están formadas por unidades simples conectadas entre sí llamadas nucleótidos (A (Adenina), C (Citosina), G (Guanina) y T (Timina)).

De esta manera, uno de los campos más importantes en la Bioinformática está relacionado con la identificación de motivos o regiones conservadas en las secuencias en las proteínas. Este cubre un conjunto de objetivos, tales como la búsqueda de genes específicos en los genomas, identificar regiones en las proteínas, predecir dominios proteicos, detectar los lugares de conexión [4]. Los problemas que emergen en este sector son interesantes porque requieren el uso de métodos matemáticos e informáticos, con tiempos de ejecución de gran complejidad algorítmica.

El ADN está constituido por cuatro bases: adenina (A), timina (T), citosina (C) y guanina (G), lo que significa que existen 16 posibles combinaciones de dinucleótidos (AA, AT, AC, AG, TT, TA, TC, TG, CC, CA, CT, CG, GG, GA, GT, y GC). Los dinucleótidos en ocasiones se refieren como NpN, donde N es la base y p representa el enlace entre las dos bases.

CpG, significa Citosina fosfo Guanina. Esto señala que la posibilidad de encontrar algún dinucleótido en una secuencia dada de ADN será 1/16, o sea aproximadamente 6%. En los seres humanos, sin embargo, la frecuencia de CpG es relativamente baja, un fenómeno denominado supresión CG. Esto es evidente a través de todo el genoma humano, excepto en pequeñas áreas, donde la frecuencia de CpG tiene valores esperados o mucho más altos. Estas áreas se llaman Islas de CpG y se acepta que representan alrededor del 1% del genoma (aproximadamente unas 45000 islas en el genoma humano). La razón de por qué estas islas CpG escapan al fenómeno de supresión CG en la evolución es que típicamente no están metiladas[1] y, por lo tanto, escapan a las presiones mutacionales.

Así, el proceso de metilación es suprimido en los alrededores de los genes, motivo por el cual estas áreas tienen una relativa alta concentración de CpG (superior al 65% en humanos, frente al promedio genómico que se sitúa en torno al 40 %). Estas regiones son llamadas islas CpG y su longitud puede variar de algunos cientos a unos pocos miles de bases [5], [6]. Debido a este fenómeno, la presencia de las islas CpG en una secuencia de ADN puede ser un indicio del comienzo de un gen, situación que ayuda a determinar la ubicación de los genes a lo largo del ADN.

Existen algunos trabajos en esta área. En [5] se ha desarrollado un algoritmo para el reconocimiento de regiones promotoras en el genoma humano. Éste extrae características de la composición e información de las islas CpG desde secuencias genómicas. Utiliza una red neuronal híbrida y aplica la predicción En [6] se desarrolló una máquina de soporte vectorial (SVM) para predecir las islas CpG. En nuestro caso, se utiliza una red neuronal para reconocer las islas CpG y no para predecir en qué lugar puede aparecer.

En este trabajo proponemos desarrollar una red neuronal que permita reconocer si un fragmento de secuencia de ADN es o forma parte de una isla CpG. Para ello, se van a realizar dos tareas.

- Encontrar el conjunto de dinucleótido CpG presente en las secuencias de ADN. Para ello, se necesita crear un modelo en el que la probabilidad de que aparezca un nucleótido en una posición de la secuencia dependa del nucleótido que se encuentra en la posición anterior.
- Reconocer los patrones en la secuencia de ADN, si pertenece a una isla CpG o a un océano.

Las secuencias de ADN contiene un alfabeto formado por 4 nucleótidos $\alpha = \{A, C, T, G\}$. Un punto muy importante en la aplicación de redes neuronales para el reconocimiento de islas CpG, es la manera de codificar las secuencias de ADN como entrada de la red neuronal, ya que dichas secuencias no es la mejor representación. Por lo tanto, la forma de representar la entrada es crucial para el éxito del aprendizaje de la red neuronal.

