{xy} la covarianza entre ambas variables.

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \tag{1.1}$$

1.4.2. Prueba de los signos de Wilcoxon

Esta prueba estadística se utiliza para comparar dos muestras relacionadas y determinar si existe una diferencia significativa entre sus medias, sin asumir necesariamente que las muestras sigan una distribución normal.

Para la prueba se utiliza como hipótesis nula que ambas muestras tienen la misma media. La hipótesis nula es aceptada o descartada según el resultado del valor llamado *p-value* arrojado por la prueba. Si este valor es menor al criterio de corte predefinido, por ejemplo $1 \cdot 10^{-2}$, se puede descartar la hipótesis nula y asumir que las muestras presentan una diferencia significativa.

1.5. Minería de datos

La minería de datos es un área en las ciencias de la computación dedicada a la extracción de patrones y conocimiento desde un grupo de datos, siendo una mezcla interdisciplinaria entre estadística, inteligencia artificial y manejo de bases de datos (Han y Kamber, 2005).

Dentro de esta área se encuentran los algoritmos clasificadores de aprendizaje automático, los cuales sirven para discriminar entre clases de datos al "aprender" las características particulares de cada clase mediante entrenamiento con un subconjunto de datos conocidos.

Existen diversos algoritmos clasificadores. A continuación se describen los que produjeron mejores resultados en el presente trabajo.

Para ejemplificar, se detallarán los algoritmos en función de su capacidad para discriminar la pertenencia de un elemento con n variables $V_1, \dots, V_n$ entre las clases C_0 y C_1.

1.5.1. Clasificador bayesiano ingenuo

Este clasificador es un algoritmo probabilístico simple, el cual asume que todas las variables son independientes entre sí, motivo por el cual se denomina ingenuo (Han y Kamber, 2005).

Aunque la suposición de independencia es bastante fuerte y pocas veces se cumple en problemas reales, el algoritmo ha demostrado ser bastante confiable y eficiente, permitiendo su aplicación en problemas complejos. Una teoría que explica su buen funcionamiento es que las dependencias entre variables se cancelan entre sí al estar distribuidas uniformemente en las clases (Zhang, 2004).

El algoritmo está basado en el Teorema de Bayes (Ecuación 1.2), el cual establece un método para calcular las probabilidades condicionales, donde $P(A)$ es la probabilidad del evento A y $P(A|B)$ es la probabilidad condicional del evento A dado el evento B con probabilidad distinta de cero.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.2}$$

Utilizando como base el Teorema de Bayes y el supuesto de independencia entre las variables, se obtiene la fórmula mostrada en la Ecuación 1.3. Esta ecuación permite calcular la probabilidad condicional de que un dato pertenezca a la clase C dadas n variables $V_1, \dots, V_n$. Las probabilidades, $P(C)$ y $P(V_i|C)$ se calculan mediante el entrenamiento del clasificador con un conjunto de datos conocidos, siendo además Z un parámetro constante.

$$P(C|V_1, \dots, V_n) = \frac{P(V_1, \dots, V_n|C)P(C)}{P(V_1, \dots, V_n)} = \frac{1}{Z}\ P(C) \prod_{i=1}^{n} P(V_i|C) \tag{1.3}$$

Aplicando la Ecuación 1.3 para la clase C_0 y la clase C_1, es posible construir mediante una división el clasificador bayesiano ingenuo, el cual se muestra en la Ecuación 1.4. Este clasificador predice que un dato pertenece a la clase C_0 si toma un valor mayor o igual a 1, o que pertenece a la clase C_1 en caso contrario (Zhang, 2004).

$$Clasificador\ Bayesiano\ Ingenuo = \frac{P(C_0)}{P(C_1)} \prod_{i=1}^{n} \frac{P(V_i|C_0)}{P(V_i|C_1)} \tag{1.4}$$

1.5.2. Regresión logística

Este método utiliza una regresión lineal generalizada, ajustando los datos a la función logística (Ecuación 1.5), en la cual z representa una combinación lineal de las variables $V_1, ... , V_n$ ponderadas por los coeficientes de la regresión $\beta_0, ... , \beta_n$, lo que se muestra en la Ecuación 1.6 (Han y Kamber, 2005).

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1.5}$$

$$z = \beta_0 + \beta_1 V_1 + ... + \beta_n V_n \tag{1.6}$$

La función logística toma valores entre 0 y 1, como se muestra en la Figura 1.4, representando la probabilidad de que el dato con variables $V_1, ... , V_n$ pertenezca a la clase C_0 si toma un valor cercano a 0 ó a la clase C_1 si toma un valor cercano a 1 (Han y Kamber, 2005).

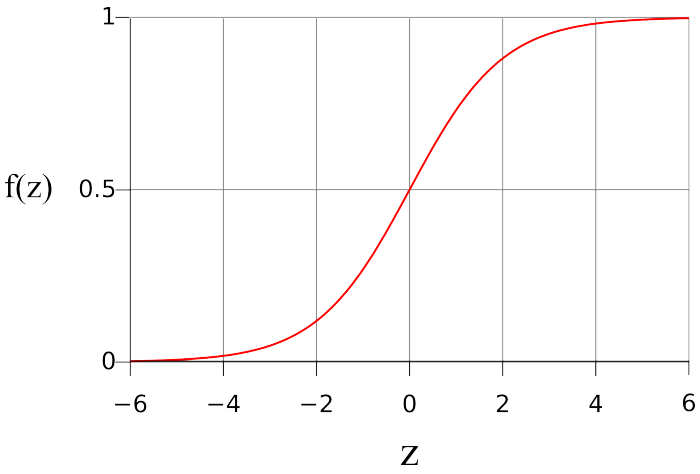


Figura 1.4: Función logística.

1.5.3. Perceptrón multicapa

Este clasificador es un algoritmo del tipo red neuronal artificial, el cual está formado por múltiples capas que le permiten resolver problemas más complejos que el de capa simple (Han y Kamber, 2005).

El perceptrón cuenta con al menos 3 capas: la capa de entrada que recibe las variables, la capa de salida que genera el resultado final y una o más capas ocultas que conectan las capas de entrada y salida.

Cada nodo o elemento de capa simula el funcionamiento de una neurona, activándose al momento de recibir un estímulo que sobrepase cierto umbral y mandando esta señal a cada neurona de la capa siguiente.

En este caso cada neurona recibe las señales de las neuronas de la capa anterior, las pondera por coeficientes y aplica una función de activación no lineal, generando así un resultado que transmite a la siguiente capa.

Un tipo de función de activación es la función logística que genera resultados entre 0 y 1, la cual se muestra en la Ecuación 1.7, donde y_i es la suma ponderada por coeficientes de los valores que ingresan a la neurona i desde la capa anterior y λ un parámetro que regula la pendiente de la parte lineal de la función logística.

$$\phi(y_i) = \frac{1}{1 + e^{-\lambda y_i}} \tag{1.7}$$

A través del aprendizaje del clasificador con el conjunto de entrenamiento se definen los coeficientes, con lo cual se genera un modelo que ante las variables de entrada genera como salida un valor entre 0 y 1, permitiendo clasificar la instancia en la clase C_0 si el valor es más cercano a 0, mientras que en el caso contrario pertenece a C_1.

En la Figura 1.5 se muestra un ejemplo de perceptrón con 3 capas y 4 variables de entrada.

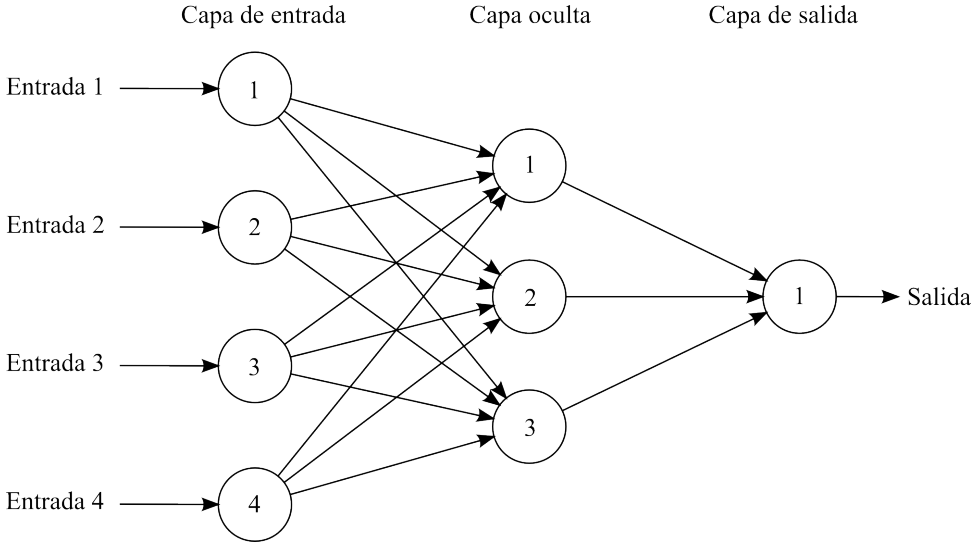


Figura 1.5: Perceptrón Multicapa.

1.5.4. Máquina de vectores de soporte

Este método de clasificación se caracteriza por encontrar el hiperplano que separa de forma óptima los datos de las clases C_0 y C_1 en el espacio dimensional de las variables medidas (Han y Kamber, 2005).

Un ejemplo para el caso bidimensional, en que existen sólo 2 variables x_1 y x_2, se muestra en la Figura 1.6.

En el ejemplo se muestran 3 hiperplanos (líneas rectas para el caso bidimensional), H_1 y H_2 separan completamente a las clases C_0 (puntos negros) y C_1 (puntos blancos), mientras que H_3 no separa completamente las clases.

Sin embargo, el hiperplano H_2 tiene una mayor separación al dato más próximo de ambas clases, por lo cual es el hiperplano óptimo.

1.5.5. Bosque aleatorio

Este método de clasificación utiliza un conjunto de otros clasificadores denominados árboles de decisión, mejorando el desempeño individual de cada uno al tomar como resultado la clase que fue más veces escogida por el conjunto de árboles de decisión (Han y Kamber, 2005).

Un árbol de decisión es un método de clasificación que a través de una serie de ramas que se bifurcan, siendo éstas decisiones binarias respecto a las variables medidas, determina la pertenencia de una instancia a una de dos clases.

En la Figura 1.7 se muestra un árbol de decisión binario. Para este caso existen 2 variables x e y, y a través de decisiones binarias se desciende desde la raíz (nodo superior), a través de las ramas (decisiones binarias) y se llega a una hoja que determina la pertenencia de la instancia a la clase C_0 ó C_1.

El bosque aleatorio es entonces un conjunto de varios árboles de decisión, cada uno con un subconjunto aleatorio de variables para lograr tener variabilidad, y se escoge la clase que fue seleccionada por la mayor cantidad de árboles.

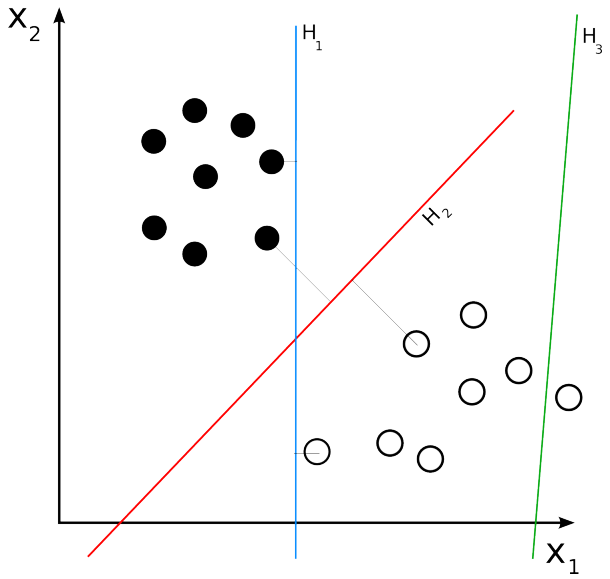


Figura 1.6: Máquina de Vectores de Soporte.

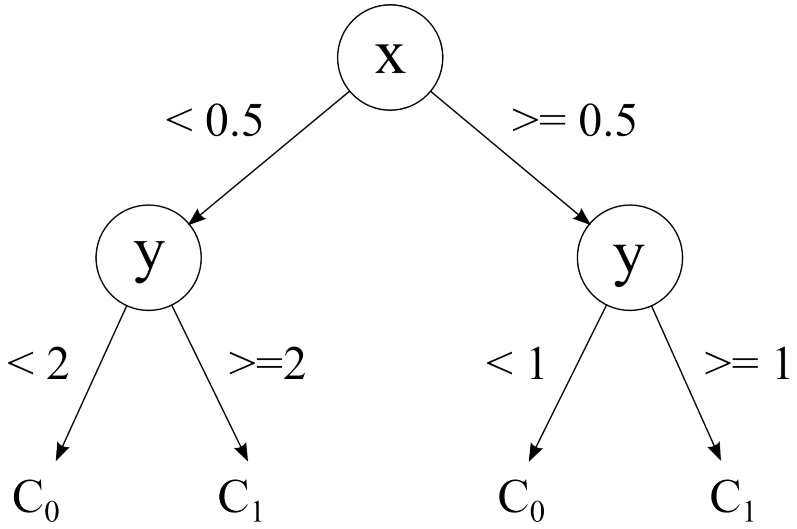


Figura 1.7: Árbol de decisión.

1.5.6. Comparación del desempeño de los clasificadores

Existen múltiples estadísticas para comparar los resultados entregados por un método clasificador, todas ellas calculadas a partir de los resultados básicos de la clasificación expresados por la matriz de confusión.

Esta matriz muestra los resultados de la clasificación separando las instancias clasificadas correcta e incorrectamente, lo que se muestra ejemplificado en la Tabla 1.1 para la clase C_0 (clase positiva) y la clase C_1 (clase negativa).

Tabla 1.1: Matriz de confusión.

Clase Real	**Clasificado como** C_0	**Clasificado como** C_1
C_0	Verdadero Positivo	Falso Negativo
C_1	Falso Positivo	Verdadero Negativo

Estadísticas del desempeño

A continuación se muestran las estadísticas calculadas directamente desde la matriz de confusión. Para las ecuaciones se considera VP como verdadero positivo, VN como verdadero negativo, FP como falso positivo y FN como falso negativo.

Tasa de Verdaderos Positivos: También llamada *Sensibilidad*, mide la cantidad de casos clasificados correctamente como positivos sobre los casos positivos reales totales, la forma de calcularla se muestra en la Ecuación 1.8 (Fawcett, 2006).

$$TVP = \frac{VP}{VP + FN} \tag{1.8}$$

Tasa de Falsos Positivos: Mide la cantidad de casos clasificados incorrectamente como positivos sobre los casos negativos reales totales, la forma de calcularla se muestra en la Ecuación 1.9 (Fawcett, 2006).

$$TFP = \frac{FP}{FP + VN} \tag{1.9}$$

Tasa de Verdaderos Negativos: También llamada *Especificidad*, mide la cantidad de casos clasificados correctamente como negativos sobre los casos negativos reales totales, la forma de calcularla se muestra en la Ecuación 1.10 (Fawcett, 2006).

$$TVN = \frac{VN}{VN + FP} \tag{1.10}$$

Precisión: También llamado *Valor Predictivo Positivo*, mide la proporción de casos clasificados correctamente entre todos los casos clasificados como positivos, la forma de calcularla se muestra en la Ecuación 1.11 (Fawcett, 2006).

$$Precisión = \frac{VP}{VP + FP} \tag{1.11}$$

Exactitud: Mide la proporción de casos clasificados correctamente respecto al total de casos, la forma de calcularla se muestra en la Ecuación 1.12 (Fawcett, 2006).

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN} \tag{1.12}$$

Curvas *ROC*

La curva *ROC* (Característica Operativa del Receptor) corresponde a la curva que se forma al graficar la *Tasa de Verdaderos Positivos* versus la *Tasa de Falsos Positivos* de una clase, y sirve para representar el conjunto de puntos que puede escoger un clasificador para clasificar las clases teniendo distinta tolerancia a la *Tasa de Falsos Positivos* (Fawcett, 2006).

La característica más importante de esta curva es que se ha demostrado que la mejor forma de comparar métodos de clasificación es a través del valor *AUROC*, que corresponde al área cubierta bajo la curva *ROC* (Ling et al., 2003), la cual varía entre 0 y 1.

Por ende, será el valor *AUROC* el que se utilice en última instancia para definir el clasificador con mejor desempeño.

