

**MODELO DE CLASIFICACIÓN DE PROMOTORES BASADO EN REDES
NEURONALES**

SANTIAGO J. BUSTAMANTE MUÑOZ

UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA SISTEMAS Y COMPUTACIÓN
PROGRAMA ACADÉMICO DE INGENIERÍA DE SISTEMAS
SANTIAGO DE CALI
2010

**MODELO DE CLASIFICACIÓN DE PROMOTORES BASADO EN REDES
NEURONALES**

SANTIAGO J. BUSTAMANTE MUÑOZ
Código 0630707

Trabajo de grado para optar al título de
Ingeniero de Sistemas

DIRECTOR DE TESIS
OSCAR BEDOYA LEYVA
Máster en Ingeniería de Sistemas

UNIVERSIDAD DEL VALLE
ESCUELA DE INGENIERÍA SISTEMAS Y COMPUTACIÓN
PROGRAMA ACADÉMICO DE INGENIERÍA DE SISTEMAS
SANTIAGO DE CALI
2010

AGRADECIMIENTOS

A mi familia por estar siempre a mi lado apoyándome incondicionalmente, motivándome a terminar con esta meta que me he trazado.

A mi director Msc. Oscar Bedoya, por su colaboración y guía en el desarrollo de este proyecto.

Y por supuesto, a Dios.

RESUMEN

Actualmente para la predicción de señales promotoras se usan diferentes métodos computacionales que utilizan como estrategia de predicción métodos intrínsecos probabilísticos tales como secuencias consenso, matrices de score o de puntuación y modelos ocultos de Markov; y métodos intrínsecos como redes neuronales artificiales (RNA), árboles de decisión y estrategias integradoras. No obstante, estos métodos computacionales tienen exactitudes que aun se pueden mejorar causando que no exista consenso entre ellas.

Este proyecto plantea el desarrollo de dos modelos de predicción de señales promotoras utilizando como estrategia el uso de redes neuronales artificiales con el propósito de evaluar y comparar dichos modelos. En el primer modelo se emplea la estrategia de utilizar como valores de entrada para la RNA resultados obtenidos por diferentes modelos ocultos de Markov. En el segundo modelo de RNA se emplea la estrategia que utiliza como valores de entrada los resultados obtenidos por otras RNAs de predicción de secuencias de caja consenso ya conocidas.

Para este proyecto los modelos propuestos están enfocados sólo a la predicción de promotores reconocidos por el ARN polimerasa tipo II o B de las especies *Rattus norvegicus* y *Mus musculus*. Los resultados de las pruebas realizadas indican que los modelos propuestos alcanzan una sensibilidad hasta del 74.54% y una especificidad del 82.72% superando a muchos de los predictores de promotores existentes.

TABLA DE CONTENIDO

CAPÍTULO 1. INTRODUCCIÓN	1
1.1 PLANTEAMIENTO DEL PROBLEMA	2
1.2 ALCANCE DEL PROYECTO	2
1.3 OBJETIVOS	3
1.3.1 OBJETIVOS GENERALES	3
1.3.2 OBJETIVOS ESPECÍFICOS	3
1.4 JUSTIFICACIÓN	4
CAPÍTULO 2. MARCO TEÓRICO	5
2.1 BIOLOGÍA MOLECULAR	5
2.1.1 ESTRUCTURA DEL PROMOTOR	8
2.2 REDES NEURONALES ARTIFICIALES	9
CAPÍTULO 3. ESTADO DEL ARTE	14
3.1 Red neuronal para la predicción de secuencias de promotores en microbacterias	14
3.2 Red neuronal para la predicción de secuencias de promotores en humanos	15
3.3 Comparación entre redes neuronales y árboles de decisión en el análisis de secuencias	17
3.4 Predictor de promotor de e. Coli usando redes neuronales feed-foward	18
3.5 Identificación de promotores en el genoma del hongo crinipellis usando redes neuronales feed-forward	20
CAPÍTULO 4. MEODELOS PROPUESTOS	23
4.1 MODELO 1	23
4.1.1 Especificaciones de RED1	23
4.2 MODELO 2	24

4.2.1 RNAs para el reconocimiento de la caja TATA, caja CAAT y caja GC.....	25
4.2.2 Especificaciones de RED2.....	26
CAPÍTULO 5. PRUEBAS Y RESULTADOS.....	29
5.1 Especificaciones de los conjuntos de entrenamiento y prueba.....	26
5.2 CRITERIOS DE COMPARACIÓN.....	32
5.3 PRUEBAS AL MODELO 1.....	33
5.3.1 PRUEBA 1.1.....	33
5.3.2 PRUEBA 1.2.....	33
5.3.3 PRUEBA 1.7.....	34
5.3.4 PRUEBA 1.8.....	34
5.4 PRUEBAS AL MODELO 2.....	35
5.4.1 PRUEBA 2.1.....	35
5.4.2 PRUEBA 2.3.....	35
5.4.3 PRUEBA 2.6.....	36
5.4.4 PRUEBA 2.8.....	36
5.4.5 PRUEBA 2.11.....	37
5.4.6 PRUEBA 2.13.....	37
5.5 DISCUSIÓN DE RESULTADOS.....	38
5.6 COMPARACIÓN CON ALGUNOS MODELOS EXISTENTES.....	42
CAPÍTULO 6. CONCLUSIONES.....	44
6.1 TRABAJO FUTURO.....	45
REFERENCIAS.....	47

LISTA DE IMÁGENES

Img 1. Anatomía de un gen de los organismos eucariotes.....	7
Img 2. Regiones del núcleo del promotor.....	8
Img 3. De la neurona biológica a la neurona artificial.....	9
Img 4. Red unidireccional (FeedFoward).....	10
Img 5. Red recurrente (FeedBack).....	10
Img 6. Funciones de activación más utilizadas.....	11
Img 7. Gráfica de la función XOR, discriminación mediante un elipsoide.....	12
Img 8. Gráfica de la función XOR, discriminación mediante dos rectas.....	12
Img 9. Gráfica de la red XOR, modelo perceptrón.....	13
Img 10. Gráfica de la red XOR, modelo perceptrón multicapa.....	13
Img 11. Esquema de la red neuronal artificial optimizada.....	15
Img 12. Red neuronal de retardo temporal (TDNN).....	16
Img 13. Esquema de la red neuronal artificial.....	18
Img 14. Topología de red para el esquema de 2 dimensiones.....	19
Img 15. Topología de red para el esquema de 4 dimensiones.....	20
Img 16. Topología de red que obtuvo el mejor desempeño.....	22
Img 17. Topología de la RED1.....	24
Img 18. Topología de la RNA que reconoce la caja TATA.....	25
Img 19. Topología de la RNA que reconoce la caja CAAT.....	26
Img 20. Topología de la RNA que reconoce la caja GC.....	26
Img 21. Esquema de la lectura y trazo arrogado que realiza las RNAs.....	27
Img 22. Topología de la RED2.....	28

LISTA DE TABLAS

Tabla 1. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata con 120 nucleótidos de tamaño.....	30
Tabla 2. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores del ratón con 120 nucleótidos de tamaño.....	30
Tabla 3. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata y ratón con 120 nucleótidos de tamaño.....	30
Tabla 4. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata con 250 nucleótidos de tamaño.....	30
Tabla.5. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores del ratón con 250 nucleótidos de tamaño.....	31
Tabla 6. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata y ratón con 250 nucleótidos de tamaño.....	31
Tabla 7. Relación entre el resultado de una prueba.....	31
Tabla 8. Resultados de la prueba 1.1.....	32
Tabla 9. Resultados de la prueba 1.2.....	32
Tabla 10. Resultados de la prueba 1.7.....	33
Tabla 11. Resultados de la prueba 1.8.....	33
Tabla 12. Resultados de la prueba 2.1.....	34
Tabla 13. Resultados de la prueba 2.3.....	34
Tabla 14. Resultados de la prueba 2.6.....	35
Tabla 15. Resultados de la prueba 2.8.....	35
Tabla 16. Resultados de la prueba 2.11.....	36
Tabla 17. Resultados de la prueba 2.13.....	36

Capítulo 1.

INTRODUCCIÓN

En el conjunto de material genético de un organismo, los genes tienen secuencias específicas identificables de señales promotoras, señal localizada justo donde se encuentra el punto de inicio de la transcripción del ARN (Ácido ribonucleico) y que contiene la información necesaria para regular la activación del gen que codifica. Estas señales son comúnmente determinadas mediante análisis experimental de forma sistemática y, por este motivo, este proceso es extremadamente laborioso y demorado. Por lo anterior, la predicción mediante métodos computacionales de promotores es especialmente interesante, pues permitiría reemplazar costosos y demorados procesos de análisis experimental. No obstante, entre los métodos computacionales que actualmente se utilizan para la predicción de señales promotoras no existe consenso debido a que no alcanzan a tener una buena tasa de acierto en los resultados de sus predicciones.

La intención de este proyecto es desarrollar un modelo de predicción de señales promotoras que sea más acertado que otros modelos ya propuestos.

1.1. PLANTEAMIENTO DEL PROBLEMA

Actualmente para la predicción de señales promotoras se usan diferentes métodos computacionales que utilizan como estrategia de predicción métodos intrínsecos probabilísticos tales como secuencias consenso, matrices de score o de puntuación y modelos ocultos de Markov; y métodos intrínsecos basados en simulación e inteligencia artificial tales como redes neuronales, árboles de decisión y estrategias integradoras.

Sin embargo, el problema es que hasta la fecha presente ninguna de las metodologías propuestas para la predicción de las señales promotoras tienen exactitudes que aun se pueden mejorar [1, 2, 6], causando que no exista consenso entre ellas y dificultando la elección de un software de análisis para la predicción de promotores a los biólogos, médicos e investigadores en general sobre la genética humana.

En este trabajo de grado se trata la siguiente pregunta de investigación: ¿Será posible conseguir un modelo más acertado para la predicción de promotores?

1.2. ALCANCE DEL PROYECTO

El proyecto tiene dos límites claramente identificables:

- La predicción de secuencias promotoras, aunque existen otros tipos, sólo se centrará en genes Eucariotas y se seleccionará un determinado organismo.
- Es importante tener claro que el proyecto es investigativo, por lo que no arrojará resultado del tipo de desarrollo de software, a cambio se obtendrán aportes y valiosas conclusiones acerca de la problemática abarcada.

1.3. OBJETIVOS

1.3.1. OBJETIVOS GENERALES

- Desarrollar un modelo clasificador de promotores usando predicción mediante redes neuronales.

1.3.2. OBJETIVOS ESPECÍFICOS

- Seleccionar de la base de datos E.P.D. (Eukaryotic Promoter Database) conjuntos de entrenamiento, validación y prueba.
- Seleccionar los modelos¹ de predicción de promotores que se van a integrar.
- Desarrollar un módulo de software para la clasificación de promotores usando la red neuronal propuesta.
- Validar el modelo obtenido para el conjunto de señales promotoras a partir del grupo de organismos eucarióticos seleccionado. Se debe comparar los niveles de sensibilidad y especificidad del modelo propuesto con alguno de los modelos existentes.

¹ Modelos integrados: WMM, WAM y WAMM, ver glosario.

1.4. JUSTIFICACIÓN

La aplicación está enfocada al estudio de las señales promotoras del genoma² de determinado tipo de organismos eucarióticos³ y, por esta misma razón, ayudaría a los biólogos, médicos e investigadores de la genética a ganar precisión y tiempo en el análisis de la predicción de las variadas y cuantiosas señales promotoras porque se dispone de una enorme cantidad de datos de estructuras primarias de multitud de genes de todo tipo de organismos.

Por otra parte, para minimizar el problema de la precisión de predicción, se propone emplear redes neuronales artificiales⁴ porque presentan ventajas a la hora de predecir señales promotoras; entre dichas ventajas se pueden enumerar el aprendizaje adaptativo, la auto-organización y la tolerancia a fallos [8].

Además, si el impacto de la aplicación resulta ser muy positivo, no sólo habrá un beneficio para el estudio de la estructura genética de organismos eucarióticos, sino que ayudaría al entendimiento del misterioso y complejo genoma humano; facilitando así el avance de la medicina en pro de la salud y el bienestar humano.

² Ver glosario.

³ Ibídem.

⁴ Ibídem.

Capítulo 2.

MARCO TEÓRICO

En este capítulo se presenta un marco teórico básico que abarca los temas relacionados con Bioinformática, la predicción de promotores y las redes neuronales artificiales.

2.1. BIOLOGÍA MOLECULAR

ÁCIDO DESOXIRRIBONUCLEICO (ADN). [15]

Es el material genético de todos los organismos celulares y casi todos los virus. Cada molécula de ADN está constituida por dos cadenas o bandas formadas por un elevado número de compuestos químicos llamados nucleótidos. Estas cadenas forman una especie de escalera retorcida que se llama doble hélice. Cada nucleótido está formado por tres unidades: una molécula de azúcar llamada desoxirribosa, un grupo fosfato y uno de cuatro posibles compuestos nitrogenados llamados bases: adenina (abreviada como A), guanina (G), timina (T) y citosina (C). La molécula de desoxirribosa ocupa el centro del nucleótido y está flanqueada por un grupo fosfato a un lado y una base al otro. El grupo fosfato está a su vez unido a la desoxirribosa del nucleótido adyacente de la cadena. Estas subunidades enlazadas desoxirribosa-fosfato forman los lados de la escalera; las bases están enfrentadas por parejas, mirando hacia el interior, y forman los travesaños.

