



Universidad
Zaragoza

Trabajo Fin de Grado

Estudio y análisis de la red neuronal del gusano
Caenorhabditis elegans

Study and analysis of the *Caenorhabditis elegans*'
neural network

Autor

Javier Fumanal Idocin

Directores

Miguel Aguilera Lizarraga
Manuel González Bedia

Escuela de Ingeniería y Arquitectura
2017

Estudio y análisis de la red neuronal del gusano *Caenorhabditis elegans*



DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./D^a. Javier Fumanal Idocin

con nº de DNI 25205329C en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) Grado, (Título del Trabajo) Estudio y análisis de la red neuronal del gusano *Caenorhabditis elegans*

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, 19 de Junio de 2017

Jan
Dowd

Edo· Javier Fumana Idocin

1 RESUMEN DEL ESTUDIO Y ANÁLISIS DE LA RED NEURONAL DEL GUSANO *C.ELEGANS*

En este trabajo se ha analizado la actividad neuronal de una parte del cerebro de un gusano *C.elegans* en movimiento utilizando modelos de máxima entropía para simular el comportamiento del mismo. Aunque este tipo de modelos ya se han utilizado previamente en el estudio de otros datos de actividad neuronal, en este caso los datos se han extraído de un gusano en libre movimiento (Nguyen et al., 2016). Por ello, mientras que los modelos de máxima entropía generalmente capturan únicamente la distribución de probabilidades del sistema en estado estacionario, en este caso se han modelado las transiciones entre unos y otros, para poder capturar la dinámica del sistema (Tyrcha, Roudi, Marsili, & Hertz, 2013).

Para conseguir todo esto, primero se ha estudiado la forma óptima de discretizar los datos de actividad neuronal, buscando robustez frente a artefactos en las medidas originales. A partir de esta actividad, se ha generado un modelo de máxima entropía, siendo utilizado un modelo de Ising cinético. Debido al alto coste computacional de entrenar estos modelos, se han implementado varias técnicas para acelerar el proceso y para garantizar que el sistema llegue siempre a un error aceptable.

Existe un debate hoy en día acerca de si los tejidos neuronales funcionan en lo que en mecánica estadística se denomina un punto de criticalidad. Esto significaría que estos no funcionarían ni en un régimen de actividad ordenado, donde el comportamiento del sistema se estabiliza en patrones más estructurados; ni desordenado, donde cada neurona es más independiente de las demás. Los sistemas que operan en punto de criticalidad tienen características de ambas fases, y poseen una serie de propiedades adicionales (Salinas, 2001). Operar en punto de criticalidad permitiría al sistema optimizar el flujo de información que pasa por el mismo, lo que supondría para el individuo una ventaja evolutiva con respecto de otros con tejidos neuronales funcionando a otro régimen (Beggs, 2008).

Para buscar signos de criticalidad se ha estudiado el modelo generado en busca de: punto de máxima capacidad calorífica, comparación con una distribución de Zipf e índice de magnetismo para diferentes temperaturas. Se han encontrado signos positivos de criticalidad a la hora de estudiar la entropía de las muestras sintéticas en función de la temperatura del sistema, índice de magnetismo y distribución de Zipf. No obstante, no ha sido posible encontrar signos de divergencia de la capacidad calorífica, aunque la falta de capacidad computacional para crear muestras de tamaños significativos podría ser el motivo real de algunos de los resultados negativos.

Adicionalmente, se ha intentado reconstruir la topología de la red neuronal real a partir de los datos obtenidos y de las conexiones reales de todas las neuronas, lo que podría permitir la utilización de modelos más sencillo. También se ha creado un clasificador utilizando una red neuronal para detectar el comportamiento del gusano a partir de la actividad neuronal.

2 CONTENIDO

| | | |
|-------|---|----|
| 1 | Resumen del estudio y análisis de la red neuronal del gusano <i>C.elegans</i> | 4 |
| 3 | Introducción..... | 7 |
| 3.1 | Modelos de máxima entropía: Ising y criticalidad..... | 7 |
| 3.2 | Discretización de la actividad neuronal..... | 8 |
| 3.3 | Inferencia del modelo..... | 8 |
| 3.4 | Ánálisis de criticalidad | 9 |
| 3.5 | Otros análisis..... | 9 |
| 4 | Datos de la actividad neuronal | 10 |
| 4.1 | Picos de entropía | 11 |
| 4.2 | Discretización de datos | 12 |
| 4.2.1 | Filtro de paso alto | 12 |
| 5 | Inferencia del modelo..... | 14 |
| 5.1 | Modelo de Ising Cinético | 14 |
| 5.1.1 | Descripción cualitativa | 14 |
| 5.1.2 | Energía del sistema y activación de cada neurona | 14 |
| 5.1.3 | Entrenamiento del sistema..... | 15 |
| 5.2 | Resultados de la simulación..... | 19 |
| 6 | Ánálisis de criticalidad | 21 |
| 6.1 | Distribución de probabilidad..... | 21 |
| 6.2 | Entropía de muestras generadas | 22 |
| 6.3 | Magnetismo..... | 24 |
| 7 | Otros análisis..... | 26 |
| 7.1 | Reconocimiento del comportamiento..... | 26 |
| 7.1.1 | Validación de modelos..... | 26 |
| 7.2 | Reconstrucción de la red original | 27 |
| 7.2.1 | Detección de conexiones a partir de actividad neuronal | 29 |
| 7.2.2 | Construcción de la red | 30 |
| 8 | Conclusiones | 33 |
| 9 | Bibliografía..... | 34 |
| 10 | Anexos..... | 36 |
| 9.1 | Picos de entropía según umbral | 36 |
| 9.2 | Ánálisis de Valores Propios | 38 |
| 9.1 | Ánálisis de Zipf orientado a estados | 39 |
| 9.2 | Ánálisis de criticalidad orientado a estados | 40 |
| 9.3 | Transferencia de entropía a lo largo del tiempo | 42 |

3 INTRODUCCIÓN

El estudio de los sistemas neuronales naturales desde una perspectiva computacional comenzó en la segunda mitad del siglo XX (Hodgkin & Huxley, 1952). No obstante, los análisis y los resultados obtenidos siempre habían tenido que limitarse a trabajar con datos provenientes de simulaciones o de seres vivos en estado estacionario, y por tanto, muy lejos de un comportamiento normal. Sin embargo, recientemente se ha conseguido monitorizar la actividad del cerebro completo de un gusano *C.elegans* en movimiento, y captar la actividad de hasta 77 de las 302 neuronas que lo componen (Nguyen et al., 2016). Esto da la oportunidad de trabajar con datos reales de un ser vivo moviéndose a voluntad, y por tanto, mucho más representativos de la actividad real del sistema neuronal del sujeto de estudio.

Los modelos de máxima entropía se caracterizan por hacer el mínimo número de asunciones posibles sobre los datos, lo que permite ajustar el modelo a estos minimizando las asunciones a priori sobre los mismos. Este tipo de modelos han sido utilizados con éxito en otras áreas como el estudio de anticuerpos (Mora, Walczak, Bialek, & Callan, 2010) y de bandadas de pájaros (Bialek et al., 2012). Es posible encontrar ya resultados aplicados con Ising estático al mismo conjunto de datos utilizado en esta ocasión (Aguilera, Alquézar, & Izquierdo, 2017).

Sin embargo, esto implica a su vez que el modelo necesitará una muestra estadísticamente significativa para modelarlos correctamente, y el entrenamiento tendrá un alto coste computacional. Se han utilizado varias técnicas para acelerar todo lo posible este proceso y para minimizar la memoria necesaria para conseguir muestras significativas, o en caso de no poder hacerlo, buscar alternativas que no comprometan la validez de los cálculos.

A partir del modelo, se han estudiado diversos rasgos de criticalidad, y se ha hecho un estudio de la topología y del comportamiento dinámico del mismo para buscar posibles trazas de conexiones reales en la actividad del modelo y de la muestra original del gusano.

El trabajo ha sido realizado dentro del grupo Isaac Lab. El departamento facilitó un código inicial en Python que permitía crear y entrenar modelos de Ising cinéticos, cargar las muestras de los gusanos en memoria, y calcular la transmisión de entropía de dos secuencias. El resto del código desarrollado ha sido hecho tomando este como punto de partida. Todos las herramientas de análisis y modelos creados están disponibles de forma pública, y han sido diseñados tanto para ser utilizados en este mismo conjunto de datos como para cualquier otro.

Repositorio: <https://github.com/Fuminides/wormbrain>

El grueso del trabajo queda dividido en las siguientes secciones:

3.1 MODELOS DE MÁXIMA ENTROPÍA: ISING Y CRITICALIDAD

Los sistemas con muchos grados de libertad tienen un número de estados muy elevado, que crece exponencialmente con el tamaño del sistema. Esto hace que estimar satisfactoriamente la probabilidad de un estado concreto a partir de datos reales sea prácticamente imposible. La utilización del principio de máxima entropía resuelve en parte este problema, ya que asume un modelo que es tan aleatorio como sea posible, pero que impone algunas restricciones impuestas a partir de algunas medidas de las obtenidas en los datos (Mora & Bialek, 2011).

El modelo de Ising es un modelo físico propuesto para estudiar, en un principio, el comportamiento de materiales ferromagnéticos. No obstante, es posible utilizarlo también para otras áreas del conocimiento, como es en este caso.