1.6. Objetivos

El objetivo general de esta Tesis es la creación de un algoritmo computacional que permita predecir genes sARN directamente de genomas bacterianos con alta sensibilidad y especificidad, identificando secuencias conocidas en la literatura, descartando secuencias codificantes y aleatorias, además de proponer buenos candidatos de genes nóveles que puedan ser comprobados en laboratorio.

1.6.1. Objetivos específicos

- Identificar las características más relevantes que han sido utilizadas por otros métodos para distinguir a los genes sARN.
- Seleccionar las variables de mayor significancia estadística para la identificación de genes sARN en genomas de origen bacteriano.
- Crear un método computacional que permita predecir genes sARNs en genomas bacterianos con alta especificidad y sensibilidad.
- Programación de un método original de mejor o similar desempeño respecto a los existentes.
- Proposición de genes nóveles de sARNs en el genoma de *Escherichia coli*.

Capítulo 2

Metodología

2.1. Recopilación de Variables en la literatura

Del estudio de los métodos utilizados en la literatura, se recopiló un total de 40 variables que miden los atributos utilizados para identificar genes sARN. A continuación se detalla cada una de las variables agrupadas según sus propiedades.

2.1.1. Variables relacionadas al ensamble de estructuras secundarias

Las moléculas transcritas por los genes sARN adoptan una estructura secundaria mediante complementariedad de bases. Sin embargo, esta estructura no es única.

Al conjunto de todas las estructuras que pueden ser adoptadas por una molécula de ARN se le llama ensamble de estructuras secundarias y contiene todas las configuraciones posibles de apareamiento entre los distintos nucleótidos, contando cada estructura con una energía de plegamiento propia.

En este ensamble, existen dos estructuras cuyo estudio resulta particularmente interesante. La primera es el *centroide*, el cual corresponde a la estructura más parecida a todas las demás estructuras del ensamble y, la segunda, es la estructura que posee la energía mínima de plegamiento (EMP).

Las variables a continuación se calculan a partir de estadísticas acerca del ensamble de estructuras secundarias para una secuencia particular y del cálculo de las energías de plegamiento, con el uso del programa *RNAfold* (Hofacker, 2004).

dG: Corresponde a la EMP de la secuencia normalizada por su largo (Freyhult et al., 2005).

dG del ensamble: Corresponde a la energía libre de Gibbs del ensamble normalizada por el largo de la secuencia, calculado directamente por el programa *RNAfold* (Hofacker, 2004).

Diferencia de energía entre EMP y ensamble: Se calcula como la resta entre *dG* y *dG del ensamble* (Ng y Mishra, 2007).

EMP$_I$: Corresponde a la EMP dividida por el número de nucleótidos G y C (Ng y Mishra, 2007).

EMP$_{II}$: Corresponde a la EMP dividida por el número de subestructuras Tallo presentes en la estructura secundaria que tiene la EMP (Ng y Mishra, 2007).

dD: Corresponde a la distancia promedio en pares de bases entre las estructuras del ensamble, normalizadas por el largo de la secuencia (Freyhult et al., 2005). La fórmula para calcular esta variable se muestra en la Ecuación 2.1, donde d_{ij} es la distancia de pares de bases entre la estructura i y la estructura j del ensamble, mientras que n es el número total de estructuras en el ensamble.

$$dD = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} \tag{2.1}$$

dQ: Corresponde a la entropía de Shannon aplicada al apareamiento de bases, normalizada por el largo de la secuencia (Freyhult et al., 2005). La fórmula para calcular esta variable se muestra en la Ecuación 2.2, donde P_{ij} es la probabilidad de que el par (i, j) forme enlace (estimada de su frecuencia en el ensamble), mientras que n es el número de bases de la secuencia.

$$dQ = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\log(P_{ij}) \tag{2.2}$$

Número de pares frecuentes en el ensamble: Corresponde al número de pares de bases que están apareados en más del 50 % de las estructuras del ensamble (Chan y Ding, 2008).

2.1.2. Variables relacionadas a la comparación estadística de la energía mínima de plegamiento frente a secuencias aleatorias

Estas variables se calculan comparando las energías mínimas de plegamiento (EMP) entre la secuencia original y un grupo de 1000 secuencias aleatorias del mismo largo y que mantienen la misma frecuencia dinucleotídica.

Valor Z: Corresponde a una medida de la distancia en desviaciones estándar entre la EMP de la secuencia original y el promedio de las EMP de secuencias aleatorias (Washietl, 2005; Freyhult et al., 2005). En la Ecuación 2.3 se muestra la forma de calcular este valor, siendo EMP la energía mínima de plegamiento de la secuencia original, mientras que μ y σ representan respectivamente el promedio y la desviación estándar del conjunto de EMP de las 1000 secuencias aleatorias.

$$Valor\ Z = \frac{EMP - \mu}{\sigma} \tag{2.3}$$

Valor de partición: Corresponde al número de secuencias aleatorias que poseen una EMP menor que la secuencia original normalizado por el número total de secuencias aleatorias (Freyhult et al., 2005). La forma de calcularlo se muestra en la Ecuación 2.4 .

$$Valor\ de\ partición = \frac{N^{o}\ de\ secuencias\ con\ menor\ EMP}{1000} \tag{2.4}$$

2.1.3. Variables relacionadas a la clusterización del ensamble de estructuras secundarias

Estas variables se calculan a partir de una clusterización, o agrupación de elementos por similitud, de las estructuras presentes en el ensamble de estructuras secundarias. Debido a que el tamaño del ensamble aumenta exponencialmente respecto al largo de la secuencia, se toma una muestra representativa de 1000 estructuras distintas del ensamble.

El algoritmo de agrupamiento utilizado corresponde al programa *RNACluster* (Liu et al., 2008), el cual calcula las distancias en pares de bases entre las estructuras secundarias del ensamble de una secuencia, y agrupa las estructuras más cercanas entre sí en un número variable de subgrupos o *clusters*, calculado automáticamente según la diversidad del ensamble.

Una de las estadísticas base para analizar los resultados de la clusterización es la medida de compactación del *cluster*. La forma de calcular esta estadística se muestra en la Ecuación 2.5, donde d_{ij} representa la distancia en pares de bases entre la estructura i y la estructura j del *cluster*, mientras que n es el número total de estructuras en el *cluster* (Liu et al., 2008; Tran, 2009).

$$Compactación = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}}{n(n-1)} \tag{2.5}$$

A continuación se describen las variables estadísticas que han sido utilizadas en la literatura para identificar los genes sARN, a partir de los resultados del algoritmo de agrupamiento.

Número de *clusters*: Corresponde al número de *clusters* o subgrupos en los cuales se separan las estructuras secundarias, definido automáticamente por el algoritmo de agrupamiento según la diversidad presente en el ensamble (Chan y Ding, 2008).

Tamaño del mayor *cluster*: Corresponde al número de elementos del *cluster* o subgrupo que agrupa la mayor cantidad de estructuras secundarias del ensamble (Chan y Ding, 2008).

Compactación promedio: Corresponde al valor promedio de compactación entre todos los *clusters*, normalizado por el largo de la secuencia (Tran, 2009).

Compactación máxima: Corresponde al valor máximo de compactación entre todos los *clusters*, normalizado por el largo de la secuencia (Tran, 2009).

Compactación mínima: Corresponde al valor mínimo de compactación entre todos los *clusters*, normalizado por el largo de la secuencia (Tran, 2009).

Compactación del mayor *cluster*: Corresponde al valor de compactación del *cluster* con mayor número de elementos, normalizado por el largo de la secuencia (Tran, 2009).

Número promedio de pares frecuentes por *cluster*: Corresponde al promedio del número de bases apareadas en más del 50 % de las estructuras en cada *cluster*, normalizado por el largo de la secuencia (Chan y Ding, 2008).

Suma entre *clusters* (SEC): Corresponde a una medida de la cercanía entre los *clusters* (Chan y Ding, 2008). La forma de calcular esta medida se muestra en la Ecuación 2.6, donde n_i es el número de elementos en el *cluster* i, $D(CE, CC_i)$ es la distancia en pares de bases entre el centroide del ensamble y el centroide del *cluster* i, mientras que k es el número total de *clusters* y L es el largo de la secuencia.

$$SEC = \frac{1}{L}\sum_{i=1}^{k}(n_i\ D(CE, CC_i)) \tag{2.6}$$

Suma dentro del *cluster* (SDC): Corresponde a una medida de la compactación total de los *clusters* (Chan y Ding, 2008). La forma de calcular esta medida se muestra en la Ecuación 2.7, donde n_i es el número de elementos en el *cluster* i, $D(CC_i, I_{ij})$ es la distancia en pares de bases entre el centroide del *cluster* i y la estructura I_{ij} que es el elemento j del *cluster* i, mientras que k es el número total de *clusters* y L es el largo de la secuencia.

$$SDC = \frac{1}{L}\sum_{i=1}^{k}\sum_{j=1}^{n_i} D(CC_i, I_{ij}) \tag{2.7}$$

2.1.4. Variables relacionadas a las características de la estructura secundaria

Estas variables se obtienen a partir de la estructura secundaria con la energía mínima de plegamiento (EMP) mediante el programa *RNAfold*, desde la que se puede calcular el número de subestructuras presentes.

Las características de cada tipo de subestructura secundaria se muestran en la Figura 1.3 en la página 5.

Estadísticas de bucles en forma de horquilla: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a bucles en forma de horquilla presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas de bucles internos: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a bucles internos presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas de bucles múltiples: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a bucles múltiples presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas de bultos: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a bultos presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas de tallos: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a tallos presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas de estructuras externas: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a estructuras externas presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

Estadísticas del total de bucles: Corresponde al número, porcentaje de secuencia y cantidad de bases promedio asociadas a la suma de los tres tipos de bucles (en forma de horquilla, internos y múltiples) presentes en la estructura secundaria que posee la EMP, normalizado por el largo de la secuencia (Tran, 2009).

2.2. Consulta a Bases de datos

2.2.1. Obtención de genomas bacterianos

Para la obtención de secuencias genómicas completas se recurre a la página de la *NCBI* (Centro Nacional de Información sobre Biotecnología de Estados Unidos) (Sayers y Barrett, 2010), desde donde se accede a la base de datos *GenBank* (Benson et al., 2011) debido a su confiabilidad y recopilación de múltiples secuencias genómicas.

Para la aplicación del clasificador se elige el genoma de *Escherichia coli*, debido a la gran cantidad de información de calidad que se encuentra disponible.

Por ende se descarga la secuencia *U00096.2* correspondiente específicamente al genoma completo de la bacteria *Escherichia coli*, cepa *K-12*, subcepa *MG1655*.

Además, se descarga también desde la base de datos *Ecocyc* (Keseler y Collado-Vides, 2011) un listado con la ubicación dentro de este genoma de los genes que codifican proteínas. Aunque el listado completo cuenta con un total de 4169 secuencias, se descartan los genes hipotéticos, con lo cual quedan 3057 secuencias de genes comprobados experimentalmente. A partir de este listado de genes se definirán las regiones intergénicas dentro del genoma, correspondiendo aproximadamente un 65,0 % del genoma a genes que codifican proteínas y un 35,0 % a regiones intergénicas.

2.2.2. Recopilación de secuencias de genes sARN

Para la obtención de los genes de sARN con los cuales entrenar el modelo, se recurre a la base de datos *Rfam* (Griffiths-Jones, 2003; Griffiths-Jones et al., 2005; Gardner et al., 2009, 2010), ya que se considera una fuente confiable al presentar los datos curados y con las referencias bibliográficas de cada gen descrito en la literatura.

Se descargó la base de datos completa en su versión *10.0* (última actualización al 22/02/2011), desde la cual se escogieron las secuencias de genomas bacterianos categorizadas como genes, verificándose manualmente la validez de cada gen en la literatura.

De esta forma se obtuvo un total de 1846 secuencias distintas separadas en 107 familias de genes (tipos de genes distintos). La Tabla 6.1 del anexo muestra un listado de estos genes y su descripción.

2.2.3. Genes sARN en genoma de *Escherichia coli*

De acuerdo a la base de datos *Rfam*, en este genoma se encuentran un total de 58 copias de genes sARN dentro de las cuales hay 49 familias distintas de genes. Esto se muestra en la Tabla 6.2 en el anexo.

Además, al comparar estas secuencias con los genes que codifican proteínas de la base de datos *EcoCyc*, se observa que la mayor parte de ellos se encuentran en regiones intergénicas del genoma como se muestra en la Tabla 2.1, lo cual se condice con la literatura ya que la mayoría de genes sARN identificados han sido encontrados mediante búsquedas en regiones intergénicas.

Tabla 2.1: Posiciones de los genes sARN en el genoma de *Escherichia coli*, respecto a genes que codifican proteínas.

Posición en el genoma	**Número de genes**	**Porcentaje de genes**
Regiones intergénicas	37	63,8 %
Traslape con gen en la misma hebra	11	19,0 %
Traslape con gen en la hebra complementaria	10	17,2 %
Total	58	100,0 %

2.2.4. Promotores y Terminadores de *Escherichia coli*

Se descarga un listado de las posiciones de promotores σ^{70} y terminadores intrínsecos (independientes del factor ρ) desde la base de datos *RegulonDB* (Gama-Castro et al., 2008), la cual publica información validada experimentalmente y específica a la cepa bacteriana utilizada en este estudio.

2.3. Construcción de los conjuntos de datos

El método clasificador requiere del ingreso de dos conjuntos de datos, el primer conjunto es de entrenamiento con el cual se le "enseña" al método clasificador a identificar las diferencias entre la clase "sARN" (grupo positivo) y la clase "aleatoria" (grupo negativo).

El segundo conjunto necesario es de prueba, con el cual se verifica la eficacia del clasificador ya entrenado en el conjunto anterior. Ambos conjuntos de datos cuentan con igual número de secuencias positivas y negativas (clases "sARN" y "aleatoria").

Al separar el total de los datos en estos dos grupos, se evita sobreajustar los parámetros del método clasificador para no sobreestimar su capacidad predictiva real.

Para construir el conjunto de datos se tomaron inicialmente las 1846 secuencias de genes obtenidas desde la base de datos *Rfam*. Sin embargo, había diferente número de secuencias identificadas para cada gen, por lo que, para evitar asimetrías que afectaran el entrenamiento del método clasificador, se escogió sólo una secuencia representante de cada familia de genes, eligiéndose aquella secuencia cuya estructura secundaria de energía mínima de plegamiento (EMP) fuera la más parecida a todas las demás mediante el *script select_local_centroid.pl*, representando esta secuencia el centroide local de cada familia.

En base a esto se obtuvieron 107 secuencias representantes de las 107 familias con lo cual se construyó el conjunto de entrenamiento positivo.

El conjunto de entrenamiento negativo se obtuvo reordenando al azar los nucleótidos de las secuencias del conjunto de entrenamiento positivo, de modo de obtener secuencias aleatorias del mismo largo pero manteniendo la frecuencia dinucleotídica.

Mantener la frecuencia dinucleotídica resulta necesario ya que en la literatura (Clote et al., 2005) se muestra que es un método más robusto de comparación. Para esto se utilizó el *script uShuffle.pl*, que utiliza al programa uShuffle (Jiang et al., 2008) para generar las secuencias aleatorias que mantienen la frecuencia dinucleotídica.

De esta forma se obtuvo el conjunto de entrenamiento negativo con 107 secuencias aleatorias con la misma frecuencia dinucleotídica que el conjunto positivo.

El conjunto de prueba positivo se construyó removiendo las 107 secuencias del conjunto de entrenamiento positivo del grupo de las 1846 secuencias de *Rfam*, quedando 1739 secuencias en este conjunto.

El conjunto de prueba negativo se obtuvo reordenando al azar las secuencias del conjunto positivo, de igual forma que para el caso del conjunto de entrenamiento.

De esta forma quedaron dos conjuntos, de entrenamiento y prueba, con la distribución de secuencias resumida en la Tabla 2.2.

Tabla 2.2: Conjuntos de entrenamiento y prueba.

Conjunto	**Secuencias Positivas** (Clase sARN)	**Secuencias Negativas** (Clase aleatoria)
Entrenamiento	107	107
Prueba	1739	1739

En la Figura 2.1 se muestra un histograma del largo de las secuencias presentes en el conjunto de entrenamiento positivo, siendo una muestra representativa de las 107 familias de genes sARN en la base de datos *Rfam*.

