ISSN 0121-1129

Predicción de proteínas con redes neuronales: Redes Feedforward vs. Redes Recurrentes

Protein Prediction with Neural Networks: Feedforward Networks vs. Recurring Networks

Beitmantt Geovanni Cárdenas Quintero *

Resumen

La función de una proteína depende de la estructura tridimensional que adopte, es por esto que el reto de la comunidad científica es encontrar un método eficiente para hallar la estructura de las proteínas. Los métodos de laboratorio existentes actualmente para la predicción estructural están limitados a un grupo muy pequeño de proteínas. Teniendo en cuenta lo anterior y conociendo que la estructura tridimensional de una proteína está definida por la secuencia de aminoácidos que la constituyen, las investigaciones han migrado de los laboratorios hacia las ciencias de la información, donde hallar la estructura se convierte en un problema de predicción a partir de una cadena de símbolos (secuencia de aminoácidos). En este artículo se evalúa y se compara el desempeño de las redes neuronales Feedforward vs. las recurrentes ante la predicción de la estructura secundaria de una proteína.

Abstract

The proteins functions depend of the three dimensional structure that it adopts. Therefore, the scientific community's challenge is to find an efficient method to predict the proteins structure. The laboratory methods existing at the moment for structural prediction are limited to a very small protein group. Considering the previous thought and knowing that the three-dimensional protein structure is defined by the amino acids sequence that constitutes it, the investigations have migrated from the laboratories towards the information sciences, where to find the protein structure is a problem of prediction based on a symbols chain (amino acids sequence). Here is evaluated and compared the performance of the Feed Forward Neural Networks vs. the Recurrent Neural Network for the prediction of the protein secondary structure.

* Universidad Distrital "Francisco José de Caldas", Bogotá. beitmantt@yahoo.com

Palabras clave: Proteínas, Redes neuronales artificiales. Redes feedforward, Redes recurrentes.

Key words: Protein, Artificial Neural Network, Feedforward Neural Networks, Recurrent Neural Network.

1. Introducción

El crecimiento acelerado del descubrimiento de la información biológica abre las puertas a "milagrosos" avances en el sector de la medicina; sin embargo, para llegar a estos avances no solo se requiere obtener la información, sino procesarla de manera eficiente. El volumen y la complejidad de la información exigen que las técnicas de procesamiento sean muy eficientes. Actualmente el descubrimiento de información biológica avanza a pasos muchos más grandes que el desarrollo de técnicas de procesamiento para analizar y dar uso a esta información. Una de las principales razones por las cuales se ha dado esta diferencia es que la mayoría de problemas de análisis de información biológica no tienen solución algorítmica.

Una de las áreas donde más se ha marcado la carencia de recursos de procesamiento de información ante la avalancha de información biológica descubierta es la de las proteínas. Descubrir la secuencia de aminoácidos que constituyen una proteína es un problema prácticamente solucionado, sin embargo, utilizar esta secuencia para predecir la estructura terciaria y la funcionalidad de las proteínas es un reto que al día de hoy se ha cumplido en una muy pequeña parte. Ante los pocos resultados conseguidos con técnicas algorítmicas convencionales, los investigadores han visto en las técnicas de aprendizaje de máquina un camino a seguir.

De las diferentes técnicas que existen, la que ha alcanzado mejores niveles de precisión han sido las redes neuronales. Sin embargo, estas son muy diversas y no todas las arquitecturas son eficientes en un dominio determinado; es por esto que a pesar de que se han conseguido esperanzadores resultados, aún no se tiene conocimiento cierto de qué arquitectura es más eficiente en la predicción de proteínas. Por tradición y por los altos niveles de precisión alcanzados, las redes neuronales más utilizadas en herramientas de predicción de proteínas actualmente son las típicas redes hacia delante, sin embargo, debido a su naturaleza estática presentan limitantes ante las longitudes grandes y el dinamismo de los datos de entrada. Ante estos limitantes algunos investigadores han presentado las arquitecturas dinámicas, especialmente las redes neuronales recurrentes, como una posible solución y como un nuevo horizonte en la lucha diaria por conseguir arquitecturas de redes cada vez más eficientes.

2. Predicción estructural de proteínas

Actualmente existen dos métodos de laboratorio que permiten hallar la estructura terciaria de las proteínas: la cristalografía de rayos x y la espectrografía de Resonancia Nuclear Magnética (NMR). La cristalografía de rayos x era, hasta hace pocos años, el único método existente; este, a pesar de que es muy preciso, tiene fuertes limitantes, como: lentitud, altos costos y aplicable solo a algunas pocas proteínas. La espectrografía de NMR es un método reciente muy preciso, pero aún limitado a proteínas de tamaño pequeño.

Ante las limitaciones de los métodos de laboratorio y ante la certeza de que la estructura terciaria está definida por la secuencia de aminoácidos, encontrar la estructura de una proteína ha dejado de ser un problema exclusivamente de los laboratorios de biología, para convertirse en un problema de predicción de las ciencias de la información.