II. Marco Teórico

A. *Redes Neuronales Artificiales*

Las Redes Neuronales Artificiales (RNA) emulan las redes neuronales biológicas. Pueden ser consideradas como un sistema de procesamiento de información que tiene las siguientes características [6], [7]:

1. Adaptabilidad: es la capacidad para aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.
2. Auto-organización: una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.
3. Tolerancia a Fallas: es la propiedad que permite a un sistema continuar operando adecuadamente en caso de una falla en alguno de sus componentes.
4. Robustez: un sistema es robusto si puede ejecutar diversos procesos de manera simultánea sin generar fallos o bloquearse

Cada neurona artificial se caracteriza por los siguientes elementos (ver Fig. 1) [8]:

- Un valor o estado de activación inicial (a_{t-1}), anterior a la recepción de los estímulos
- Unos estímulos o entradas a la neurona (x_i), con unos pesos asociados (w_{ij}).
- Una función de propagación, que determina la entrada total a la neurona (Net_j).
- Una función de activación o transferencia (f), que combina las entradas a la neurona con el estado de activación inicial para producir un nuevo valor de activación.
- Una función de salida (F), que transforma el estado inicial de activación en la señal de salida

[1] Modifica la función del ADN y generalmente actúa para reprimir la transcripción genética.

- Una señal de salida que se transmite, a otras neuronas artificiales o es la respuesta del sistema (y_i).
- Una regla de aprendizaje, que determina la forma de actualización de los pesos de la red (aprendizaje).

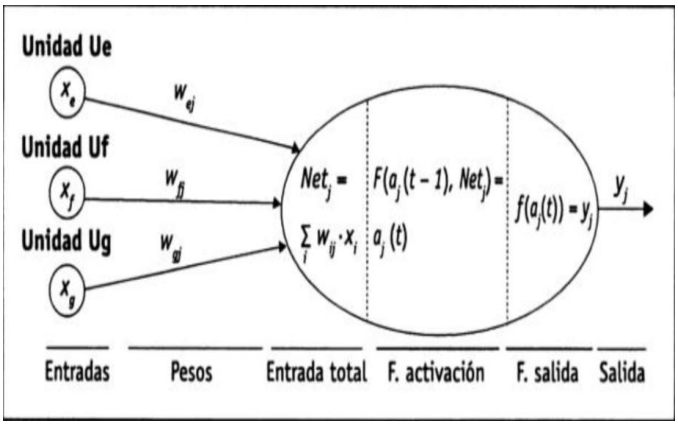

Fig. 1. Esquema de una neurona artificial

B. Red Neuronal de Retropropagación

En la Red Neuronal de Retropropagación las salidas de las neuronas en una capa pueden estar interconectadas a las entradas de las neuronas de la misma capa o a entrada de neuronas en capas precedentes (Fig. 2). Este hecho le proporciona al arreglo neuronal características de procesamiento dinámico, así las salidas de la red dependen no solo de sus entradas en un instante dado, sino también de sus entradas y salidas en instantes anteriores [6], [7].

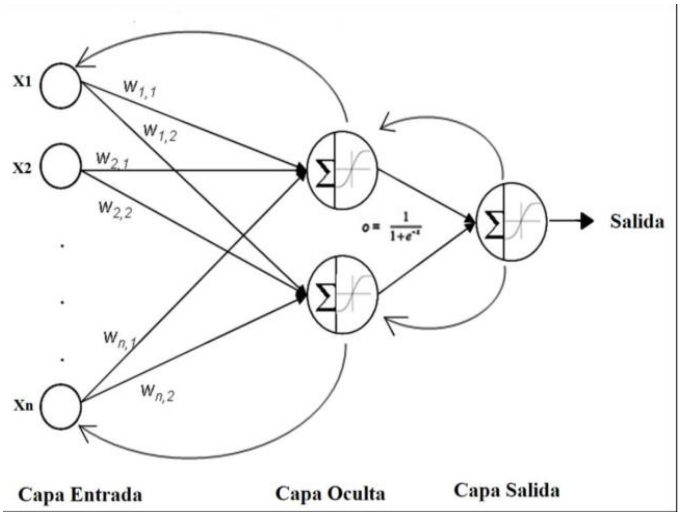

Fig. 2. Diagrama de una Red neuronal de Retropropagación [7]

Este tipo de red neuronal utiliza el aprendizaje supervisado. El algoritmo consiste en el aprendizaje de un número pre-definido de patrones de entrada-salida, empleando un ciclo "propagación-adaptación" con dos fases diferenciadas (Fig. 3).