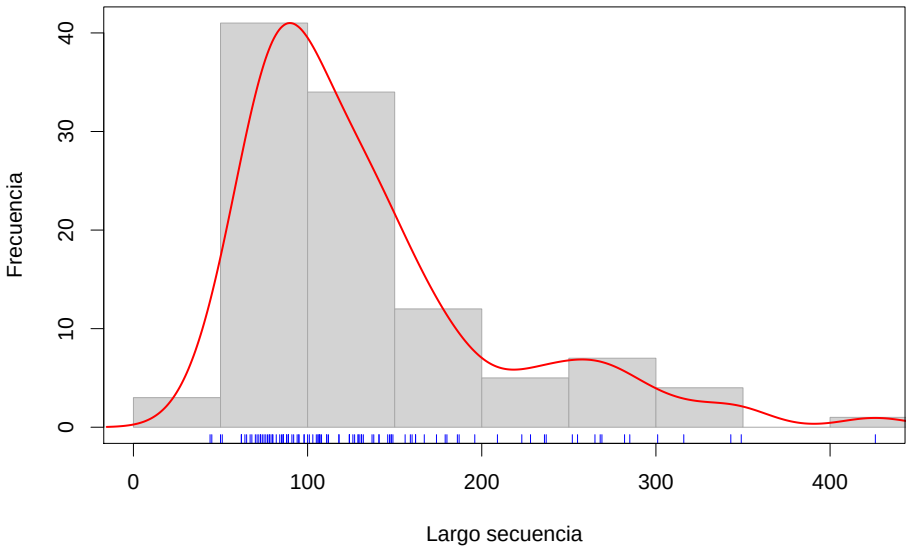


Figura 2.1: Histograma del largo de la secuencia de genes sARN en conjunto de entrenamiento.

2.4. Cálculo de variables

2.4.1. Programación de *scripts*

Se decide utilizar el lenguaje de programación *Perl* debido a su eficiencia para trabajar con cadenas de texto, ser de uso gratuito, tener una gran cantidad de documentación disponible y el hecho de que es comúnmente utilizado en bioinformática (Dwyer, 2003). Además, este lenguaje cuenta con la existencia de los módulos *Bioperl* (Stajich et al., 2002), un conjunto de paquetes bioinformáticos disponibles en forma gratuita.

También se utiliza el conjunto de paquetes *Vienna RNA* (Hofacker, 2004) para calcular la estructura secundaria y energía mínima de plegamiento (EMP) de las secuencias, integrándose estos cálculos a los *scripts* para calcular las variables.

Para calcular el total de las variables se utiliza el *script RNA_statistics.pl*, el cual recibe un archivo con las secuencias nucleotídicas de ARN y calcula para cada una las variables correspondientes, devolviendo un archivo tabulado con los resultados.

2.5. Clasificación

Para el entrenamiento y prueba de diversos clasificadores se utiliza el programa *WEKA* (Hall et al., 2009; Han y Kamber, 2005). Este corresponde a un conjunto de algoritmos de minería de datos de código abierto que cuenta con múltiples herramientas.

El programa acepta como entrada un archivo con los resultados de las variables para cada clase, permitiendo aplicar diversos filtros a los datos, realizar selección de variables y utilizar diversos algoritmos de clasificación.

En este programa se introducen los resultados del cálculo de variables mediante el *script RNA_statistics.pl* para el conjunto de datos de entrenamiento, para luego realizar una seleccion de variables. Luego se aplican los métodos clasificadores utilizando el conjunto de entrenamiento y de prueba, rescatando los resultados de aquéllos que muestran un buen desempeño.

Para cada clasificador se utiliza el valor semilla igual a 1 para generar los números aleatorios y se utilizan los conjuntos de entrenamiento y prueba antes descritos.

2.5.1. Ajuste de parámetros

Cada clasificador tiene un grupo de parámetros que se pueden fijar manualmente para obtener mejores resultados. A continuación se detallan los parámetros utilizados en cada método con tal de que los resultados obtenidos sean reproducibles.

La forma de determinar los parámetros óptimos es a través de inspección manual junto al método *CVParameterSelection* (Kohavi, 1995) de *WEKA* (Hall et al., 2009), el cual permite realizar una optimización de parámetros dentro de un intervalo determinado.

Clasificador bayesiano ingenuo

Se utiliza como verdadero el parámetro *Estimador kernel*, el cual no asume que las variables tienen una distribución normal. Esto se selecciona ya que por inspección manual se observa que mejora los resultados.

Regresión logística

Este método tiene sólo un parámetro ajustable, llamado *Ridge*, determinándose mediante una búsqueda con el método *CVParameterSelection* que su valor óptimo es 1000 para la clasificación sobre el conjunto de entrenamiento.

Perceptrón multicapa

Este método tiene múltiples parámetros ajustables, cuya optimización se determina mediante inspección manual y el método *CVParameterSelection* sobre el conjunto de entrenamiento.

Mediante el método *CVParameterSelection* se determinan dos parámetros numéricos, la *Tasa de aprendizaje* con valor óptimo 0.2, y el *Número de capas ocultas* con valor óptimo 2.

Mediante inspección manual se determinan dos parámetros binarios, se elige como verdadero el *Decaimiento de la Tasa de Aprendizaje* y falso la *Normalización de Atributos*.

Máquina de vectores de soporte

Para este método se utiliza el clasificador *LibSVM* (Chang y Lin, 2001), el cual cuenta con múltiples parámetros cuya optimización se determina mediante inspección manual y el método *CVParameterSelection* sobre el conjunto de entrenamiento.

Mediante el método *CVParameterSelection* se determinan dos parámetros numéricos, el parámetro *Costo* con valor óptimo 10, y el parámetro *Gamma* con valor óptimo de 2.2.

Mediante inspección manual se determina como óptima la utilización del *Tipo de SVM* correspondiente a *C-SVC* y un *Tipo de Kernel* lineal, además de fijar como verdadero el uso de *Estimación Probabilística*.

Bosque aleatorio

Este método tiene múltiples parámetros ajustables, cuya optimización se determina mediante inspección manual y el método *CVParameterSelection* sobre el conjunto de entrenamiento.

Mediante el método *CVParameterSelection* se determinan dos parámetros numéricos, se selecciona el parámetro *Número de Árboles* igual a 100 y el *Número de atributos* igual a 2.

Mediante inspección manual se determina el parámetro *Máxima profundidad del árbol* igual a 3.

2.5.2. Comparación del desempeño de los clasificadores

Para determinar el clasificador con mejor desempeño se comparan los valores *AUROC* de cada método, ya que en la literatura se considera como la mejor forma de comparar métodos de clasificación (Ling et al., 2003).

2.5.3. Predicción en el genoma de *Escherichia coli*

A partir del mejor clasificador encontrado en la sección anterior, se procede a aplicar el método sobre el genoma de la bacteria *Escherichia coli*, cepa *K-12*, subcepa *MG1655*.

El genoma bacteriano es explorado en ambos sentidos (directo y secuencia inversa complementaria) por secciones de diferentes longitudes de largo fijo, dividiéndolo en intervalos de largo 50, 100, 150, 200, 250 y 300 pares de bases. Avanzando en la lectura cada 25, 50, 75, 100, 125 y 150 pares de bases respectivamente. Este intervalo de ventanas se utiliza debido a que se encontró que el largo de los genes sARN se encuentra aproximadamente entre 50 y 300, como se observa en la Figura 2.1.

Si bien al avanzar la lectura en una distancia igual a la mitad de la ventana puede reducir la detección de secuencias positivas, al leer un genoma completo esto permite optimizar tiempo y es compensado con el hecho de usar ventanas de distinto largo. Además, esta misma técnica ha sido utilizada con buenos resultados por estudios anteriores (Tran, 2009).

Para las predicciones de cada intervalo de distinto largo, si existen predicciones que se traslapan más de un 50 % de su largo en hebras distintas, se conserva sólo aquella con mayor valor de confianza entregado por el clasificador. Luego, si existe traslape de predicciones en la misma hebra, se unen en una sola predicción promediando sus valores de confianza.

Posteriormente se combinan las predicciones con intervalos de distinto largo, repitiéndose el procedimiento anterior de remover predicciones traslapadas en hebras distintas, filtrando el resultado por distintos niveles de confianza.

Es importante notar que como un 36,2 % de los genes sARN conocidos en *Escherichia coli* se traslapa con genes que expresan proteínas validados experimentalmente, como se observa en la Tabla 2.1, en este estudio no se busca únicamente en regiones intergénicas sino que en el genoma completo.

Capítulo 3

Resultados y Discusión

A continuación se presentan los resultados obtenidos, inicialmente relacionados con la obtención de un clasificador óptimo para identificar secuencias de genes sARN, y posteriormente su aplicación sobre el genoma de la bacteria *Escherichia coli*.

3.1. Identificación de variables significativas

Se ha encontrado que eliminar atributos poco significativos mejora el desempeño de los métodos de clasificación y aprendizaje automático, ya que datos con mucho ruido o irrelevantes pueden confundir al clasificador (Han y Kamber, 2005), por lo que se procedió a realizar una prueba estadística para determinar y eliminar del conjunto de datos las variables poco significativas.

La prueba estadística utilizada corresponde a la *prueba de los signos de Wilcoxon* (Hollander y Wolfe, 1999) sobre el conjunto de entrenamiento, para determinar aquellas variables en las que se presenta una diferencia estadística significativa entre los valores de las clases sARN y aleatoria (positivo y negativo), usando como hipótesis nula que las variables para ambas clases tienen la misma media sin seguir necesariamente una distribución normal.

Las variables se muestran en la Tabla 3.1 ordenadas según su *p-value*, que corresponde a la probabilidad de obtener el resultado entregado por la prueba estadística si la hipótesis nula fuera cierta.

Aceptando un 1 % de error de tipo I, lo que corresponde al error asociado a mantener en el conjunto una variable que no es realmente significativa, se procedió a descartar todas las variables con un *p-value* mayor o igual a 0.01, con lo que el nuevo conjunto de datos sólo contiene aquellas variables en las que existe una diferencia estadística significativa entre los valores del conjunto positivo y negativo, con un nivel de confianza del 99 %.

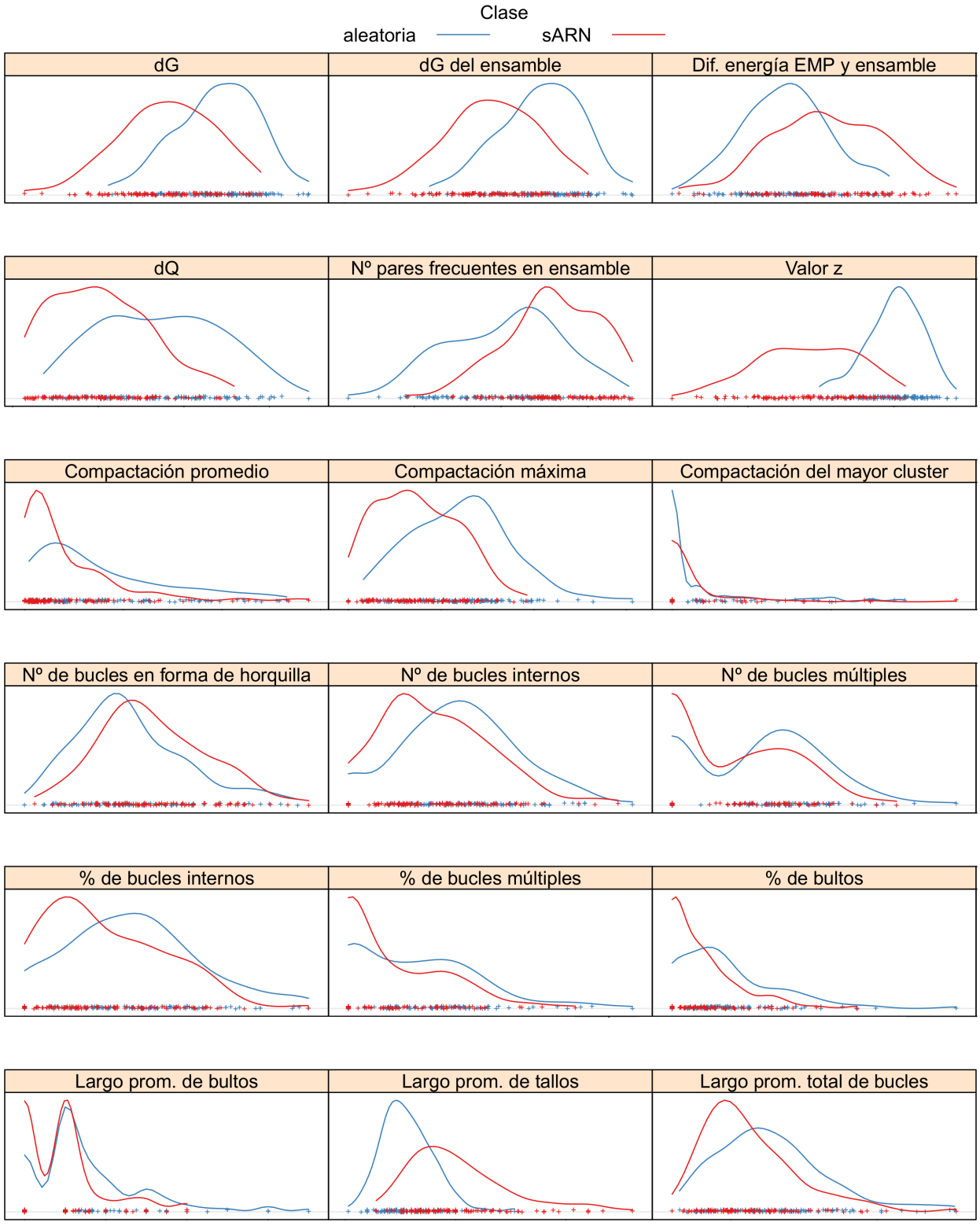
Con esto, se eliminan las últimas 7 variables de la Tabla 3.1 que están bajo el criterio de corte, quedando un conjunto de 33 variables significativas.

El hecho de que la mayoría de las variables medidas tenga un *p-value* bastante pequeño y sólo 7 sean descartadas según el criterio de corte, se explica porque estas variables han sido previamente utilizadas por otros métodos con resultados positivos, por lo que esta prueba estadística confirma que efectivamente estas variables son significativas en mayor o menor medida para discriminar a los genes sARN de secuencias aleatorias.

Además, se muestran las distribuciones de cada una de las 33 variables para ambas clases en la Figura 3.1. Como se observa en la figura, las diferencias en las distribuciones de cada variable son dispares. Es decir, mientras que en algunos casos hay una clara diferencia entre las distribuciones de las clases sARN y aleatoria (por ejemplo *Valor z*), en otras hay una diferencia despreciable (por ejemplo *Nº de bucles múltiples*). Por esto se hace necesario disminuir el número de variables, y así reducir el ruido en los datos para mejorar el funcionamiento del clasificador.

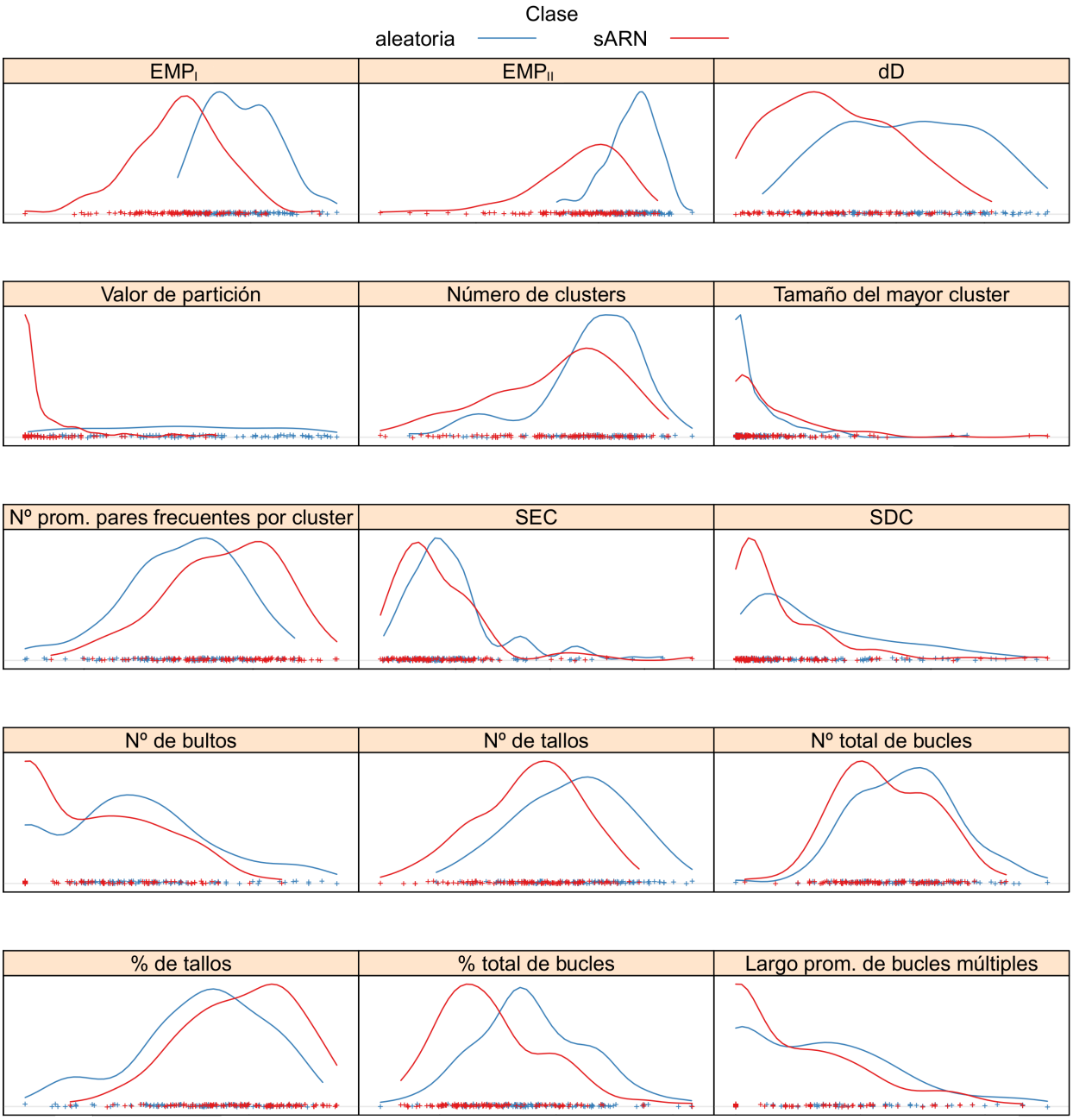
Tabla 3.1: Ranking de variables según prueba de los signos de Wilcoxon.