2.1 Predicción de la estructura tridimensional

Existen muchos métodos para la predicción estructural tridimensional de proteínas, que se clasifican en tres categorías principales, de acuerdo con la técnica que utilizan:

Ab inicio

Estos métodos se basan en la suposición de que la información necesaria para conocer la estructura tridimensional de una proteína está en su secuencia de aminoácidos. Mediante la minimización de la energía potencial derivada de las consideraciones físico-químicas y estáticas se busca imitar el proceso de doblaje que tendría una proteína. El método y servidor más exitoso en esta categoría ha sido:

- ROSETTA http://robetta.bakerlab.org/

Reconocimiento de Doblez

Los investigadores creen que el doblez natural de las proteínas forma un diccionario finito de clases con solamente algunos miles de palabras. El método más utilizado de esta categoría es aquel que se basa en "roscado" o Threading. A continuación se listan algunos de los servidores de predicción basados en métodos de reconocimiento de Doblez:

- 3DPSSM http://www.sbg.bio.ic.ac.uk/%7E3dpssm/
- SAMT99 http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
- SAMT02 http://www.soe.ucsc.edu/research/compbio/HMM-apps/T02-query.html
- GenTHREADER http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
- FUGUE http://wwwcryst.bioc.cam.ac.uk/%7Efugue/prfsearch.html
- FFAS03 http://bioinformatics.ljcrf.edu/pages/
- PHDthreader http://www.embl-heidelberg.de/predictprotein/
- T3P2 http://www.mbi.ucla.edu/people/frsvr/frsvr.html

Molado de proteínas por homología

Estos métodos se basan en el hecho de que todas las parejas de proteínas que presentan una identidad de secuencia mayor al 30% tienen estructura tridimensional similar. De este modo se puede construir el modelo tridimensional de una proteína de estructura desconocida, partiendo de la semejanza de secuencia con proteínas de estructura conocida. Esta técnica es actualmente la más utilizada. Existe un gran número de servidores de predicción basados en homología, entre los cuales los principales son:

- Swiss-Model http://www.expasy.ch/swissmod/SWISS-MODEL.html
- TITO/Modeller http://bioserv.infobiosud.univ-montp1.fr/
- CPHmodels http://www.cbs.dtu.dk/services/CPHmodels/
- SDSC1 http://cl.sdsc.edu/hm.html
- 3D-JIGSAW http://www.bmm.icnet.uk/servers/3djigsaw/
- Loops Database http://www.bmm.icnet.uk/loop/

El éxito de estos métodos radica en el crecimiento de las bases de datos de proteínas en las cuales se pueda buscar proteínas homólogas de una proteína desconocida. En los últimos diez años el crecimiento de las bases de datos de proteínas ha sido prácticamente exponencial, de ahí la explicación de que los métodos por homología han tenido tanta aceptación. En la figura 1 se grafica el crecimiento de la base de datos de proteínas PDB (Protein Data Bank).

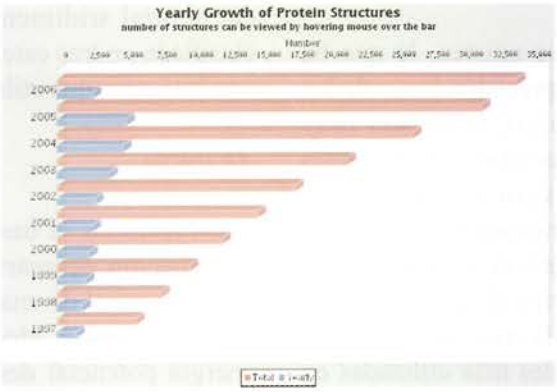

Figura 1. Crecimiento de la base de datos de proteínas PDB (Protein Data Bank).
Imagen tomada de http://www.rcsb.org/pdb/Welcome.do

En la figura 2 se muestran dos proteínas que tienen secuencias similares y estructuras terciarias similares.

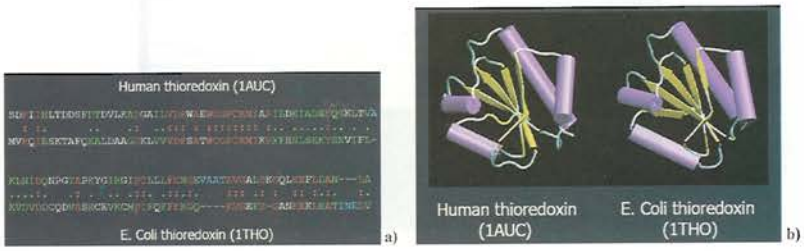

Figura 2. Similitud entre dos proteínas. a) Similitud en las secuencias. b) Similitud en la estructura terciaria. Imágenes tomadas de la presentación "Protein Structure Prediction from A-Z" del MMTSB NIH Research Resource PSC Workshop 2003.

2.2 Predicción de características 1D y 2D

Una gran parte de la comunidad científica cree que los métodos de predicción de la estructura tridimensional de las proteínas podrían avanzar mucho si no se basaran solo en la secuencia de aminoácidos, sino que también se alimentaran de otras características 1D y 2D de las proteínas. Debido a esto han surgido una gran cantidad de métodos que buscan predecir características unidimensionales y bidimensionales de las proteínas.

Las características 1D de una secuencia son aquellas que pueden ser representadas por un solo valor asociado a cada aminoácido [11], tales como: propiedades de los residuos, accesibilidad, estructura secundaria, hélices transmembrana, péptidos de señal, entre otros.