Los nucleótidos de cada una de las dos cadenas que forman el ADN establecen una asociación específica con los correspondientes de la otra cadena. Debido a la afinidad química entre las bases, los nucleótidos que contienen adenina se acoplan siempre con los que contienen timina, y los que contienen citosina con los que contienen guanina. Las bases complementarias se unen entre sí por enlaces químicos débiles llamados enlaces de hidrógeno.

ÁCIDO RIBONUCLEICO (ARN) [15, 17]

El ARN (ácido ribonucleico) está constituido por un azúcar ribosa, el ácido fosfórico y las bases nitrogenadas adenina, guanina, citosina y uracilo. En la célula aparecen cuatro tipos de ARN, con distintas funciones, que son el ARN mensajero, el ARN ribosómico y el ARN transferente. El ARN

mensajero (ARNm), contiene la información copiada del ADN para sintetizar una proteína. Se forma en el núcleo celular a partir de una secuencia de ADN, y cuando sale del núcleo se asocia a ribosomas, donde se construye la proteína. El ARN ribosómico (ARNr), unido a proteínas de carácter básico, forma los ribosomas. Finalmente, el ARN transferente (ARNt), tiene a función de llevar un aminoácido específico al ribosoma. En él se une a la secuencia complementaria del ARNm para luego transferirle el aminoácido correspondiente a la secuencia de aminoácidos que está formándose en el ribosoma.

ARN POLIMERASA [19]

El ARN polimerasa es la enzima encargada de la síntesis de ARN a partir de templado de ADN. Todas las células contienen a la ARN polimerasa. En procariontes esta enzima sintetiza todo el ARN excepto el necesario para el cebador en la replicación del ADN. En los eucariontes, el núcleo contiene tres tipos de enzima ARN polimerasas:

- ARN polimerasa tipo I o A: se encuentra en nucleolo (cuerpo granular y denso que se localiza en el núcleo y contiene los genes ribosomales), sintetiza los precursores de la mayorías de los ARNr.
- ARN polimerasa tipo II o B: se encuentra en el nucleoplasma, sintetiza precursores de ARNm.
- ARN polimerasa tipo III o C: se encuentra en el nucleoplasma, sintetiza precursores de ARNr, de ARNt y una variedad de otros ARN pequeños nucleares y citoplásmicos.

DOGMA CENTRAL DE LA BIOLOGIA MOLECULAR. [16]

El Dogma Central es un concepto fundamental de la biología molecular, donde expone los tres procesos más importantes a nivel celular, que son, Replicación, Transcripción y Traducción.

- La replicación consiste en la copia del ADN de una célula, antes de la división celular, para que la célula hija tenga el mismo ADN que la madre.

- La transcripción consiste en convertir la información contenida en el ADN en un formato "legible" para la maquinaria celular de síntesis de proteínas, el ARN.
- La traducción es el mecanismo por el que el mensaje que lleva el ARN se utiliza para sintetizar proteínas.

Con los tres mecanismos se consigue extraer de la información genética (ADN) los materiales (proteínas) necesarios tanto funcional como estructuralmente para que una célula funcione.

GENES. [14, 15]

Un gen es una sección de la cadena de ADN que lleva las instrucciones de una función específica. En términos moleculares puede definirse como la secuencia lineal de nucleótidos considerada como unidad de almacenamiento de información. Los genes están localizados en los cromosomas en el núcleo celular y se disponen en línea a lo largo de cada uno de ellos.

La biología actualmente ha clasificado regiones o secuencias de los genes de organismos eucarióticos de la siguiente manera: El *promotor* es la parte de la molécula de ADN que tiene la información para hacer que el gen se active y se desactive. Los *exones* e *intrones* van de la mano en la genética de los organismos eucariotes u organismos superiores. Los exones representan la parte del gen que codifica la proteína durante la traducción. Y los intrones son básicamente espaciadores, pues son fragmentos de ADN que se encuentran entre dos exones adyacentes.

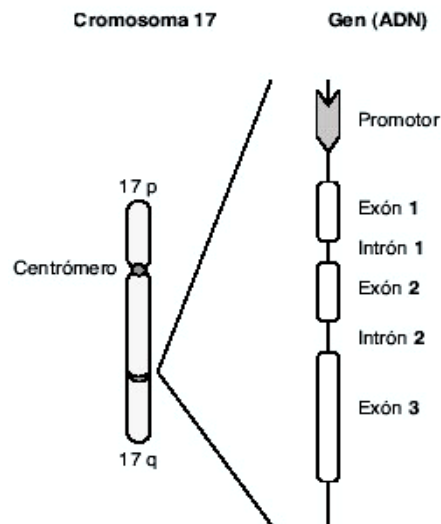


Imagen 1: Anatomía de un gen de los organismos eucariotes. Tomado de [9].

SEÑALES CARACTERÍSTICAS DE LOS GENES EUCARIOTAS

- Región reguladora: Cajas TATA, GC, CAAT, sitios de unión de proteínas reguladoras, enhancers y silencers
- Sitio de inicio de transcripción
- Zona 5'UTR
- Codón⁵ de inicio
- Intrones y exones alternados, con sitios de procesamiento dadores
- Codón de parada
- Zona 3'UTR

2.1.1. ESTRUCTURA DEL PROMOTOR [18]

El promotor de un gen es la región de ADN que controla la iniciación de la transcripción de dicho gen a ARN. El promotor está compuesto por una secuencia específica de ADN localizado justo donde se encuentra cerca al punto de inicio de la transcripción del ADN y contiene la información necesaria para activar o desactivar el gen que regula mediante la maquinaria asociada a la ARN polimerasa II.

El núcleo del promotor puede extenderse 35 pares de base (bp) hacia arriba y/o hacia abajo del sitio de iniciación de la transcripción. La mayoría de los elementos del núcleo del promotor parecen interactuar directamente con componentes de maquinaria básica de transcripción.

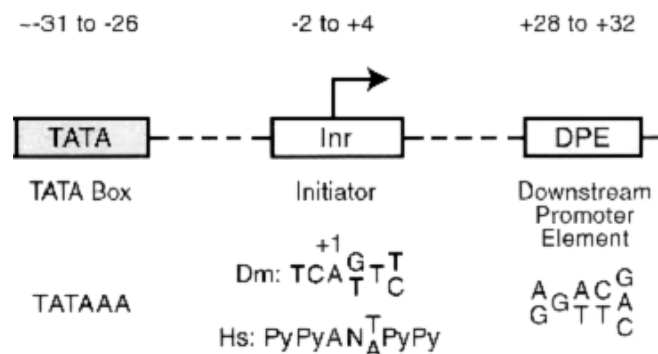


Imagen 2: Regiones del núcleo del promotor. Tomado de [20].

El promotor de un organismo eucariote contiene varias

⁵ Ver glosario.

regiones diferenciadas que son:

- TATA box (Caja TATA). Es una secuencia de 7 nucleótidos (TATAAAA) ubicada a -30 pares de bases (pb) antes del inicio de la transcripción.
- CAAT box, (Caja CAT). Secuencia CAAATC ubicada a -90 pares de bases del inicio de la transcripción a la cual se unen los factores de la transcripción.
- GC box. (Caja rica en GC). Secuencia de consenso GG CCGG ubicada a -110 pares de bases antes del inicio de la transcripción. Se asocia también a la unión de factores de la transcripción.

2.2. REDES NEURONALES ARTIFICIALES [11]

2.2.1. GENERALIDADES

En una red neuronal artificial el modelo se basa en la imitación del proceso de una neurona biológica (ver imagen 1) donde las entradas X_i representan las señales que provienen de otras neuronas y que son capturadas por las dendritas. Los pesos W_i son la intensidad de la sinapsis que conecta dos neuronas; tanto X_i como W_i son valores reales y θ es la función umbral que la neurona debe sobrepasar para activarse. Este proceso ocurre biológicamente en el cuerpo de la célula.

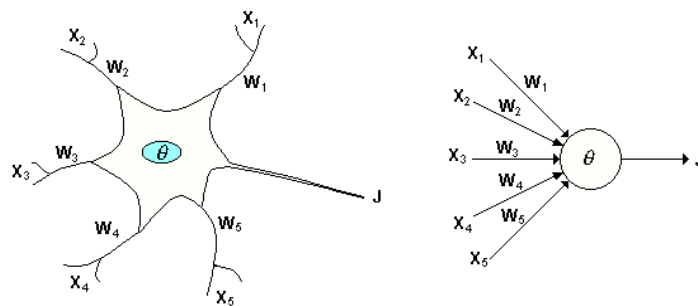


Imagen 3: De la neurona biológica a la neurona artificial. Tomado de [11a].

Las señales de entrada a una neurona artificial $X_1, X_2, \dots, X_n$ son variables continuas en lugar de pulsos discretos, como se presentan en una neurona biológica. Cada señal de entrada pasa a través de una ganancia o peso. Los pesos pueden ser

positivos (excitatorios), o negativos (inhibitorios), el nodo sumatorio acumula todas las señales de entradas multiplicadas por los pesos o ponderadas y las pasa a la salida a través de una función umbral o función de transferencia.

Una red neuronal puede verse como un grafo dirigido, donde los nodos son las neuronas y las conexiones son las sinapsis, en estos grafos se distinguen tres tipos de nodos los de entrada, los de salida y los intermedios. Estos nodos cumplen las siguientes características:

- A cada nodo i se asocia una variable de estado x_i .
- A cada conexión (i, j) de los nodos i y j se asocia un peso $W_{ij} \in \mathbf{R}$.
- A cada nodo i se asocia un umbral θ_i .
- Para cada nodo se define una función de activación f , que depende de los pesos sinápticos de sus conexiones, del umbral y de los estados de los nodos a él conectados. Esta función proporciona el nuevo estado del nodo.

Cada uno de estos nodos representa una neurona artificial, una conexión de estas neuronas, ordenada en capas y con un sistema definido para procesar la información es lo que se conoce como una red neuronal artificial. Una red neuronal se divide en tres capas: capa de entrada, a las neuronas ordenadas sin sinapsis entrantes; capa de salida, a las neuronas ordenadas sin sinapsis salientes; y capas ocultas, a las neuronas ordenadas que no son de entrada ni de salida.

Una red es unidireccional cuando no presenta bucles o ciclos en sus conexiones, es decir las conexiones entre las neuronas son en un solo sentido y se denomina recurrente cuando el flujo de información puede encontrar unos bucles, es decir, una retroalimentación (ver imágenes 2 y 3).

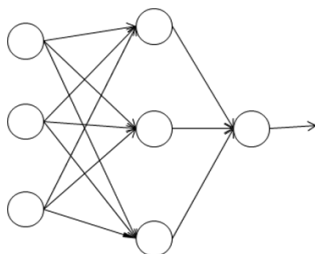


Imagen 4. Red unidireccional (FeedForward).

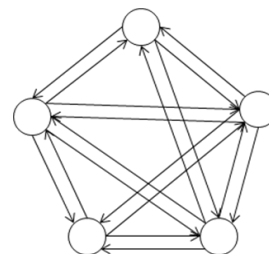


Imagen 5. Red recurrente (FeedBack).

Las funciones de activación las que más se utilizan son:

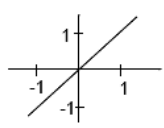
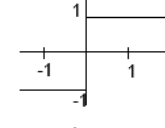
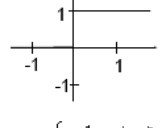
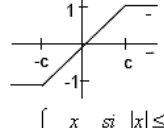
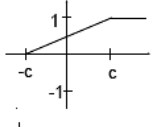
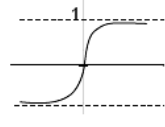
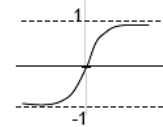
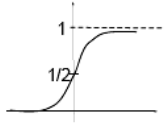
<i>Identidad</i>  $f(x) = x$	<i>Escalón</i>  $f(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{si } x < 0 \end{cases}$	<i>Escalón</i>  $f(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$	<i>Lineal Mixta</i>  $f(x) = \begin{cases} x & \text{si } x \leq c \\ 1 & \text{si } x > c \\ -1 & \text{si } x < -c \end{cases}$
<i>Lineal Mixta</i>  $f(x) = \begin{cases} \frac{x}{2c} + \frac{1}{2} & \text{si } x \leq c \\ 1 & \text{si } x \geq c \\ 0 & \text{si } x \leq -c \end{cases}$	<i>Tan hiperbolica</i>  $f(x) = \tan(ax)$	<i>Sigmoidal</i>  $f(x) = \frac{2}{1 + \exp(-ax)} - 1$	<i>Sigmoidal</i>  $f(x) = \frac{1}{1 + \exp(-ax)}$

Imagen 6. Funciones de activación más utilizadas. Tomado de [11b].