Este se compone de un conjunto de N partículas, que pueden tener un espín a -1 ó +1, y que interactúan todas con todas, y un campo magnético externo. Interpretado dentro del ámbito de las redes neuronales, el modelo de Ising sería una red donde cada partícula es una neurona que tiene una conexión bidireccional con cada una de las otras, el campo magnético externo que afecta a cada elemento se traduciría en el sesgo de cada neurona, y el espín de cada elemento del sistema, +1 ó -1, sería el equivalente a que una neurona estuviese activada o no.

A determinadas temperaturas, es conocido que los materiales ferromagnéticos se imantan espontáneamente, pero si se calientan por encima del punto de criticalidad pierden su imantación. Durante décadas la fenomenología relacionada con los puntos de criticalidad ha sido ampliamente estudiada en mecánica estadística, debido a lo interesante de sus propiedades.

Las transiciones de fase y el punto de criticalidad son fenómenos ampliamente estudiados (Salinas, 2001), al igual que el modelo de Ising en sí. Una diferencia notable entre el modelo estático y el cinético, es que el primero tiene solución analítica bajo ciertas condiciones, mientras que el segundo, no. Esto hace necesario utilizar técnicas no analíticas, como descenso de gradiente, para poder resolver el sistema.

3.2 DISCRETIZACIÓN DE LA ACTIVIDAD NEURONAL

A pesar de que el modelo necesita la actividad neuronal discretizada, las neuronas sólo pueden estar apagadas o excitadas, los datos reflejan la fluorescencia de una neurona en cada momento (valor real). Escoger un umbral adecuado para decidir cuáles están apagadas y cuáles no, es crucial para que el resto de resultados sean fiables, y un valor adecuado no es tan evidente como sería deseable.

Diversos criterios han sido aplicados, dando resultados dispares cada uno de ellos. Para eliminar artefactos surgidos al realizar las medidas originales, se ha utilizado un filtro de paso alto.

3.3 INFERENCIA DEL MODELO

Aunque la implementación para crear y entrenar uno de estos modelos estaba facilitada por el propio departamento, existían problemas más allá de la implementación en sí del sistema a resolver:

- El número de muestras para cada gusano es muy distinto e insuficiente.
- El coste computacional y en memoria de entrenar un sistema de estas características es muy elevado.
- El modelo de Ising cinético puede modelar transiciones de estados en varias escalas temporales, hay que escoger la que mejor capture las interacciones entre las neuronas.

3.4 ANÁLISIS DE CRITICALIDAD

Existen resultados que indican que los tejidos neuronales operan en punto crítico o cerca del mismo (Tetzlaff, Okujeni, Egert, Wrgtter, & Butz, 2010), al igual que otros sistemas que se pueden encontrar en la naturaleza. Si el sistema se encuentra en dicho estado, esto podría presentar algunas ventajas para maximizar el flujo de información entre ellas (Beggs, 2008). Los sistemas operando en punto de criticalidad poseen además otras características: máxima sensibilidad a estímulos externos, presencia de correlaciones de largo alcance, etc. (Salinas, 2001)

Los indicadores de criticalidad estudiados son los siguientes:

1. En el límite termodinámico, el régimen de actividad tiende a seguir una distribución de Zipf.
2. En el límite termodinámico, la capacidad calorífica del sistema tiende a infinito. Si el sistema cumple esta propiedad, el sistema está necesariamente en punto de criticalidad, pero su no cumplimiento no prueba que no lo esté.
3. El índice de magnetismo a partir del punto de criticalidad es muy cercano a 0.

3.5 OTROS ANÁLISIS

Con los datos obtenidos del modelo y de la muestra original, se realizan:

- Un clasificador que permita reconocer el comportamiento del gusano a partir de su actividad neuronal.
- Una reconstrucción de las conexiones reales de las neuronas a partir de los datos originales y el modelo generado.

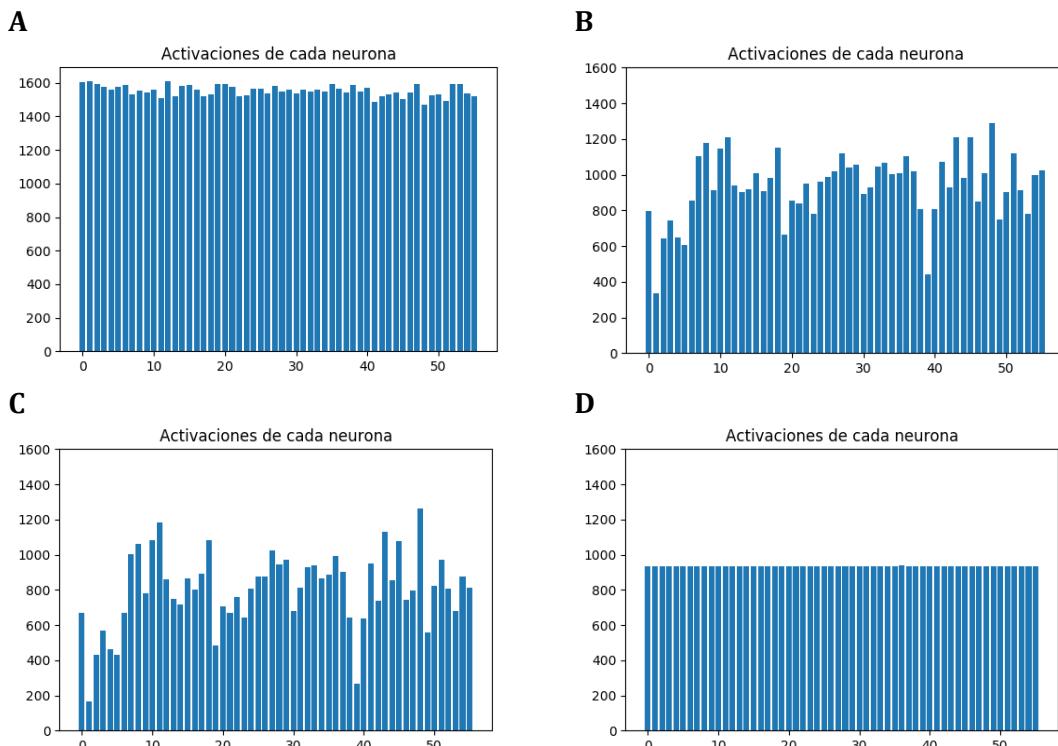


Figura 1. Estudio del número de activaciones de las neuronas del gusano 1. Umbral igual a 0 (A), mediana de todas las neuronas (B), media de todas las neuronas (C) y mediana de cada neurona (D). La amplia diferencia de activaciones entre los distintos métodos sugiere que un análisis más profundo es necesario para encontrar un umbral fiable.

4 DATOS DE LA ACTIVIDAD NEURONAL

En esta sección se explican el proceso y las técnicas empleadas para discretizar los datos de actividad neuronal originales, con el objetivo de discernir qué neuronas están apagadas y cuáles están excitadas.

Los datos de actividad neuronal han sido obtenidos de una serie de gusanos *C.elegans* moviéndose a tiempo real que han sido modificados genéticamente para convertir a sus neuronas en fluorescentes con el calcio. De esta manera, es posible medir las transmisiones del mismo que se producen entre estas debido a su brillo, lo que permite medir la actividad neuronal del gusano (Nguyen et al., 2016).

Esto presenta una serie de inconvenientes a la hora de modelar el sistema, pues la información necesita ser discretizada para poder trabajar con ella. Es preciso buscar un umbral que permita decidir si una neurona está activada o no. Elegir este umbral no es una tarea trivial, y elegir un valor u otro puede dar lugar a resultados muy distintos (Figura 1).

Diferentes soluciones al problema de la elección del umbral han sido estudiadas:

1. Utilizar un valor fijo precalculado.
2. Utilizar la media de todas las neuronas.
3. Utilizar la mediana de todas las neuronas.
4. Utilizar la media de cada neurona como su umbral.
5. Utilizar la mediana de cada neurona con su propio umbral.
6. Utilizar algún tipo de filtro de la actividad neuronal.

| Gusano | Neuronas | Muestras | Entropía Estados | Entropía Transiciones |
|--------|----------|----------|------------------|-----------------------|
| 1 | 56 | 1872 | 4.98 | 6.74 |
| 2 | 77 | 1516 | 10.43 | 6.33 |
| 3 | 68 | 398 | 3.89 | 5.32 |
| 4 | 57 | 999 | 2.92 | 3.92 |
| 5 | 63 | 2055 | 4.12 | 5.7 |

Tabla 1. Estudio de la entropía de los estados y de las transiciones entre estados para cada gusano (igual a entropía de pares estados de consecutivos), utilizando el filtro de paso alto y con umbral 0, con la fórmula de Shannon. Bien sea con la entropía de los estados del gusano, o con sus transiciones, que parecen tener un comportamiento más robusto; esta no crece ni con el número de muestras ni con el número de neuronas del gusano, lo que podría indicar que la actividad del gusano se mueve en un espacio muy pequeño de todo el espacio de estados posible.

Para elegir un umbral se sigue el criterio de utilizar aquel que maximice la entropía de los estados resultantes del gusano. Se utiliza este criterio ya que un umbral mal escogido, bien sea muy bajo o muy alto, tiende a llevar a esta a 0. Luego, por contraposición, se sospecha que el umbral óptimo debería aumentarla todo lo posible.

4.1 PICOS DE ENTROPÍA

El primer paso a realizar es estudiar la entropía que presenta cada gusano para cada umbral, buscando un pico común de la misma para todos ellos.