Predicción de la estructura secundaría. De las características 1D de las proteínas, la estructura secundaría ha sido la más investigada. Como se mencionó anteriormente, esta estructura informa de la disposición de los aminoácidos en el espacio. Las herramientas de predicción de estructura secundaria cumplen una labor de clasificación, toman como entrada los residuos de los aminoácidos y los clasifican en 8 ó 3 clases, de acuerdo con el DSSP.
Los principales métodos y servidores de predicción de estructura secundaría son los siguientes:

- APSSP2 http://www.imtech.res.in/raghava/apssp2/
- JPred http://jura.ebi.ac.uk:8888/
- JUFO http://www.jens-meiler.de/jufo.html
- Mlrprdsc http://mlrprdsec.cbi.pku.edu.cn/
- PHD http://cubic.bioc.columbia.edu/predictprotein
- PHDpsi http://cubic.bioc.columbia.edu/predictprotein
- Porter http://distill.ucd.ie/porter/
- PROF_king http://www.aber.ac.uk/~phiwww/prof/
- PROFsec http://cubic.bioc.columbia.edu/predictprotein
- Prospect http://compbio.ornl.gov/PROSPECT/PROSPECT-Pipeline/cgi-bin/proteinpipeline_form.cgi
- PSIpred http://insulin.brunel.ac.uk/psiform.html
- SABLE http://sable.cchmc.org/
- SABLE2 http://sable.cchmc.org/
- SAM-T99sec http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
- SCRATCH (SSpro3) http://www.igb.uci.edu/tools/scratch/

En la figura 3 se observa la predicción de la estructura secundaria a partir de una secuencia por parte de varios servidores, además se relaciona la precisión de la predicción.

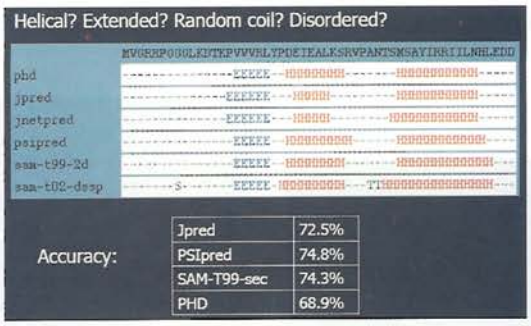

Figura 3. Predicción de la estructura secundaria a partir de una secuencia por parte de varios servidores. Imagen tomada de la presentación "Protein Structure Prediction from A-Z" del MMTSB NIH Research Resource PSC Workshop 2003.

3. Evaluación de la predicción de proteínas

Existen tres organizaciones que tienen como objetivo la evaluación de la calidad de la predicción de proteínas:

- CASP (Critical Assessment of Techniques for Protein Structure Prediction) http://predictioncenter.org
- EVA (Evaluation of Automatic protein structure prediction) http://cubic.bioc.columbia.edu/eva
- LiveBench http://bioinfo.pl/meta/livebench.pl

EVA y LiveBench son herramientas en línea de evaluación de servidores de predicción, mientras que CASP es una organización encargada de realizar un experimento bianual para la evaluación de herramientas y expertos para la predicción de proteínas.

4. Redes neuronales artificiales para la predicción de proteínas

La implementación de cualquiera de los métodos de predicción de proteínas mencionados en el numeral 3 requiere técnicas computacionales que los soporten. El común denominador de estos métodos es el reconocimiento y la clasificación de patrones. Es conocido y comprobado que en materia de reconocimiento y clasificación, el aprendizaje de máquina ha ofrecido excelentes resultados.

En la predicción de proteínas se han probado una gran variedad de técnicas de aprendizaje de máquina, sin embargo las más utilizadas, por sus notorios resultados, han sido las redes neuronales artificiales. Actualmente muchos de los servidores y herramientas de predicción de proteínas se basan en redes neuronales artificiales.

4.1 Evolución de las redes neuronales en la predicción de proteínas

El uso de las redes neuronales en la predicción de proteínas, según Rost [13], ha pasado por cuatro etapas:

Etapa 1: Inicialmente las redes neuronales artificiales se aplicaban como cajas negras y las investigaciones se enfocaban en optimizar los parámetros internos, tales como la velocidad de aprendizaje y la arquitectura.

Etapa 2: Los investigadores abren las cajas negras para extraer o implementar reglas y para fijar conocimientos específicos en las redes.

Etapa 3: La combinación de las redes neuronales artificiales y la información evolutiva (bases de datos de proteínas en constante crecimiento) dejan al descubierto el verdadero potencial de las redes neuronales, convirtiéndose estas en una de las principales técnicas para la implementación de herramientas de predicción de proteínas. Una de las áreas donde las redes neuronales se sobrepusieron totalmente sobre las técnicas existentes fue en la predicción de la estructura secundaria.

Etapa 4: En los últimos años, además de utilizar todo lo aprendido en las etapas anteriores, las herramientas de predicción se están implementando de manera híbrida: Las redes neuronales se aplican a problemas específicos y el resto del sistema se implementa con metodologías expertas diferentes a redes neuronales o metodologías de software no expertos.