2.2.2. APRENDIZAJE O ENTRENAMIENTO DE UNA RED NEURONAL

El proceso de aprendizaje o entrenamiento en una red neuronal puede ser visto como un problema de actualización de los pesos sinápticos de las neuronas de manera tal que la red neuronal mejore la forma en que realiza cierta tarea específica. Estas mejoras en el desempeño de una red neuronal son logradas a lo largo de este proceso de entrenamiento y mediante la actualización iterativa de los pesos sinápticos.

El aprendizaje de las redes neuronales artificiales es un proceso basado en ejemplos o patrones, estos ejemplos deben tener dos características fundamentales, primero, ser significativo, debe haber un número suficiente de ejemplos para que la neurona sea capaz de adaptar sus pesos en forma eficaz y ser representativo, es decir los ejemplos deberán ser diversos y reflejar todas los posibles estados de la red. El proceso de aprendizaje o actualización de los pesos se puede realizar de dos maneras: utilizando información previamente establecida o sin utilizar alguna información. A estos tipos de aprendizaje se les denomina supervisado y no supervisado o auto-organizado, respectivamente.

Redes con aprendizaje supervisado. Se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento

controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor comprueba la salida generada por el sistema y en el caso de que no coincida con la esperada, se procederá a modificar los pesos de las conexiones.

Redes con aprendizaje no supervisado. La red no recibe ninguna información por parte del entorno que le indique si la salida generada en respuesta a una determinada entrada es o no correcta; son capaces de auto-organizarse. Estas redes deben encontrar las características, regularidades, correlaciones o categorías que se pueden establecer entre los datos de la entrada.

Uno de los ejemplos más estudiados consiste en crear una red que sea capaz de codificar la función lógica "O exclusivo" o XOR, cómo sabemos, la tabla de verdad para esta función lógica es:

A	B	XOR
0	0	1
0	1	0
1	0	0
1	1	1

Tabla 1. Tabla de verdad para la función lógica XOR.

Y si llevamos a la gráfica esta función, podemos apreciar que es imposible obtener una recta que sea capaz de discriminar entre estas dos categorías. Una solución a este problema podría ser un elipsoide o dos rectas. (Ver imágenes 5 y 6).

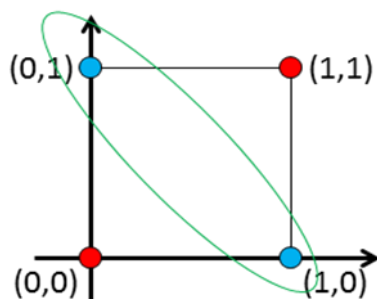


Imagen 7. Gráfica de la función XOR, discriminación mediante un elipsoide.

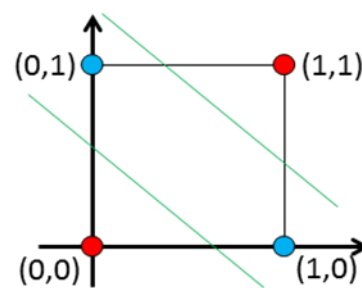


Imagen 8. Gráfica de la función XOR, discriminación mediante dos rectas.

Sin embargo, estos tipos de solución son imposibles de generar con una **perceptrón**, red neuronal que solo tiene capa de entrada y capa de salida (Ver imagen 7). Para el caso del XOR este problema puede ser resuelto con la introducción de la capa oculta, es decir con una red **perceptrón multicapa** (Ver imagen 8).

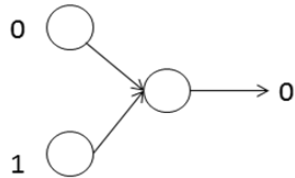


Imagen 9. Gráfica de la red XOR, modelo perceptrón.

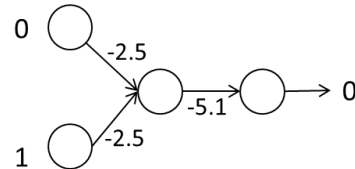


Imagen 10. Gráfica de la red XOR, modelo perceptrón multicapa.

El perceptrón multicapa es una red neuronal artificial formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón. El perceptrón multicapa puede ser totalmente o localmente conectado.

Los elementos de procesamiento del perceptrón multicapa se encuentran agrupados por las siguientes capas:

CAPA DE ENTRADA: Recibe las señales de la entrada de la red, algunos autores no consideran el vector de entrada como una capa pues allí no se lleva a cabo ningún proceso.

CAPAS OCULTAS: Estas capas son aquellas que no tienen contacto con el medio exterior, sus elementos pueden tener diferentes conexiones y son estas las que determinan las diferentes topologías de la red.

CAPA DE SALIDA: Recibe la información de la capa oculta y transmite la respuesta al medio externo.

Capítulo 3.

ESTADO DEL ARTE

En este capítulo se presenta un estado del arte de algunas de las estrategias que han sido aplicadas para el problema de la predicción de promotores usando redes neuronales. Se hace énfasis en mostrar la topología de la red planteada y de los resultados obtenidos.

3.1. Red neuronal para la predicción de secuencias de promotores en microbacterias [3].

En el trabajo realizado en [3] se utilizó la base de datos que consiste en una compilación de 125 señales promotoras de microbacterias. Para la capa de entrada no se realiza un ajuste de posición para un alineamiento perfecto entre las secuencias de señales promotoras, en cambio, todas fueron alineadas a la izquierda dando como ventaja que permite el análisis de las secuencias en las que la adaptación es difícil o imposible de definir. Además se usó, como parámetros de entrada, la estrategia de codificación CODE-4 que consiste en representar cada nucleótido como 4 dígitos binarios de la siguiente manera: C=0001, G=0010, A=0100, y T=1000. Se utiliza sólo una capa oculta donde sólo hay un nodo debido a que el autor argumenta que es la red "perfecta" porque evita el sobre-entrenamiento.

Se propone para el entrenamiento de la red neuronal el algoritmo de retropropagamiento de error (EPB) porque se argumenta que estos se han utilizado con éxito para diversas aplicaciones en la bioinformática. Brevemente, la metodología que se plantea es la siguiente:

- (1) Cálculo de la salida de la red mediante la propagación de un patrón de entrada a través de las capas oculta y de salida.
- (2) Cálculo del error de predicción (La diferencia entre la salida deseada y la salida arrogada por la red neuronal).
- (3) la utilización del valor del error de predicción para actualizar las ponderaciones de la red con el fin de minimizar la función de ajuste de error, (generalmente la

desviación cuadrática media del error (RMSE)).

(4) Finalmente, para una secuencia de entrada dada si la salida de la red está entre 0.5 y 1.0 entonces la secuencia se supone que debe ser un promotor, en caso contrario (es decir, la salida de la red < 0.5) es un no promotor.

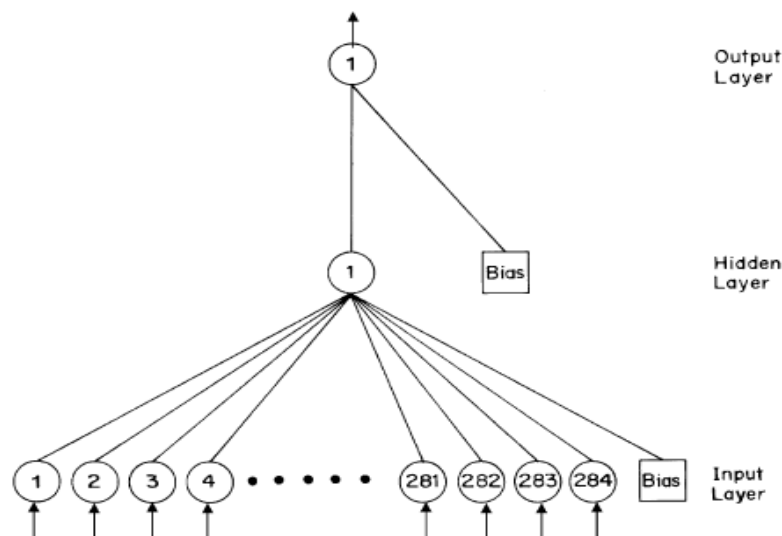


Imagen 11: Esquema de la red neuronal artificial optimizada, contiene 284 neuronas en la capa de entrada y una sola neurona en la capa oculta y de salida. Tomado de [3].

El resultado fue que la red podría de hecho diferenciar entre 95 secuencias promotoras y 285 secuencias aleatorias. Por otra parte, la red clasificó correctamente 155 secuencias de la prueba que comprende 160 secuencias (96.9% precisión de la clasificación). Los resultados indican que el modelo optimizado de la red posee buena capacidad de diferenciar entre una secuencia de promotor de una microbacteria y una secuencia aleatoria.

3.2. Red neuronal para la predicción de secuencias de promotores en humanos [4].

En el trabajo realizado en [4] como datos de entrada se seleccionó dos conjuntos representativos de señales promotoras de organismos eucarióticos, uno de la base de datos EPD, de diciembre de 1994, y otro de la base de datos de GenBank, de 1995. El conjunto de señales promotoras se compone de 429 regiones promotoras de vertebrados no

relacionados. Para el conjunto de genes humanos se seleccionaron todos los genes humanos conocidos en Genbank.

Los autores proponen entrenar dos redes neuronales distintas para luego combinarlas con una red neuronal de retardo temporal (TDNN) porque según su estudio, las TDNNs son apropiadas para el reconocimiento de elementos promotores, ya que son capaces de combinar múltiples características que aparecen en diferentes posiciones relativas. Como método de desarrollo se procede así:

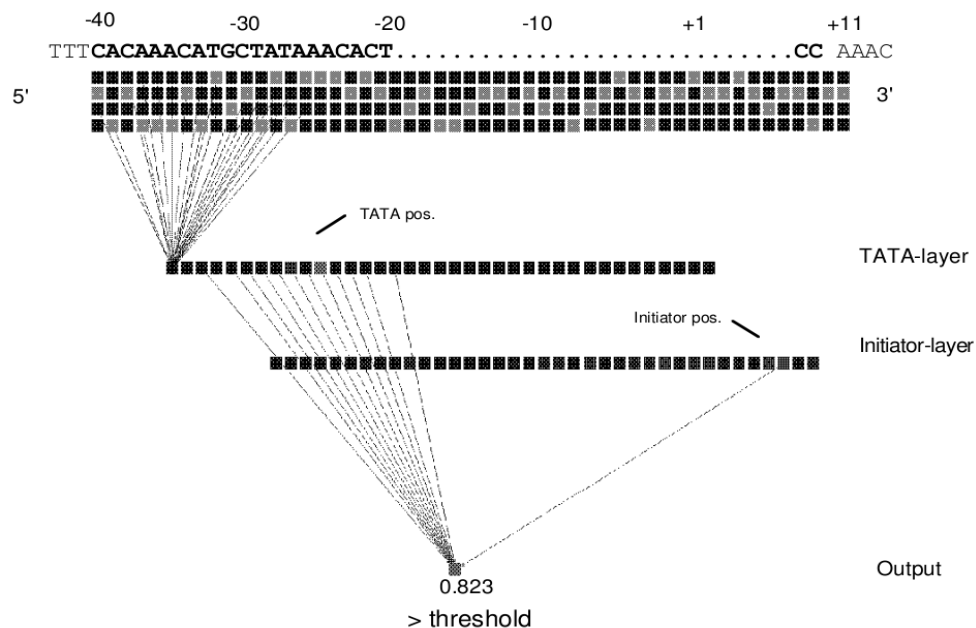


Imagen 12: Red neuronal de retardo temporal (TDNN). Tomado de [4].

(1) Se analiza la estructura de los elementos individuales dentro de los promotores utilizando una técnica que combina las redes neuronales con poda de peso. Esta técnica consiste en entrenar una red neuronal para reconocer los elementos de un promotor, hasta encontrar el mínimo local. Luego, el procedimiento de poda elimina los pesos en la red que suman el menor valor de predicción de la predicción total. Después de la poda, la red neuronal se vuelve a entrenar, hasta que otra vez encuentre un nuevo mínimo. Este procedimiento se repite hasta que se alcance el nivel del error establecido.

(2) Con el procedimiento anterior se entrenan dos redes neuronales distintas, uno para la caja TATA y otra para el iniciador. Cada red neuronal se entrena de forma independiente usando como entrada la secuencia que se encuentra alrededor del potencial sitio de interés.

(3) Se combinan las dos redes neuronales con una red neuronal de retardo temporal. La capa de entrada a la TDNN consiste en 51 nodos, usando la estrategia CODE-4 (explicado en la referencia anterior del estado del arte), que abarca el inicio de la transcripción de la posición de -40 a 11, incluyendo la caja TATA y el iniciador.

Los resultados obtenidos fueron, con un conjunto de 284 secuencias de genes humanos, predijo el 29% de los promotores existentes con una tasa 1/2955 de falsos positivos (el 0.03% por cada base). Pero aplicándole un umbral de 0.5 cambió al 58% de predicciones correctas y una tasa 1/717 de falsos positivos (el 0.14% por cada base).