La entropía es una magnitud que mide la incertidumbre de una fuente de información y que permite relacionar cada símbolo con la cantidad real de información que contiene. Para calcularla, se ha utilizado la fórmula de Shannon:

$$S = - \sum p_i \log_2 p_i \quad (1)$$

Sin embargo, el cálculo de la entropía utilizando esta fórmula presenta un problema de significancia estadística para las muestras a analizar. Para obtener un conjunto significante que esté caracterizado por n dimensiones, se necesita disponer de un orden de magnitud similar a 2^n muestras.

Siendo que en ningún caso se dispone de una muestra del tamaño adecuado (Tabla 1), se utiliza una técnica alternativa para calcular dónde se encuentra el pico de entropía: en vez de calcular la entropía para todas las muestra entera, se calculan las entropías de todas las 2-combinaciones de neuronas, y luego se suman.

Aunque esta técnica no da una medida real de la entropía, sí que es posible utilizarla para compararla en cada umbral de cada gusano, lo que nos permite tener una idea de cómo de similares son los gusanos entre sí y de dónde debería encontrarse el umbral correcto (Figura 16).

Un estudio de los picos de entropía de cada gusano para un rango de umbrales de 0 a 2, indica que el mejor umbral orbita en torno a 0.15 (Figura 18). Este estudio indica que todos los gusanos responden de forma parecida a los mismos umbrales, y la ausencia de picos adicionales indica que no existen comportamientos extraños. No obstante, estos picos, aunque en posiciones parecidas, tienden a moverse en cada gusano. Esto parece indicar que la medida óptima para escoger el umbral no es un valor fijo, sino que debe ser alguna medida única para cada gusano.

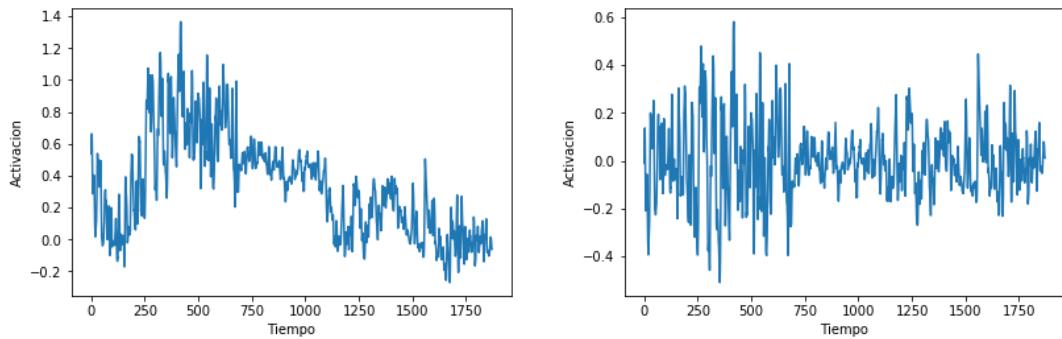


Figura 2. Activaciones de una neurona. A la izquierda sin utilizar ningún filtro, a la derecha con un filtro de paso alto. El uso del filtro permite eliminar los artefactos creados por el movimiento del gusano.

4.2 DISCRETIZACIÓN DE DATOS

Dado que un valor precalculado parece una mala solución, se ha procedido a estudiar las siguientes técnicas de umbralización:

1. **Media global y media de cada neurona:** aunque en teoría son unas buenas medidas, ya que colocan el umbral en un punto siempre cercano al 0.15 calculado con anterioridad, son poco resistentes a valores atípicos y a artefactos que sabemos que están presentes en los conjuntos de datos.
2. **Mediana global:** se mueve en torno a los valores esperados y es más resistente a datos atípicos y a artefactos que la media. Es en principio una buena medida, aunque aún es mejorable.
3. **Mediana de cada neurona:** es en principio resistente a datos atípicos y artefactos. Sin embargo, si se escoge la mediana de cada neurona como su umbral, necesariamente estará activada y desactivada la mitad de la veces; lo que supone introducir un sesgo innecesario.

4.2.1 Filtro de paso alto

La aplicación de un filtro de paso alto ha sido estudiada previamente para este conjunto de datos (Aguilera et al., 2017) con resultados positivos. El filtro de paso alto permite eliminar la mayoría de los artefactos encontrados en el set de datos, que son inducidos por el movimiento del gusano, que hace que las neuronas se expandan o contraigan, modificando de forma artificial las medidas resultantes (Figura 2).

Una vez aplicado el filtro, es posible utilizar cualquiera de las técnicas anteriores de umbralización, o sencillamente utilizar 0 como umbral.

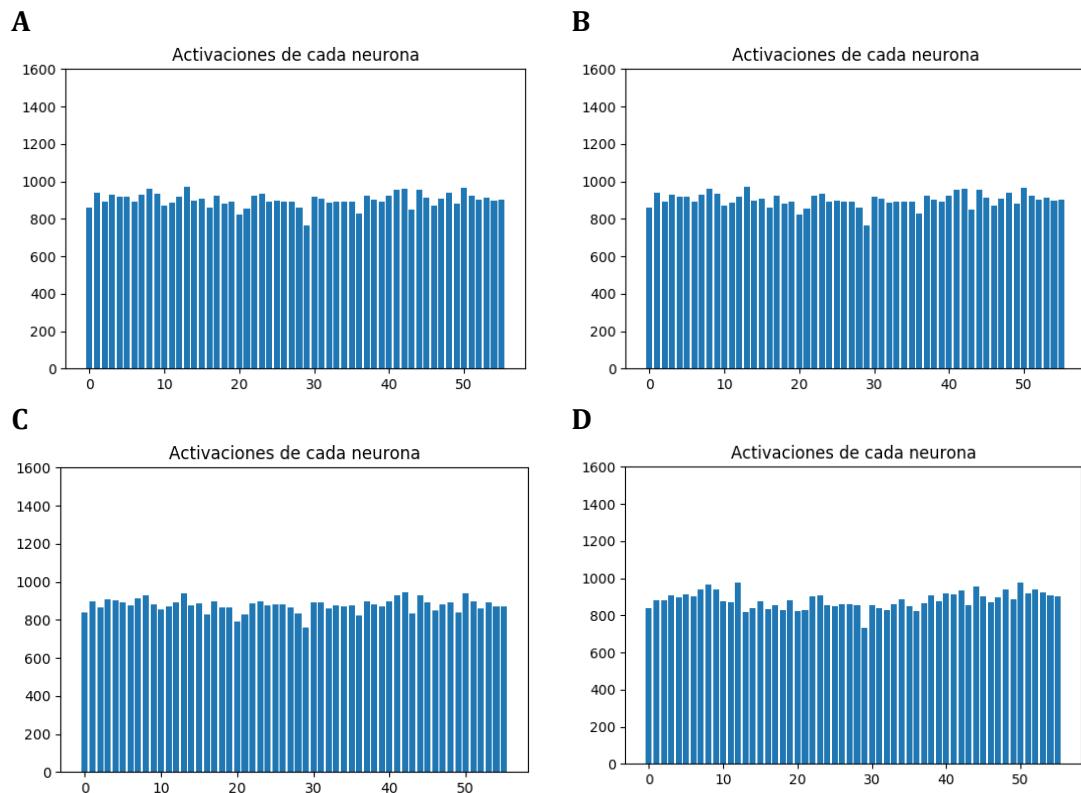


Figura 3. Estudio del número de activaciones de las neuronas del gusano 1, después de aplicar un filtro de paso alto. Umbral igual a 0 (A), mediana global (B), media global (C) y media de cada neurona (D).

El filtro de paso alto mejora dramáticamente los resultados obtenidos, ya que al eliminar los artefactos del movimiento del gusano, todas las técnicas estudiadas dan resultados muy similares (Figura 3). Esto hace que elegir cualquiera de las mismas sea una decisión secundaria.

Finalmente, se ha escogido como umbral la aplicación de un filtro de paso junto con la mediana global. No obstante, en la práctica, la elección entre uno y otro una vez aplicado el filtro es prácticamente nula.

5 INFERENCIA DEL MODELO

Para simular el comportamiento del *C.elegans* se ha utilizado un modelo de máxima entropía, Ising cinético. Modelos de máxima entropía ya han sido aplicados en situaciones similares y con resultados satisfactorios, bien sea en el caso de una salamandra (Tkačik et al., 2015) o incluso para este mismo conjunto de datos (Aguilera et al., 2017). Este modelo no tiene solución analítica, y por tanto, es necesario utilizar descenso de gradiente para resolver el sistema.

Utilizar el modelo cinético permite capturar la dinámica del sistema, ya que permite modelar las transiciones del modelo en lugar de la probabilidad de cada estado. No obstante, tiene problemas añadidos: es necesario escoger un espacio de tiempo adecuado entre el estado actual y el anterior del gusano, de manera que se evite coger transiciones sin completar de un estado a otro, y el entrenamiento del modelo es muy costoso computacionalmente.

5.1 MODELO DE ISING CINÉTICO

5.1.1 Descripción cualitativa

El modelo consiste en N neuronas conectadas todas con todas de forma bidireccional. Cada neurona posee únicamente dos estados o espines: +1 ó -1. El estado de cada neurona a lo largo del tiempo depende de su interacción con las demás y de la temperatura del sistema. Cada neurona está caracterizada por un sesgo y un peso para cada conexión con cada otra neurona.

En este sistema no se modela la probabilidad de cada estado, $s(t)$, sino la probabilidad de las transiciones de estos estados entre un momento y otro: $s(t+T|t)$, donde T es el intervalo de tiempo de distancia entre un estado y otro.