4.2 Evolución de los métodos de predicción basados en redes neuronales artificiales

La primera aplicación de las redes neuronales a la predicción de proteínas aparece en 1988; Bohr [22] y Qian [45] proponen métodos para predecir la estructura secundaría de las proteínas basados en redes neuronales. La siguiente década se caracteriza por la proposición de un gran número de métodos basados en redes neuronales para predecir características 1D (estructura secundaria, hélices transmembrana, filamentos transmembrana, accesibilidad), por parte de autores muy reconocidos como: Andrade [42], Rost [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12], Sander [10] y Casadio [53, 54], entre otros. A mediados de los noventa las redes neuronales comenzaban a constituirse como un método estándar en la predicción de proteínas; en 1996, Rost [12] propone un método para la predicción de características 1D de proteínas basado en redes neuronales, el cual dio origen a uno de los servidores de predicción de estructura secundaría más importantes que existe en la actualidad, llamado PDH; en 1997 aparece el primer servidor de predicción de estructura terciaria por homología basado en redes neuronales CPH Models, propuesto por Lund, Frimand, Gorodkin, H. Bohr, J. Bohr, Hansen y Brunak [46]; a partir de 1997 se comienzan a proponer métodos que van más allá de predecir características 1D, aparecen métodos para descubrir y predecir motifs [16, 17, 36, 37, 38], actividad biológica [43, 52], modificaciones postranslacionales [23, 24, 25, 32, 33, 39, 40, 44, 55], tipos particulares de proteínas [1], dominios [35], desórdenes en proteínas [15, 41, 50, 51], entre otros.

A partir del 2002 han aparecido métodos de predicción muy eficientes basados en redes neuronales recurrentes: SSPro y SSPro8 [18]; son métodos de predicción de estructura secundaria en 3 y 8 clases respectivamente. ACCpro [19] es un servidor para la predicción de la accesibilidad de los residuos de una proteína; CONpro [49, 20] es un servidor que predice si el número de contactos de cada residuo en una proteína es mayor o menor del promedio; CMAPpro [21, 47] es un servidor para la predicción de mapas de contactos entre residuos de una proteína. Los servidores mencionados se pueden encontrar en un Meta Servidor llamado SCRATCH [26].

En el 2005 y en el presente año se ha consolidado el uso de las redes neuronales recurrentes, se han creado servidores muy eficientes basados en este tipo de red, tales como: Mupro 1.1 [27], que es un servidor que predice cambios de estabilidad de mutaciones de sitios simples de secuencias de proteínas; DOMpro 1.0 [28], que es un servidor para predecir dominios de proteínas; DIpro II [29, 48], que es un servidor para predecir enlaces de disulfuro en las proteínas; Betapro 1.0 [30], que es un servidor para la predicción de pares Beta-residuos, pares Beta-filamentos, filamentos alineados, dirección de apareamiento y topología beta-sheet; DISpro 1.0 [31], que es un servidor para predecir regiones desordenadas en secuencias de proteínas.

5. Redes Feedforward vs. Redes Recurrentes

Utilizando una versión de prueba del software NeuroSolution 5.0 se diseñaron e implementaron dos redes neuronales (feedforward y recurrente) para medir su desempeño frente al problema de la predicción de la estructura secundaria de las proteínas a partir de su secuencia de aminoácidos.

5.1 Set de datos de entrenamiento y evaluación

Se utilizó el reconocido conjunto de 106 proteínas usadas por Ning Qian y Terrence J. Sejnowski (1989) en la investigación que inició la era de las redes neuronales en la predicción de proteínas.

Tabla 1. Conjunto de 106 proteínas utilizadas para entrenar y evaluar las redes neuronales.

Código	Nombre proteína
labp	1 -Arabinose-binding protein
laCX	Actinoxanthin
lapr	Acid protease
laza	Azurin
lazu	Azurin
lbp2	Phospholipase A2
1 cat	Carbonic anhydrase form c
lcc5	Cytochrome c5 (oxidized)
1 ccr	Cytochrome c (rice)
lcpv	Calcium-binding parvalbumin b
lcrn	Crambin
1ctx	a-Cobratoxin
1 cy3	Cytochrome c3
lCYC	Ferrocytochrome c
lecd	Haemoglobin (deoxy)
lest	Tosyl-elastase
1 fc2	Immunoglobulin FC-Frag B complex
lfdh	Haemoglobin (deoxy, human fetal)
lfdx	Ferredoxin
lfxl	Flavodoxin
lgcn	Glucagon (pH 6-pH 7 form)
lgcr	y-Crystallin
lgfl	Insulin-like growth factor
I&@	Insulin-like growth factor
lgpl lhds	Glutathione peroxidase Haemoglobin (sickle cell)
1 hip	High potential iron protein
lhmq	haemerythrin (met)
1 ig2	Immunoglobunlin Gl
1 ige	Fc fragment (model)
lins	Insulin
lldx	Lactate dehydrogenase
llzl	Lysozyme
llzm	Lysozyme
llzt	Lysozyme, triclinic cristal form
lmbd	Myoglobin (deoxy, pH 8.4)
lmbs	Myoglobin (met)
lmlt	Melittin
lnxb	Neutrotoxin b
' P2P	Phospholipase A2
lpfc	Fragment of IgG
lppd	2-hydroxyethylthiopapain
1PPt	Avian pancreatic polypeptide
'PYP	Inorganic pyrophosphatase
1 rei	Immunoglobulin B-J fragment v
lrhd	Rhodanese
lrn3	Ribonuclease A
lsn3	Scorpion neurotoxin (variant 3)
ltim	Triose phosphate isomerase
1 tgs	Trypsinogen complex
2act	Actinidin (sulphhydryl proteinase)
2adk	Adenylate kinase
2tllp	a-Lytic protease