3.3. Comparación entre redes neuronales y árboles de decisión en el análisis de secuencias. [12].

En el trabajo realizado en [12] se compara la capacidad de aprendizaje entre árboles de decisión con las redes neuronales de propagación hacia adelante con retro propagación, para la predicción de estructuras secundarias para cada aminoácido en una cadena de proteínas. Se utilizó como algoritmo de comparación el paquete PSI-BLAST y se seleccionó un conjunto de 1156 cadenas que contienen un total de 209.529 aminoácidos. El conjunto de pruebas consiste en 63 cadenas (que abarca 17.731 aminoácidos) que son variados para cada conjunto de entrenamiento. Brevemente el procedimiento utilizado consistió en:

(1) Para el desarrollo y entrenamiento del árbol de decisión se usó el software usfC4.5 que utiliza un método basado en la poda de error para simplificar el árbol construido a partir de todos los datos. El monto del árbol que se poda se controló con un parámetro de factor de certidumbre de 25 (entre [0-100]).

(2) Para el desarrollo y entrenamiento de la red neuronal se utilizó una red neuronal de propagación hacia adelante con retro propagación con sólo una capa oculta de 75 nodos, cuyo vector de entrada se compone de 315 valores, esto es porque hay 21 aminoácidos que aparecen en las cadenas proteínicas y porque se utilizó una ventana de tamaño 15.

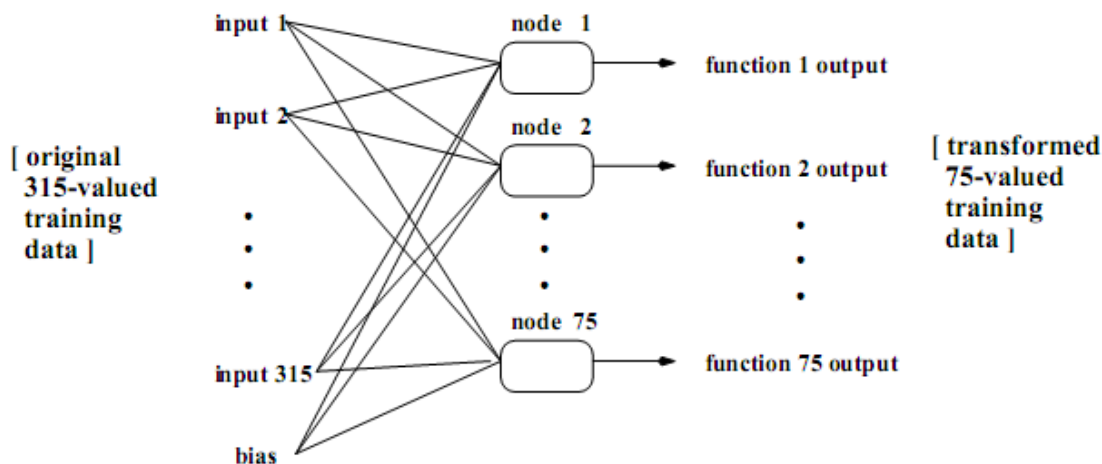


Imagen 13: Esquema de la red neuronal artificial. Tomado de [12].

Con respecto al árbol de decisión construido se obtuvo para un conjunto de entrenamiento y de pruebas, una precisión del 52.8% por aminoácido y 52.22% por cadena. La red neuronal mejor entrenada, después de un entrenamiento de 150 etapas, logró una precisión de 74.2% por aminoácido y 73.01% por la cadena. Esta es una diferencia de más del 20%. El mejor resultado obtenido con el árbol de decisión, bajando el factor de certidumbre a 0.01, fue una precisión de 68.8% por aminoácido y el 67.8% por la cadena. De este modo, al comparar los resultados, el árbol de decisión ha aumentado alrededor del 8% de precisión reduciendo la diferencia de rendimiento con la red neuronal.

Los autores finalmente concluyen que "Las redes neuronales superan significativamente a los árboles de decisión en el ámbito de la predicción de la estructura secundaria de proteínas a partir de su secuencia de aminoácidos. Una posible explicación para la diferencia en el rendimiento es que los nodos ocultos crean nuevos tipos de características que la red neuronal utiliza para diferenciar clases".

3.4. Predictor de promotor de E. coli usando redes neuronales Feed-Foward [20]

En el trabajo realizado en [20] para el entrenamiento de la red neuronal se seleccionaron dos conjuntos de datos con patrones de longitud de 60 pares de bases. El primero consiste en un conjunto positivo de promotores de 471

patrones obtenidos a partir de bases de datos de secuencias promotoras en las cuales se han determinado y validado experimentalmente los puntos de inicio de transcripción. El segundo consiste en un conjunto negativo de promotores de 365 patrones de la región codificante de *E. coli* K12 disponibles en la base de datos NCBI.

El procedimiento propuesto por los autores consiste en realizar una red neuronal con datos de entrada transformados en datos binarios o numéricos, donde se prueban dos tipos de esquemas de codificación de los ácidos nucleicos. El primer esquema de codificación consta de cuatro dimensiones: A = 1000, C = 0100, G = 0010, T = 0001 (Ver imagen 7). El segundo esquema de codificación consta de dos dimensiones: A = 00, C = 01, G = 10, T = 11. En el esquema de codificación de cuatro dimensiones, la capa de entrada dispone de 240 nodos, la capa oculta de 80 nodos, y la capa de salida de un nodo (Ver imagen 8). En el esquema de codificación de dos dimensiones, la capa de entrada consta de 120 nodos, la capa oculta de 80 nodos, y la capa de salida de un nodo. Para los conjuntos de entrenamiento positivos la salida es 1, y para las series de entrenamiento negativos la salida es -1. Si la salida es mayor o igual a cero la entrada es una secuencia promotora, y si la salida es menor que cero la entrada no es una secuencia promotora.

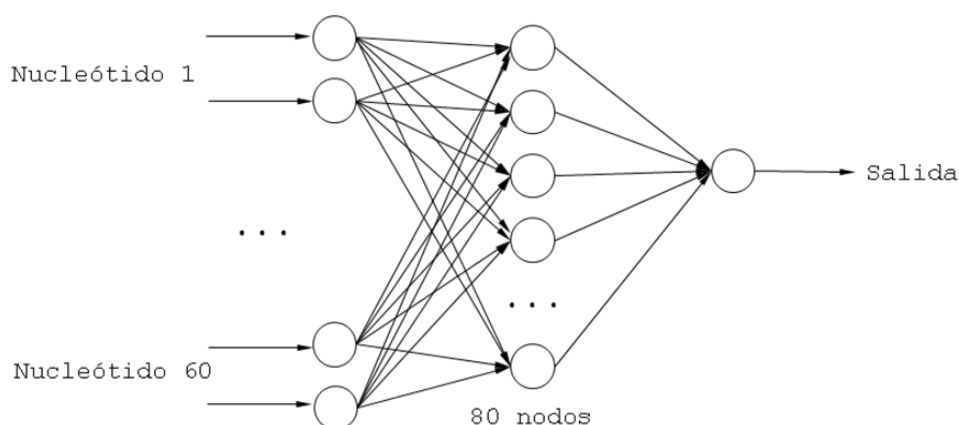


Imagen 14. Topología de red para el esquema de 2 dimensiones.

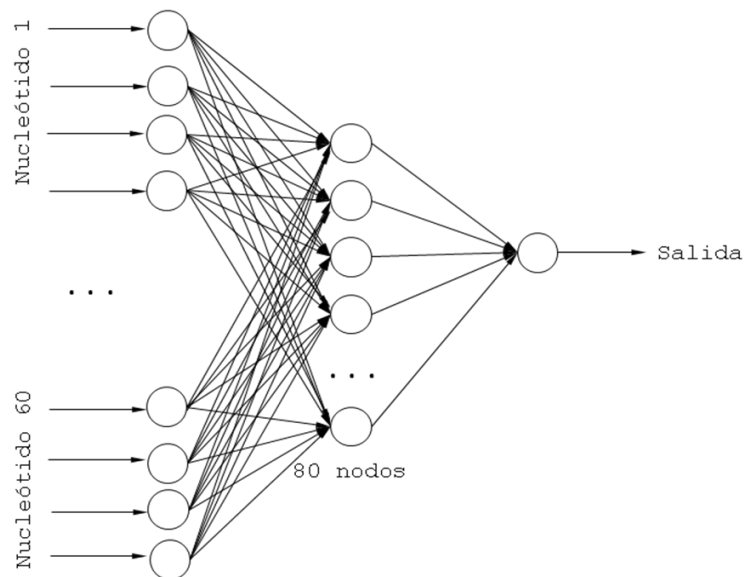


Imagen 15. Topología de red para el esquema de 4 dimensiones.

Los resultados obtenidos fueron que la tasa de aprendizaje para el sistema de codificación de cuatro dimensiones fue de 0.5, mientras que para el segundo esquema fue de 0.25 y que al cambiar el número de nodos de la capa oculta para ambos esquemas de codificación, los parámetros de entrenamiento de la red se mantuvieron iguales.

Se interpreta que al comparar la tasa de error de ambos esquemas de codificación, el sistema de entrenamiento de cuatro dimensiones es superior al de dos dimensiones. Segundo, que no influye el número de nodos en la capa oculta para la precisión de la predicción de la secuencia promotora del *E. coli*. Finalmente, se concluye que la red neuronal parece ser una opción adecuada y suficientemente precisa para detección automática de la ubicación de promotores para el *E. coli*.

3.5. Identificación de promotores en el genoma del hongo *Crinipellis perniciosus* usando redes neuronales feed-forward [21]

En el trabajo realizado en [21] se decidió utilizar los datos ya validados experimentalmente de Eukaryotic Promoter Database (<http://my.epd.isb-sib.ch/>) no solamente de las secuencias promotoras del hongo *Crinipellis perniciosus*, sino que también de otros promotores de muchas especies, incluyendo plantas, vertebrados y el *homo sapiens*. El

procedimiento se desarrolló de la siguiente manera:

(1) Se seleccionó la secuencia de la caja-TATA como subestructura de referencia interna para usarla como indicador de la presencia de un promotor hipotético.

(2) Segundo, a partir de diferentes muestras de secuencias de caja-TATA se determinó usar el tamaño medio de 10 pares de bases como ancho definido para hacer los siguientes análisis, ya que el tamaño de la caja-TATA varía según la especie.

(3) Se eligió 8 medidas características de las 80 posibles que el autor podía realizar para la caja-TATA de tamaño ya definido. Algunas de las medidas elegidas se determinó según fuera más repetida la combinación de dos o tres nucleótidos en determinada región a partir del sitio de inicio de la transcripción, por ejemplo AA, TA, AAA, TAA, TAT, etc., que se presentaba a lo largo de los 250 pares de bases de entre las diferentes secuencias promotoras.

(4) Se construyó una red neuronal feed-forward con 8 neuronas de entrada, una por cada característica, una capa oculta y una neurona de salida. La salida varía dentro del intervalo [0,1] que indica la probabilidad de que el patrón de entrada corresponde a una caja-TATA. El número de neuronas ocultas se determinó mediante pruebas de diferentes configuraciones, dando como mejor topología dos neuronas ocultas (Ver Imagen 9).

(5) Se formó dos patrones de vectores de dimensión 8; el primero, el patrón "caja-TATA", a partir de todas las secuencias promotoras que se calcularon que presentaban positivamente las 8 características escogidas; el segundo, el patrón "no caja-TATA" a partir de secuencias de promotores válidas pero con un recorte, que se elige aleatoriamente entre los 200 pares de bases fuera de la caja-TATA, de 10 pares de bases. A partir de estos dos patrones se crean tres conjuntos diferentes: dos para entrenamiento de la red y el tercero para la validación de la misma.

(6) Los conjuntos de entrenamiento fueron contruidos por una selección al azar de la mitad de los patrones "caja-TATA" y "no caja-TATA", dejando el resto de los patrones para efectos de validación y formar el tercer conjunto.

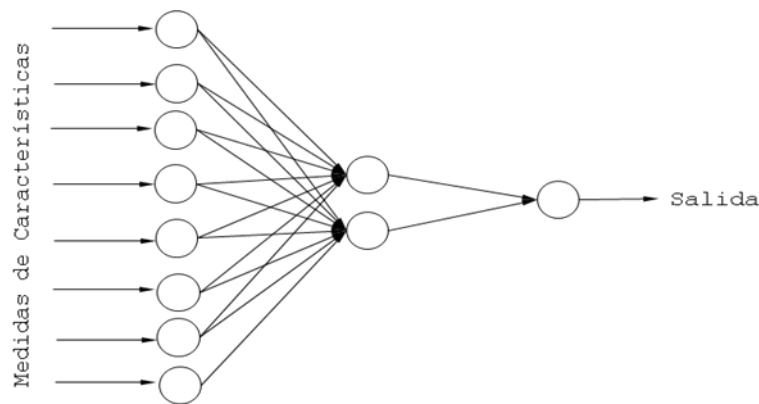


Imagen 16. Topología de red que obtuvo el mejor desempeño.