Variable	***p-value***
Valor z	$1{,}01 \cdot 10^{-18}$
Valor de partición	$1{,}15 \cdot 10^{-18}$
dG	$1{,}16 \cdot 10^{-18}$
dG del ensamble	$1{,}16 \cdot 10^{-18}$
EMP_I	$1{,}29 \cdot 10^{-18}$
EMP_{II}	$9{,}14 \cdot 10^{-17}$
Largo promedio de tallos	$4{,}49 \cdot 10^{-15}$
dQ	$7{,}79 \cdot 10^{-12}$
dD	$9{,}47 \cdot 10^{-12}$
Número de pares frecuentes en el ensamble	$1{,}59 \cdot 10^{-11}$
Número promedio de pares frecuentes por *cluster*	$1{,}09 \cdot 10^{-10}$
Compactación promedio	$1{,}45 \cdot 10^{-10}$
Compactación máxima	$2{,}49 \cdot 10^{-10}$
SDC	$1{,}22 \cdot 10^{-09}$
Porcentaje de tallos	$1{,}92 \cdot 10^{-08}$
Número de tallos	$2{,}29 \cdot 10^{-08}$
Porcentaje total de bucles	$3{,}36 \cdot 10^{-08}$
Número de *clusters*	$6{,}39 \cdot 10^{-08}$
Diferencia de energía entre EMP y ensamble	$9{,}72 \cdot 10^{-08}$
Tamaño del mayor *cluster*	$1{,}70 \cdot 10^{-07}$
Porcentaje de bultos	$8{,}87 \cdot 10^{-05}$
Número de bultos	$1{,}49 \cdot 10^{-04}$
Número de bucles múltiples	$1{,}74 \cdot 10^{-04}$
Largo promedio total de bucles	$2{,}56 \cdot 10^{-04}$
Porcentaje de bucles múltiples	$2{,}69 \cdot 10^{-04}$
Número total de bucles	$4{,}02 \cdot 10^{-04}$
Largo promedio de bultos	$6{,}43 \cdot 10^{-04}$
Porcentaje de bucles internos	$1{,}10 \cdot 10^{-03}$
Largo promedio de bucles múltiples	$1{,}20 \cdot 10^{-03}$
Compactación del mayor *cluster*	$1{,}29 \cdot 10^{-03}$
SEC	$2{,}72 \cdot 10^{-03}$
Número de bucles en forma de horquilla	$3{,}39 \cdot 10^{-03}$
Número de bucles internos	$6{,}21 \cdot 10^{-03}$
Criterio de corte: *p-value* $< 10^{-02}$	
Largo promedio de bucles en forma de horquilla	$1{,}20 \cdot 10^{-02}$
Largo promedio de estructuras externas	$4{,}60 \cdot 10^{-02}$
Porcentaje de estructuras externas	$6{,}20 \cdot 10^{-02}$
Largo promedio de bucles internos	$1{,}20 \cdot 10^{-01}$
Compactación mínima	$1{,}42 \cdot 10^{-01}$
Porcentaje de bucles en forma de horquilla	$5{,}09 \cdot 10^{-01}$
Número de estructuras externas	1,00



(a)

Figura 3.1: Distribuciones de las variables significativas. En (a) se muestran las primeras 18 y en (b) las 15 restantes (página siguiente). En rojo se muestra la clase sARN y en azul la clase aleatoria, además en el eje vertical se grafica la frecuencia de la variable y en el eje horizontal el valor de la variable.



(b)

Figura 3.1: Distribuciones de las variables significativas. En (a) se muestran las primeras 18 (página anterior) y en (b) las 15 restantes. En rojo se muestra la clase sARN y en azul la clase aleatoria, además en el eje vertical se grafica la frecuencia de la variable y en el eje horizontal el valor de la variable.

3.2. Selección de Atributos

Se ha encontrado que los métodos presentan un mejor desempeño si los atributos utilizados son independientes entre sí, por lo cual se procede a eliminar las variables que presentan un alto nivel de correlación (Han y Kamber, 2005).

Por este motivo, además de eliminar las variables que no son significativas estadísticamente, se procede a seleccionar el subconjunto de variables que mejora el comportamiento de los algoritmos de clasificación. Esto es, aquellas variables que tienen mayor correlación con la clase y la menor correlación entre sí.

Para la selección se utilizan las herramientas de selección de atributos del programa *WEKA* (Hall et al., 2009) sobre el conjunto de entrenamiento. En primera instancia, el método *CfsSubsetEval*, el cual selecciona el subconjunto de variables con mayor correlación con respecto a la clase y la menor correlación entre sí. Posteriormente se utiliza el método *WrapperSubsetEval*, el cual utiliza un método de clasificación sobre los datos para elegir el subconjunto óptimo de variables que maximice el $AUROC$ (área cubierta bajo la curva ROC).

3.2.1. Método *CfsSubsetEval*

Este método encuentra el subconjunto de variables con mayor mérito, definido como el de aquellas variables con mayor capacidad predictiva individual y que tienen la menor redundancia entre sí (Hall, 1999).

De esta forma se obtiene un subconjunto de variables altamente correlacionadas con la clase, intentando mantener un bajo nivel de correlación lineal entre ellas.

Debido a que existen 33 variables, y con ello existen $2^{33} \approx 8{,}6 \cdot 10^9$ subconjuntos posibles (según coeficiente binomial), se utiliza una heurística de búsqueda para evaluar el espacio de soluciones.

Para la búsqueda se utilizan 2 heurísticas distintas, *BestFirst* y *GeneticSearch*, entregando ambas el mismo resultado de 7 variables que se muestran en la Tabla 3.2.

Tabla 3.2: Selección de Atributos por método *CfsSubsetEval*.

Variable
EMP_I
EMP_{II}
Valor z
Valor de partición
Compactación máxima
Porcentaje de bultos
Largo promedio de tallos

3.2.2. Método *WrapperSubsetEval*

El método selecciona el subconjunto de atributos usando un método de clasificación (Kohavi y John, 1997) sobre el conjunto de entrenamiento. En este caso, al haber sólo 7 variables, se explora el espacio completo de subconjuntos mediante una búsqueda exhaustiva con el método *ExhaustiveSearch*, utilizando como término a maximizar el valor $AUROC$ (área bajo la curva ROC).

La búsqueda se repite para los clasificadores *Bayesiano Ingenuo*, *Regresión Logística*, *Perceptrón Multicapa*, *Máquina de Vectores de Soporte* y *Bosque Aleatorio*.

Los resultados de cada clasificador, junto al subconjunto escogido y el valor $AUROC$, se muestran en la Tabla 3.3.

Tabla 3.3: Valores *AUROC* de los clasificadores sobre el subconjunto de entrenamiento con distintos grupos de variables seleccionadas mediante el método *WrapperSubsetEval*.

Método clasificador	**Variables escogidas**	**AUROC**
Bayesiano Ingenuo	*Valor z*, *Porcentaje de bultos*	93,7 %
Regresión Logística	*Valor z*, *Porcentaje de bultos*, *Valor de partición*	94,2 %
Perceptrón Multicapa	*Valor z*	93,0 %
Máquina de Vectores de Soporte	*Valor z*, *Porcentaje de bultos*	93,6 %
Bosque Aleatorio	*Valor z*, *Porcentaje de bultos*, *Valor de partición*, EMP_I	93,5 %

Tomando en cuenta estos resultados, se decide eliminar las variables no seleccionadas por ningún clasificador, las cuales son EMP_{II}, *Compactación máxima* y *Largo promedio de tallos*. El subconjunto de variables resultante se muestra en la Tabla 3.4.

Recalculando la eficiencia de cada método con este nuevo subconjunto de variables, se observa que se logra mantener un valor *AUROC* alto para todos los casos sin comprometer demasiado el valor máximo obtenido anteriormente usando subconjuntos distintos. Esto se muestra en la Tabla 3.5.

Tabla 3.4: Selección de Atributos por método *WrapperSubsetEval*.

Variable
EMP_I
Valor z
Valor de partición
Porcentaje de bultos

Tabla 3.5: Valores *AUROC* de los clasificadores sobre el subconjunto de entrenamiento con las variables seleccionadas mediante el método *WrapperSubsetEval*.

Método clasificador	**AUROC**
Bayesiano Ingenuo	93,5 %
Regresión Logística	93,6 %
Perceptrón Muticapa	92,8 %
Máquina de Vectores de Soporte	93,3 %
Bosque Aleatorio	93,5 %

3.2.3. Atributos Seleccionados

Luego de la selección de atributos, quedan un total de 4 variables de las 40 iniciales, las que se muestran en la Tabla 3.4. A continuación se analizan estas variables mediante distintos gráficos.

En la Figura 3.2 se muestran las variables seleccionadas representadas por diagramas de caja. En este gráfico se observa claramente que el valor de la mediana para cada variable es mayor en el caso de las secuencias aleatorias, pero aun así existen zonas de traslape entre ambas clases.

Un caso interesante se observa para la variable *Valor de Partición*, ya que la clase aleatoria muestra una distribución a lo largo de todo el intervalo $(0, 1)$, mientras que la distribución de la clase *sARN* se concentra cerca de 0. Por otro lado, la variable *Porcentaje de bultos* muestra una diferencia relativa menor entre ambas clases.

En la Figura 3.3 se muestran las variables representadas mediante histogramas. En este gráfico se pueden ver claramente las diferencias entre las distribuciones de las clases sARN y aleatoria para cada variable.

En el caso de la variable EMP_I se observa que la clase sARN sigue una distribución normal centrada en $-0{,}7$, mientras que la clase aleatoria pareciera ser la unión de dos normales centradas en $-0{,}6$ y $-0{,}45$ aproximadamente. Si bien la clase sARN se distribuye en un rango mayor de valores de energía $(-1{,}1, -0{,}4)$ en promedio tiene una energía de plegamiento menor que las secuencias al azar, lo cual es un resultado esperable ya que le atribuye un carácter más estable a las moléculas de sARN en comparación con secuencias aleatorias.

En el caso de la variable $Valor\ z$ se observa una gran diferencia entre las distribuciones, teniendo la clase aleatoria una distribución normal en torno a 0 y la clase sARN una distribución más amplia entre $(-8, 1)$ aproximadamente. Esto muestra que las secuencias sARN tienen en su gran mayoría varias desviaciones estándar de diferencia entre su valor de energía mínima de plegamiento y los valores de una muestra de secuencias aleatorias.

En el caso de la variable $Valor\ de\ partición$ se observa también una clara diferencia, ya que la clase aleatoria se distribuye uniformemente por todo el intervalo $(0, 1)$ mientras que la clase sARN está principalmente concentrada cercana a 0. Esto muestra que la gran mayoría de secuencias sARN tiene menor energía mínima de plegamiento que una muestra de secuencias aleatorias.

Por último, para la variable *Porcentaje de bultos*, se observa una diferencia un poco menos marcada, pero de todas formas se puede ver que la clase sARN tiene un número de bultos cercano a 0, lo cual es significativamente menor que la clase aleatoria, por lo que se deduce que sería una estructura secundaria poco común en las moléculas sARN.

En la Figura 3.4 se muestra una matriz de gráficos de dispersión junto a los coeficientes de correlación de *Pearson*. En este gráfico se observa que 3 de las variables seleccionadas tienen un alto nivel de correlación lineal, estas son EMP_I, $Valor\ z$ y $Valor\ de\ partición$.

Esto se explica ya que las tres variables corresponden a un cálculo sobre la energía mínima de plegamiento (EMP) de la secuencia sARN, por lo que parten de la misma información. Por esto existe el gráfico de dispersión entre EMP_I y $Valor\ z$ que se asemeja a una línea recta. Esto significa que hay una fuerte correlación lineal entre ambas variables.

Por otro lado el gráfico de dispersión entre las variables $Valor\ z$ y $Valor\ de\ partición$ muestra una fuerte correlación no lineal entre las variables, lo que se explica debido a que ambas son calculadas a partir de la misma información. Sin embargo, al calcularse de distinta forma la correlación no es lineal. Esto también explica por qué existe una alta correlación entre las variables $Valor\ de\ partición$ y EMP_I.

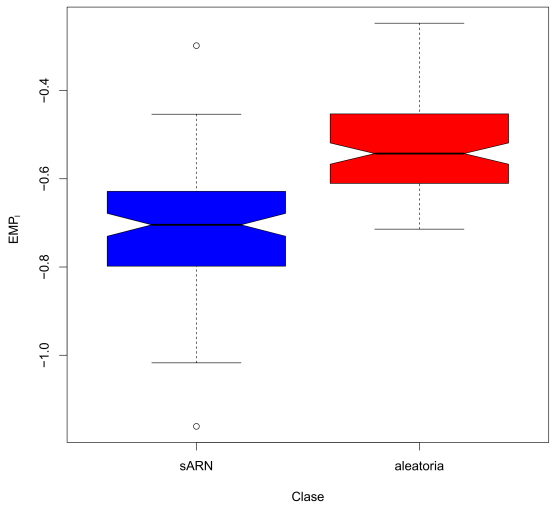
Por último, la variable *Porcentaje de bultos* tiene muy baja correlación con todas las demás variables, lo que evidencia que no existe una correlación entre esta estructura y la energía mínima de plegamiento.

Tomando en cuenta todo el conjunto de variables seleccionadas, 3 de ellas muestran un alto nivel de correlación, lo cual se explica debido a que todas se relacionan con la energía mínima de plegamiento (EMP) de la secuencia sARN.

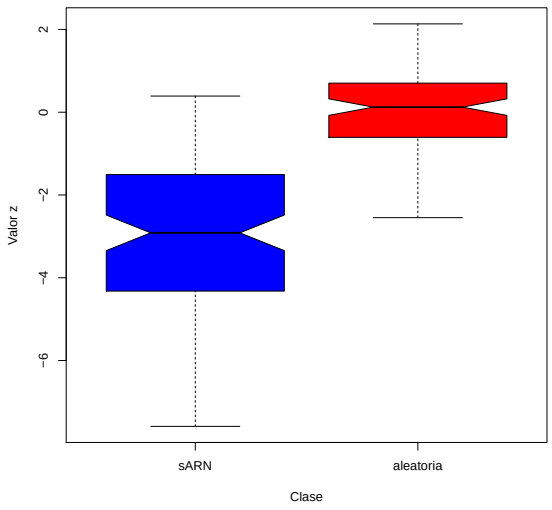
Se seleccionaron las 3 variables puesto que se consideraron dos condiciones para encontrar el mejor subconjunto. La primera es que la variable tenga distribuciones significativamente distintas entre las clase sARN y aleatoria, mientras que la segunda es que la variable tenga un bajo nivel de correlación con las demás del subconjunto.