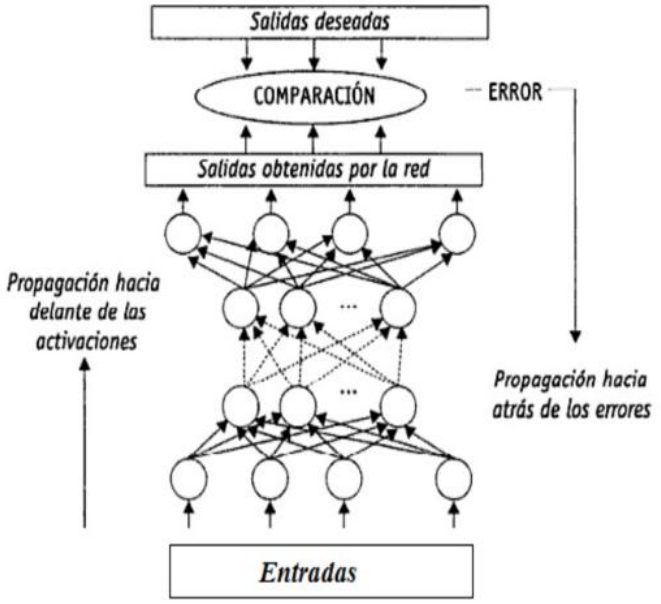

Fig. 3. Funcionamiento de una Red neuronal de Retropropagación [7]

Para realizar este proceso se debe inicialmente tener definida la topología de la red: número de neuronas de la capa de entrada (depende del número de componentes del vector de entrada), cantidad de capas ocultas y número de neuronas de cada una de ellas, número de neuronas en la capa de salida [6], [7]. Las dos fases son descritas a continuación:

- Fase de aprendizaje "hacia adelante": se aplica un patrón de entrada como estímulo para la primera capa de neuronas de la red, se va propagando a través de todas las capas superiores hasta generar una salida, se compara el resultado en las neuronas de salida con la salida que se desea obtener y se calcula un valor de error para cada neurona de salida.
- Fase de aprendizaje "hacia atrás": los errores obtenidos en la fase anterior se transmiten hacia atrás, partiendo de la capa de salida hacia todas las neuronas de la capa intermedia que contribuyan directamente a la salida. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido un error. Luego se procede al reajuste de los pesos de las neuronas. Este proceso se repite por un número de iteraciones, o hasta que el error sea el deseado por el usuario.

El algoritmo que se emplea para entrenar una Red Neuronal de Retropropagación es la *Regla Delta Generalizada*. Este algoritmo utiliza una función de error asociada a la red, buscando el mínimo error a través del gradiente descendiente. Los pasos del algoritmo son [7]:

1. Paso 1: Inicializar los pesos de la red con valores aleatorios.
2. Paso 2: Presentar un patrón de entrada $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ con n componentes, a la capa de entrada de la red y especificar la salida deseada (d_p) que debe generar ésta.
3. Paso 3: Calcular el valor de entrada a cada una de las neuronas en las capas ocultas (1)

$$Net_{pj}^{h} = \sum_{i=1}^{n} w_{ji}^{h} * x_{pi} + \theta_{j}^{h} \quad (1)$$

Donde w_{ji}^{h} es el peso en la conexión de la unidad de entrada i a la unidad de la capa oculta j, θ_{j}^{h}es el valor del bias, y el índice h representa a la capa oculta.

4. Paso 4: Calcular la salida de la capa oculta (2)

$$y_{pj} = f_{j}^{h}(Net_{pj}^{h}) \quad (2)$$

5. Paso 5: En la capa de salida, calcular el valor de entrada a cada neurona (3)

$$Net_{pk}^{O} = \sum_{j=1}^{L} W_{kj}^{O} * I_{pj} + \theta_{k}^{O} \quad (3)$$

Donde L es el número de neuronas de la capa oculta, y el índice o representa a la capa de salida

6. Paso 6: Calcular la salida (4)

$$y_{pk} = f_{k}^{O}(Net_{pk}^{O}) \quad (4)$$

7. Paso 7: Calcular el error para las neuronas de la capa de salida (5)

$$\delta_{pk}^{O} = (y_{pk} - d_{pk}) * f_{k}^{O'}(Net_{pk}^{O}) \quad (5)$$

La función f debe ser derivable. Las funciones de salida más utilizadas son (6) y (8):

La función lineal: $f_{k}^{O}(Net_{jk}^{O}) = Net_{jk}^{O}$ donde $f_{k}^{O}(Net_{jk}^{O}) = 1$ (6)

El error para las neuronas de salida para la función lineal es (7):

$$\delta_{pk}^{O} = (y_{pk} - d_{pk}) * f_{k}^{O'}(Net_{pk}^{O}) = (d_{pk} - y_{pk}) \quad (7)$$