5.1.2 Energía del sistema y activación de cada neurona

Cada neurona del sistema recibe una entrada que se calcula a partir de la fórmula:

$$H_i(t) = h_i(t) + \sum J_{ij} s_j(t) \quad (2)$$

Donde:

- J_{ij} : es el peso de la conexión que va de la neurona i a la j
- s_j : es el valor de la neurona j -ésima en el momento t .
- h_i : es el sesgo de la neurona i -ésima en el momento t .

La probabilidad de que una neurona esté activada queda entonces definida por las siguientes fórmulas:

$$f(x) = \frac{1}{1 + e^{\frac{-2x}{T}}} \quad (3)$$

$$\Pr(s_i(t+T) = 1 | \{s_j(t)\}) = f(H_i(t)) \quad (4)$$

Donde:

- Pr : es la probabilidad de que una neurona esté activada.
- T : es una constante artificial que representa la temperatura del sistema.
- T : es el intervalo de tiempo de diferencia entre un estado y otro.

Como puede verse, las neuronas tienden a aceptar con mayor probabilidad cambios que minimicen su energía, y por tanto, que disminuyan la energía total del sistema. La temperatura del sistema afecta a la velocidad de estos cambios, ya que una mayor temperatura hace que las neuronas cambien de signo con más frecuencia.

5.1.3 Entrenamiento del sistema

Para entrenar el modelo de forma correcta, se requiere de una cantidad de muestras elevada, debido a la dimensionalidad del problema. Del modelo hay que calcular un sesgo para cada neurona, y un peso para cada conexión. Esto significa que en un sistema de N neuronas hay que calcular N sesgos y N^2 pesos de las conexiones. Como este proceso se realiza en cada iteración del descenso de gradiente, hay que tener en cuenta que es necesario recalcular todos estos en cada una. Además, existe cierta relación (que no ha sido posible precisar con exactitud) entre el número de neuronas y el número de iteraciones necesarias para entrenar de forma satisfactoria el sistema, lo que aumenta muchísimo el tiempo de entrenamiento en modelos más grandes de los aquí estudiados.

Para entrenar el sistema se utiliza descenso de gradiente. Para empezar el proceso, todos los pesos se igualan a 0, los sesgos se calculan de acuerdo con el modelo independiente, la temperatura se establece en 1 (arbitrario) y el estado de las neuronas del sistema es aleatorio.

Este entrenamiento tiene un enfoque distinto a un entrenamiento supervisado o no supervisado. El objetivo no es crear un clasificador, ni hacer un agrupamiento, tampoco hay un único valor como salida. En este caso, lo que se busca es que las muestras que genera el sistema se parezcan a la original.

Los sesgos de cada neurona se calculan a partir de la media de la muestra y los sesgos de cada conexión se calculan a partir de las correlaciones entre cada neurona. Al utilizar Ising cinético, estas correlaciones se calculan utilizando las neuronas en el estado anterior $t-T$.

Los gradientes de los pesos y las medias se calculan como:

$$\delta J_{ij} = \mu \langle [s_i(t + T) - \tanh(H_i(t))] s_j(t) \rangle_t \quad (5)$$

$$\delta h_i = \mu \langle s_i(t + T) - \tanh(H_i(t)) \rangle_t \quad (6)$$

Donde:

- δJ_{ij} : es el gradiente del peso de la conexión de la neurona i -ésima a la j -ésima.
- δh_i : es el gradiente del sesgo de la neurona i -ésima.
- μ : es el factor de aprendizaje.

Para calcular el error actual del sistema, se calculan las diferencias entre las correlaciones y medias actuales del sistema y las de la muestra. A continuación bien se puede escoger la máxima diferencia encontrada, bien sea en la matriz de muestras o de medias; o el máximo entre el error medio en las correlaciones y el de las medias. Ambos métodos ofrecen resultados similares, pero el primero es más conservador y devuelve siempre el mayor error posible.

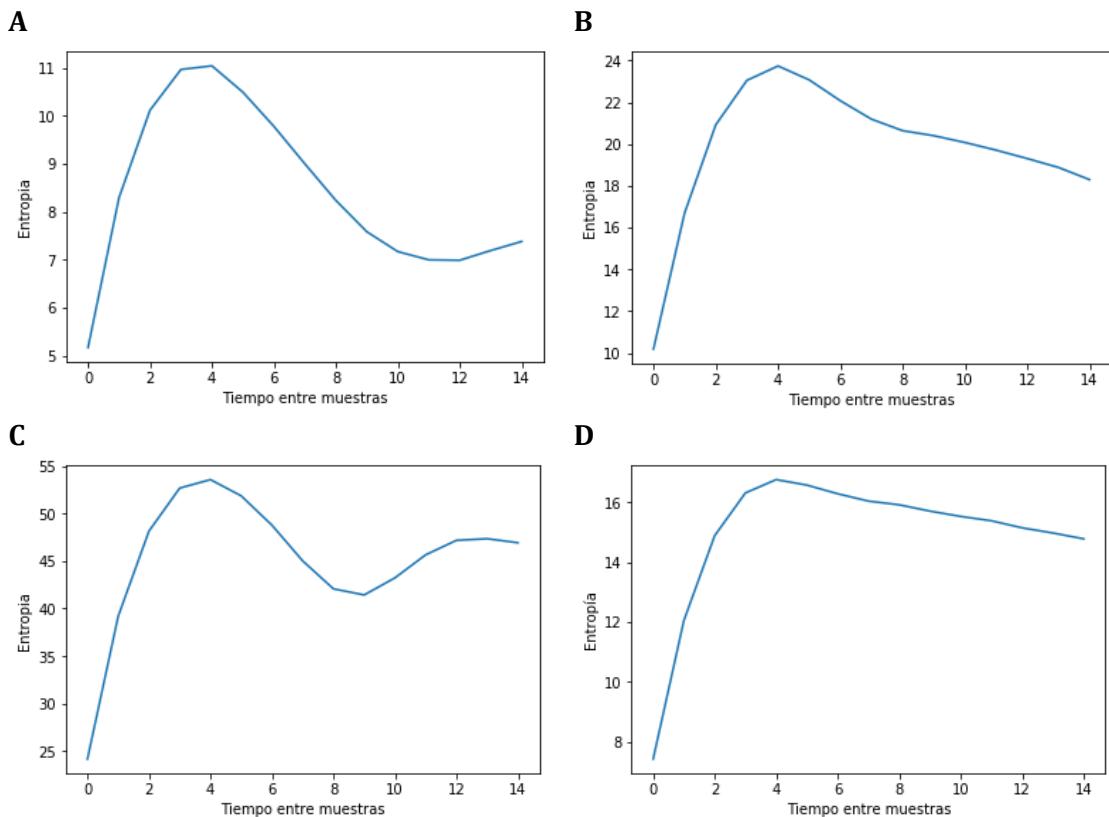


Figura 4. La suma total de la transmisión de entropía entre pares de neuronas en distintos intervalos de tiempo. Gusano 1 (A), gusano 2 (B), gusano 3 (C) y gusano 4 (D). Como se puede observar en los cuatro casos, existe un pico entorno a $T = 4$ en los 4 casos.

5.1.3.1 Elección de $t+T$

En el modelo utilizado, la probabilidad de que una neurona tenga su espín a +1 ó -1 viene determinado por el estado del sistema en un momento anterior (Ecuación 4). En los datos de activación neuronal, cada medida está tomada con 0.15 segundos de diferencia, pero ese tiempo no tiene por qué tener ningún significado real en el gusano. Es posible que este tiempo sea demasiado grande, y que por tanto, no se puedan observar interacciones de interés entre neuronas, o bien que sea demasiado pequeño.

Para estimar el intervalo de tiempo óptimo, se utilizan los flujos de entropía entre neuronas. Esta magnitud nos indica la cantidad de información que se transmite de una neurona a otra. El cálculo de los flujos en entropía sigue la siguiente fórmula:

$$H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(x|y) \log p(y|x) \quad (7)$$

$$T_{X \rightarrow Y} = H(Y_t|Y_{t-1:t-L}) - H(Y_t|Y_{t-1:t-L}, X_{t-1:t-L}) \quad (8)$$

Donde:

- X, Y: la serie de activaciones de dos neuronas.
- $H(Y|X)$: es la fórmula de Shannon de la entropía condicionada.

Se estudia la transferencia de entropía para distintos intervalos de tiempo (Figura 4). Con lo que se espera encontrar un pico en un intervalo de tiempo que nos permita maximizar la información que recibimos entre el estado t y $t+T$. Escoger un valor de T donde estos flujos fueran mínimos haría que las activaciones neuronales en un momento t y otro fueran muy parecidas, y cada neurona tendería a darle demasiada importancia a sus autoconexiones.

| Número de neuronas | Tiempo Entrenamiento |
|--------------------|----------------------------|
| 4 | 0.44 segundos |
| 8 | 3.72 segundos |
| 12 | 13.13 segundos |
| 20 | 26.75 segundos |
| 28 | 47.31 segundos |
| 36 | 1 minuto y 7.17 segundos |
| 56 | 55 minutos y 9.58 segundos |

Tabla 2. Tamaño del modelo y tiempo de entrenamiento. Tres variables afectan principalmente al mismo: número de neuronas, número de conexiones (que depende directamente del anterior) y número de iteraciones. El número de iteraciones necesarias en el sistema para llegar al error que buscamos también depende del tamaño del sistema, lo que provoca un aumento muy pronunciado del tiempo de entrenamiento para el sistema completo.