Código	Nombre proteína
Zape	Acid proteinase, endothiapepsin
lapp	Acid proteinase, penicillopepsin
2b5c	Cytochrome b5 (oxidized)
2rab	Carbonic anhydrase form b
Zrcg	Cytochrome c (prime)
Pcdv	Cytochrome c3
2cyp	Cytochrome e peroxidase
%dhh	Haemoglobin (horse, deoxy)
%fdl	Ferredoxin
dgch	y-Chymotrypsin a
Ign.5	Gene 5/DNA binding protein
zgrs	Glutathione reductase
2icb	Calcium-binding protein
2kai	Kallikrein a
%Ih 1	Leghaemoglobin (acetate, met)
2Ihb	Haemoglobin V (cyano, met)
2mcp	Ig Fab mcpc603/phosphocholine
2mdh	Cytoplasmic malate dehydrogenase
2mt:!	Cd, Zn metallothionein
Zpab	Prealbumin (human plasma)
2rhr	Immunoglobulin B-J fragment V-MN
lsbt	Subtilisin novo
"sga	Proteinase A
"sns	Staphylococcal nuclease complex
"sod	Cu.& superoxide dismutase
2ssi	Streptonzyces subtilisin inhibit0
zstv	Satellite tobacco necrosis virus
2taa	Taka-amylase a
"tb\	Tomato bushy stunt virus
3r2t	('ytochrome c2 (reduced)
3rna	Concanavalin A
3fXC	Ferredoxin
%pd	Glyceraldehyde-3-P-dehydrogenase
3hhb	Haemoglobin (deoxy)
3pq	Plastocyanin (Hg" substituted)
3p&	Phosphoglycerate kinase complex
3pfzm	Phosphoglycerate mutase
3@	Rat mast cell protease
3sgb	Proteinase R
3t1n	Thermolysin
451c	Cytochrome ~551 (reduced)
M's	Citrate synthase complex
4dfr	Dihydrofolate reductase
Ifx 11	Flavodoxin (semiquinone form)
Isbv	Southern bean mosaic virus coat protein
5atc	Aspartate carbamoyltransferase
5cpa	Jarboxypeptidase
5ldh	Lactate dehydrogenase complex
5pt i	Trypsin inhibitor
5rx n	Rubredoxin (oxidized)
6adh	Alcohol dehydrogenase complex
6api	Modified a-1-antitrypsin
Xcat	Catalase

La secuencia y la segunda estructura de estas proteínas se obtuvieron de la base de datos del PDB (Protein Data Bank).

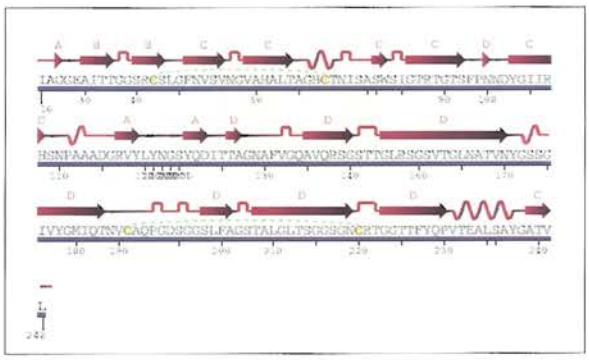

Figura No. 4. Secuencia y estructura secundaria de la proeinasa A obtenida del PDB

5.2 Redes neuronales a evaluar

Se evaluaron dos tipos de redes, descritas a continuación:

Red Feedforward

Arquitectura: Feedforward generalizada
Capa de entrada: 7 neuronas
Capas internas: 1 capa interna de 5 neuronas
Capa de salida: 1 de una neurona
Algoritmos de aprendizaje: Gradiente descendente
Función de activación: Sigmoidea
Algoritmo de parada: Número de iteraciones = 1000

Red Recurrente

Arquitectura: Recurrente generalizada bidireccional
Capa de entrada: 7 neuronas
Capas internas: 1 capa interna de 5 neuronas
Capa de salida: 1 de una neurona
Algoritmos de aprendizaje: Gradiente descendente
Función de activación: Sigmoidea
Algoritmo de parada: Número de iteraciones = 1000

5.3 Entrenamiento y resultados

A continuación se muestran los resultados tanto en la fase de entrenamiento como en la fase de predicción.