La función sigmoidea: $\frac{1}{(1+e^{-Net_{jk}^{O}})}$ donde $f_{k}^{O'}(Net_{pk}^{O}) = f_{k}^{O}(1 - Net_{pk}^{O}) = y_{pk}(1 - y_{pk})$ (8)

El error para las neuronas de salida para la función sigmoidea es (9):

$$\delta_{pk}^{O} = (y_{pk} - d_{pk}) * f_{k}^{O'}(Net_{pk}^{O}) = (d_{pk} - y_{pk}) y_{pk}(1 - y_{pk}) \quad (9)$$

8. Paso 8: Calcular el error en las neuronas de la capa oculta (10)

$$\delta_{pk}^{O} = f_{j}^{h'}(Net_{pj}^{h}) \sum \delta_{pk}^{O} \quad (10)$$

El error en la capa oculta depende de todos los términos del error de la capa de salida. De aquí surge el término retropropagación o propagación hacia atrás.

9. Paso 9: Actualizar los pesos de la capa de salida (11)

$$w_{kj}^{O}(t+1) = w_{kj}^{O}(t) + \propto \delta_{pk}^{O} * y_{pj} \quad (11)$$

Dónde α representa la tasa de aprendizaje, el cual es un valor en el intervalo [0, 1] que permite aumentar la velocidad de convergencia del error. A mayor tasa de aprendizaje, mayor es la modificación de los pesos en cada iteración, pero puede dar lugar a oscilaciones en este valor. Por lo tanto, se puede agregar un término que tiende a mantener los cambios de los pesos en una misma dirección, llamado momento β (12)

$$w_{kj}^{O}(t+1) = w_{kj}^{O}(t) + \propto \delta_{pk}^{O} * y_{pj} + \beta(w_{kj}^{O}(t) - w_{kj}^{O}(t-1)) \quad (12)$$

Donde β es un valor en el intervalo [0, 1].

10. Paso 10: Actualizar los pesos en la capa oculta (13)

$$\boldsymbol{w_{ji}^{h}(t+1) = w_{ji}^{h}(t) + \propto \delta_{pj}^{h} * x_{p}} \quad (13)$$

Utilizando el término momento (14)

$$w_{ji}^{h}(t+1) = w_{ij}^{h}(t) + \propto \delta_{pj}^{h} * x_{p} + \beta(w_{ji}^{h}(t) - w_{ji}^{h}(t-1)) \quad (14)$$

11. Paso 11: El proceso se repite hasta que el error para el patrón de entrada sea mínimo o se cumpla un número máximo de iteraciones (15)

$$E_{p} = \frac{1}{2} \sum_{k=1}^{M} \delta_{pk}^{2} \quad (15)$$

C. Acido Desoxirribonucleico (ADN)

Los Ácidos Nucleicos son las biomoléculas portadoras de la información genética. Son biopolímeros, de elevado peso molecular, formados por otras subunidades estructurales o monómeros, denominados Nucleótidos.

Un nucleótido es un compuesto formado por la combinación de ácido fosfórico con azúcar y una base nitrogenada. La molécula de ADN contiene cuatro tipos distintos de nucleótidos (Fig. 4).

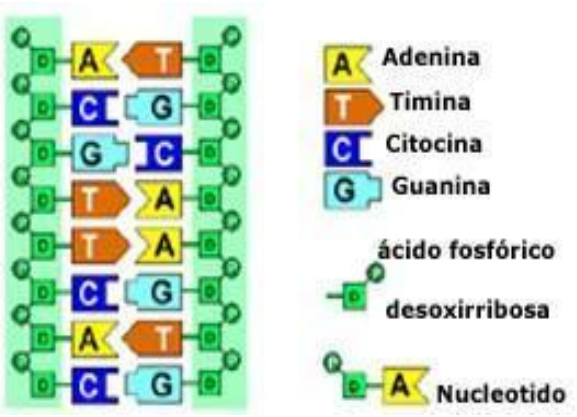

Fig. 4. Modelo de una Molecula de ADN [9]

Cada uno de ellos se compone de un grupo fosfato asociado al azúcar desoxirribosa el cual se une a una de cuatro bases. En consecuencia, los nucleótidos se diferencian por la base nitrogenada adenina (A) o guanina (G) (bases purínicas) o bien citosina (C) o timina (T) (bases pirimidínicas). El termino acido desoxirribonucleico significa que el compuesto contiene desoxirribosa, aparece en el núcleo y se trata de un ácido [9], [10].

