Obtención de dimensión fractal para series de tiempo de expresión génica mediante una red neuronal artificial

Marco Antonio Esperón Pintos, Jorge Velázquez Castro, Benito de Célis Alonso

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas

marco.ep44@gmail.com
jorge.velazquezcastro@correo.buap.mx
bdca_buap@yahoo.com.mx

https://meet.google.com/nxf-gswf-ztu

Resumen

En este trabajo se utiliza un modelo dinámico estocástico de una red de regulación genética mínima para simular la dinámica característica de la concentración de proteínas. Al cambiar los parámetros del modelo se podrán simular distintas condiciones celulares que posteriormente serán de utilidad para el entrenamiento de algunas redes neuronales artificiales que reconozcan las propiedades de las series en el estado celular correspondiente. En particular se analizará el exponente de hurst y la dimensión fractal de la señal. El exponente de hurst es relevante en el diagnóstico pues determina la autocorrelación de las series de tiempo y permite etiquetar diferentes estados celulares. Esta investigación evalúa la eficiencia y factibilidad de emplear una red neuronal artificial que diagnostique estados celulares por medio de la dinámica de concentraciones proteicas.

Introducción

Las células son parte de un sistema complejo que es afectado por parámetros físicos como la temperatura y la presión osmótica. También perciben nutrientes y sustancias químicas perjudiciales. Las células responden a estas señales externas produciendo un tipo de proteína. Para representar estos estados ambientales, la célula utiliza como insignia unas proteínas llamadas factores de transcripción (FT). Los factores de transcripción [1, 2] están diseñados para transitar rápidamente entre estados moleculares activos e inactivos a una velocidad modulada por una señal ambiental específica. Cada factor de transcripción activo puede unirse al ADN para regular el ritmo de lectura de determinados genes. A la traducción y producción de proteínas se les conoce como expresión génica, es decir, el proceso de síntesis de proteínas a partir de la activación de un gen específico. Podemos concluir que la expresión génica es el proceso por el cual la información del genotipo da lugar al fenotipo (características observables).

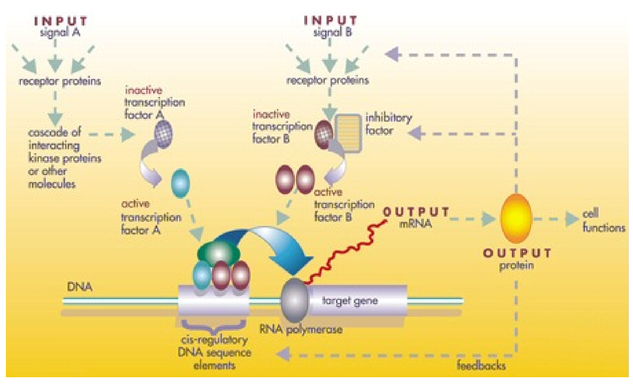

Figura 1: Red de regulación de expresión génica

Por lo tanto, para poder entender la dinámica de expresión génica, es necesario investigar las propiedades estadísticas de los datos que se miden experimentalmente (series de tiempo) [3].

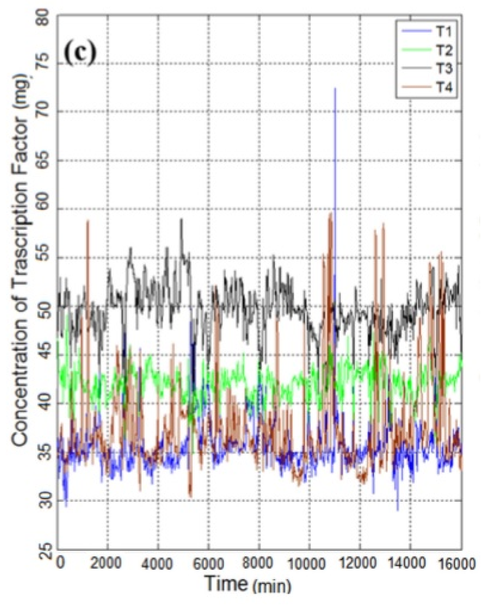

Figura 2: Series de tiempo de FT de E. Coli obtenidos de Marbach et al. (2012)