Después de analizar los datos de 4 de los gusanos, y de un modelo creado inicialmente con $T = 1$, se estima que el mejor valor para este parámetro es $T=4$. No obstante, el sistema es capaz de ofrecer buenos resultados también con valores menores que este, y para los análisis aquí realizados, la diferencia entre uno y otro es muy pequeña.

5.1.3.2 Mejorar el rendimiento del entrenamiento

Como se ha mencionado anteriormente, el coste computacional de entrenar todo el sistema es muy elevado. Los tiempos de ejecución crecen de forma muy rápida a medida que el número de neuronas aumenta (Tabla 2). Es también importante para el tiempo de cómputo el número de muestras de las que se dispone para entrenar el modelo, el error que se busca y las correlaciones presentes en los datos de entrenamiento (a más correlacionados los datos, más rápido el entrenamiento). Por ejemplo, el gusano 1 teniendo un número de neuronas muy similar al del gusano 4 es mucho más fácilmente entrenable que este último, en el que no ha sido posible llegar a un resultado satisfactorio.

Aunque no es posible reducir el coste teórico del entrenamiento del sistema, sí que es posible adoptar diversas estrategias de optimización para disminuir un porcentaje del coste de ejecución del algoritmo.

1. **Vectorizar todas las operaciones posibles:** esto significa minimizar bucles y sustituirlos por operaciones facilitadas por la librería *Numpy* siempre que sea posible. Además, se han implementado algunas operaciones matriciales con la librería *Numexpr*, no obstante, se desaconseja su uso para sistemas que no sean de tamaño muy elevado, ya que el coste de crear los hilos es mayor que la ganancia de utilizarlo en operaciones matriciales pequeñas.
2. **Utilizar Numba:** Numba es una librería para Python que permite convertir código Python en código de máquina virtual de bajo nivel (LLVM), lo que permite eliminar el intérprete en tiempo de ejecución en las funciones más costosas, además de utilizar código más eficiente. Para utilizar Numba en las funciones más costosas únicamente hay que añadir el decorador `@jit` a estas. Utilizar Numba restringe algunas de las posibilidades que ofrece Python en cuanto a tipado de variables, pero en condiciones normales apenas requiere modificaciones con respecto de un código que no la utilice.

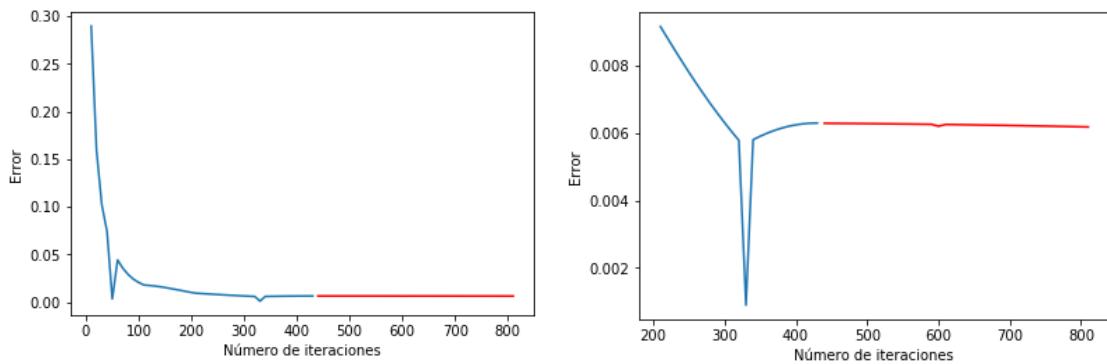


Figura 5. Evolución del error del sistema durante las primeras iteraciones de un entrenamiento. En azul, se entrena con factor de aprendizaje 0.1, y en rojo, se ha reducido a 0.01. La imagen de la derecha representa la evolución global, la de la izquierda es un zoom al momento en el que este se reduce. En el momento en el que el sistema detecta que el error ha subido por tercera vez, espera a que termine de subir para reducir el coeficiente de aprendizaje. Luego, el sistema sigue descendiendo hasta 0.001 (sólo se muestra una pequeña parte del entrenamiento).

3. **Maximizar el coeficiente de aprendizaje:** permite no sólo acelerar el entrenamiento, sino que asegura que el entrenamiento acabará en el error buscado: mientras el entrenamiento avanza (Ecuaciones 5 y 6), se buscan síntomas que indiquen que este factor no es adecuado, y en ese caso se reduce en un orden. Es importante tener en cuenta que cuando el sistema se ha saltado un mínimo local y el error comienza a subir, este lo hace independientemente de los ajustes que se realicen mientras sube, luego, si se detecta que es necesario una disminución del coeficiente de aprendizaje se espera a que el sistema termine de subir para que esta inevitable subida sea lo más rápida posible. Para decidir si disminuir el coeficiente de aprendizaje se utilizan dos criterios:

- El sistema ha convergido antes de tiempo:** esto quiere decir que el sistema ha convergido a un valor, pero este es mayor que el error que queremos, probablemente porque tiene un mínimo local próximo al que se acerca siempre, pero dando pasos demasiado grandes. Para decidir si el sistema se encuentra en esta situación, simplemente hay que comprobar los últimos errores y ver si la proporción entre errores adyacentes es pequeña, y si el número de aumentos y de disminuciones en el error es parejo. Esta situación, aunque posible, sólo ha sucedido 2 veces en los, aproximadamente, 150 entrenamientos del modelo realizados.
- El sistema ha divergido:** de una forma similar al caso anterior, se utilizan los últimos 100 errores para comprobar que: los errores sufriendo aumentos/disminuciones grandes y si existen varios aumentos del error. Se espera a la detección de tres aumentos del error para modificar el coeficiente de aprendizaje, ya que es posible que el sistema aumente el error pero se recupere rápidamente. Para cambiar el coeficiente de aprendizaje, se espera siempre a que los dos últimos errores indiquen que este está descendiendo, para evitar tener que subir con el coeficiente reducido (Figura 5).

En caso de no ser esto suficiente, se han implementado una serie de herramientas para reducir la dimensionalidad de la muestra, y poder utilizar un sistema más pequeño se han implementado varias técnicas: PCA, selección según correlación con el comportamiento, y elección aleatoria de neuronas.

Tanto utilizar PCA como un número determinado de neuronas aleatorias han dado buenos resultados, y son especialmente útiles para hacer pruebas en un modelo pequeño y fácil de entrenar, para luego pasar a uno de dimensionalidad completa. Escoger las neuronas por su parecido al comportamiento del gusano, en cambio, está desaconsejado, y el sistema reducido se comporta de forma muy distinta al de dimensionalidad real.

A pesar de las optimizaciones hechas, debido a la alta dimensionalidad y a la falta de muestras, sólo ha sido posible entrenar de forma satisfactoria el modelo para el gusano 1, y por tanto, el análisis del modelo se centra únicamente sobre el calculado a partir de este.

5.2 RESULTADOS DE LA SIMULACIÓN

Para comprobar si nuestro sistema funciona de forma correcta, se comprueba que las muestras sintéticas que genera este son parecidas a la muestra original, incluso cuando el tiempo que lo dejamos simular es alto y genera una muestra mucho mayor en tamaño que esta. Las medidas a comparar son correlaciones entre neuronas y medias de cada una de ellas, que nos dan una descripción buena de los datos.

El modelo se ha inferido con dos T distintas:

- T=1: es la aproximación más sencilla. Modelamos la probabilidad de cada estado a partir del anterior, espaciados 0.15 segundos en el tiempo. En principio, este modelo debería hacer que el peso mayor de la conexión de cada neurona sea consigo misma, lo que es una propiedad no deseable.
- T=4: este es el T calculado como el pico de la suma de las entropías de cada conexión entre cada par de neuronas y por tanto, es el momento en el que más información fluye entre las neuronas. Están espaciados 0.6 segundos en el tiempo.

Los modelos se entrena hasta obtener un error del 0.001. Los resultados obtenidos, tanto con T=1 como con T=4, entran dentro de lo esperado. Ambos generan muestras con correlaciones y medias de neuronas similares a los de la muestra inicial, independientemente del tamaño de la muestra que deban generar y del tiempo que lleven ya generando muestras.

Al aumentar el tiempo entre un estado y otro, la magnitud de las correlaciones desciende y las de t+4 son de media el 30% de sus homólogas en t+1. Además, la caída es más pronunciada en la diagonal, donde la suma de los pesos de las conexiones consigo mismas cae hasta reducirse a un 25% (Figura 6). Esto ocurre porque al medir los estados con tan poco tiempo de margen en t+1 se ven muy pocos cambios, y por tanto, la mayor parte de las neuronas están inalteradas respecto de un momento a otro.

Es importante notar que parece que para el modelo es más difícil modelar las transiciones entre estados más distantes en el tiempo que entre los cercanos, ya que los primeros son más parecidos entre ellos que los segundos. Un aumento del tamaño de la muestra de entrenamiento probablemente mejoraría los resultados para T=4.