Por estos criterios, las 3 variables fueron escogidas porque, a pesar de su alto grado de correlación, tienen una alta capacidad de discriminación entre las clases (ver Figura 3.3) y al sacar cualquiera de ellas del subconjunto la clasificación empeora, por lo que la información contenida en cada variable no sería totalmente redundante entre sí.

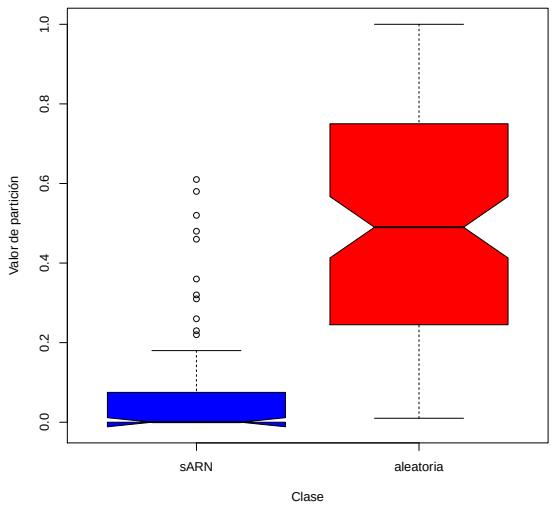
Mientras que la variable *Porcentaje de bultos*, que no esta correlacionada con ninguna otra, aunque no tiene tanta capacidad de discriminación (ver Figura 3.3), al tener baja correlación con las demás cumple el



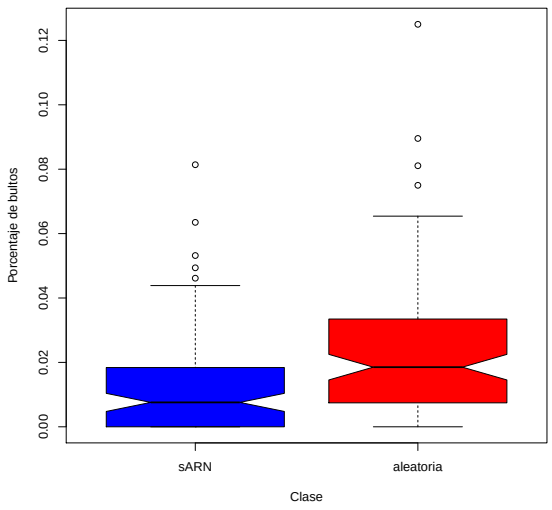
(a) Diagrama de Caja para variable EMP_I.



(b) Diagrama de Caja para variable *Valor z*.

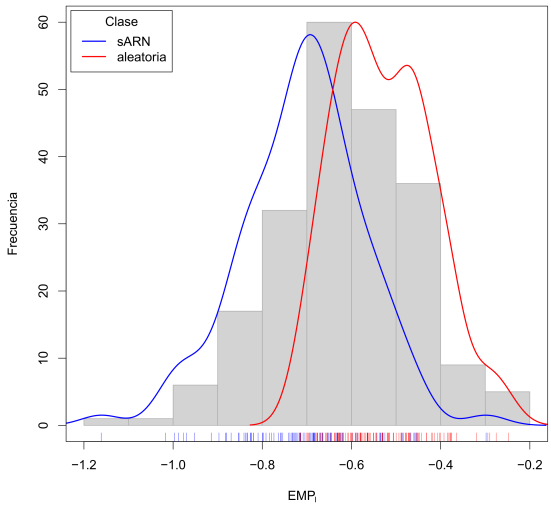


(c) Diagrama de Caja para variable *Valor de partición*.

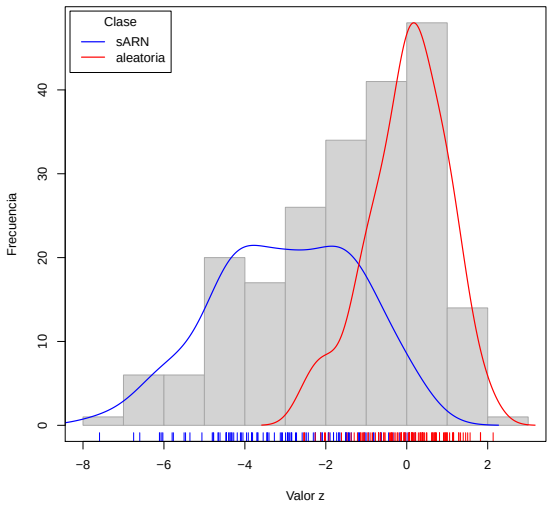


(d) Diagrama de Caja para variable *Porcentaje de bultos*.

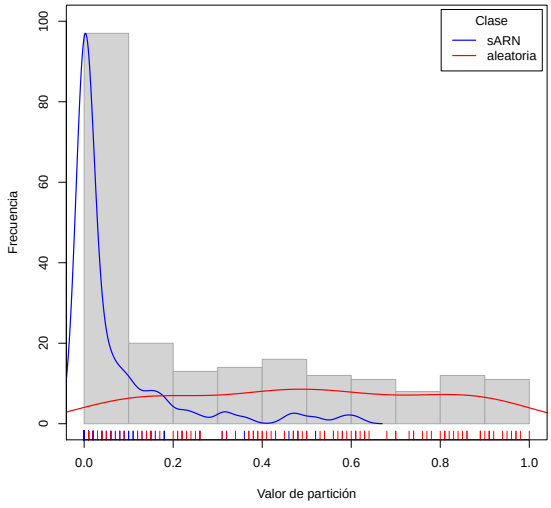
Figura 3.2: Diagramas de Caja de las variables seleccionadas. La caja azul representa a la clase sARN y la roja a la clase aleatoria. La barra superior e inferior corresponden al máximo y mínimo valor de la variable, el extremo inferior y superior de la caja son el 1º y 3º cuartil respectivamente, siendo la barra en el centro la mediana. Los puntos dibujados fuera de las cajas son considerados valores atípicos.



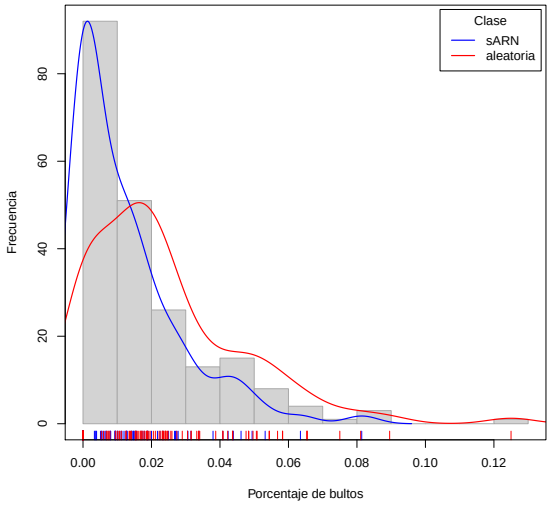
(a) Histograma para variable EMP_I.



(b) Histograma para variable *Valor z*.



(c) Histograma para variable *Valor de partición*.



(d) Histograma para variable *Porcentaje de bultos*.

Figura 3.3: Histogramas de las Variables Seleccionadas. Las barras verticales representan la frecuencia de la variable en el total de las clases, mientras que la curva azul representa la frecuencia individual de la clase sARN y la curva roja la frecuencia individual de la clase aleatoria.

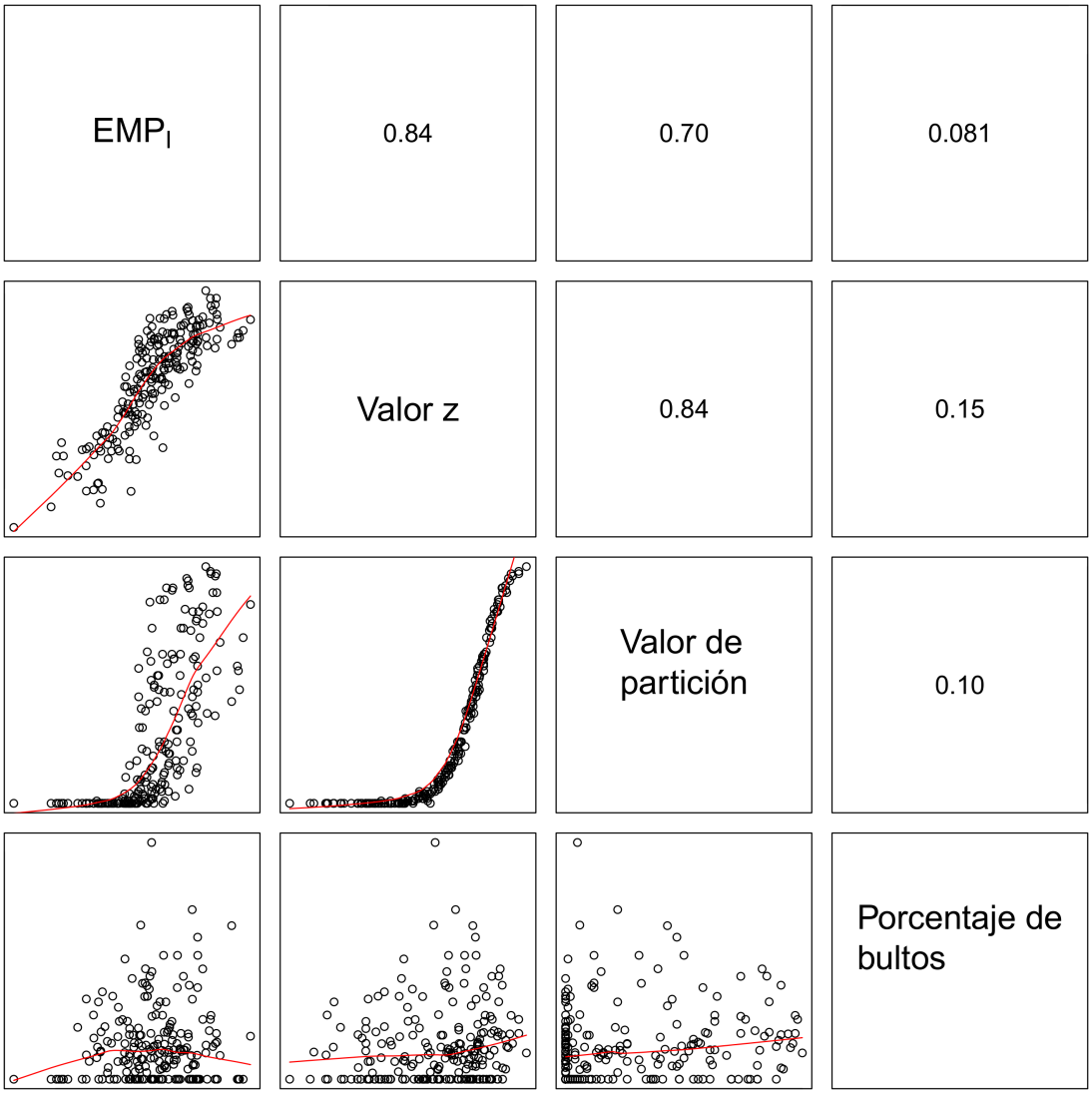


Figura 3.4: Matriz de gráficos de correlación entre las variables seleccionadas. En la diagonal inferior se muestran gráficos de dispersión, correspondiendo los ejes a los valores de las variables en la diagonal. En la diagonal superior se muestra el valor del coeficiente de correlación de *Pearson* entre las variables de la diagonal.

segundo criterio y aporta información distinta al conjunto de variables.

Con respecto a las variables seleccionadas, es importante destacar que 3 de las 4 están relacionadas con la energía mínima de plegamiento de la molécula de ARN no codificante. Del análisis de estas variables se desprende que una característica fundamental de los genes sARN es que su transcrito tiene una energía de plegamiento significativamente menor que la de secuencias aleatorias con el mismo porcentaje dinucleotídico, como se demuestra por la selección de las variables *Valor z* y *Valor de partición*. Esto quiere decir que la estabilidad de la estructura secundaria no se debe simplemente a su contenido de pares *G-C*, como demuestra la variable EMP_I, sino que el diseño de la estructura secundaria juega un rol fundamental en estabilizar la molécula.

Por otra parte, la selección de la variable *Porcentaje de bultos* indica que la ausencia de la estructura de bulto es una característica importante, al presentarse menos frecuentemente en las secuencias de sARN estudiadas que lo obtenido en las secuencias de ARN al azar. Lo anterior podría indicar la existencia de una presión selectiva en la evolución de estos genes con tendencia a eliminar los bultos. Esto puede explicarse debido a que el elemento bulto produce un efecto desestabilizante sobre la estructura de la molécula de ARN. La desestabilización se debe principalmente al mayor grado de libertad de movimiento debido al bulto, ya que esto aumenta la entropía de la conformación tridimensional y con ello la energía de plegamiento de la molécula (Hermann y Patel, 2000; Blose et al., 2007).

3.3. Clasificación

Se probaron diferentes algoritmos de clasificación, implementados por el programa de minería de datos *WEKA* (Hall et al., 2009), de los que se presentan los 5 con mejor desempeño a continuación.

Para cada método, se ajustaron los parámetros del modelo utilizando el conjunto de entrenamiento y luego se aplicó sobre el conjunto de prueba para obtener los resultados de la clasificación. De esta forma se evita un sobreajuste del modelo a los datos.

3.3.1. Estadísticas de la Clasificación

A continuación se muestra la matriz de confusión obtenida para cada clasificador en la Tabla 3.6. De las matrices se desprende que todos los métodos clasifican correctamente la mayoría de las secuencias. Sin embargo, es difícil compararlos sólo mediante esta información, por lo cual resulta necesario calcular estadísticas que describan de forma mas objetiva la eficiencia del clasificador.

Tabla 3.6: Matriz de Confusión para los distintos métodos.

Clasificador	Clase real	Clasificado como sARN	Clasificado como aleatoria
Bayesiano Ingenuo	**sARN**	1547	192
	aleatoria	282	1457
Regresión Logística	**sARN**	1567	172
	aleatoria	349	1390
Perceptrón Multicapa	**sARN**	1544	195
	aleatoria	253	1486
Bosque Aleatorio	**sARN**	1506	233
	aleatoria	215	1524
Máquina de Vectores de Soporte	**sARN**	1454	285
	aleatoria	147	1592

En la Tabla 3.7 se muestra una serie de estadísticas que describen diferentes aspectos del desempeño del clasificador. De estos datos se desprende que el método más sensible corresponde a la *Regresión Logística*. Mientras que el más específico, más exacto y con menor tasa de falsos positivos corresponde a la *Máquina de Vectores de Soporte*.

Si bien este último clasificador tiene muchas de las estadísticas más altas también tiene la sensibilidad más baja, por lo que es necesario otro tipo de análisis que tome en cuenta las características del desempeño en su conjunto.

Tabla 3.7: Estadísticas de la clasificación para la clase sARN con los distintos métodos.

Clasificador	**Sensibilidad**	**Especificidad**	**Tasa de Falsos Positivos**	**Valor Predictivo Positivo**	**Exactitud**
Bayesiano Ingenuo	89,0 %	83,8 %	16,2 %	84,6 %	86,4 %
Regresión Logística	90,1 %	79,9 %	20,1 %	81,8 %	85,0 %
Perceptrón Multicapa	88,8 %	85,5 %	14,5 %	85,9 %	87,1 %
Bosque Aleatorio	86,6 %	87,6 %	12,4 %	87,5 %	87,1 %
Máquina de Vectores de Soporte	83,6 %	91,5 %	8,5 %	90,8 %	87,6 %

3.3.2. Comparación del desempeño

Debido a que resulta difícil comparar el desempeño general de los clasificadores con las estadísticas mostradas en la Tabla 3.7, se busca otro método de comparación y se encuentra que la mejor forma de comparar métodos de clasificación es a través del valor *AUROC*, que corresponde al área cubierta bajo la curva *ROC* de cada método para la clase objetivo (Ling et al., 2003).