Entrenamiento de la red feedforward

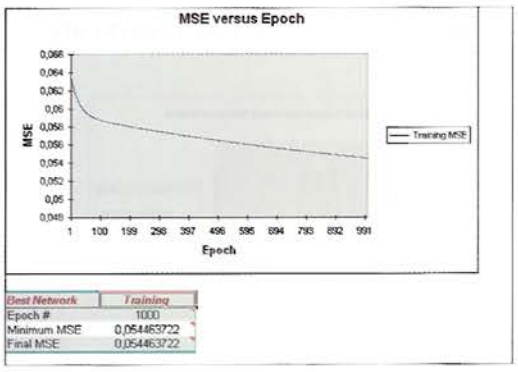

Figura 5. Desempeño del entrenamiento de la red feedforward

Desempeño de la red feedforward

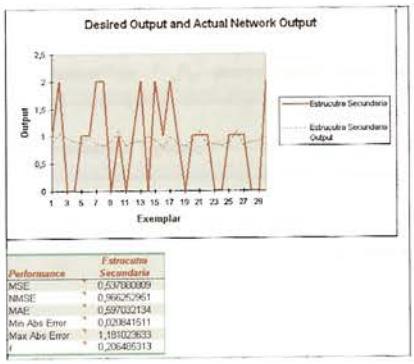

Figura 6. Resultados de la red feedforward

Entrenamiento de la red recurrente

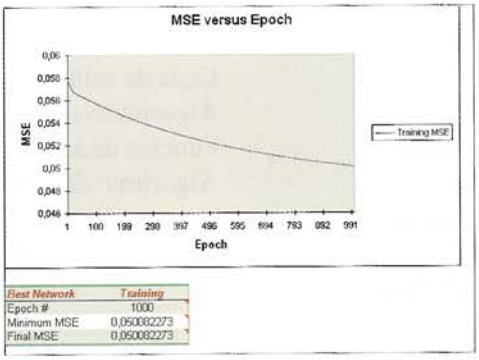

Figura 7. Desempeño del entrenamiento de la red recurrente.

Desempeño de la red recurrente

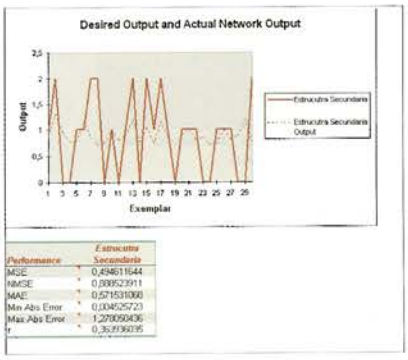

Figura 8. Resultados de la red recurrente.

La red de arquitectura recurrente tiene un procedimiento de aprendizaje más eficiente y alcanza un nivel de estabilidad más rápido que la red feedforward.

A pesar de que ambos tipos de redes presentan una efectividad no muy buena, las redes recurrentes alcanzan una efectividad del 73%, superando a la red feedforward, que presentó una efectividad del 69%.

6. Conclusiones

Las redes neuronales artificiales en la última década se han consolidado como una de las herramientas computacionales más eficientes en el procesamiento de información biológica. Una de las áreas donde más éxito han tenido las redes neuronales ha sido en la predicción de proteínas; actualmente un gran porcentaje de los servidores de predicción están basados en redes neuronales. El reto actual es encontrar arquitecturas de redes neuronales que permitan desarrollar herramientas de predicción con mayor precisión.

Las redes neuronales de arquitectura recurrente demostraron tener un mejor desempeño tanto en el entrenamiento como en su funcionamiento en la predicción de la estructura secundaria de proteínas a partir de la secuencia de aminoácidos.

Referencias

[1] A. Gurvitz, S. Langer, M. Piskacek, B. Hamilton, H. Ruis and A. Hartig. "Predicting the Function and Subcellularlocation of Caenorhabditis Elegans Proteins Similar to Saccharomyces Cerevisiae Beta-oxidation Eenzymes". *Yeast* 17(2000): 188-200.

[2] B. Rost and C. Sander. "Prediction of protein secondary structure at better than 70% accuracy". *J. Mol. Biol.* 232 (1993): 584-599.

[3] B. Rost and C. Sander. "Improved Prediction of Protein Secondary Structure by Use of Sequence Profiles and Neural Networks". *Proc. Natl. Acad. Sci. USA.* 90 (1993): 7558-7562.

[4] B. Rost and C. Sander. "Secondary Structure Prediction of All-Helical Proteins in Two States". *Prot. Engin.* 6 (1993): 831-836.

[5] B. Rost and C. Sander. "Combining Evolutionary Information and Neural Networks to Predict Protein Secondary Structure". *Proteins* 19 (1994): 55-72.

[6] B. Rost, C. Sander and R. Schneider. "PHD - an Automatic Server for Protein Secondary Structure Prediction". *CABIOS* 10 (1994): 53-60.

[7] B. Rost, R. Casadio, P. Fariselli and C. Sander. "Prediction of Helical Transmembrane Segments at 95% Accuracy". *Prot. Sci.* 4 (1995): 521-533.

[8] B. Rost, R. Casadio and P. Fariselli. "Refining Neural Network Predictions for Helical Transmembrane Proteins Bydynamic Programming". In: D. States, P. Agarwal, T. Gaasterland, L. Hunter and R. F. Smith (eds.): *Fourth International Conference on Intelligent Systems for Molecular Biology*. Menlo Park, CA: AAAI Press, St. Louis, M.O., U.S.A., 1996, pp. 192-200.

[9] B. Rost, R. Casadio and P. Fariselli. "Topology Prediction for Helical Transmembrane Proteins at 86% Accuracy. *Prot. Sci.* 5 (1996): 1704-1718.

[10] B. Rost and C. Sander. "Conservation and Prediction of Solvent Accessibility in Protein Families". *Proteins* 20 (1994): 216-226.

[11] B. Rost. "PHD: Predicting One-Dimensional Protein Structure by Profile Based Neural Networks". *Meth. Enzymol.* 266 (1996): 525-539.