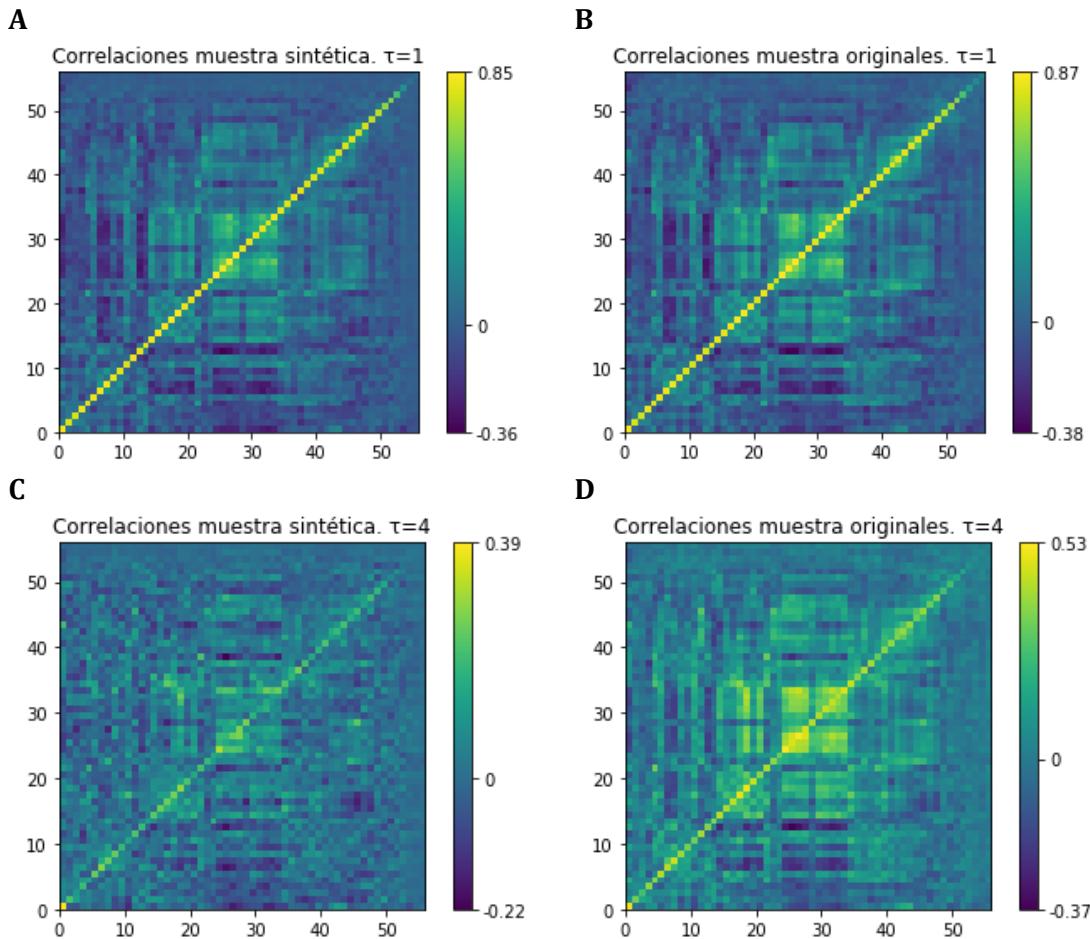


Figura 6. Correlaciones muestra calculadas en función $t+1$ (B y D) y $t+4$ (A y C). La fila de arriba corresponde con la muestra original, y la inferior con muestras generadas a partir de un modelo. Como se puede observar, las correlaciones con $T=4$ decrecen de media un 70% con respecto a $T=1$, y un 75% en la diagonal. Aumentar el tiempo entre las transiciones de estados a modelar permite eliminar correlaciones que aparecen de forma artificial al tomar estados demasiado cercanos en el tiempo entre sí.

Además, las correlaciones de la diagonal siguen siendo un 25% aproximado de todas las correlaciones en ambos casos. Una hipótesis para explicar este comportamiento, es que en la red neuronal real las activaciones de cada neurona no dependen de un momento pasado concreto, sino de un intervalo de tiempo más largo. Esto se traduciría en que nuestro sistema debería modelar la probabilidad de un estado $P(t+T)$ no únicamente en función de $P(t)$ sino que tendría que tener en cuenta también $P(t+T-1)$, $P(t+T-2)$, etc. Sin embargo, esto implicaría construir un modelo más complejo que el aquí estudiado, y por tanto, no entra dentro de los objetivos de este trabajo.

6 ANÁLISIS DE CRITICALIDAD

El punto de criticalidad de un sistema es aquel en el que el sistema se encuentra justo en la transición de una fase a otra, y por tanto, presenta propiedades de ambas. La temperatura del sistema a la que se encuentra el punto de criticalidad es denominada comúnmente como temperatura de Curie: T_c .

El punto crítico de un sistema se alcanza cuando un cambio de temperatura infinitamente pequeño produce en el sistema un cambio de fase en el mismo. En este caso, la función que describe el modelo tiene un cambio de fase de segundo orden entre la fase ordenada, donde el comportamiento del sistema tiende al determinismo, y la desordenada, donde tiende al azar.

Para comprobar si cerebro del *C.elegans* opera en un punto crítico, se pueden realizar distintos análisis. Siendo que el tamaño de las muestras de las que se dispone no es muy elevado, se realizan varios de estos con el objetivo de conseguir resultados más concluyentes. Es posible estudiar la distribución de probabilidades tanto en el modelo generado como en la muestra original (aunque con las limitaciones que tiene el tamaño de la misma), mientras que el índice de magnetismo y la capacidad calorífica sólo se pueden estudiar a partir del modelo.

6.1 DISTRIBUCIÓN DE PROBABILIDAD

La distribución de probabilidad de los distintos estados de un sistema operando en punto de criticalidad sigue una distribución de Zipf, que queda definida por esta fórmula:

$$P(n) = \frac{1}{n^a} \quad (9)$$

Donde:

- $P(n)$: es la probabilidad del estado n -ésimo, estando estas ordenadas de mayor a menor.
- a : es una constante elegida arbitrariamente.

Para comprobar si nuestro modelo funciona siguiendo esta distribución, aproximamos el parámetro a de una distribución de este tipo a una muestra sintética generada con el modelo y comprobamos la bondad del ajuste realizado. Ya que el modelo creado no modela la probabilidad de cada estado, sino la de las transiciones entre ellos, se ha calculado la probabilidad de cada transición entre uno y otro para mantener la coherencia con el análisis realizado en el modelo. Para hacer esto, lo único necesario ha sido modelar no la probabilidad de cada estado, sino de cada par consecutivo de estos.

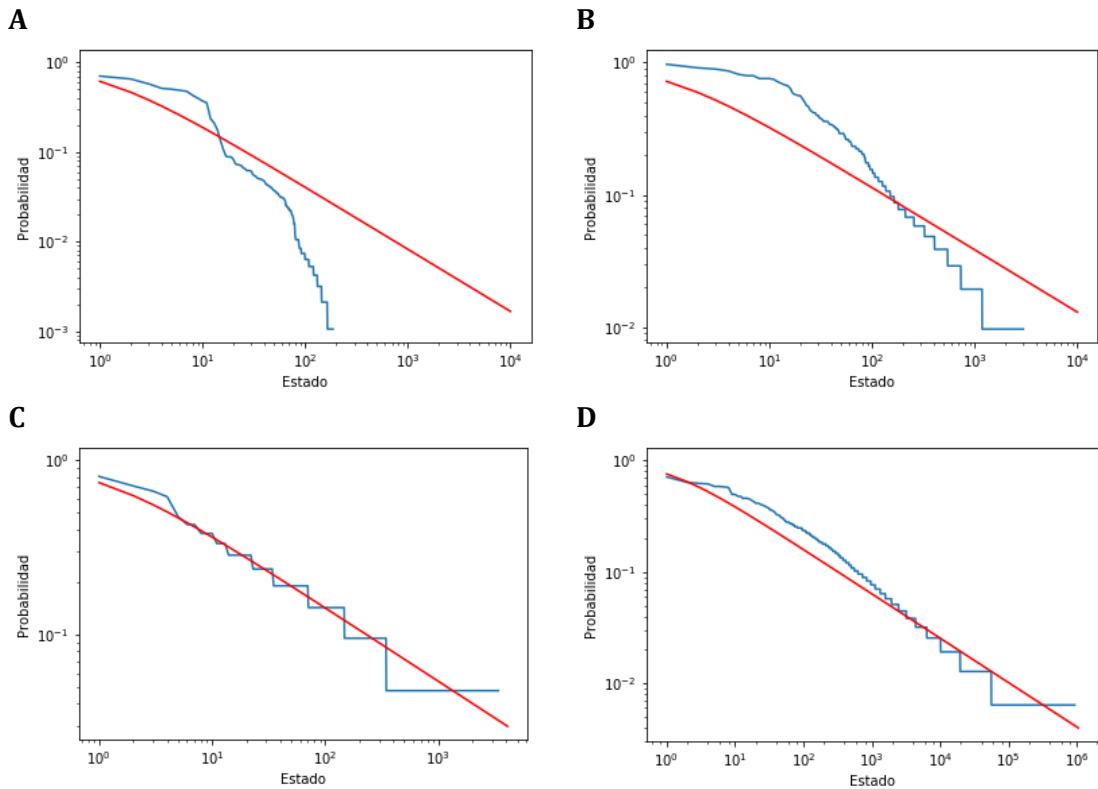


Figura 7. Probabilidad de cada transición de estados de mayor a menor para una muestra de tamaño adecuado a cada modelo, generada a partir de distintos sistemas de tamaño reducido de 4 (A), 8 (B), 12 (C) y 20 (D) neuronas. La probabilidad de cada estado de la muestra es el eje azul, la distribución de Zipf ajustada a los datos es la línea roja. Al igual que en el caso de los estados, el mejor resultado está en 12 neuronas. A medida que el tamaño del modelo aumenta, el número de estados distintos crece de forma exponencial y es imposible generar una muestra del tamaño suficiente como para tomar como relevante su resultado, por lo que no se ha podido realizar este análisis con el modelo completo.