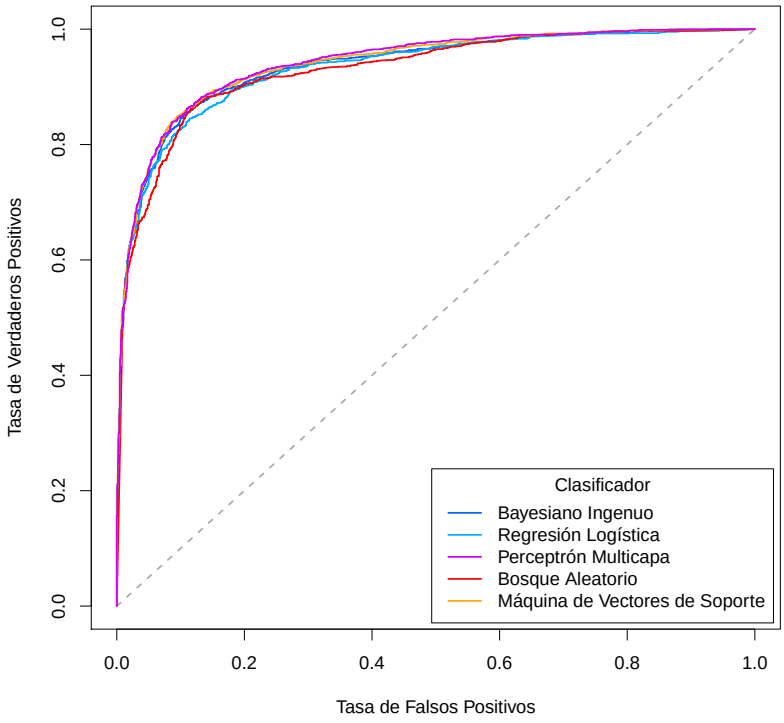
Por esto, se grafican las curvas ROC en la Figura 3.5 y se mide el área bajo su curva, con lo cual se construye la Tabla 3.8 que muestra la comparación de los resultados del valor *AUROC* para todos los métodos de clasificación.

Ya que el valor *AUROC* permite comparar el desempeño general de los clasificadores sin hacer énfasis en sólo una característica, se determina que el clasificador más eficiente corresponde al método de *Perceptrón Multicapa*, el que será utilizado para la predicción sobre el genoma de la bacteria *Escherichia coli*.

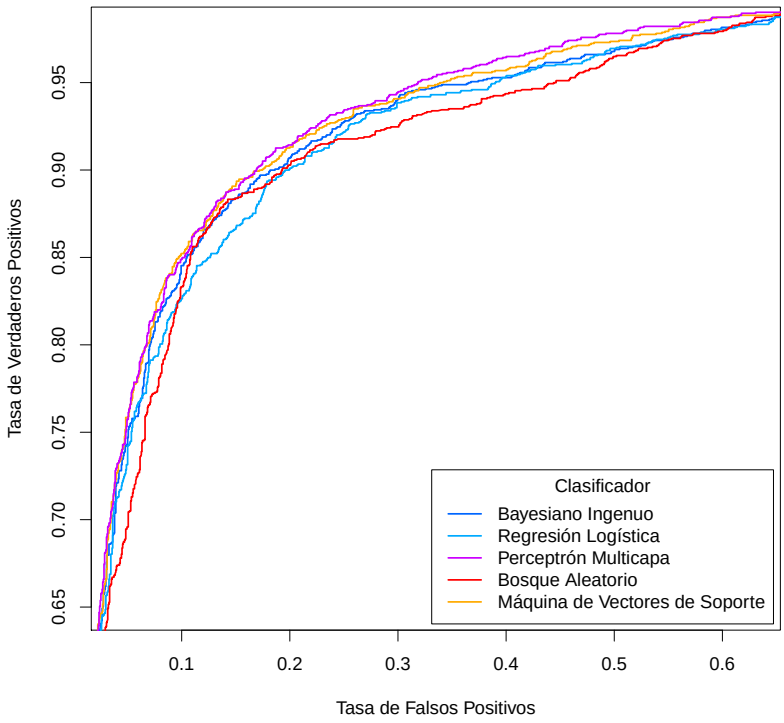
Este resultado tiene sentido al observar que si bien este método no tiene el valor más alto para ninguna de las estadísticas medidas en la Tabla 3.7, sí tiene consistentemente un valor alto para cada una de ellas, lo que concuerda con el resultado de tener el valor más alto en el análisis por la curva *ROC*.

Tabla 3.8: Ranking de Clasificadores según valor *AUROC*.

Clasificador	**AUROC**
Perceptrón Multicapa	94,2 %
Máquina de Vectores de Soporte	94,1 %
Bayesiano Ingenuo	93,7 %
Regresión Logística	93,3 %
Bosque Aleatorio	93,0 %



(a)



(b)

Figura 3.5: Curvas *ROC* para todos los clasificadores. En (a) se muestran las curvas *ROC* en su escala completa. En (b) se muestran las mismas curvas con los ejes ajustados para resaltar sus diferencias.

3.4. Justificación de la selección de atributos

Para comprobar que la reducción de variables mediante los métodos de selección de atributos contribuye a mejorar la eficacia del clasificador, se repite el método de *Perceptrón Multicapa* sobre el conjunto de datos de entrenamiento y prueba, para los subconjuntos de variables durante cada paso de la selección de atributos.

Los resultados obtenidos del valor *AUROC* (área bajo la curva *ROC*) para cada subconjunto de variables se muestran en la Tabla 3.9, mientras que en la Figura 3.6 se muestran las curvas *ROC* para los subconjuntos, siendo el de mejor resultado el subconjunto final de 4 variables seleccionadas.

Este resultado es relevante, ya que demuestra que es posible mejorar el desempeño del clasificador utilizando un menor número de variables. Esto ofrece 2 grandes ventajas. La primera es que simplifica el algoritmo y reduce el tiempo de ejecución al realizar menos cálculos para medir variables. La segunda ventaja es que permite extraer conocimiento específico acerca de cuáles son las variables más importantes a medir para diferenciar a los genes sARN de entre el total de variables medidas en la literatura, lo cual permite mejorar el entendimiento de estos genes y enfocar futuros estudios sobre características relacionadas con estas variables.

Además, este resultado es consistente con lo conocido en la literatura, ya que se sabe que la estructura secundaria y energía mínima de plegamiento son características claves para la estabilidad y función de las moléculas sARN, lo cual valida desde el punto de vista teórico el hecho de que estas 4 variables sean las más importantes entre las 40 estudiadas.

Tabla 3.9: Comparación del valor *AUROC* para el clasificador óptimo y diferentes conjuntos de variables.

Conjunto de variables	**AUROC**
Conjunto final de 4 variables seleccionadas por método *WrapperSubsetEval*	94,2 %
Conjunto de 7 variables seleccionadas por método *CfsSubsetEval*	93,9 %
Conjunto de 33 variables significativas según prueba de Wilcoxon	92,3 %
Conjunto inicial de 40 variables recopiladas en la literatura	92,0 %

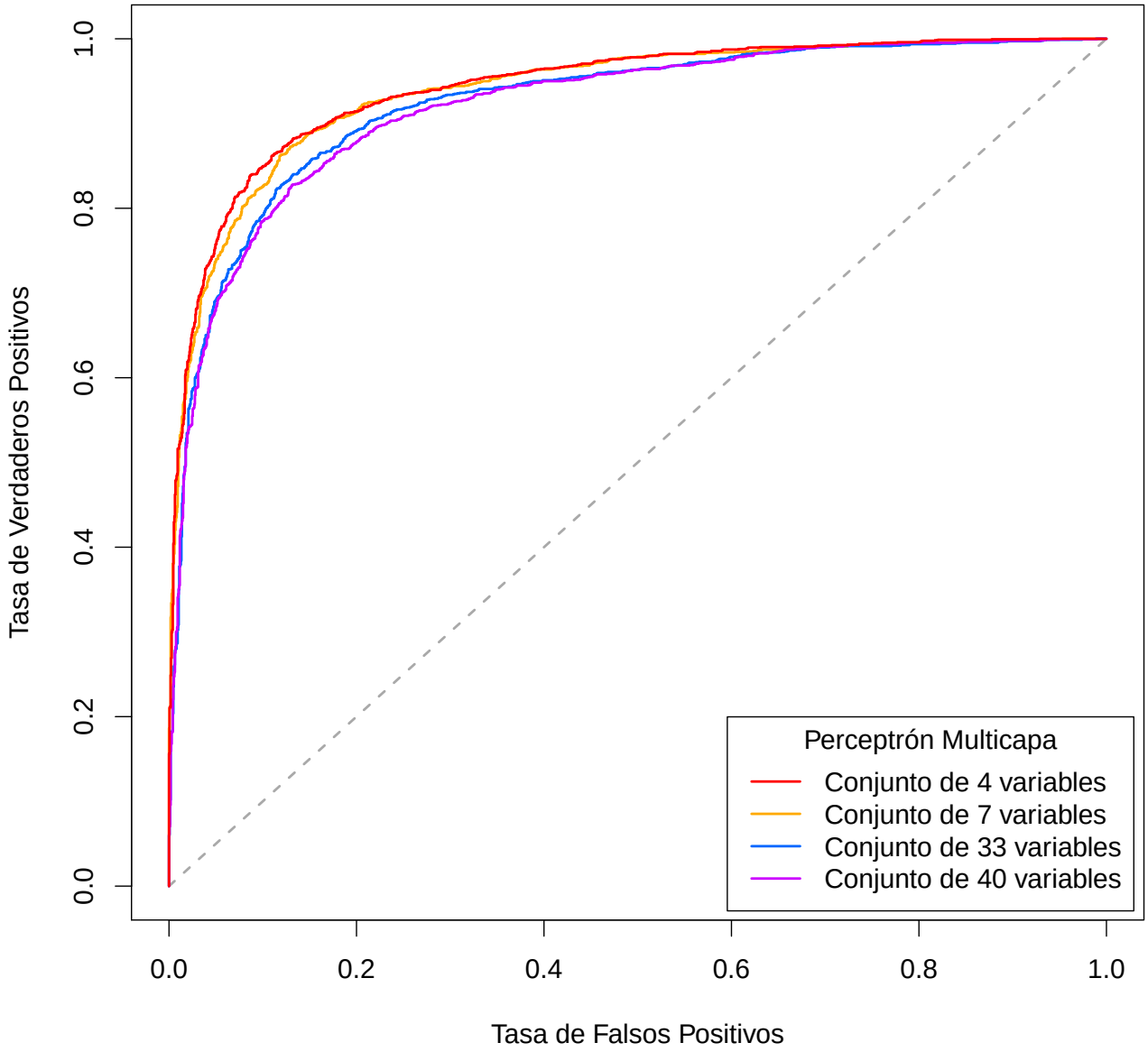


Figura 3.6: Comparación de curvas *ROC* para clasificación con método de *Perceptrón Multicapa* y distintos conjuntos de variables.

3.5. Predicción de genes sARN en genoma de *Escherichia coli*

Se aplica el método de *Perceptrón Multicapa* sobre el genoma de *Escherichia coli* utilizando las 4 variables más significativas obtenidas tras la selección de atributos (Tabla 3.4).

Para realizar las predicciones, se recorre el genoma en intervalos de 50, 100, 150, 200, 250 y 300 pares de bases. Luego se agrupan las predicciones para intervalos de distinto largo y en caso de traslape se elimina aquella que posea menor nivel de confianza entregada por el método clasificador.

Con este procedimiento se genera un conjunto de predicciones con valores de confianza entre 0,5 y 0,855. Valores menores a 0,5 son consideradas secuencias aleatorias.

Debido a la gran cantidad de predicciones generadas por la unión de las predicciones de los distintos intervalos, se decide refinar el resultado obtenido estableciendo un criterio de corte sobre el valor de confianza, y de este modo obtener un conjunto de predicciones más relevante. Por este motivo se presentan distintos niveles de cortes para el valor de confianza en la Tabla 3.10, con el objetivo de mostrar que existe una relación inversa entre la sensibilidad y el número de predicciones lo que hace crecer el porcentaje del genoma que es considerado como sARN.

Respecto a la posición de estas predicciones respecto a los genes que expresan proteínas, para el caso con confianza mínima de 0,80 se tiene que el 59,0 % de las predicciones se encuentran en regiones intergénicas del genoma, estadística que concuerda con la de los genes sARN conocidos (ver Tabla 2.1).

Tabla 3.10: Predicciones sobre el genoma de *Escherichia coli* con distintos cortes en el valor de confianza.

Confianza mínima	Número de predicciones	Genes recuperados	Sensibilidad	Valor predictivo positivo	Porcentaje del genoma
0,80	28831	45	76,27 %	0,16 %	40,28 %
0,75	42325	53	89,83 %	0,13 %	54,26 %
0,70	52526	55	93,22 %	0,10 %	62,94 %
0,50	86910	56	94,92 %	0,06 %	81,87 %

Sin embargo, como el objetivo es generar un listado pequeño de predicciones con alta probabilidad de ser genes sARN para proponer secuencias nóveles y poder ser comprobados experimentalmente, se decide cruzar las predicciones con un listado de las posiciones de promotores σ^{70} y terminadores intrínsecos (independientes del factor ρ) validados experimentalmente, lo cual produce un listado significativamente menor de predicciones las cuales son más significativas y se muestran en la Tabla 3.11.

Si bien esta predicción tiene una sensibilidad bastante menor a las anteriores, el valor predictivo positivo es considerablemente mayor al reducir el número total de predicciones. Lo anterior implica que cada predicción tiene una probabilidad mucho más alta de ser un gen sARN, por lo que resultaría más indicada para comprobarla experimentalmente.

Respecto a la posición de estas predicciones respecto a los genes que expresan proteínas, se tiene que sólo el 38,4 % de las predicciones se encuentra en regiones intergénicas del genoma, lo cual podría explicarse debido a que la mayoría de los promotores conocidos se descubren mediante su asociación a genes que codifican proteínas.

3.5.1. Comparación con otros métodos

Se compara el desempeño del método con el desempeño logrado por distintos autores en la literatura, los cuales se muestran en la Tabla 3.12 con el motivo de poder comparar eficazmente. Estos corresponden sólo a métodos que han sido previamente aplicados al genoma de *Escherichia coli*.

Según estos datos, el método propuesto genera resultados comparables a métodos existentes, teniendo

Tabla 3.11: Predicciones sobre el genoma de *Escherichia coli* cercanas a un promotor σ^{70} o terminador intrínseco.

Número de predicciones	Genes recuperados	Sensibilidad	Valor predictivo positivo	Porcentaje del genoma
1192	18	30,50 %	1,51 %	2,47 %

como beneficio que no depende de información de homología de bacterias cercanas filogenéticamente, y basta con medir sólo 4 variables, lo que simplifica y agiliza los cálculos.

Tabla 3.12: Predicciones sobre el genoma de *Escherichia coli* por distintos autores en la literatura (Tran, 2009).

Autor	Número de predicciones	Sensibilidad	Valor predictivo positivo
Carter, 2001	563	33,41 %	5,68 %
Chen, 2002	227	29,03 %	11,89 %
Rivas, 2001	275	40,86 %	13,82 %
Saestrom, 2005	306	11,83 %	3,59 %
Wang, 2006	420	7,53 %	1,67 %
Tran, 2009	601	40,86 %	6,32 %
Rogers, 2012	1192	30,50 %	1,51 %

3.5.2. Listado de secuencias propuestas

A partir del listado de predicciones, se selecciona un subconjunto de secuencias con alta probabilidad de ser genes sARN nóveles, las cuales se seleccionan en base a la existencia de un promotor σ^{70} conocido a menos de 50 pares de base del inicio de la predicción y tener un terminador intrínseco a menos de 50 pares de base del final de la predicción, además de no traslapar con genes sARN conocidos. Según este criterio se seleccionan 5 secuencias, las que se muestran en la Tabla 3.13 junto a su posición dentro del genoma (límites izquierdo y derecho). Estas secuencias se proponen como genes sARN nóveles que pueden ser verificadas en el laboratorio.

Tabla 3.13: Secuencias sARN propuestas.

Límite izquierdo	Límite derecho	Largo	Hebra	Confianza	Distancia a promotor	Distancia a terminador
83724	83625	100	−1	0,673	11	22
3316149	3316100	50	−1	0,805	48	35
3316299	3316150	150	−1	0,829	20	41
3316374	3316075	300	−1	0,854	23	10
3908549	3908500	50	−1	0,818	16	4

Capítulo 4

Conclusiones

Se recopiló un total de 40 variables utilizadas previamente en la literatura para identificar secuencias de genes sARN, de las cuales 33 mostraron diferencias significativas entre el conjunto de entrenamiento y secuencias aleatorias.

Con las 33 variables significativas se realizó una selección de atributos, lo que permitió reducir el conjunto de variables a 4, las cuales son: *Valor z*, *Valor de partición*, EMP_I y *Porcentaje de bultos*. Esta reducción de variables permitió mejorar el desempeño del clasificador respecto al conjunto inicial de 40 variables, por lo cual se concluye que son las variables más relevantes que permiten identificar efectivamente a las secuencias de genes sARN eliminando el ruido aportado por las demás variables.