Los resultados obtenidos fueron que la red neuronal feed-forward con sólo dos neuronas ocultas, tuvo el mejor desempeño alcanzando una especificidad del 99% y una sensibilidad del 95%.

Además, cuando la red neuronal se utilizó para encontrar cajas-TATA sobre un conjunto de doce potenciales secuencias promotoras de 600 bases cada una, se observó que diez de las doce posibles secuencias promotoras muestran una región bien definida rica en el patrón que presenta una caja-TATA a distancias que varían desde 90 hasta 300 bases desde el codón de inicio, lo cual es consistente con las observaciones biológicas. Según el autor, la ausencia de la supuesta caja-TATA en las otras dos secuencias parece estar relacionada con una estructura inusual de promotor o mutaciones degenerativas.

Finalmente concluye que los resultados de su trabajo sugieren que el uso de características basadas en las medidas locales de la composición tales como dinucleótidos, trinucleótidos y frecuencias, en vez de la habitual codificación binaria de nucleótidos, mejora la aplicación de redes neuronales para la caracterización estructural de los promotores y el estudio de los mecanismos reguladores de genes.

En el anexo 2 se puede ver la continuación del estado del arte.

Capítulo 4

MODELOS PROPUESTOS

En este capítulo se presentan los modelos de redes neuronales propuestos que intentan dar solución al problema de la predicción de promotores. Se propone realizar diversas redes neuronales artificiales (RNA) para la predicción de señales promotoras con el propósito de evaluar y comparar dichos modelos. Para ello se pretende crear dos tipos de modelos de RNA de diferente topología y de diferente estrategia de valores de entrada y codificación.

4.1. MODELO 1

En el primer modelo, que llamaremos RED1, se pretende emplear la estrategia de utilizar como valores de entrada para la RNA resultados obtenidos por diferentes modelos ocultos de Markov, Hidden Markov Model (HMM). La estrategia pretende integrar la salida de las medidas usadas por GeneZilla[27]. Dicho predictor tiene 4 modelos disponibles para la detección de promotores pero la decisión de clasificación se debe tomar seleccionando tan solo uno. En este trabajo de investigación se tiene la hipótesis de que el integrar estos modelos se puede tomar una decisión de clasificación más acertada.

4.1.1. Especificaciones de RED1

Los HMM que se escogieron como valores de entrada para la RED1 fueron Weight Matrix Model (WMM), Weight Array Model (WAM) y Windowed Weight Array Model (WWAM) pertenecientes al paquete de la herramienta bioinformática GeneZilla v.0.1. Para que los HMM logren un buen desempeño en la predicción de promotores éstos requieren de un entrenamiento previo. El entrenamiento de GeneZilla para predecir promotores requiere que se elijan los conjuntos de secuencias de promotores y no promotores. Para nuestro trabajo se escogieron para este entrenamiento secuencias de promotores de las especies seleccionadas, más adelante mencionadas en detalle (Ver sección 5.1), y con tamaño de 120 nucleótidos.

Además de utilizar los modelos de Markov disponibles en GeneZilla, se incluyó como valor de entrada el resultado de

Banana[28]. Esta es una aplicación bioinformática que permite calcular la elasticidad y curvatura dada una secuencia de nucleótidos basada en el algoritmo BEND[29]. Estudios relacionados con la funcionalidad de curvaturas presentes en el ADN cercano a promotores [25, 26] sugieren que la flexión del ADN puede ser un mecanismo mediante el cual se regule la transcripción eucariótica. Por este motivo, el agregar los valores de elasticidad y curvatura de una secuencia servirá para validar si la región puede ser promotora.

Debido a que los valores obtenidos por los HMMs y Banana superan el rango de $[0.0 - 1.0]$, se decidió normalizar los valores de la RED1 dividiendo cada parámetro de entrada por el mayor valor posible que podría alcanzar dicho parámetro. Finalmente, la mejor topología que se logró para la RED1 consiste en una capa de entrada con 5 nodos, una capa oculta de un nodo y como capa de salida un nodo (Ver imagen 10), donde la salida puede ser "1" o "0" en caso de ser la predicción positiva (promotor) o negativa (no promotor) respectivamente.

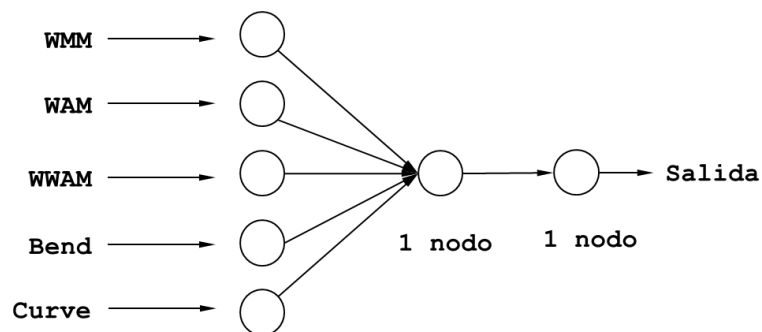


Imagen 17. Topología de la RED1.

4.2. MODELO 2.

En el segundo modelo de RNA, que llamaremos RED2, se pretende emplear la estrategia que utiliza como valores de entrada los resultados obtenidos por otras RNAs de predicción de secuencias de caja consenso ya conocidas. Como no existen para dominio público RNAs para la predicción de cajas consenso, inicialmente se crearon tres RNAs para que reconozcan la caja TATA, la caja GC y la caja CAAT respectivamente, que a continuación se describirán.

4.2.1. RNAs para el reconocimiento de la caja TATA, CAAT y GC

Para el entrenamiento de la RNA que se reconoce la caja TATA se utilizaron como conjunto de secuencias positivas secuencias de caja TATA que fueron generadas a partir de la tabla de frecuencias de bases de la caja TATA obtenidas de The Eukaryotic Promoter Database (EPD)[7b]; y como conjunto de secuencias negativas pequeñas secuencias de intrones y exones omitiendo las que se parecen demasiado a la caja TATA. Se usaron secuencias de tamaño 7 porque que el tamaño de la caja TATA en promedio es de 7 nucleótidos.

Como parámetros de entrada se utilizó la estrategia de codificación ortogonal CODE-4, que consiste en representar cada nucleótido como 4 dígitos binarios de la siguiente manera: C=0001, G=0010, A=0100, y T=1000. Esta estrategia se eligió teniendo en cuenta los resultados de estudios anteriores [3, 4, 20, 23, 24] según los cuales es más eficiente esta codificación para este tipo de problemas.

Adicionalmente a la codificación se agregó como valor incierto el N=1111. La topología de la red propuesta consta de 28 nodos en la capa de entrada, 1 nodo para la capa oculta y 1 nodo para la capa de salida (Ver imagen 11). La salida de la red será "1" cuando identifica una caja TATA y "0" en caso contrario.

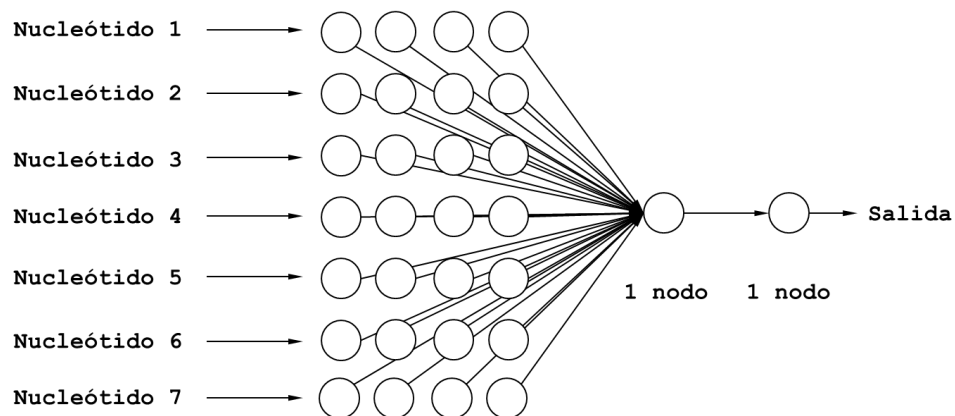


Imagen 18. Topología de la RNA que reconoce la caja TATA.

Para la creación de las RNAs para el reconocimiento de la caja CAAT y GC se utilizó el mismo procedimiento para la creación de la RNA que reconoce la caja TATA teniendo en cuenta que el conjunto de entrenamiento de secuencias positivas serán las secuencias de caja CAAT o GC según sea la especialización de la RNA y además para la caja CAAT la cantidad de nodos de entrada será de 28 por tener en promedio 7 nucleótidos (Ver imagen 12) y para la caja GC la cantidad de nodos de entrada será de 24 por tener en promedio 6 nucleótidos (Ver imagen 13).

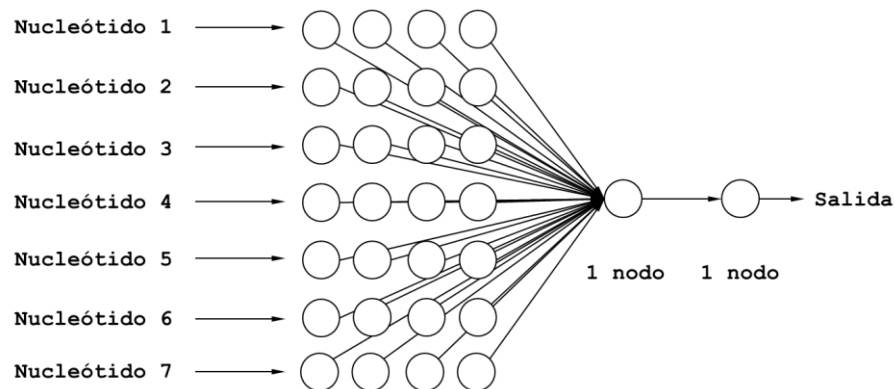


Imagen 19. Topología de la RNA que reconoce la caja CAAT.

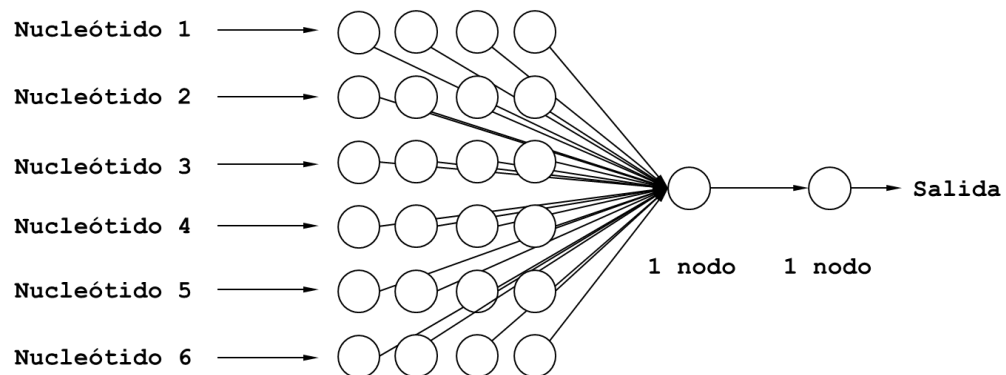


Imagen 20. Topología de la RNA que reconoce la caja GC.

4.2.2. Especificaciones de RED2

Lo que se pretende para este modelo, RED2, es utilizar los resultados arrojados por las RNAs como parámetros de entrada

a otra RNA. Ya varios estudios han utilizado como estrategia de entrada valores de otras RNAs [4, 24] y por lo general el valor de predicción arrojado por las RNAs se convierte en parámetro de entrada a un nodo de otra RNA. Sin embargo, hasta la fecha y en varios artículos que se han publicado sobre modelos de predicción por RNAs no se ha utilizado la estrategia que se va a implementar en este modelo y que a continuación se describe.

En vez de utilizar los valores de predicción de las RNAs como simples parámetros de entrada, como lo hace el modelo 1, se propone que cada RNA recorra toda la secuencia de código genético a evaluar y vaya dejando un rastro con valor positivo "1" si ha encontrado la caja consenso en determinada región de la secuencia o valor negativo "0" en caso contrario. De esta manera, al final habrán tres trazos que son los valores del rastreo de cada RNA sobre la secuencia del código (Ver imagen 14).

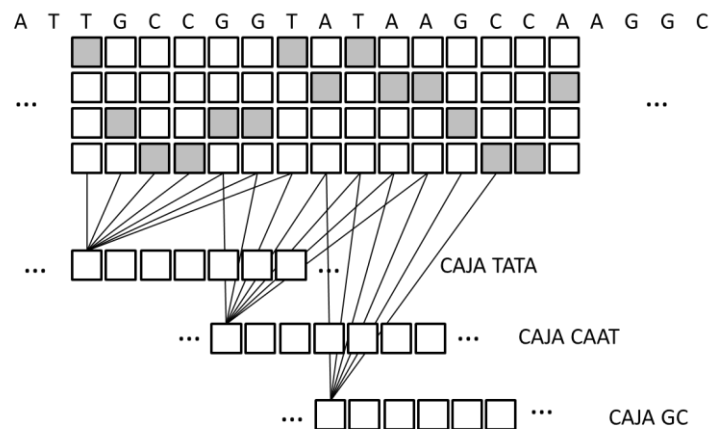


Imagen 21. Esquema de la lectura y trazo arrojado que realiza las RNAs.