[12] B. Rost. "PHD: Predicting One-Dimensional Protein Structure by Profile Based Neural Networks". Methods in Enzymology, 266, 525-539, 1996.

[13] B. Rost. "Neural Networks Predict Protein Structure: Hype or Hit?". In: Paolo Frasconi and RonShamir (eds.): Artificial Intelligence and Heuristic Methods in Bioinformatics.

[14] J. Chahine, J. R. Ruggiero, L. P. B. Scott. "Prediction of Protein Structures Using a Hopfield Network". *Dept. de Física, Ibilce, Unesp.*

[15] E. Garner, P. Cannon, P. Romero, Z. Obradovic and A. K. Dunker. "Predicting Disordered Regions from Amino Acid Sequence: Common Themes Despite Differing Structural Characterization". *Genome Inform.* 9 (1998): 201-214.

[16] F. J. Lebeda and M. A. Olson. "Predicting Differential Antigen-Antibody Contact Regions Based on Solvent Accessibility". *J. Prot. Chem.* 16 (1997): 607-618.

[17] G. Mlinsek, M. Novic, M. Hodoscek and T. Solmajer. "Prediction of Enzyme Binding: Human Thrombin Inhibition Study By Quantum Chemical And Artificial Intelligence Methods Based On X-Ray Structures". *J Chem Inf Comput Sci* 41 (2001): 1286-1294.

[18] G. Pollastri, D. Przybylski, B. Rost, P. Baldi. "Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles". *Proteins*, 47, 228-235, 2002.

[19] GPollastri, P. Baldi, P. Fariselli, R. Casadio. "Prediction of Coordination Number and Relative Solvent Accessibility in Proteins". *Proteins*, 47, 142-153, 2002.

[20] G. Pollastri, P. Baldi, P. Fariselli, R. Casadio. "Improved Prediction of the Number of Residue Contacts in Proteins by Recurrent Neural Networks". *Bioinformatics*, 17 Suppl 1, S234-S242 (2001).

[21] G. Pollastri, P. Baldi. "Prediction of Contact Maps by Recurrent Neural Network Architectures and Hidden Context Propagation from All Four Cardinal Corners". *Bioinformatics*, 18 Suppl 1, S62-S70 (2002).

[22] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Nørskov, O. H. Olsen and S. B. Petersen. "Protein Secondary Structure and Homology by Neural Networks". *FEBS Lett.* 241 (1988): 223-228.

[23] H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne. "Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites". *Prot. Engin.* 10 (1997) 1-6. IOS Press, Amsterdam 2003, ISBN 1-58603-294-1, pp. 34-50.

[24] H. Nielsen, J. Engelbrecht, S. Brunak and G. von Heijne. A Neural Network Method for Identification of Prokaryotic and Eukaroytoic Signal Peptides and Prediction of Their Cleavage Sites". *Internationl Journal of Neural Systems* 8 (1997): 581-599.

[25] H. Nielsen, S. Brunak and G. von Heijne. "Machine Learning Approaches for the Prediction of Signal Peptides and Other Protein Sorting Signals". *Prot. Engin.* 12 (1999): 3-9.

[26] J. Cheng, A. Randall, M. Sweredoski, P. Baldi. "SCRATCH: a Protein Structure and Structural Feature Prediction Server", Nucleic Acids Research, Web Server Issue, vol. 33, w72-76, 2005.

[27] J. Cheng A. Randall, P. Baldi. "Prediction of Protein Stability Changes for Single-Site Mutations Using Support Vector Machines". *Proteins*, vol. 62, no. 4, pp. 1125-1132, 2006.

[28] J. Cheng, M. Sweredoski, P. Baldi. "DOMpro: Protein Domain Prediction Using Profiles, Secondary Structure, Relative Solvent Accessibility, and Recursive Neural Networks". *Knowledge Discovery and Data Mining*, vol. 13, no. 1, pp. 1-20, 2006.

[29] J. Cheng, H. Saigo, P. Baldi. "Large-Scale Prediction of Disulphide Bridges Using Kernel Methods, Two-Dimensional Recursive Neural Networks, and Weighted Graph Matching". *Proteins,* vol. 62, no. 3, pp. 617-629, 2006.

[30] J. Cheng and P. Baldi. "Three-Stage Prediction of Protein Beta-Sheets by Neural Networks, Alignments, and Graph Algorithms". *Bioinformatics*, vol. 21, Suppl. 1, pp. 75-84, 2005.

[31] J. Cheng, M. Sweredoski, P. Baldi. "Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data". *Data Mining and Knowledge Discovery*, vol. 11, no. 3, pp. 213-222, 2005.

[32] J. Hansen, O. Lund, N. Tolstrup, A. A. Gooley, K. L. Williams and S. Brunak. "NetOglyc: Prediction of Mucin Type Oglycosylation Sites Based on Sequence Context and Surface Accessibility". *Glycoconjugate Journal* 15 (1998): 115-130.

[33] J. L. Herrmann, R. Delahay, A. Gallagher, B. Robertson and D. Young. "Analysis of Post-Translational Modification of Mycobacterial Proteins Using a Cassette Expression System. *FEBS Lett.* 473 (2000): 358-362.