El ADN tiene la información genética de los organismos y es el responsable de su transmisión hereditaria. El conocimiento de la estructura de los ácidos nucleicos permitió la elucidación del código genético, la determinación del mecanismo y control de la síntesis de las proteínas y el mecanismo de transmisión de la información genética de la célula madre a las células hijas [10].

D. Islas CpG

Las islas CpG se definen formalmente como regiones de al menos 200 pares de bases de longitud, con un contenido mínimo en G+C del 50% y una frecuencia de CpGs (CpG O/E: relación entre CpGs observados y esperados) de al menos 0,6 [11].

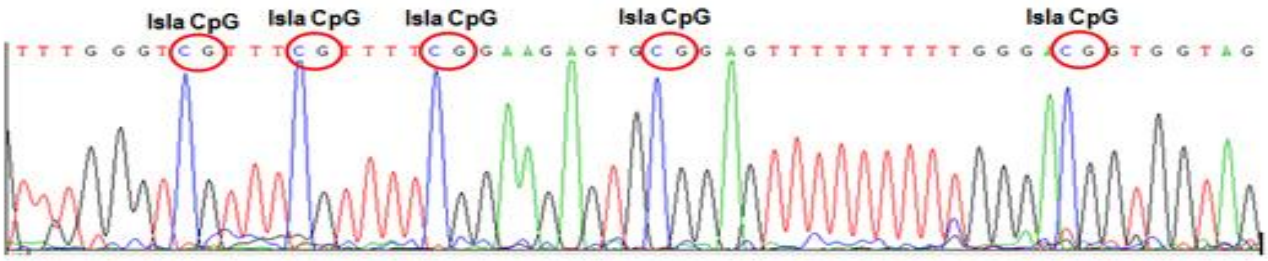

Fig. 5. Islas de CpG en una secuencia de ADN [11]

Las islas CpG constituyen la fracción no metilada [2] del genoma que generalmente se acumulan en los genes. En el resto del genoma C en CpG tiende a ser Cmetilado por un proceso llamado metilación y el Cmetilado tiene una alta tendencia a mutar a T. Entonces debido al proceso de metilación el dinucleótido CpG, que se denota de esta forma porque esta enlazado por un fosfato, es más raro que lo que se esperaría por las probabilidades independientes de C y G. El proceso de metilación es suprimido en los alrededores de los genes, motivo por el cual estas áreas tienen una relativa alta concentración de CpG [12], [13].

La Fig. 5 muestra las islas de CpG en una secuencia de ADN.

III. Problema

El ADN está conformado por largas secuencias de nucleótidos que cual como son obtenidas se almacenan pero no suelen proporcionar conocimiento en forma directa, su valor real reside en la información que podamos extraer de estás: información que nos ayude a tomar decisiones o a mejorar nuestra comprensión de lo que ocurre en las células.

Estos procesos de recolección o generación de información a partir de las secuencias de ADN producen volúmenes tales que superan las capacidades humanas para analizarlas. Esta limitación se debe a varios factores, entre los cuales tenemos la disponibilidad en tiempo, la incapacidad de relacionar grandes volúmenes de datos, entre otros. Pero para analizar los datos computacionalmente requieren de gran cantidad de procesamiento y la utilización de técnicas de computación inteligente, emergente o evolutiva además de la minería de datos, que nos permitan obtener información útil.

En este trabajo lo importante en las secuencias de ADN son los dinucleótidos, pues se desea encontrar el dinucleótido CpG presentes en estas secuencias. Para ello, se necesita crear un modelo en el que la probabilidad de que aparezca un nucleótido en una posición de la secuencia dependa del nucleótido que se encuentra en la posición anterior. De esta forma se tendrá la capacidad de obtener dos nucleótidos seguidos en una secuencia de ADN. La presencia de las islas CpG en una secuencia de ADN puede ser un indicio del comienzo de un gen, situación que ayuda a determinar la ubicación de los genes a lo largo del ADN.

Por lo tanto, es necesario definir y desarrollar un método matemático/informático utilizando una red neuronal para la identificación de islas CpG en secuencias de ADN.

IV. Diseño del Sistema

La red neuronal tendrá una primera capa de neuronas que serán las neuronas de entrada de la red, seguida de una capa oculta de neuronas y luego la capa de salida.