Los vectores con el trazo arrojado de las RNAs, que contienen ceros y unos, se alinean de manera que coincida el inicio de lectura de cada trazo para dar como estrategia de entrada valores de codificación binaria de 3 dígitos, y que finalmente son los que recibirá otra RNA como parámetros de entrada. Actualmente, la red neuronal propuesta (RED2) se diferencia de los modelos existentes para la predicción de promotores, en que utiliza rastros binarios arrojados por la predicción de las RNAs para generar una codificación binaria de 3 dígitos para la entrada de otra RNA.

Como las cajas consenso TATA, CAAT y GC, en su gran mayoría, se encuentran entre la posición -219 hasta la 10 a partir del sitio de inicio de transcripción[7b] se decidió usar secuencias de tamaño 250 nucleótidos que se encuentren entre las posición -229 a la 30. Sin embargo, el tamaño del trazo arrojado por las lecturas de las RNAs para el reconocimiento de las cajas consenso es de 244 porque la última lectura de las RNAs termina en esa posición. Por lo tanto, la topología de la RED2 consiste en una capa de entrada con 732 nodos, un nodo para la capa oculta y un nodo en la capa de salida (Ver imagen 15).

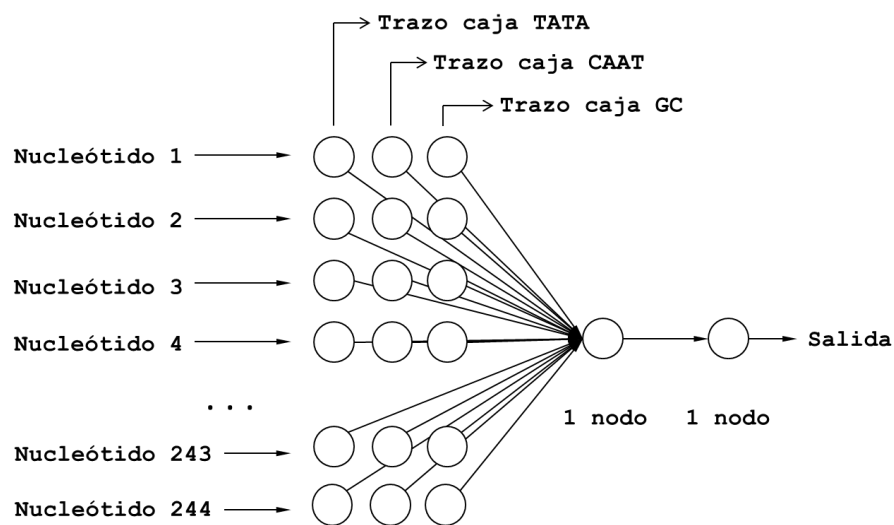


Imagen 22. Topología de la RED2.

La secuencia a evaluar deberá tener un tamaño mínimo de 250 nucleótidos para que la RED2 la recorra y después dar su resultado de predicción. Esto porque la RED2 se entrenó para reconocer la región promotora a partir de una región exacta cercana al sitio de inicio de la transcripción.

El procedimiento para la predicción final consiste en que la RED2 a medida que recorre la secuencia va dejando un rastro de uno "1" o cero "0" si encontró promotor o no respectivamente; y finalmente, como se entrenó la RED2 para reconocer la secuencia promotora en 10 posiciones seguidas cercanas al sitio de inicio de la transcripción, si encuentra más de 10 unos seguidos arroja acierto "1" en la predicción de promotor y falso "0" en caso contrario.

Capítulo 5

PRUEBAS Y RESULTADOS

En este capítulo se presentan detalles del proceso de entrenamiento de los modelos RED1 y RED2 y las pruebas realizadas para validación. Además, se hace un análisis de los resultados obtenidos en las pruebas para identificar y comparar el grado de precisión en la predicción de promotores de ambos modelos.

5.1. Especificaciones de los conjuntos de entrenamiento y prueba

Para el entrenamiento y validación de la RED1 y RED2 se utilizaron como conjunto de secuencias positivas las 355 secuencias promotoras de la especie "Rattus norvegicus" (rata) y las 192 secuencias promotoras del "Mus musculus" (ratón); disponibles en la EPD[7]. Como conjunto de secuencias negativas se utilizaron intrones y exones de diferentes especies extraídas de la GISSD[31] y de la TiGER[30] respectivamente.

Es importante aclarar que los modelos de RED1 y RED2 están enfocados sólo a la predicción de promotores reconocidos por el ARN polimerasa tipo II o B (Ver sección 5.1.1) porque las secuencias que se han obtenido de la EPD[7] para el entrenamiento y validación de las RNAs sólo contienen secuencias promotoras de esa categoría.

El proceso de entrenamiento consistió en utilizar el 80% del conjunto total de las secuencias promotoras de la especie escogida y para las no promotoras un poco más del equivalente de secuencias promotoras. El tamaño de las secuencias para el entrenamiento de la RED1, debido a las características del entrenamiento de Genezilla (Ver sección 4.1.1), serán de tamaño 120 nucleótidos entre la posición -99 hasta la 20 a partir del sitio de inicio de transcripción (Ver tablas 1, 2, 3). El tamaño de las secuencias para el entrenamiento de la RED2, debido a las características de la misma (Ver sección 4.2.4), serán de tamaño 250 nucleótidos y secuencias promotoras entre la posición -229 hasta la 30 a partir del sitio de inicio de transcripción (Ver tablas 4, 5 y 6).

Rata	Cantidad Promotores	Cantidad no promotores
Conjunto de entrenamiento	11	283
Conjunto de prueba	5	120

Tabla 1. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata con 120 nucleótidos de tamaño.

Ratón	Cantidad Promotores	Cantidad no Promotores
Conjunto de entrenamiento	23	283
Conjunto de prueba	10	120

Tabla 2. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores del ratón con 120 nucleótidos de tamaño.

Rata - Ratón	Cantidad Promotores	Cantidad no Promotores
Conjunto de entrenamiento	34	283
Conjunto de prueba	15	120

Tabla 3. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata y ratón con 120 nucleótidos de tamaño.

Rata	Cantidad Promotores	Cantidad no promotores
Conjunto de entrenamiento	160	325
Conjunto de prueba	40	40

Tabla 4. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata con 250 nucleótidos de tamaño.

Ratón	Cantidad Promotores	Cantidad no Promotores
Conjunto de entrenamiento	260	325
Conjunto de prueba	70	70

Tabla 5. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores del ratón con 250 nucleótidos de tamaño.

Rata - Ratón	Cantidad Promotores	Cantidad no Promotores
Conjunto de entrenamiento	420	325
Conjunto de prueba	110	110

Tabla 6. Número de secuencias para entrenamiento y prueba de la RNA para la detección de promotores de la rata y ratón con 250 nucleótidos de tamaño.

5.2. CRITERIOS DE COMPARACIÓN

El proceso de validación consistió en evaluar los indicadores de sensibilidad y especificidad. Primero, con las mismas secuencias del conjunto de entrenamiento para calcular el error de entrenamiento, error que permite conocer si la topología de red utilizada es adecuada para clasificar los datos que hacen parte del entrenamiento. Segundo, se calcula el error de generalización con el 20% restante de las secuencias promotoras positivas que **no** se entrenaron y un equivalente número de secuencias negativas que tampoco se habían entrenado. A partir de la tabla 7, se puede definir la sensibilidad (S_n) y la especificidad (S_p) como $Sensitividad = \frac{VP}{VP+FN}$ y la $Especificidad = \frac{VN}{VN+FP}$ respectivamente.

	Promotor	No promotor
Positivo	Verdaderos Positivos (VP)	Falsos Positivos (FP)
Negativo	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Tabla 7. Relación entre el resultado de una prueba.

5.3. PRUEBAS AL MODELO 1.

5.3.1. PRUEBA 1.1

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 1, con 74 secuencias de exones y 11 secuencias de promotores de la especie "Rattus norvegicus". La prueba permitirá conocer el comportamiento de la red empleada en el modelo 1 en la predicción de promotores de la especie rata con secuencias entrenadas de promotores y exones, es decir, se calculará el error de entrenamiento y se evaluará el comportamiento de predicción sin intrones.

Resultados obtenidos: En la tabla 8 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
11	1	73	0	1.0	0.986

Tabla 8. Resultados de la prueba 1.1.

5.3.2. PRUEBA 1.2

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 1, con 31 secuencias de exones y 5 secuencias de promotores de la especie "Rattus norvegicus". La prueba permitirá conocer si la red empleada en el modelo 1, entrenada con el conjunto de secuencias indicadas en la prueba 1.1, permite detectar promotores de la especie rata con el conjunto de secuencias no entrenadas, es decir, se calculará el error de generalización y se evaluará el comportamiento de predicción sin intrones.

Resultados obtenidos: En la tabla 9 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
2	1	30	3	0.4	0.967

Tabla 9. Resultados de la prueba 1.2.

5.3.3. PRUEBA 1.7

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 1, con 74 secuencias de exones y 23 secuencias de promotores de la especie "Mus musculus". La prueba permitirá conocer si la red empleada en el modelo 1 admite la predicción de promotores de la especie ratón con secuencias entrenadas de promotores y exones, es decir, se calculará el error de entrenamiento y se evaluará el comportamiento de predicción sin intrones.

Resultados obtenidos: En la tabla 10 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
16	0	74	7	0.695	1.0

Tabla 10. Resultados de la prueba 1.7.

5.3.4. PRUEBA 1.8

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 1, con 31 secuencias de exones y 10 secuencias de promotores de la especie "Mus musculus". La prueba permitirá conocer si la red empleada en el modelo 1, entrenada con el conjunto de secuencias indicadas en la prueba 1.7, permite detectar promotores de la especie ratón con el conjunto de secuencias no entrenadas, es decir, se calculará el error de generalización y se evaluará el comportamiento de predicción sin intrones.

Resultados obtenidos: En la tabla 11 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
2	0	31	8	0.2	1.0

Tabla 11. Resultados de la prueba 1.8.

De la misma forma se llevaron a cabo pruebas al modelo 1 donde se modificaron las cantidades de datos que hacen parte del entrenamiento y el tipo de organismo. En el anexo 3 se puede ver la documentación completa de tales pruebas.

5.4. PRUEBAS AL MODELO 2.

5.4.1. PRUEBA 2.1

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 120 secuencias de exones e intrones y 120 secuencias de promotores de la especie "Rattus norvegicus". La prueba permitirá conocer si la topología de red empleada en el modelo 2 permite detectar promotores de la especie rata a partir de un conjunto de secuencias entrenadas, es decir, se calculará el error de entrenamiento.

Resultados obtenidos: En la tabla 12 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
120	0	120	0	1.0	1.0

Tabla 12. Resultados de la prueba 2.1.

5.4.2. PRUEBA 2.3

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 40 secuencias de exones e intrones y 40 secuencias de promotores de la especie "Rattus norvegicus". La prueba permitirá conocer si la red empleada en el modelo 2, entrenada con el conjunto de secuencias indicadas en la prueba 2.2 (Ver anexo 4), permite detectar promotores de la especie rata con el conjunto de secuencias no entrenadas, es decir, se calculará el error de generalización.

Resultados obtenidos: En la tabla 13 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
27	10	30	13	0.675	0.75

Tabla 13. Resultados de la prueba 2.3.

5.4.3. PRUEBA 2.6

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 200 secuencias de exones e intrones y 200 secuencias de promotores de la especie "Mus musculus". La prueba permitirá conocer si la topología de red empleada en el modelo 2 permite detectar promotores de la especie ratón a partir de un conjunto de secuencias entrenadas, es decir, se calculará el error de entrenamiento.

Resultados obtenidos: En la tabla 14 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
200	0	200	0	1.0	1.0

Tabla 14. Resultados de la prueba 2.6.

5.4.4. PRUEBA 2.8

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 70 secuencias de exones e intrones y 70 secuencias de promotores de la especie "Mus musculus". La prueba permitirá conocer si la red empleada en el modelo 2, entrenada con el conjunto de secuencias indicadas en la prueba 2.7 (Ver anexo 4), permite detectar promotores de la especie ratón con el conjunto de secuencias no entrenadas, es decir, se calculará el error de generalización.

Resultados obtenidos: En la tabla 15 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
28	8	62	42	0.4	0.8857

Tabla 15. Resultados de la prueba 2.8.

5.4.5. PRUEBA 2.11

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 200 secuencias de exones e intrones y 200 secuencias de promotores de las especies "Rattus norvegicus" y "Mus musculus". La prueba permitirá conocer si la topología de red empleada en el modelo 2 permite detectar promotores de las especies rata y ratón a partir de un conjunto de secuencias entrenadas, es decir, se calculará el error de entrenamiento.

Resultados obtenidos: En la tabla 16 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
200	0	200	0	1.0	1.0

Tabla 16. Resultados de la prueba 2.11.