Ninguno de los ajustes realizados consigue pasar el test de Kolmogórov-Smirnov, aunque es posible que esto ocurra por la cola de la distribución, que es la parte más sensible al tamaño de la muestra, y trabajar con muestras mayores que las utilizadas se vuelve computacionalmente muy costoso. No obstante, es posible ver a mero golpe de vista las similitudes entre las distribuciones reales y las de Zipf ajustadas para diferentes tamaños del modelo (Figura 7), lo que es un buen indicativo de criticalidad.

6.2 ENTROPÍA DE MUESTRAS GENERADAS

Si la red neuronal original del gusano operaba a temperatura T_c , nuestro sistema deberá tener como T_c la misma temperatura a la que lo hayamos entrenado (En este caso es 1). La temperatura T_c se distingue porque en ella el sistema posee la máxima capacidad calorífica posible. Así que para comprobar si nuestro modelo opera en punto de criticalidad, se calcula la entropía de muestras sintéticas para un rango de temperaturas, y se calcula la capacidad calorífica normalizada del mismo, que debería divergir en el límite termodinámico.

Este análisis puede orientarse tanto a estados como a transiciones entre estados. No obstante, en este caso aplicar el análisis a estados produce efectos poco satisfactorios, ya que no es posible generar una muestra de tamaño adecuado para el modelo completo de 56 neuronas, las medidas de entropía con la fórmula de Shannon son poco fiables (Figura 21 y Figura 22). Si se orienta el análisis a las transiciones, en cambio, además de ser más

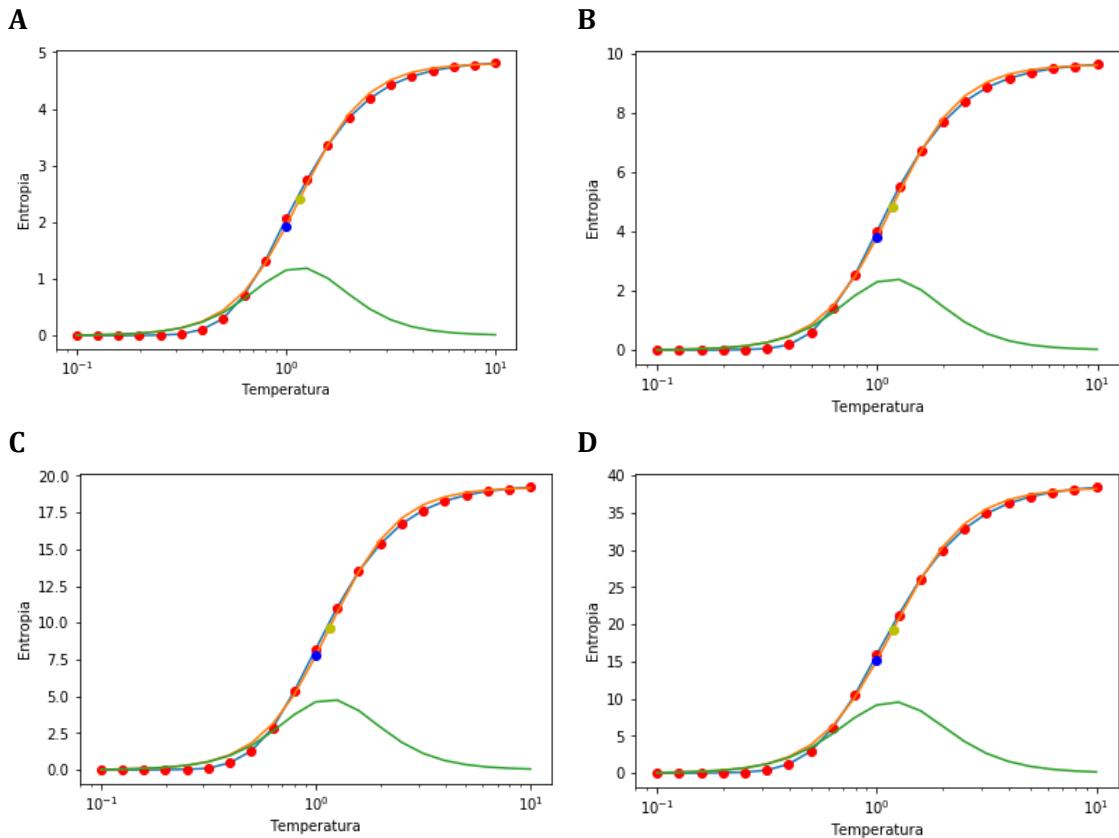


Figura 8. Análisis de entropía orientado a transiciones. Para 7 (A), 14 (B), 28 (C) y 56 (D) neuronas respectivamente. La línea naranja corresponde a la función sigmoidal estimada a partir de los datos de entropía, y la línea a la capacidad calorífica del sistema (derivada de la función sigmoidal). El punto azul indica donde se encuentra el sistema, y el punto amarillo, donde se encuentra el punto crítico del sistema, de forma aproximada. En este caso, el análisis indica que el sistema está operando en punto crítico (o muy cerca de él), independientemente del tamaño del sistema.

consecuente con lo que captura el modelo, es posible utilizar una fórmula alternativa para el cálculo de la entropía sin las limitaciones de la fórmula de Shannon.

Entonces, para calcular la entropía se utiliza la siguiente fórmula:

$$Pr(S_i(t+1) = 1 | \{s_j(t)\}) = \frac{\exp[S_i(t+1)H_i(t)]}{2\cosh H_i(t)} \quad (10)$$

$$\begin{aligned} S(s'_i|s) &= - \sum_s P(s) \sum_{s'_i} \log(P(s'_i|s)) \cdot P(s'_i|s) = \\ &= - \sum_s P(s) (B_i H_i \tanh(B_i H_i) - \log(2 \cosh(B_i H_i))) \end{aligned} \quad (11)$$

La idea detrás de la misma es que es posible calcular la entropía sin necesidad de conocer las probabilidades de cada estado del sistema, utilizando la distribución de la variable H_i (Ecuación 2).

Como el cambio de fase sigue una función sigmoidal, es posible ajustar una función de este tipo a las entropías obtenidas. La capacidad calorífica del sistema en cada temperatura se corresponde con la derivada de la función ajustada (Figura 8).

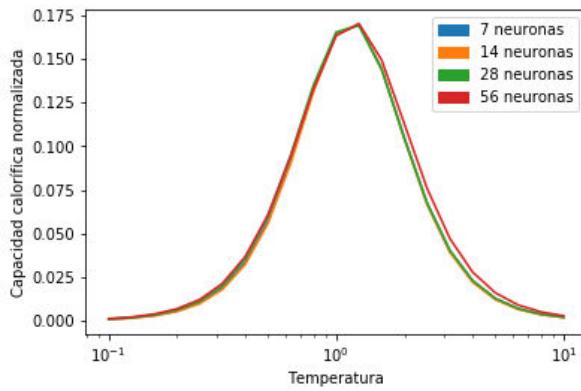


Figura 9. Evolución del punto crítico del sistema para diferentes tamaños del sistema, a lo largo de diferentes temperaturas. En todos los casos, este pico se sitúa muy cercano a la temperatura, pero no parece aumentar con el tamaño del sistema.

Utilizando estas medidas esta vez sí que se obtiene como resultado una T_c cercana a 1, que es la temperatura con la que se ha entrenado al modelo, y que por tanto, indica que este se encuentra en punto crítico, o por lo menos muy cerca de él. No obstante, el aumento de la capacidad calorífica normalizada es irregular y tan reducido que podría considerarse ruido de la medición (Figura 9).

La causa de esto es seguramente el tamaño de las muestras estudiadas, pero dado que un análisis con muestras de tamaños más apropiados con el modelo es impracticable, no se ha podido dar un resultado claro sobre si la capacidad calorífica del sistema diverge o no en el límite termodinámico.

6.3 MAGNETISMO

Es posible utilizar también los spines de las neuronas para poder detectar cómo de cerca se encuentra el modelo con respecto al cambio de fase. La idea es la siguiente: si el sistema se comporta de forma muy desordenada, cada neurona tenderá a estar en +1 ó -1 de forma aleatoria; aproximadamente un 50% de las veces en cada espín. No obstante, si el sistema se comporta de forma muy ordenada, las neuronas tenderán a estar siempre en los mismos estados.

De esta forma, la suma de todos los spines que ha tenido una neurona a lo largo de una simulación sería muy cercana a 0 si el modelo está en fase desordenada, ya que habría un número muy similar de +1 y -1. En cambio, si el modelo está en fase ordenada, habrá mayoría de +1 ó -1, y por tanto, el resultado se alejará de 0.

Para detectar el punto de criticalidad a partir de esta medida, se utiliza el hecho de que una vez el sistema entra en fase desordenada, el sistema ya debería estar muy cerca de tener un índice muy cercano a su mínimo, para luego reducirse de forma casi plana. Luego el punto de criticalidad deberá estar justo al comienzo del descenso cuando este ya apenas es apreciable. Sin embargo, el tener muestras de tamaño finito, y que no estén modeladas todas las neuronas del gusano, hace que los resultados sean menos definidos, y aparezcan picos que hacen más difícil interpretar los resultados.