Las variables seleccionadas muestran que, de todas las variables estudiadas, los atributos más eficaces para identificar a los genes sARN tienen relación directa con su estructura secundaria. En particular, se encontró que estas moléculas tienen una energía mínima de plegamiento y un porcentaje de estructuras de bultos significativamente menor que lo esperado por azar, en secuencias aleatorias del mismo largo y composición dinucleotídica, por lo que su estabilidad no estaría dada sólo por su composición nucleotídica sino por la complejidad del diseño de su estructura y la reducción de elementos desestabilizantes como bultos.

Estas características fundamentales muestran la existencia de una presión selectiva en la evolución de los genes sARN que apunta a aumentar la estabilidad de la molécula de ARN, modificando su estructura para disminuir la energía de plegamiento y eliminar subestructuras desestabilizantes no funcionales.

De los 5 métodos de clasificación estudiados, el mejor resultado se obtuvo con el método *Perceptrón Multicapa* basado en redes neuronales, según el criterio de valor $AUROC$ (94,2 %). Este método obtuvo un buen desempeño generalizado, con una alta sensibilidad ($88,8\,\%$) y alta especificidad ($85,5\,\%$), teniendo una tasa de falsos positivos moderadamente baja ($14,5\,\%$).

Se aplicó el método sobre el genoma completo de la bacteria *Escherichia coli* y se cruzaron estas predicciones con un listado de promotores σ^{70} y terminadores intrínsecos validados experimentalmente, obteniendo así resultados similares al de otros métodos en la literatura. Se obtuvo un total de 1192 predicciones lo que corresponde a un 2,47 % del total del largo del genoma, con un valor de sensibilidad de 30,5 % y un valor predictivo positivo de 1,51 % respecto a los genes sARN conocidos en la bacteria.

Los resultados obtenidos son similares a el de los métodos existentes, con la ventaja de no requerir previamente con información de organismos cercanos filogenéticamente, lo que permite utilizar este método en organismos de los que no se tiene información de genes sARN homólogos.

Por último se proponen 5 secuencias con alta probabilidad de corresponder a genes sARN nóveles en *Escherichia coli*, las cuales pueden ser comprobadas experimentalmente.

Capítulo 5

Bibliografía

Aarons, S., Abbas, A., Adams, C., Fenton, A., y O'Gara, F. (2000). A regulatory rna (prrb rna) modulates expression of secondary metabolite genes in pseudomonas fluorescens f113. *J. Bacteriol.*

Ahmed, A. M. y Shimamoto, T. (2003). msdna-st85, a multicopy single-stranded dna isolated from salmonella enterica serovar typhimurium lt2 with the genomic analysis of its retron. *FEMS Microbiology Letters.*

Altuvia, S. (2007). Identification of bacterial small non-coding RNAs: experimental approaches. *Current opinion in microbiology*, 10(3):257–61.

Altuvia, S., Zhang, A., Argaman, L., Tiwari, A., y Storz, G. (1998). The *Escherichia coli* OxyS regulatory RNA represses fhla translation by blocking ribosome binding. *EMBO J*, (20).

Antal, M., Bordeau, V., Douchin, V., y Felden, B. (2005). A small bacterial rna regulates a putative abc transporter. *Journal of Biological Chemistry.*

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E., Margalit, H., y Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Current Biology*, 11(12):941–950.

Axmann, I., Kensche, P., Vogel, J., Kohl, S., Herzel, H., y Hess, W. (2005). Identification of cyanobacterial non-coding rnas by comparative genome analysis. *Genome Biology*, 6(9):R73.

Backofen, R., Bernhart, S. H., Flamm, C., Fried, C., Fritzsch, G., Hackermüller, J., Hertel, J., Hofacker, I. L., Missal, K., Mosig, A., Prohaska, S. J., Rose, D., Stadler, P. F., Tanzer, A., Washietl, S., y Will, S. (2007). RNAs everywhere: genome-wide annotation of structured RNAs. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 308(1):1–25.

Bae, T., Kozlowicz, B. K., y Dunny, G. M. (2004). Characterization of cis-acting prgq mutants: evidence for two distinct repression mechanisms by qa rna and prgx protein in pheromone-inducible enterococcal plasmid pcf10. *Molecular Microbiology.*

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L., y Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA.*

Bejerano-Sagie, M. y Xavier, K. (2007). The role of small RNAs in quorum sensing. *Current opinion in microbiology*, pages 2–11.

Benito, Y., Kolb, F. A., Romby, P., Lina, G., Etienne, J., y Vandenesch, F. (2000). Probing the structure of rnaiii, the staphylococcus aureus agr regulatory rna, and identification of the rna domain involved in repression of protein a expression. *RNA.*

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., y Sayers, E. W. (2011). Genbank. *Nucleic Acids Research.*

Blose, J. M., Manni, M. L., Klapec, K. A., Stranger-Jones, Y., Zyra, A. C., Sim, V., Griffith, C. A., Long, J. D., y Serra, M. J. (2007). Non-nearest-neighbor dependence of the stability for rna bulge loops based on the complete set of group i single-nucleotide bulge loops. *Biochemistry*, 46(51):15123–15135.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V., y van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes.

Brown, J. W. (1999). The ribonuclease p database. *Nucleic acids research*.

Chan, C. Y. y Ding, Y. (2008). Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. *Journal of mathematical biology*, 56(1-2):93–105.

Chang, C. C. y Lin, C. J. (2001). *LIBSVM: a library for support vector machines*.

Chen, S., Lesnik, E. A., Hall, A. T., Sampath, R., Griffey, R. H., Ecker, D. J., y Blyn, L. B. (2002). A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *BioSystems*, 65:157–177.

Chen, S., Zhang, A., Blyn, L. B., y Storz, G. (2004). Micc, a second small-rna regulator of omp protein expression in escherichia coli. *J. Bacteriol.*

Chen, X., Quinn, A. M., y Wolin, S. L. (2000). Ro ribonucleoproteins contribute to the resistance of deinococcus radiodurans to ultraviolet irradiation. *Genes & Development*.

Christiansen, J. K., Nielsen, J. S., Ebersbach, T., Valentin-Hansen, P., Sø gaard Andersen, L., y Kallipolitis, B. H. (2006). Identification of small Hfq-binding RNAs in Listeria monocytogenes. *RNA (New York, N.Y.)*, 12(7):1383–96.

Clote, P., Ferré, F., Kranakis, E., y Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA (New York, N.Y.)*, 11(5):578–91.

Delihas, N. y Forst, S. (2001). Micf: an antisense rna gene involved in response of escherichia coli to global stress factors. *Journal of Molecular Biology*.

Dinger, M. E., Pang, K. C., Mercer, T. R., y Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS computational biology*, 4(11):e1000176.

Duan, K., Liu, C.-Q., Supple, S., y Dunn, N. W. (1998). Involvement of antisense rna in replication control of the lactococcal plasmid pnd324. *FEMS Microbiology Letters*.

Dwyer, R. (2003). *Genomic perl: From bioinformatics basics to working code*, volume 1. Cambridge Univ Pr.

Eddy, S. (2002). Computational genomics of noncoding RNA genes. *Cell*, 109(2):137–140.

Eddy, S. R. (1999). Noncoding RNA genes. *Current Opinion in Genetics and Development*.

Faubladier, M. y Bouche, J. P. (1994). Division inhibition gene dicf of escherichia coli reveals a widespread group of prophage sequences in bacterial genomes. *J. Bacteriol.*

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*.

Flamm, C., Bompfunewerer, A., Fried, C., Fritzsch, G., Hofacker, I. L., Missal, K., Mosig, A., Prohaska, S. J., Stadler, P. F., Tanzer, A., Washietl, S., y Witwer, C. (2005). Evolutionary Patterns of Non-Coding RNAs. *Applied Sciences*, (January).

Freyhult, E., Gardner, P. P., y Moulton, V. (2005). A comparison of RNA folding measures. *BMC bioinformatics*, 6:241.

Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz Rascado, L., Martínez-Flores, I., Salgado, H., Bonavides-Martínez, C., Abreu-Goodger, C., Rodríguez-Penagos, C., Miranda-Ríos, J., Morett, E., Merino, E., Huerta, A. M., Treviño Quintanilla, L., y Collado-Vides, J. (2008). RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic acids research*, 36(Database issue):D120–4.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., y Bateman, A. (2010). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Research*.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., y Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic acids research*, 37(Database issue):D136–40.

Gerhart, E., Wagner, H., y Nordström, K. (1986). Structural analysis of an rna molecule involved in replication control of plasmid rl. *Nucleic Acids Research*.

Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet.*, 21:399–404.

Griffiths-Jones, S. (2003). Rfam: an RNA family database. *Nucleic Acids Research*, 31(1):439–441.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., y Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research*, 33(Database issue):D121–4.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., y Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18.

Hall, M. A. (1999). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Han, J. y Kamber, M. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

He, L., Soderbom, F., Wagner, E. G., Binnie, U., Binns, N., y Masters, M. (1993). Pcnb is required for the rapid degradation of rnai, the antisense rna that controls the copy number of cole1-related plasmids. *Molecular microbiology*, (6). Research Support, Non-U.S. Gov't,.

Heidrich, N. y Brantl, S. (2003). Antisense-rna mediated transcriptional attenuation: Importance of a u-turn loop structure in the target rna of plasmid pip501 for efficient inhibition by the antisense rna. *Journal of Molecular Biology*.

Hermann, T. y Patel, D. J. (2000). RNA bulges as architectural and recognition motifs. *Structure*, 8(3).

Hershberg, R., Altuvia, S., y Margalit, H. (2003). A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res.*, 31:1813–1820.

Hofacker, I. (2004). RNA secondary structure analysis using the Vienna RNA package. *Structure*, Vi(1994):1–12.

Hollander, M. y Wolfe, D. A. (1999). *Nonparametric Statistical Methods, 2nd Edition*. Wiley-Interscience, 2 edition.

Jerome, L. J., van Biesen, T., y Frost, L. S. (1999). Degradation of finp antisense rna from f-like plasmids: the rna-binding protein, fino, protects finp from ribonuclease e. *Journal of Molecular Biology*.

Jiang, M., Anderson, J., Gillespie, J., y Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, 9(i):192.

Johansson, J. (2003). RNA-mediated control of virulence gene expression in bacterial pathogens. *Trends in Microbiology*, 11(6):280–285.

Kaikkonen, M. U., Lam, M. T. Y., y Glass, C. K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular research*, pages 1–41.

Kavanaugh, L. a. y Dietrich, F. S. (2009). Non-coding RNA prediction and verification in Saccharomyces cerevisiae. *PLoS genetics*, 5(1):e1000321.

Keseler, I. M. y Collado-Vides (2011). Ecocyc: a comprehensive database of escherichia coli biology. *Nucleic Acids Research*.

Kim, K. y Meyer, R. J. (1986). Copy-number of broad host-range plasmid r1162 is regulated by a small rna. *Nucleic Acids Research*.

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T., y Asai, K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic acids research*, 35(Database issue):D145–8.

Kittle, J. D., Simons, R. W., Lee, J., y Kleckner, N. (1989). Insertion sequence is10 anti-sense pairing initiates by an interaction between the 5' end of the target rna and a loop in the anti-sense rna. *Journal of Molecular Biology*.

Klein, R. J., Misulovin, Z., y Eddy, S. R. (2002). Noncoding rna genes identified in at-rich hyperthermophiles. *Proceedings of the National Academy of Sciences*.

Kohavi, R. (1995). Wrappers for performance enhancements and oblivious decision graphs. Technical report, Stanford, CA, USA.

Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*.

Kulkarni, R. V. y Kulkarni, P. R. (2007). Computational approaches for the discovery of bacterial small RNAs. *Methods (San Diego, Calif.)*, 43(2):131–9.

Lee, M. T. y Kim, J. (2008). Self containment, a property of modular RNA structures, distinguishes microRNAs. *PLoS computational biology*, 4(8):e1000150.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S., y Bassler, B. L. (2004). The Small RNA Chaperone Hfq and Multiple Small RNAs Control QuorumSensing in Vibrio harveyi and Vibrio cholerae. *Cell*, 118:69–82.

Ling, C., Huang, J., y Zhang, H. (2003). AUC: a statistically consistent and more discriminating measure than accuracy. In *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 18, pages 519–526. Citeseer.

Liu, J., Gough, J., y Rost, B. (2006). Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS genetics*, 2(4):e29.

Liu, Q., Olman, V., Liu, H., Ye, X., Qiu, S., y Xu, Y. (2008). Rnacluster: An integrated tool for rna secondary structure comparison and clustering. *Journal of Computational Chemistry*.

Livny, J., Brencic, A., Lory, S., y Waldor, M. K. (2006). Identification of 17 pseudomonas aeruginosa srnas and prediction of srna-encoding genes in 10 diverse pathogens using the bioinformatic tool srnapredict2. *Nucleic Acids Research*.

Livny, J. y Waldor, M. K. (2007). Identification of small RNAs in diverse bacterial species. *Current Opinion in Microbiology*, 10:96–101.

Machado-Lima, A., del Portillo, H. a., y Durham, A. M. (2008). Computational methods in noncoding RNA research. *Journal of mathematical biology*, 56(1-2):15–49.

Majdalani, N., Chen, S., Murrow, J., St John, K., y Gottesman, S. (2001). Regulation of rpos by a novel small rna: the characterization of rpra. *Molecular Microbiology*.

Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T., y Gottesman, S. (1998). DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proceedings of the National Academy of Sciences of the United States of America*.

Masse, E. (2003). Regulatory roles for small RNAs in bacteria. *Current Opinion in Microbiology*, 6(2):120–124.

Massé, E. y Gottesman, S. (2002). A small rna regulates the expression of genes involved in iron metabolism in escherichia coli. *Proceedings of the National Academy of Sciences*.

Matera, a. G., Terns, R. M., y Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature reviews. Molecular cell biology*, 8(3):209–20.

Mellin, J. R., Goswami, S., Grogan, S., Tjaden, B., y Genco, C. A. (2007). A novel fur- and iron-regulated small rna, nrrf, is required for indirect fur-mediated regulation of the sdha and sdhc genes in neisseria meningitidis. *J. Bacteriol.*

Mimouni, N. K., Lyngso, R. B., Griffiths-Jones, S., y Hein, J. (2009). An analysis of structural influences on selection in RNA genes. *Molecular biology and evolution*, 26(1):209–16.

Moller, T., Franch, T., Udesen, C., Gerdes, K., y Valentin-Hansen, P. (2002). Spot 42 rna mediates discoordinate expression of the e. coli galactose operon. *Genes & Development.*

Nawrocki, E. y Eddy, S. (2009). Computational identification of functional RNA homologs in metagenomic data. *selab.janelia.org.*

Ng, K. L. S. y Mishra, S. K. (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics (Oxford, England)*, 23(11):1321–30.

Novick, R. P., Iordanescu, S., Projan, S. J., Kornblum, J., y Edelman, I. (1989). *Cell*, chapter pT181 plasmid replication is regulated by a countertranscript-driven transcriptional attenuator. Number 2. Cell Press.

Opdyke, J. A., Kang, J.-G., y Storz, G. (2004). Gady, a small-rna regulator of acid response genes in escherichia coli. *J. Bacteriol.*

Padalon-Brauch, G., Hershberg, R., Elgrably-Weiss, M., Baruch, K., Rosenshine, I., Margalit, H., y Altuvia, S. (2008). Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence. *Nucleic acids research*, 36(6):1913–27.

Perez, N., Treviño, J., Liu, Z., Ho, S. C. M., Babitzke, P., y Sumby, P. (2009). A genome-wide analysis of small regulatory RNAs in the human pathogen group A Streptococcus. *PloS one*, 4(11):e7668.

Perkins, D. O., Jeffries, C., y Sullivan, P. (2005). Expanding the 'central dogma': the regulatory role of nonprotein coding genes and implications for the genetic liability to schizophrenia. *Molecular psychiatry*, 10(1):69–78.

Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V., y Vogel, J. (2007). A small non-coding rna of the invasion gene island (spi-1) represses outer membrane protein synthesis from the salmonella core genome. *Molecular Microbiology.*

Pichon, C. y Felden, B. (2008). Small RNA gene identification and mRNA target predictions in bacteria. *Bioinformatics (Oxford, England)*, 24(24):2807–13.

Regalia, M., Rosenblad, M. A., y Samuelsson, T. (2002). Prediction of signal recognition particle rna genes. *Nucleic Acids Research.*

Reichenbach, B., Maes, A., Kalamorz, F., Hajnsdorf, E., y Görke, B. (2008). The small rna glmy acts upstream of the srna glmz in the activation of glms expression and is subject to regulation by polyadenylation in escherichia coli. *Nucleic Acids Research.*

Repoila, F. y Darfeuille, F. (2009). Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biology of the cell / under the auspices of the European Cell Biology Organization*, 101(2):117–31.