5.4.6. PRUEBA 2.13

Objetivo de la prueba: Calcular los indicadores de exactitud S_n y S_p , para el modelo 2, con 110 secuencias de exones e intrones y 110 secuencias de promotores de las especies "Rattus norvegicus" y "Mus musculus". La prueba permitirá conocer si la red empleada en el modelo 2, entrenada con el conjunto de secuencias indicadas en la prueba 2.12 (Ver anexo 4), permite detectar promotores de la especie rata y ratón con el conjunto de secuencias no entrenadas, es decir, se calculará el error de generalización.

Resultados obtenidos: En la tabla 17 se muestra un reporte de las medidas obtenidas como resultado de las pruebas.

VP	FP	VN	FN	S_n	S_p
82	19	91	28	0.7454	0.8272

Tabla 17. Resultados de la prueba 2.13.

De la misma forma se llevaron a cabo pruebas al modelo 2 donde se modificaron las cantidades de datos que hacen parte del entrenamiento y el tipo de organismo. En el anexo 4 se puede ver la documentación completa de tales pruebas.

5.5. DISCUSIÓN DE RESULTADOS

Considerando los resultados obtenidos en las pruebas realizadas a los modelos, se evidencian los siguientes aspectos:

- Dados los resultados obtenidos por los indicadores de sensibilidad y especificidad en las pruebas 1.1, 1.7 y 1.13 el modelo propuesto para la RED1 presenta sus valores más altos en el aprendizaje y reconocimiento de secuencias promotoras y no promotoras de tipo exón, con valores de sensibilidad de 1.0, 0.695, y 0.882 respectivamente y con valores de especificidad de 0.986, 1.0 y 0.972 respectivamente.

Sin embargo, en las pruebas 1.4, 1.10 y 1.16 que contienen, además de exones, intrones en el conjunto de secuencias negativas muestran que el comportamiento de la RED1 en el aprendizaje y reconocimiento de secuencias promotoras baja considerablemente, con valores de sensibilidad de 0.909, 0.478 y 0.323 respectivamente.

Lo anterior significa que el modelo de RED1 no es sensible al reconocimiento y diferenciación entre promotores e intrones. Esto se debe a que los resultados arrojados por los HMM de GeneZilla y banana no muestran suficiente diferencia entre los valores obtenidos para promotores y para intrones.

- Los resultados obtenidos de sensibilidad en las pruebas 1.2, 1.8 y 1.14 muestran que el comportamiento de la RED1 para la predicción de promotores de la rata y el ratón con el conjunto de prueba de secuencias negativas sin intrones bajaron hasta un 50%, con valores de sensibilidad de 0.4, 0.2 y 0.6 respectivamente. Así mismo, en las pruebas 1.5, 1.11 y 1.17 con promotores de la rata y el ratón con el conjunto de prueba de secuencias negativas con intrones y sin entrenar bajaron también hasta un 60%, con valores de Sn de 0.4, 0.10 y 0.133 respectivamente.

Es muy probable, como se explicará más detalladamente en el análisis para la RED2, que esta dificultad para predecir promotores no entrenados se deba a que las especies de la rata y el ratón posean variados motivos de promotores y no tan rígidos como en especies inferiores, como las procariotas.

- Según los resultados obtenidos por los indicadores de sensibilidad y especificidad en las pruebas 1.3, 1.6, 1.9, 1.12, 1.15 y 1.18 se evidencia la incapacidad de la RED1 para la predicción de promotores de la rata y ratón con secuencias de mayor tamaño, en estas pruebas se usaron 250 nucleótidos de tamaño.

Esto ocurre porque el modelo propuesto para la RED1 está limitado a las características de los HMM de GeneZilla, que para la predicción de promotores todavía está en etapa de prueba y el entrenamiento es muy rígido aceptando como un máximo de secuencias promotoras de 120 nucleótidos de tamaño (Ver sección 4.1.1).

- Como el modelo 1 se desarrolló para el reconocimiento de promotores de especies eucariotas y como las especies de prueba en este trabajo fueron la rata y ratón, que son genéticamente muy similares, ninguna especie mostró ventajas frente a la otra con respecto a la facilidad de modelarse en las redes propuestas. Por este motivo, las pruebas 1.1, 1.7 y 1.13 muestran que el tipo de organismo afecta en lo mínimo el error de entrenamiento de la RED1, es decir, es muy poco sensible al organismo en el conjunto de entrenamiento.
- Dados los resultados obtenidos por los indicadores de sensibilidad y especificidad en las pruebas 2.1, 2.2, 2.6, 2.7, 2.11 y 2.12 el modelo propuesto para la RED2 se comporta de manera perfecta en el aprendizaje de secuencias. Esto es debido a la estrategia de codificación binaria que se utilizó como entrada de los parámetros tanto para codificar los nucleótidos como para codificar la predicción de las cajas consenso TATA, CAAT y GC.

Sin embargo, los indicadores de sensibilidad y especificidad en las pruebas 2.3 y 2.8 muestran que el comportamiento de la RED2 para la predicción de promotores de la rata y el ratón no entrenados en promedio acierta el 50%. Este porcentaje comparado con otros estudios relacionados [3, 21, 23, 24] y basados también en redes neuronales es un tanto más bajo, pero debe considerarse que en dichos estudios se analizan secuencias de código genético de especies más simples como bacteria *E. coli*, el hongo *Crinipellis perniciosus* y la mosca *Drosophila*. Se sabe que entre más compleja sea la especie su genoma posee más cantidad de código que no se transcribe y por lo tanto mayor será la dificultad para encontrar la zona donde se

encuentra el código promotor.

Por este mismo motivo el estudio realizado en [4] para secuencias promotoras humanas muestra que los indicadores de especificidad y sensibilidad son menores al 40%. Otro factor a considerar es que en algunos casos de estudio se calcularon los indicadores de exactitud usando secuencias que hicieron parte del entrenamiento.

- Los resultados obtenidos de sensibilidad y especificidad en las pruebas 2.2, 2.7 y 2.12 muestran que el modelo de RED2 no es sensible a la cantidad de datos para entrenamiento, al aumentar la cantidad, el modelo es capaz de mantener su sensibilidad y especificidad.
- Por otra parte, los resultados obtenidos de sensibilidad y especificidad en las pruebas 2.4, 2.9 y 2.14, con conjunto de secuencias de mayor cantidad de nucleótidos a las pruebas 2.3, 2.8 y 2.13, son menores porque detecta erróneamente promotores en secuencias de intrones. En los entrenamientos con las especies de la rata, el ratón y conjuntamente la rata y el ratón, la cantidad de falsos positivos aumentó cuando las secuencias eran de mayor cantidad de nucleótidos, en proporción con el número de secuencias, de 25% a 42.9%, de 11.4% a 65% y de 17.3% a 79.9% respectivamente según el conjunto de datos de entrenamiento. Esto sugiere la posibilidad de que existan prototipos o copias de trazos de secuencia promotora en regiones donde no hay transcripción a proteínas.
- Al comparar los resultados obtenidos en las pruebas 2.5, 2.10 y 2.15, cuyos valores de S_n superan 0.9, con los resultados de las pruebas 2.3, 2.8 y 2.13, cuyos valores de S_n oscilaban entre 0.4 a 0.7, muestran que la RED2 es sensible al reconocimiento de una misma secuencia en diferentes posiciones del sitio de inicio de transcripción. Es probable que este aumento de predicción de promotores se deba a que la RED2 reconoce pequeños patrones de código diferentes a las cajas consenso ya reconocidas en regiones cercanas al sitio de inicio de transcripción, posiblemente no identificados hasta la fecha, que no se encuentran exactamente en la misma posición y en el mismo orden.
- Las pruebas 2.11 y 2.12, que se realizaron con un entrenamiento de secuencias promotoras de dos especies diferentes al mismo tiempo, muestra el buen comportamiento

que posee el modelo 2 en la capacidad de escalabilidad, ya que no disminuyó la eficacia en la predicción según los indicadores especificidad y sensibilidad. Además, dado el resultado de la prueba 2.13, el modelo es más efectivo si es entrenado con secuencias de diferentes especies, seguramente porque la RED2 aprende más posibilidades de motivos de promotores.

- Como el modelo 2 se desarrolló para el reconocimiento de promotores de especies eucariotas y como las especies de prueba en este trabajo fueron la rata y ratón, que son genéticamente muy similares, ninguna especie mostró ventajas frente a la otra con respecto a la facilidad de modelarse en el modelo propuesto. Por este motivo, las pruebas 2.2, 2.7 y 2.12 muestran que el tipo de organismo no afecta el error de entrenamiento de la RED2, es decir, no es sensible al organismo en el conjunto de entrenamiento.
- Los tiempos de construcción de los modelos 1 y 2, una vez identificadas la mejor topología para las RNAs, se hace en un periodo corto de tiempo porque se utiliza la herramienta JavaNNS.

En promedio el proceso de construcción de la topología y el entrenamiento de RNA se hace en menos de 1 minuto sin importar la cantidad de nodos que se van a utilizar. Por este motivo el tiempo de construcción de ambos modelos, a pesar que la topología de la RNA del modelo 1 utiliza sólo 5 nodos de entrada frente a los 732 nodos de entrada del modelo 2, se realizan en equivalente periodo de tiempo.

- Finalmente, comparando los resultados de los modelos propuestos, el modelo de RED2 es más exacto que el modelo RED1 debido a que es más preciso en la predicción tanto en los conjuntos de entrenamiento como en los de prueba.

Además, como el modelo de RED2 no está limitado a las capacidades técnicas de otra herramienta bio-informática puede predecir, con su actual precisión, secuencias de mayor tamaño a las entrenadas.

Por otra parte, el modelo de RED2 ha mostrado su fortaleza en la etapa de entrenamiento, con unos niveles de especificidad y sensibilidad superiores al 0.99.

5.6. COMPARACIÓN CON ALGUNOS MODELOS EXISTENTES

Se debe tener en cuenta que la comparación de los modelos propuestos con los modelos que existen actualmente no será muy rigurosa porque el conjunto de secuencias de entrenamiento no son los mismos entre los estudios realizados. Considerando lo anterior, a partir de los resultados que se han documentado de los modelos presentados en [20] y [21], y comparándolos con los resultados obtenidos en los modelos propuestos se observa lo siguiente:

- En el trabajo desarrollado en [20] se utilizaron secuencias promotoras de la bacteria E. Coli con 60 nucleótidos de tamaño y se desarrollaron dos modelos de RNAs con diferente estrategia de codificación de nucleótidos para la entrada de los datos. La primera utiliza codificación binaria de cuatro dimensiones configurando su topología con 240 nodos en la capa de entrada, 80 nodos en la capa oculta y un nodo en la capa de salida. La segunda utiliza codificación binaria de dos dimensiones configurando su topología con 120 nodos en la capa de entrada, 80 nodos en la capa oculta y un nodo en la capa de salida.

Los resultados documentados fueron que la tasa de aprendizaje para el sistema de codificación de cuatro dimensiones fue de 0.5, mientras que para el segundo esquema fue de 0.25, es decir, el sistema de entrenamiento de cuatro dimensiones es superior al de dos dimensiones.

Se puede apreciar que la propuesta del modelo de RED2 es más exacta que el modelo desarrollado en [20], porque alcanza una tasa de aprendizaje superior al 0.99, es decir, se obtuvo unos valores de sensibilidad y especificidad para el conjunto de entrenamiento superiores a 0.99. Sin embargo, la propuesta del modelo de RED1 no aventaja mucho a la del modelo desarrollado en [20] porque obtuvo valores de sensibilidad entre 0.4 y 0.9.

Por otra parte, se sabe que el genoma de la bacteria E. Coli es menos complejo que el genoma de la rata y del ratón porque posee en proporción más código codificante que no codificante. Lo anterior hace que el modelo propuesto en [20] no sea tan efectivo por ser menos preciso que el modelo de RED2 a pesar de que se haya entrenado para especies de menor complejidad.

- En el trabajo desarrollado en [21] se utilizaron secuencias promotoras del hongo *Crinipellis* y se desarrolló un modelo de RNA que emplea como estrategia de entrada valores de medidas características. La topología de la RNA consistió de 8 nodos en la capa de entrada, 2 nodos en la capa intermedia y un nodo para la capa de salida.

Los resultados documentados fueron que el modelo propuesto en [21] alcanzó una especificidad del 99% y una sensibilidad del 95% en el conjunto de pruebas.

Se puede apreciar que las propuestas del modelo de RED1 y RED2 son menos exactas que el modelo desarrollado en [21] porque presentan valores de sensibilidad de 0.4 y 0.67 y valores de especificidad de 0.98 y 0.75 respectivamente para el conjunto de pruebas. Sin embargo, se debe tener en cuenta que el trabajo desarrollado en [21] en realidad sólo está reconociendo el patrón de la caja TATA como tal y no la región promotora del gen, por eso concluye, acertadamente, que el patrón de la caja TATA también se presenta a distancias que varían desde 90 hasta 300 bases desde el codón de inicio.