- Capa de entrada: dado que las islas CpG presentan la peculiaridad de tener más presencia de C (Citosina) y G (Guanina) que de A (Adenina) y T (Timina), entonces la probabilidad de hallar una G después de un nucleótido será mayor en una isla que en un océano, si en la posición actual hay una C. Entonces se calculará las probabilidades a partir de un conjunto de secuencias de ADN que servirán para entrenar la red. Las probabilidades dependerán de la frecuencia de cada par posible para cada nucleótido. Se calculará la frecuencia de cada par posible y se dividirá entre las suma total de dinucleótidos en la secuencia, de este manera se obtiene la tabla I.

  $p(i|j)$ = probabilidad de hallar una j en la siguiente posición si en la posición actual hay una i

Tabla I. Matriz d Probabilidades p(i|j)

	C	**A**	**G**	**T**
C	P(C|C)	P(A|C)	P(G|C)	P(T|C)
A	P(C|A)	P(A|A)	P(G|A)	P(T|A)
G	P(C|G)	P(A|G)	P(G|G)	P(T|G)
T	P(C|T)	P(A|T)	P(G|T)	P(T|T)

Los segmentos de secuencias de ADN pertenecientes a islas CpG las denotaremos como grupo "+" y los segmentos pertenecientes a océanos como grupo "-".

$C^*_{i,j}$ representa el número de veces que el nucleótido j sigue a i en una secuencia, siendo * ∈ {+, -} e i, j {A, C, T, G} entonces:

$$a^+_{i,j} = \frac{C^+_{i,j}}{\sum C^+_{i,j}} \qquad (16)$$

$$a^-_{i,j} = \frac{C^-_{i,j}}{\sum C^-_{i,j}} \qquad (17)$$

[2] Es la adición de un grupo metilo (-CH3) a una molécula

Dada la matriz de probabilidades mostrada en la Tabla I compuesta por (16) o (17) (dependiendo si la secuencia pertenece a una isla o un océano) entonces se define como valores de entrada de la red neuronal a cada uno de los valores que componen la matriz, siendo entonces el tamaño de esta capa de 16 neuronas, producto de la cantidad de permutaciones que se pueden obtener de los cuatro nucleótidos en dinucleótidos.

- Capa Oculta: se utilizaron 8 neuronas, ya que con este valor se obtuvo mejor reconocimiento de las islas CpG en los entrenamientos y las pruebas. Es importante mencionar que no existe una técnica para determinar el número de capas ocultas, ni el número de neuronas por capa que debe contener una red en un problema específico, esto dependerá de las experiencias y los resultados obtenidos en las pruebas realizadas.
- Capa de Salida: la capa de salida estará conformada por 2 neuronas, ambas neuronas pueden tomar valores en un rango de -1 a 1. Si la entrada suministrada a la red pertenece a una isla CpG la salida de la primera neurona deberá tender a 1 y la salida de la segunda neurona deberá tender a -1. Por el contrario si la entrada suministrada pertenece a un océano la salida de la primera neurona deberá tender a -1 y la salida de la segunda neurona deberá tender a 1. Si la entrada no pertenece a los casos anteriores la salida de la primera neurona tenderá a -1 y la segunda neurona tenderá a -1

V. Pruebas

El sistema se diseñó en lenguaje C, se utilizó la biblioteca fann [14] que contiene las funciones y métodos de la red neuronal de retrorpopagación. Se utilizaron datos para el entrenamiento y pruebas extraídos de [15] de genomas de vertebrados [13]. Para el entrenamiento se utilizaron 15 fragmentos de secuencias de ADN pertenecientes a islas CpG y 18 no pertenecientes a islas y para las pruebas se utilizaron 15 y 17 fragmentos respectivamente, cada fragmento contiene 14 valores de probabilidad dado según la tabla I, y valores para conocer si pertenece o no a una isla CpG.