[34] John Hawkins, Mikael Bodén. "The Applicability of Recurrent Neural Networks for Biological Sequence Analysis". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 3, pp. 243-253, July-September, 2005.

[35] J. Murvai, K. Vlahovicek, C. Szepesvari and S. Pongor. "Prediction of Protein Functional Domains from Sequences Using Artificial Neural Networks". *Genome Res.* 11 (2001): 1410-1417.

[36] J. S. Fetrow, M. J. Palumbo and G. Berg. "Patterns, Structures, and Amino Acid Frequencies in Structural Building Blocks, a Protein Secondary Structure Classification Scheme". *Proteins* 27 (1997): 249-271.

{37] K. Gulukota and C. DeLisi. "Neural Network Method for Predicting Peptides that Bind Major Histocompatibility Complex Molecules". *Meth. Mol. Biol.* 156 (2001): 201-209.

[38] K. Gulukota, J. Sidney, A. Sette and C. DeLisi. "Two Complementary Methods for Predicting Peptides Binding Major Histocompatibility Complex Molecules". *J. Mol. Biol.* 267 (1997): 1258-1267.

[39] K. Nakai. "Protein Sorting Signals and Prediction of Subcellular Localization". *Adv Protein Chem* 54 (2000): 277-344.

[40] K. Nakai. "Review: Prediction of in Vivo Fates of Proteins in the Era of Genomics and Proteomics". *J. Struct. Biol.* 134 (2001): 103-116.

[41] L. M. Iakoucheva, A. L. Kimzey, C. D. Masselon, J. E. Bruce, E. C. Garner, C. J. Brown, A. K. Dunker, R. D. Smith and E. J. Ackerman. "Identification of intrinsic order and disorder in the DNA repair protein XPA". *Prot. Sci.* 10 (2001): 560-571.

[42] M. A. Andrade, P. Chacón, J. J. Merelo and F. Morán. "Evaluation of Secondary Structure of Proteins from UV Circular Dichroism Spectra Using an Unsupervised Learning Neural Network". *Prot. Engin.* 6 (1993): 383-390.

[43] M. C. Honeyman, V. Brusic, N. L. Stone and L. C. Harrison. "Neural Network-Based Prediction of

Candidate T-Cell Epitopes". *Nat. Biotechnol.* 16 (1998): 966-969.

[44] N. Blom, J. Hansen, D. Blaas and S. Brunak. "Cleavage Site Analysis in Picornaviral Polyproteins: Discovering Cellular Targets by Neural Networks. *Prot. Sci.* 5 (1996): 2203-2216.

[45] N. Qian and T. J. Sejnowski. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models". *J. Mol. Biol.* 202 (1988): 865-884.

[46] O. Lund, K. Frimand, J. Gorodkin, H. Bohr, J. Bohr, J. Hansen and S. Brunak. "Protein Distance Constraints Predicted by Neural Networks and Probability Density Functions". *Protein Engineering*, 10, 1241-1248, 1997.

[47] P. Baldi, G.Pollastri. "Machine Learning Structural and Functional Proteomics". *IEEE Intelligent Systems* (Intelligent Systems in Biology II), March/April 2002.

[48] P. Baldi, J. Cheng, A. Vullo. "Large-Scale Prediction of Disulphide Bond Connectivity". In: *Advances in Neural Information Processing Systems* 17 (NIPS 2004), L. Saul, Y. Weiss and L. Bottou (editors), MIT Press, pp. 97-104, Cambridge, MA, 2005.

[49] P. Baldi and G. Pollastri. "The Principled Design of Large-Scale Recursive Neural Network Architectures—DAG-RNNs and the Protein Structure Prediction Problem", *Journal of Machine Learning Research*, 4, 575-602, 2003.

[50] P. Romero, Z. Obradovic, C. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot and A. K. Dunker. "Thousands of Proteins Likely to Have Long Disordered Regions". *Pac. Symp. Biocomput.* 3 (1998) 437-448.

[51] P. Romero, Z. Obradovic and A. K. Dunker. "Folding Minimal Sequences: the Lower Bound for Sequence Complexity of Globular Proteins". *FEBS Lett.* 462 (1999): 363-367.

[52] P. Wrede, O. Landt, S. Klages, A. Fatemi, U. Hahn and G. Schneider. "Peptide Design Aided by Neural Networks: Biological Activity of Artificial Signal Peptidase I Cleavage Sites". *Biochem.* 37 (1998): 3588-3593.

[53] R. Casadio, P. Fariselli, C. Taroni and M. Compiani. "A Predictor of Transmembrane ?-Helix Domains of Proteins Based on Neural Networks". *European Journal of Biophysics* (1994) submitted, 8/94.

[54] R. Casadio, P. Fariselli, C. Taroni and M. Compiani. "A Predictor of Transmembrane ?-Helix Domains of Proteins Based on Neural Networks". *European Journal of Biophysics* 24 (1996): 165-178.

[55] R. Gupta, E. Jung, A. A. Gooley, K. L. Williams, S. Brunak and J. Hansen. "Scanning the Available Dictyosteliumdiscoideum Proteome for O-linked GlcNAc Glycosylation Sites Using Neural Networks". *Glycobiology* 9 (1999): 1009-1022.

Fecha de recepción: 20 de febrero de 2007
Fecha de aprobación: 14 de agosto de 2007