Además, se sabe que el genoma del hongo *Crinipellis* es de menor complejidad que el genoma de la rata y del ratón porque posee en proporción más código codificante que no codificante. Lo anterior hace que el modelo propuesto en [21] tenga mayor posibilidad de alcanzar mayor precisión que el modelo de RED2 porque se ha entrenado para una especie de menor complejidad.

Capítulo 6

CONCLUSIONES

A continuación se presentan algunas conclusiones y recomendaciones derivadas del desarrollo del proyecto.

- Los modelos propuestos han mostrado que las redes neuronales artificiales (RNA) resultan apropiadas para el problema de la predicción de promotores. Los indicadores de sensibilidad y especificidad para las secuencias de validación de la RED1 fueron de 0.6 y 0.96 respectivamente; y para la RED2 fueron de 0.74 y 0.82 respectivamente.
- La estrategia de utilizar la codificación binaria ortogonal Code-4 en el proceso de entrenamiento de las RNAs con secuencias de nucleótidos resultó ser muy efectiva. Esto se pudo notar porque con pocas iteraciones de entrenamiento las RNAs lograron reconocer la gran mayoría de las secuencias.
- Los modelos propuestos han mostrado que las RNAs, aplicadas en el contexto de la predicción de promotores, no son sensibles a la naturaleza de los datos porque pueden aprender a reconocer especies genéticamente parecidas con sólo entrenar las nuevas secuencias sin necesidad de cambiar la topología de la RNA.

De acuerdo a los resultados obtenidos en las pruebas de entrenamiento de la RED1 en los diferentes conjuntos de secuencias de rata y ratón se alcanzan valores de sensibilidad 1.0 y 0.7 respectivamente y valores de especificidad de 0.98 y 1.0 respectivamente. Así mismo, los resultados obtenidos en las pruebas de entrenamiento de la RED2 en los diferentes conjuntos de secuencias de rata y ratón con valores de sensibilidad 1.0 y 1.0 respectivamente y valores de especificidad de 1.0 y 1.0 respectivamente.

- La estrategia empleada en la RED2 que consistió en utilizar codificación binaria para representar el acierto de las tres cajas consenso, que en su gran mayoría presentan los organismos vertebrados, resultó ser muy efectiva en el proceso de entrenamiento y prueba. Con esta red se obtienen valores superiores a 0.99 en los indicadores de sensibilidad y especificidad en el proceso

de entrenamiento y valores de 0.74 y 0.82 en los indicadores de sensibilidad y especificidad respectivamente en el proceso de prueba.

- Comparando los resultados de los modelos propuestos, el modelo de RED2, que utiliza la estrategia del trazo con las marcas de las cajas consenso, es más exacto que el modelo RED1, que utiliza la estrategia de integración de los algoritmos basados en modelos ocultos de Markov y el algoritmo banana. Esto ocurre porque la RED2 se desarrolló específicamente para la predicción de promotores de vertebrados y no como la RED1 que está limitada a las características de dichos algoritmos que no están diseñados específicamente para la predicción de promotores.
- A partir del análisis de los resultados (sección 5.5 Discusión de resultados) es muy posible que se encuentren patrones parecidos a las cajas consenso ya reconocidas en regiones cercanas al sitio de inicio de transcripción en las regiones de intrones, ya que se pudo observar que los modelos aumentan su tasa de falsos positivos cuando el conjunto de entrenamiento de no promotores tiene en su mayoría intrones. Esto se da porque la RED2 encuentra similitudes en la configuración de las señales de promotores con la configuración de algunos intrones.

6.1. TRABAJO FUTURO

El presente trabajo se desarrolló utilizando las secuencias de la base de datos SIB-EPD[7] cuyas secuencias promotoras se caracterizan porque son reconocidas por la enzima ARN Polimerasa II y son secuencias experimentales o computacionales suficientemente similares a las experimentales⁶. Para validar y generalizar aún más los modelos propuestos se podría continuar el proceso de investigación al entrenarlos y validarlos con secuencias promotoras reconocidas por los diferentes tipos de enzima ARN polimerasa y con secuencias totalmente experimentales.

Como este trabajo se dirigió en la predicción de secuencias promotoras de las especies rata y ratón, se puede probar con

⁶ Referencia: SIB-EPD. User Manual: 2. Promoter selection:
http://www.epd.isb-sib.ch/current/usrman.html#PROMOTER_SELECTION

otras especies eucariotas genéticamente más complejas como el ser humano y menos complejas como la mosca *Drosophila* para validar la adaptabilidad del modelo al tipo de organismo.

Asimismo, resulta interesante agregarle al modelo de RED2 que no sólo reconozca y realice el trazo de las cajas TATA, CAAT y GC sino también el codón de inicio y otras señales que en promedio aparezcan cerca del codón de inicio.

REFERENCIAS

- [1]. Qicheng Ma, Jason T. L. Wang, Dennis Shasa y Cathy H. Wu; *DNA SEQUENCE CLASSIFICATION VIA EXPECTATION MAXIMIZATION ALGORITHM AND NEURAL A CASE STUDY*, Septiembre 2009. <http://web.njit.edu/~wangj/publications/ARTICLES/smc2001.pdf>. Consulta: 4/10/2009.
- [2]. Romesh Ranawana y Vasile Palade, *ADVANCES IN EVOLUTIONARY COMPUTING FOR SYSTEM DESIGN*, capítulo 4; Editorial Springer, 2007.
- [3]. Kalate, R.N., Tambe, S.S. and Kulkarni, B.D. (2003) ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF MYCOBACTERIAL PROMOTER SEQUENCES. *Comput. Biol. Chem.*, 27, 555-564.
- [4]. M. Reese and F. Eeckman, *NOVEL NEURAL NETWORK PREDICTION SYSTEMS FOR HUMAN PROMOTERS AND SPLICE SITES*, Gene-Finding and Gene Structure Prediction Workshop, 1995.
- [5]. Sun Kim, *PROBABILISTIC MODELS FOR BIOLOGICAL SEQUENCE ANALYSIS*, Indiana University, Bloomington, Seminar in Bioinformatics, spring 2002.
- [6]. Manuel Calaza, *BÚSQUEDA AUTOMÁTICA DE GENES*, 2005, <http://mathgene.usc.es/cursoverano/cv2005/materiales/genes/autogenes.pdf>. Consulta: 4/10/2009.
- [7]. *THE EUKARYOTIC PROMOTER DATABASE* (E.P.D.) Release 100, June 2009. <http://www.epd.isb-sib.ch/>. Consulta: 21/10/2009.
- [7b]. EPD Promoter Elements. *THE EUKARYOTIC PROMOTER DATABASE*. Release 100, June 2009. http://www.epd.isb-sib.ch/promoter_elements/. Consulta: 17/07/2010.
- [8]. Elizabeth Gutierrez Kafati. INNOVACIÓN BAJO INCERTIDUMBRE, artículo "Redes neuronales ventajas y desventajas". http://egkafati.bligoo.com/content/view/184582/Redes_neuronales_Ventajas_y_Desventajas.html, Consulta: 13 de noviembre del 2009.
- [9]. Avendaño Luis Hernando, *NEFROLOGÍA CLÍNICA*, Segunda edición. Enero 2003. http://ens.xochicalco.edu.mx/medicina/archivos/ebooks/Nefrologia%20y%20Urologia/nefrologia%20clinica/nefro/ch10.htm#f_3.2.3.1, Consulta: 09/11/2009.

[10]. Wikimedia Foundation, WIKIPEDIA THE FREE ENCYCLOPEDIA, references Hidden Markov model, http://en.wikipedia.org/wiki/Hidden_Markov_model, Consulta: 10/11/2009.

[11]. Ben Krose and Patrick van der Smagt. AN INTRODUCTION TO NEURAL-NETWORKS. University of Amsterdam, January 1996.

[11a]. Acosta María Isabel, Salazar Harold y Zuluaga Camilo, TUTORIAL DE REDES NEURONALES, Universidad Tecnológica de Pereira, <http://ohm.utp.edu.co:16080/neuronales/main.htm>, Consulta: 23/10/2009.

[11b]. Isai Orlando Morales González. COMPARACIÓN TEÓRICO PRÁCTICA ENTRE MODELOS ESTADÍSTICOS Y EL PERCEPTRÓN MULTICAPA. Pontificia Universidad Católica De Valparaíso. Valparaíso, 2010.

[12]. Lawrence O. Hall, Xiaomei Liu, Kevin W. Bowyer, Robert Banfield. "WHY ARE NEURAL NETWORKS SOMETIMES MUCH MORE ACCURATE THAN DECISION TREES: AN ANALYSIS ON A BIO-INFORMATICS PROBLEM". In the Proc. of the IEEE International Conference on Systems, Man & Cybernetics, 2003.

[14]. National Human Genome Research Institute (NHGRI), GENOME.GOV, <http://www.genome.gov/sglossary.cfm>, Consulta: 13/11/2009.

[15]. Microsoft® Encarta® 2009 Student Edition. ©1993-2008 Microsoft Corporation, Multimedia DVD-ROM, 2009.

[16]. Ciencia y Biología. Blog de la ciencia y biología, www.cienciaybiologia.com/bggeneral/transcripcion-arn.htm, Consulta: 24/01/2010.

[17]. Proyecto Biósfera, Ministerio de Educación del gobierno de España, <http://recursos.cnice.mec.es/biosfera/alumno/2bachillerato/biomol/contenidos19.htm>, Consulta: 24/01/2010.

[18]. Smale ST, Kadonaga JT., THE RNA POLYMERASE II CORE PROMOTER. Howard Hughes Medical Institute and Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, California, Annu Rev Biochem. 72, 449-479. 2003.

[19]. Dr. Vásquez-Contreras Edgar, BIOQUÍMICA Y BIOLOGÍA MOLECULAR EN LINEA, Instituto de Química, UNAM,

<http://laguna.fmedic.unam.mx/~evazquez/0403/rna%20polimerasa.html>, Consulta: 26/01/2010.

[20]. Fan Zhang, Michael D. Kuo, Adrian Brunkhorns, E. COLI PROMOTER PREDICTION USING FEED-FORWARD NEURAL NETWORKS, Department of Radiology and Center for Translational Medical Systems, University of California, San Diego, 2006.

[21]. Diego Frias, Ftamon Vidd and Julio C.M. Cascardo, FINDING GENE PROMOTERS IN THE GENOME OF THE FUNGUS CRINIPELLIS PERNICIOSA USING FEED-FORWARD NEURAL NETWORKS, Bioinformatic Lab, Universida de Estadual de Santa Cruz, Bahia, 2004.

[22]. Hong-Mei Zhu, Jia-Xin Wang, PREDICTING EUKARYOTIC PROMOTER USING BOTH INTERPOLATED MARKOV CHAINS AND TIME-DELAY NEURAL NETWORKS, Department of Computer Science and Technology, Tsinghua University, Beijing, China, 2006.

[23]. Qicheng Ma, Jason T. L. Wang, Dennis Shasha, and Cathy H. Wu, DNA SEQUENCE CLASSIFICATION VIA AN EXPECTATION MAXIMIZATION ALGORITHM AND NEURAL NETWORKS: A CASE STUDY, IEEE transactions on systems, Man, and cybernetics, Applications and reviews, Vol. 31, No. 4, November 2001.

[24]. Martin G. Reese, APPLICATION OF A TIME-DELAY NEURAL NETWORK TO PROMOTER ANNOTATION IN THE DROSOPHILA MELANOGASTER GENOME, Berkeley Drosophila Genome Project, Department of Molecular and Cell Biology, University of California, Berkeley, may 2001.

[25]. Ujvari A and Martin CT (2000). EVIDENCE FOR DNA BENDING AT THE T7 RNA POLYMERASE PROMOTER. J Mol Biol, 295, 1173-1184.

[26]. TenHarmsel, A. and Biggin, M. D. (1995). BENDING DNA CAN REPRESS A EUKARYOTIC BASAL PROMOTER AND INHIBIT TFIID BINDING. Mol. Cell. Biol. 15, 5492-5498.

[27]. The BioInformatics Organizations. GeneZilla 2003. <http://www.genezilla.org/>. Consulta: 17/07/2010.

[28]. EMBOSS: The European Molecular Biology Open Software Suite (2000). Rice, P. Longden, I. and Bleasby, A. Trends. <http://emboss.sourceforge.net/apps/release/5.0/emboss/apps/banana.html>. Consulta: 17/07/2010.

[29]. David S. Goodsell and Richard E. Dickerson. BENDING AND CURVATURE CALCULATIONS IN B-DNA. Molecular Biology Institute, Department of Chemistry and Biochemistry, and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, 1994.

[30]. Bioinformatics Lab. TISSUE - SPECIFIC GENE EXPRESSION AND REGULATION (TiGER). Wilmer Institute, Johns Hopkins University. <http://bioinfo.wilmer.jhu.edu/tiger/>. Consulta: 17/07/2010.

[31]. RNA Research Group. GROUP I INTRON SEQUENCE AND STRUCTURE DATABASE (GISSD). Wuhan University, China. <http://www.rna.whu.edu.cn/gissd/>. Consulta: 17/07/2010.