Un Ejemplo de una instancia del problema se muestra a continuación:

Se tiene un fragmento de una secuencia de ADN:
TTAATGCCTGAGACTGTGTGAAGTAAGAGATGGATCA
GAGGCCGGGCGCGGGGGCTCGCGCCTGTCATCCCAGC
ACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAG
GAGATCGAGACCATCCTGGCTAACACGGGGAAACCCC
GTCTCCACTAAAAATACAAAAAGTTAGCCGGGCGCGG
TGGCGGGCGCCTGCGGTCCCAGCTGCTGGGGAGGCCG
AGGCGGGAGCATGGCGGGAACCGGGAGGCGGAGCCT
GCAGTGAGCCGAGATGGCGCCACCGCACTCCAGCCTG
GGCGACCCAGCGAGACTCCGCCTCAAAAAAAAAAAAA
GAAGA

Se procesa para calcular las probabilidades donde se obtienen los siguientes valores que representan la entrada a la red neuronal:

0.0833333 0.0416667 0.0863095 0.0297619 0.0535714 0.0833333 0.0863095 0.0505952 0.0892857 0.107143 0.133929 0.0297619 0.0178571 0.0416667 0.0535714 0.0119048

La salida de la red neuronal son dos valores:

0.941154 -0.990584

Se procesan estos valores y se obtiene si la secuencia es o no una isla CpG

Para el entrenamiento de la red neuronal se utilizó una función de activación sigmoidal simétrica para la capa oculta y de salida, además se entrenó hasta conseguir un error inferior a 0.00001 y un umbral para reconocimiento de 0.6 o mayor que está definido en [11. Si la salida de la primera neurona era igual o superior a 0.6 y la salida de la segunda neurona era inferior o igual a -0.6, se reconocía el fragmento como parte de una isla CpG, si la salida de la primera neurona era inferior o igual a -0.6 y la salida de la segunda neurona era superior o igual a 0.6, se reconocía el fragmento se reconocía como no isla CpG, si alguna de las salidas no superaba el umbral de 0.6 o -0.6 la red no era capaz de reconocer la secuencia.

En los resultados de las pruebas realizadas con las secuencias pertenecientes a islas CpG, 13 secuencias fueron identificadas correctamente, 2 secuencias no reconocidas y 0 secuencias de forma incorrecta. En los resultados de las pruebas realizadas con las secuencias no pertenecientes a islas CpG, 14 secuencias fueron identificadas correctamente, 3 secuencias no reconocidas y 0 secuencias de forma incorrecta. Con estos datos se obtuvieron entonces un reconocimiento exitoso de aproximadamente 87 %, un porcentaje de no reconocidas de aproximadamente 13% y 0% de reconocimientos no exitosos en el caso de los fragmentos de islas CpG, por otra parte en los fragmentos no pertenecientes a islas CpG se obtuvo un reconocimiento exitoso de aproximadamente 82 %, un porcentaje de no reconocidas de aproximadamente 18% y 0% de reconocimientos no exitosos. Si se dispone de mayor cantidad de datos de entrenamiento es posible que se puedan mejorar los resultados obtenidos, también se podrían utilizar otras medidas para determinar la calidad de la clasificación, tales como sensibilidad y especificidad.

En total se probaron 32 secuencias entre islas CpG y no islas CpG de las cuales el prototipo de red logró identificar correctamente 27 secuencias, no hubo ningún reconocimiento incorrecto y tan solo 5 secuencias no pudieron ser reconocidas por la red. Los datos de comparación se muestran en el gráfico de barras de la Fig. 6.

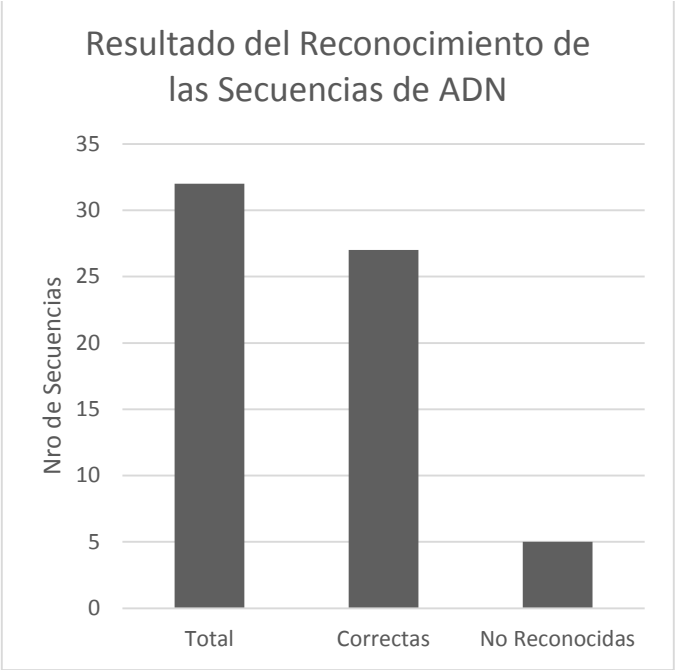

Fig. 6. Resultado del reconocimiento de las secuencias de ADN utilizando la Red Nuronal propuesta.