Diversos estudios han demostrado que las series de tiempo de los factores de transcripción tienen un comportamiento fractal, es decir que son invariantes ante cambios de escala. Estas series de tiempo cuentan con propiedades estadísticas particulares como el exponente de hurst y la dimensión fractal. El exponente de hurst es un parámetro acotado entre 0 y 1. El valor de este parámetro nos da una idea de la correlación que tiene cada serie de tiempo.

Exponente de Hurst

- $0 \leq H < 0{,}5$ **(Ruido Rosa)** Corresponde a un comportamiento de anti-persistencia o anti-correlación en la serie de tiempo (un periodo de crecimiento es seguido de otro de decrecimiento) que se caracteriza por un mayor contenido de alta frecuencia.
- $0{,}5 < H \leq 1$ **(Ruido Negro)** Series de tiempo que muestran procesos persistentes o correlacionados (un periodo de crecimiento es seguido de otro análogo) y presenta un aspecto "suave".
- $H = 0{,}5$ **(Ruido Blanco)** Proceso completamente aleatorio e independiente, con ausencia de correlación entre los incrementos de la señal.

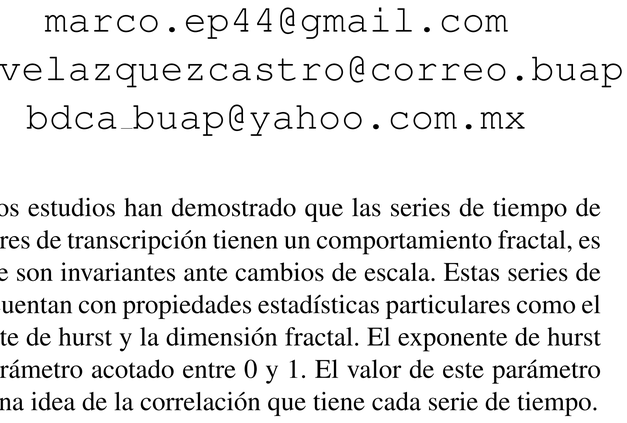

Figura 3: Series de tiempo con diferentes exponentes de hurst

Algunos estudios han demostrado que bacterias como E.Coli y hongos como S.Cerevisiae muestran un estado celular sano cuando las series de tiempo de los factores de transcripción tienen un exponente de hurst mayor a 0.5 [2].

Objetivos

- Simular series que emulen las series de tiempo de expresión génica.
- Implementar y entrenar redes neuronales artificiales que predigan el exponente de hurst de las señales (series de tiempo) generadas en las simulaciones.

Metodología

- Se elaboró un algoritmo capaz de generar series de tiempo basados en el Movimiento Browniano Fractal (emulando un modelo mínimo estocástico de la dinámica de concentración de proteínas en una célula) [4].
- Para la reproducción de los datos se computó un script con ayuda de la librería "FBM" que se encuentra indexada dentro del repositorio PyPI de Python. El método "fbm" de la librería "FBM" permite generar series cronológicas con diferentes exponentes de hurst.
- Se crearon 2 listas de series de tiempo, cada lista contiene 1000 series y cada serie representa una realización experimental conformada por 501 valores discretos generados por la función "fbm"mediante un dominio de 500 incrementos (pasos) equiespaciados que representan el tiempo. De las 2 listas, una de ellas contiene las series temporales, la siguiente contiene las series en un espacio de frecuencias.

Resultados

Hasta ahora se tiene un registro completo de tres redes neuronales dentro del dominio temporal y de frecuencias. La red neuronal densa y LSTM que se utilizaron cuentan con 8 neuronas en la capa de entrada, 8 neuronas en la capa intermedia y una neurona en la capa de salida. Para la red neuronal convolucional (CNN) se utilizaron 4 filtros o ventanas unidimesionales de 10 entradas que posteriormente se enlazan a una red densa de 8 neuronas en la capa intermedia y 1 neurona en la capa de salida. La función de costo utilizada fue el error absoluto medio y los resultados obtenidos se muestran en la siguiente tabla.