Rivas, E. y Eddy, S. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC bioinformatics*, 19:1–19.

Rivas, E., Klein, R., Jones, T., y Eddy, S. (2001). Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology*, 11(17):1369–1373.

Romeo, T. (1998). Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Molecular microbiology.*

Saito, S., Kakeshita, H., y Nakamura, K. (2009). Novel small RNA-encoding genes in the intergenic regions of Bacillus subtilis. *Gene*, 428(1-2):2–8.

Sayers, E. W. y Barrett (2010). Database resources of the National Center for Biotechnology Information. *Nucleic acids research.*

Schaefer, K. L. y McClure, W. R. (1997). Antisense rna control of gene expression in bacteriophage p22. i. structures of sar rna and its target, ant mrna. *RNA*.

Sharma, C. M. y Vogel, J. (2009). Experimental approaches for the discovery and characterization of regulatory small RNA. *Current opinion in microbiology*, pages 1–11.

Sittka, A., Lucchini, S., Papenfort, K., Sharma, C. M., Rolle, K., Binnewies, T. T., Hinton, J. C. D., y Vogel, J. (2008). Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq.

Song, T. y Wai, S. N. (2009). A novel srna that modulates virulence and environmental fitness of vibrio cholerae. *RNA Biology*, (3).

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehva, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., y Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, pages 1611–1618.

Storz, G., Opdyke, J. a., y Zhang, A. (2004). Controlling mRNA stability and translation with small, noncoding RNAs. *Current opinion in microbiology*, 7(2):140–4.

Szymanski, M., Erdmann, V. a., y Barciszewski, J. (2007). Noncoding RNAs database (ncRNAdb). *Nucleic acids research*, 35(Database issue):D162–4.

Tabei, Y. y Asai, K. (2009). A local multiple alignment method for detection of non-coding RNA sequences. *Bioinformatics (Oxford, England)*, 25(12):1498–505.

Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E., y Rosenow, C. (2002). Transcriptome analysis of escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Research*.

Toledo-Arana, A., Repoila, F., y Cossart, P. (2007). Small noncoding RNAs controlling pathogenesis. *Current opinion in microbiology*, 10(2):182–8.

Tran, T. (2009). *Genomic data mining for the computational prediction of small non-coding rna genes*. PhD thesis, Georgia Institute of Technology.

Tran, T. T., Zhou, F., Marshburn, S., Stead, M., Kushner, S. R., y Xu, Y. (2009). De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics (Oxford, England)*, 25(22):2897–905.

Urbanowski, M. L., Stauffer, L. T., y Stauffer, G. V. (2000). The gcvb gene encodes a small untranslated rna involved in expression of the dipeptide and oligopeptide transport systems in escherichia coli. *Molecular Microbiology*.

Uzilov, A. V., Keegan, J. M., y Mathews, D. H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:101186/1417–2105–7–173.

Valverde, C., Heeb, S., Keel, C., y Haas, D. (2003). Rsmy, a small regulatory rna, is required in concert with rsmz for gaca-dependent expression of biocontrol traits in pseudomonas fluorescens cha0. *Molecular Microbiology*.

Vanderpool, C. K. (2007). Physiological consequences of small rna-mediated regulation of glucose-phosphate stress. *Current Opinion in Microbiology*. Cell regulation (RNA special issue).

Varani, G. y McClain, W. H. (2000). The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1(1):18–23.

Venkova-Canova, T., Patek, M., y Nesvera, J. (2003). Control of rep gene expression in plasmid pga1 from corynebacteriumglutamicum. *J. Bacteriol.*

Venkova-Canova, T., Soberón, N. E., Ramírez-Romero, M. A., y Cevallos, M. A. (2004). Two discrete elements are required for the replication of a repabc plasmid: an antisense rna and a stem–loop structure. *Molecular Microbiology*.

Vogel, J., Argaman, L., Wagner, E. G. H., y Altuvia, S. (2004). *The Small RNA IstR Inhibits Synthesis of an SOS-Induced Toxic Peptide*. Number 24. Cell Press.

Vogel, J. y Bartels, V. (2003). RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Research*, 31(22):6435–6443.

Vogel, J. y Papenfort, K. (2006). Small non-coding RNAs and the bacterial outer membrane. *Current opinion in microbiology*, 9(6):605–11.

Vogel, J. y Sharma, C. M. (2005). How to find small non-coding RNAs in bacteria. *Biological chemistry*, 386(12):1219–38.

Vogel, J. y Wagner, E. G. H. (2007). Target identification of small noncoding RNAs in bacteria.

Voss, B., Georg, J., Schön, V., Ude, S., y Hess, W. R. (2009). Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC genomics*, 10:123.

Washietl, S. (2005). Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, 102:2454–2459.

Wassarman, K. M. (2007). 6S RNA: a small RNA regulator of transcription. *Current opinion in microbiology*, 10(2):164–8.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., y Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes & Development*, pages 1637–1651.

Waters, L. S. y Storz, G. (2009). Regulatory RNAs in Bacteria. *Cell*, 136(4):615–628.

Wilderman, P. J., Sowa, N. a., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. a., y Vasil, M. L. (2004). Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9792–7.

Williams, K. P. y Bartel, D. P. (1996). Phylogenetic analysis of tmrna secondary structure. *RNA*.

Xu, X., Ji, Y., y Stormo, G. D. (2009). Discovering cis-regulatory RNAs in Shewanella genomes by Support Vector Machines. *PLoS computational biology*, 5(4):e1000338.

Zhang, H. (2004). The Optimality of Naive Bayes. *Machine Learning*.

Capítulo 6

Anexo

Anexo A. Tablas

Tabla 6.1: Listado de genes sARN en base de datos *Rfam*

Acceso de *Rfam*	**Nombre del gen**	**Largo Promedio**	**Referencia**
RF00010	RnaseP bact a	385	(Brown, 1999)
RF00011	RnaseP bact b	360	(Brown, 1999)
RF00013	6S	178	(Barrick et al., 2005)
RF00014	DsrA	84	(Majdalani et al., 1998)
RF00018	CsrB	282	(Romeo, 1998)
RF00021	Spot 42	120	(Moller et al., 2002)
RF00022	GcvB	201	(Urbanowski et al., 2000)
RF00023	tmRNA	369	(Williams y Bartel, 1996)
RF00033	MicF	91	(Delihas y Forst, 2001)
RF00034	RprA	107	(Majdalani et al., 2001)
RF00035	OxyS	111	(Altuvia et al., 1998)
RF00039	DicF	52	(Faubladier y Bouche, 1994)
RF00042	CopA	90	(Gerhart et al., 1986)
RF00043	Plasmid R1162	73	(Kim y Meyer, 1986)
RF00057	RyhB	67	(Massé y Gottesman, 2002)
RF00062	HgcC	126	(Klein et al., 2002)
RF00077	SraB	134	(Argaman et al., 2001)
RF00078	SraD	76	(Argaman et al., 2001)
RF00079	SraE / OmrA / OmrB	86	(Argaman et al., 2001)
RF00081	SraH	107	(Argaman et al., 2001)
RF00082	SraG	170	(Argaman et al., 2001)
RF00083	SraJ / GlmZ	199	(Argaman et al., 2001)
RF00084	CsrC	256	(Argaman et al., 2001)
RF00101	SraC / RyeA	144	(Argaman et al., 2001)
RF00106	RNAI	105	(He et al., 1993)
RF00107	FinP	77	(Jerome et al., 1999)
RF00110	RybB	76	(Wassarman et al., 2001)
RF00111	RyeB	99	(Wassarman et al., 2001)

Acceso de *Rfam*	**Nombre del gen**	**Largo Promedio**	**Referencia**
RF00112	RyeE / CyaR	88	(Wassarman et al., 2001)
RF00113	QUAD	148	(Wassarman et al., 2001)
RF00115	IS061	136	(Chen et al., 2002)
RF00116	C0465	78	(Tjaden et al., 2002)
RF00117	C0719	176	(Tjaden et al., 2002)
RF00118	rydB	68	(Wassarman et al., 2001)
RF00119	C0299	70	(Tjaden et al., 2002)
RF00120	C0343	75	(Tjaden et al., 2002)
RF00121	MicC	121	(Chen et al., 2004)
RF00122	GadY	107	(Opdyke et al., 2004)
RF00124	IS102	158	(Tjaden et al., 2002)
RF00125	IS128	200	(Tjaden et al., 2002)
RF00126	ryfA	298	(Wassarman et al., 2001)
RF00127	t44	110	(Tjaden et al., 2002)
RF00128	GlmY / tke1	155	(Reichenbach et al., 2008)
RF00166	PrrB / RsmZ	163	(Aarons et al., 2000)
RF00169	SRP bact	100	(Regalia et al., 2002)
RF00170	msr	72	(Ahmed y Shimamoto, 2003)
RF00195	RsmY	121	(Valverde et al., 2003)
RF00235	Plasmid RNAIII	130	(Heidrich y Brantl, 2003)
RF00236	ctRNA pGA1	78	(Venkova-Canova et al., 2003)
RF00238	ctRNA pND324	84	(Duan et al., 1998)
RF00240	RNA-OUT	70	(Kittle et al., 1989)
RF00242	ctRNA pT181	96	(Novick et al., 1989)
RF00262	sar	68	(Schaefer y McClure, 1997)
RF00368	sroB	78	(Vogel y Bartels, 2003)
RF00369	sroC	160	(Vogel y Bartels, 2003)
RF00370	sroD	87	(Vogel y Bartels, 2003)
RF00371	sroE	83	(Vogel y Bartels, 2003)
RF00372	sroH	152	(Vogel y Bartels, 2003)
RF00378	Qrr	107	(Lenz et al., 2004)
RF00388	QaRNA	92	(Bae et al., 2004)
RF00444	PrrF	147	(Wilderman et al., 2004)
RF00489	ctRNA p42d	45	(Venkova-Canova et al., 2004)
RF00503	RNAIII	426	(Benito et al., 2000)
RF00505	RydC	64	(Antal et al., 2005)
RF00534	SgrS	221	(Vanderpool, 2007)
RF00615	LhrA	264	(Christiansen et al., 2006)
RF00616	LhrC	111	(Christiansen et al., 2006)
RF00623	P1	173	(Livny et al., 2006)
RF00624	P9	77	(Livny et al., 2006)
RF00625	P11	138	(Livny et al., 2006)
RF00627	P15	121	(Livny et al., 2006)
RF00628	P16	195	(Livny et al., 2006)
RF00629	P24	252	(Livny et al., 2006)
RF00630	P26	64	(Livny et al., 2006)

Acceso de *Rfam*	**Nombre del gen**	**Largo Promedio**	**Referencia**
RF01053	Deinococcus Y RNA	126	(Chen et al., 2000)
RF01116	Yfr1	62	(Axmann et al., 2005)
RF01384	InvR	91	(Pfeiffer et al., 2007)
RF01385	isrA	121	(Padalon-Brauch et al., 2008)
RF01386	isrB	89	(Padalon-Brauch et al., 2008)
RF01387	isrC	104	(Padalon-Brauch et al., 2008)
RF01388	isrD	52	(Padalon-Brauch et al., 2008)
RF01389	isrF	269	(Padalon-Brauch et al., 2008)
RF01390	isrG	272	(Padalon-Brauch et al., 2008)
RF01391	isrH	223	(Padalon-Brauch et al., 2008)
RF01392	isrI	98	(Padalon-Brauch et al., 2008)
RF01393	isrJ	71	(Padalon-Brauch et al., 2008)
RF01394	isrK	78	(Padalon-Brauch et al., 2008)
RF01395	isrL	344	(Padalon-Brauch et al., 2008)
RF01396	isrN	142	(Padalon-Brauch et al., 2008)
RF01397	isrO	192	(Padalon-Brauch et al., 2008)
RF01398	isrP	111	(Padalon-Brauch et al., 2008)
RF01399	isrQ	161	(Padalon-Brauch et al., 2008)
RF01400	istR	134	(Vogel et al., 2004)
RF01401	rseX	98	(Sittka et al., 2008)
RF01402	STnc150	189	(Sittka et al., 2008)
RF01403	STnc290	65	(Sittka et al., 2008)
RF01404	STnc440	76	(Sittka et al., 2008)
RF01405	STnc490k	107	(Sittka et al., 2008)
RF01406	STnc500	285	(Sittka et al., 2008)
RF01407	STnc560	224	(Sittka et al., 2008)
RF01408	sraL	142	(Sittka et al., 2008)
RF01409	STnc250	95	(Sittka et al., 2008)
RF01410	BsrC	86	(Saito et al., 2009)
RF01411	BsrF	111	(Saito et al., 2009)
RF01412	BsrG	243	(Saito et al., 2009)
RF01416	NrrF	159	(Mellin et al., 2007)
RF01456	VrrA	132	(Song y Wai, 2009)

Tabla 6.2: Genes sARN en genoma de *Escherichia coli* según base de datos *Rfam*

Acceso de *Rfam*	**Nombre del gen**	**Inicio**	**Fin**	**Largo**	**Hebra**
RF00010	RNaseP_bact_a	1371437	1371061	376	−1
RF00013	6S	3054005	3054188	183	+1
RF00014	DsrA	2616424	2616338	86	−1
RF00018	CsrB	1717497	1717138	359	−1
RF00021	Spot_42	4047912	4048030	118	+1
RF00022	GcvB	2940718	2940923	205	+1
RF00023	tmRNA	2753615	2753976	361	+1
RF00033	MicF	2311105	2311198	93	+1

Acceso de *Rfam*	**Nombre del gen**	**Inicio**	**Fin**	**Largo**	**Hebra**
RF00034	RprA	1768396	1768503	107	+1
RF00035	OxyS	483367	483258	109	−1
RF00039	DicF	3439252	3439199	53	−1
RF00039	DicF	3223241	3223186	55	−1
RF00039	DicF	1647407	1647458	51	+1
RF00057	RyhB	1060729	1060665	64	−1
RF00077	SraB	1145812	1145980	168	+1
RF00078	SraD	2812823	2812897	74	+1
RF00079	SraE / OmrA / OmrB	1665551	1665464	87	−1
RF00079	SraE / OmrA / OmrB	1665349	1665268	81	−1
RF00081	SraH	3348599	3348706	107	+1
RF00082	SraG	3309249	3309420	171	+1
RF00083	GlmZ / SraJ	1950498	1950314	184	−1
RF00083	GlmZ / SraJ	3984455	3984661	206	+1
RF00084	CsrC	4049059	4049312	253	+1
RF00101	SraC / RyeA	1921124	1921268	144	+1
RF00110	RybB	3752476	3752398	78	−1
RF00111	RyeB	2718547	2718448	99	−1
RF00112	CyaR / RyeE	2165136	2165221	85	+1
RF00113	QUAD	2151299	2151447	148	+1
RF00113	QUAD	2151634	2151776	142	+1
RF00113	QUAD	3054837	3054987	150	+1
RF00113	QUAD	1446902	1446753	149	−1
RF00113	QUAD	1446527	1446378	149	−1
RF00115	IS061	1403654	1403833	179	+1
RF00116	C0465	1970763	1970840	77	+1
RF00117	C0719	3119380	3119601	221	+1
RF00118	rydB	1762737	1762804	67	+1
RF00119	C0299	1229852	1229930	78	+1
RF00120	C0343	1407387	1407461	74	+1
RF00121	MicC	1435142	1435263	121	+1
RF00122	GadY	3662884	3662997	113	+1
RF00124	IS102	2069339	2069542	203	+1
RF00125	IS128	2651537	2651745	208	+1
RF00126	ryfA	2651877	2652180	303	+1
RF00127	t44	189754	189847	93	+1
RF00128	GlmY / tke1	1950460	1950313	147	−1
RF00128	GlmY / tke1	3984457	3984605	148	+1
RF00169	SRP / bact	475679	475778	99	+1
RF00368	sroB	506429	506511	82	+1
RF00369	sroC	3953771	3953609	162	−1
RF00370	sroD	2753634	2753549	85	−1
RF00371	sroE	2001058	2000967	91	−1
RF00372	sroH	451325	451165	160	−1
RF00505	RydC	3150208	3150145	63	−1
RF00534	SgrS	77367	77593	226	+1

Acceso de *Rfam*	**Nombre del gen**	**Inicio**	**Fin**	**Largo**	**Hebra**
RF00630	P26	4178953	4179014	61	+1
RF01400	istR	788534	788405	129	−1
RF01407	STnc560	3018916	3018703	213	−1
RF01408	sraL	363726	363586	140	−1