VI. CONCLUSIONES

La red neuronal diseñada para el reconocimiento de secuencias pertenecientes a islas CpG generó resultados adecuados, obteniendo porcentajes de aciertos mayores de 85% en las pruebas realizadas y un porcentaje nulo de fallos.

El modelo permite el reconocimiento de las islas CpG con redes neuronales, para así, descubrir conocimiento, que se encuentra implícito en las secuencias de ADN y que sin el uso del modelo propuesto no se tiene.

Los experimentos recogen un conjunto de datos del ADN, estos son almacenados en un repositorio de datos. Estos contienen información importante que se encuentra oculta, por lo que es necesario idear y utilizar mecanismos que permitan identificarla, extraerla e interpretarla como los propuestos en este trabajo.

El uso de Redes Neuronales permite el procesamiento de las secuencias de ADN, significa que los expertos pueden automáticamente experimentar con más modelos e información para entender datos complejos, también va a hacer que sea práctico analizar inmensas cantidades de datos extraídos del ADN.

Como trabajo futuro, se deben continuar realizando el reconocimiento de islas CpG en las secuencias de ADN, ya que existe mucha información oculta en ellas, que es necesario descubrir.

Por otro lado, es necesario demostrar que las islas CpG encontradas forman parte de los genes o no.

En las proteínas se utiliza el término "hot spots", que son segmentos cortos de aproximadamente 6 aminoácidos que se asocian para formar fibras o forman parte de una función importante dentro de éstas.

AGRADECIMIENTO

Al Proyecto CDCHTA I – 1407 – 14 – 02 – B de la Universidad de Los Andes por su apoyo financiero.

REFERENCIAS

[1] J. Altamiranda, J. Aguilar, L. Hernandez, *"Sistema de reconocimiento de patrones de sustancias químicas cerebrales basado en minería de datos"* Computación y Sistemas, Vol. 19, N°1, pp. 89 – 107, 2015

[2] J. Altamiranda, J. Aguilar, C. Delamarche, "*Comparison and fusion model in protein motifs*". Proceedings XXXIX Latin America Computing Conference (CLEI 20013),. pp. 1–12,. 2013

[3] J. Altamiranda, J. Aguilar, C. Delamarche,"*Similarity of Amyloid Protein Motif using an Hybrid Intelligent System*". Latin AmericaTransactions IEEE, Vol. 9, No. 5, pp. 700–710, 2011

[4] L. Buehler., H. Rashidi "*Bioinformatics basics: applications in biological science and medicine*" . Segunda Edición. CRC Press. 2005

[5] C. Chen, T. Li, "*A hybrid neural netwoork system for prediction and recognition of prometer regions in human genome"* J. Zhejiang Univ. SCIENCE. Vol. 6B, No 5, pp. 408 – 414, 2005.

[6] J. Aguilar, F. Rivas, "*Introducción a las técnicas de Computación Inteligente* ". Meritec. 2001

[7] J. Hilera, V. Martinez "*Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones".* Addison – Wesley. 1995.

[8] R. Flórez, J. Fernández, *"Las Redes Neuronales Artificiales"* Editorial NETBIBLO. 2008

[9] M. Bhasin., H. Zhang, E. Reinherz, P. Reche "*Prediction of methylated CpGs in DNA sequences using a support vector machine"* FEBS Letters 579, pp. 4302 – 4308. 2005

[10] J. Oriola, J. Claria, R. Oliva, F. Ballesta. Genética médica. Publicacions I Edicions de la Universitat de Barcelona, Barcelona, España, 2004..

[11] K. Patton, G. Thibodeau, Anatomía y fisiología. Elsevier. España, SL, Barcelona, España, 2013.

[12] V. Burriel V. "Estructura y propiedades de los ácidos nucléicos". Disponible en: http://www.uv.es/tunon/pdf_doc/AcidosNucleicos_veronica.pdf

[13] M. Gardiner-Garden, M. Frommer, Cpg islands in vertebrate genomes. Journal of Molecular Biology, 196:261– 282, July 1987.

[14] S. Nissen Fast artificial neural network library, 2014.

[15] Genome Bioinformatics Group. UCSC genome browser website, 2014. http://genome.ucsc.edu/