Tipo de red	Tiempo de computo para entrenamiento (segundos)	Error absoluto medio de los datos de prueba	Épocas	Dominio de las series
Densa	129.0	0.2	100	Temporal
Densa	226.3	0.13	100	Frecuencias
LSTM	675.9	0.061	20	Temporal
LSTM	673.1	0.026	20	Frecuencias
CNN	82.5	0.03	15	Temporal
CNN	30.9	0.35	15	Frecuencias
CNN	326.8	0.073	200	Temporal
CNN	333.1	0.38	200	Frecuencias

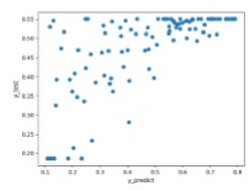

Figura 4.1: Datos predichos vs datos de prueba para una red neuronal Densa

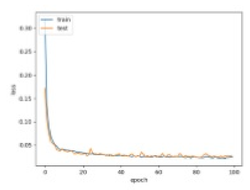

Figura 4.2: número de épocas vs Función de costo de una red neuronal Densa

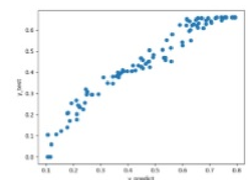

Figura 4.3: Datos predichos vs datos de prueba para una red neuronal Convolucional

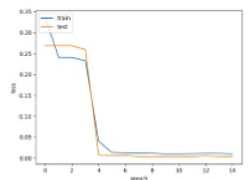

Figura 4.4: número de épocas vs Función de costo de una red neuronal Convolucional

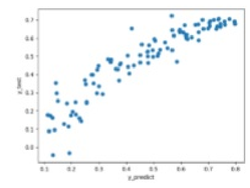

Figura 4.5: Datos predichos vs datos de prueba para una red neuronal LSTM

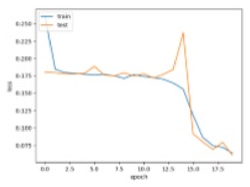

Figura 4.6: número de épocas vs Función de costo de una red neuronal LSTM

Figura 4

Conclusiones

Se puede concluir que las redes neuronales artificiales son eficientes para analizar series de tiempo de expresión génica y poder predecir el exponente de hurst. Considerando el tiempo de cómputo para entrenar a las redes neuronales y el error absoluto medio de cada una, se puede observar que la más eficiente es la red neuronal convolucional (CNN) al momento de ser entrenada con las series en el dominio temporal. Por lo tanto aparenta ser la mejor candidata para futuras predicciones para las cuales podría utilizarse más de una capa intermedia (deep learning).

Referencias

[1] Uri Alon. *An introduction to systems biology: design principles of biological circuits*. Second edition. Boca Raton, Fla: CRC Press, 2019. ISBN: 978-1-4398-3717-7 978-1-138-49011-6.

[2] Mahboobeh Ghorbani, Edmond A. Jonckheere y Paul Bogdan. "Gene Expression Is Not Random: Scaling, Long-Range Cross-Dependence, and Fractal Characteristics of Gene Regulatory Networks". En: *Frontiers in Physiology* 9 (oct. de 2018), pág. 1446. ISSN: 1664-042X. DOI: 10.3389/fphys.2018.01446. URL: https://www.frontiersin.org/article/10.3389/fphys.2018.01446/full (visitado 27-09-2021).

[3] *Serie temporal*. es. Page Version ID: 131528954. Dic. de 2020. URL: https://es.wikipedia.org/w/index.php?title=Serie_temporal&oldid=131528954 (visitado 28-09-2021).

[4] Lyudmyla Kirichenko, V. Bulakh y T. Radivilova. "Machine learning classification of multifractional Brownian motion realizations". En: *CMIS*. 2020.