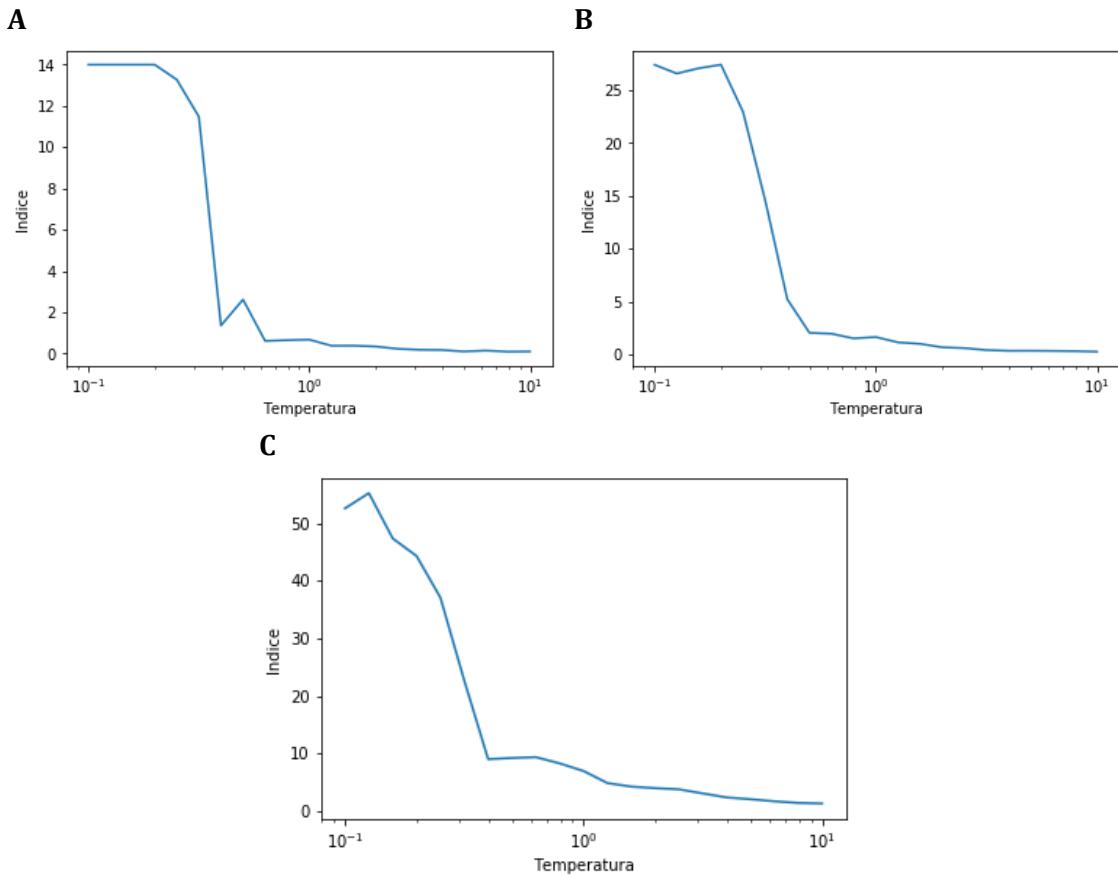


Figura 10. Índice de magnetismo para 14 (A), 28 (B) y 56 (C) neuronas, a partir de una muestra sintética de 15000 muestras. Como puede observarse, las temperaturas bajas hacen que las neuronas se mantengan más estables y por tanto aumente el sumatorio de sus espines. A medida que se aumenta la temperatura, estas van equilibrando el porcentaje de tiempo que pasan en cada spin, y este se reduce. En los tres casos, el índice de magnetismo queda reducido a su nivel mínimo (con un número ilimitado de muestras este acabaría siendo 0) cerca de la temperatura de entrenamiento del sistema, lo que es un buen indicativo de que ese es el punto de criticalidad del sistema.

El índice de magnetismo desciende rápidamente a medida que se aumenta la temperatura en la fase ordenada (Figura 10). Es difícil distinguir donde está exactamente el punto de criticalidad debido a los picos cercanos a la temperatura 1. No obstante, en los tres casos el comienzo del decrecimiento suave se encuentra cercano a dicha temperatura, quedando siempre esta ligeramente escorada en la fase desordenada.

Es importante notar que estudios recientes afirman que el punto óptimo podría no ser exactamente el punto de criticalidad, sino en la fase desordenada (Botcharova, Farmer, & Berthouze, 2014), que es donde este análisis sitúa al modelo.

7 OTROS ANÁLISIS

7.1 RECONOCIMIENTO DEL COMPORTAMIENTO

El objetivo de esta sección es deducir el comportamiento del gusano a partir de su actividad neuronal. En los datos originales, los análisis realizados por el propio equipo que los tomó (Nguyen et al., 2016) indican que existen cuatro comportamientos distintos.

Para ello, se crea una clasificador que tiene como entrada la actividad neuronal de las N neuronas del gusano, y como salida el comportamiento detectado a partir de las mismas. Para implementar el clasificador, se utiliza una red neuronal implementada en la librería Keras.

Para entrenar la red neuronal se necesitan como datos de entrenamiento la activación neuronal en cada momento, y como etiquetas, el comportamiento asociado al gusano en ese mismo instante. Se escoge un 20% aleatorio de las muestras como conjunto de test y el resto se utiliza para entrenamiento y validación. La propia librería escoge qué valores utilizará para entrenar y cuáles para validar el entrenamiento.

7.1.1 Validación de modelos

Aunque construir una red para cualquiera de los gusanos presentes es relativamente sencillo, se aspira, tanto en este aspecto como en todos los demás, a crear herramientas que permitan escalar a conjuntos de datos donde se dispongan de un mayor número de muestras y de un mayor número de neuronas.

Así pues, lo que se ha hecho es automatizar el entrenamiento y la validación de modelos de la red, de forma que ella misma a la hora de entrenar sea capaz de escoger el tamaño idóneo para el problema a solucionar. El algoritmo que sigue la selección del tamaño de la red es el siguiente:

1. Se crea una red neuronal con una capa de entrada y otra de salida.
2. Se entrena la red con la activación neuronal y los comportamientos del gusano, y se comprueba su tasa de acierto.
 - a. Si su tasa de acierto es mayor que la mayor encontrada hasta ahora, se guarda el número de capas utilizado y el número de neuronas utilizado.
 - b. Si su tasa de acierto es mayor que la de otros modelos con el mismo número de capas, se guarda el número de neuronas idóneo para la última capa oculta de la red. (No se realiza esta comprobación en el primer paso, cuando no hay capa oculta)
3. Se aumenta en 10 el número de neuronas en la última capa oculta.
 - a. Si se han sobrepasado las 100 neuronas en la última capa oculta: se establece como número de neuronas de esta capa el que mejor resultado haya dado hasta ese momento, y se añade una nueva capa oculta con 10 neuronas.
4. Si el número de capas ocultas es menor que 6, volver al paso 2.
5. Reentrenar el clasificador con el mejor número de capas y neuronas por capa.

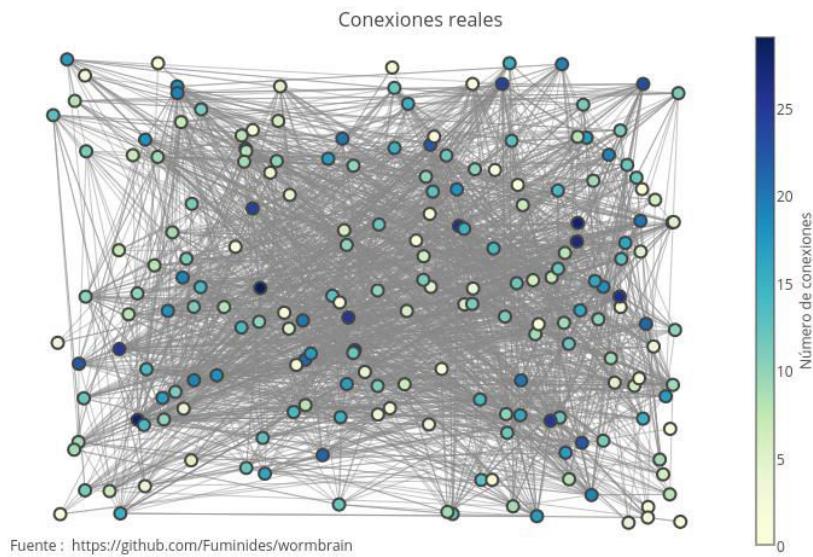


Figura 11. Grafo con las 2230 conexiones reales de las 302 neuronas del gusano *C.elegans* (Durbin, 1987). Existe un alto número de conexiones por neurona, que se reparten de forma desigual. A pesar de que es difícil de observar debido al alto número de conexiones, no existe ningún ciclo en toda la red.

Los resultados del clasificador con las muestras indican un 100% de acierto, lo que lleva a sospechar de sobreajuste en su entrenamiento. En cualquier caso, es difícil obtener un clasificador realmente fiable para los estados del gusano con las muestras del tamaño actual. Además, medir la validez de los resultados en muestras sintéticas es difícil, sino directamente imposible. Luego aunque este funcione correctamente para las muestras actuales, sería muy recomendable disponer de otras de más tamaño para eliminar el sobreajuste y disponer de un clasificador más fiable.

7.2 RECONSTRUCCIÓN DE LA RED ORIGINAL

Como se ha visto en secciones previas, entrenar un modelo de Ising cinético es muy costoso computacionalmente, y el coste crece de forma muy rápida con el número de neuronas a calcular. Aunque el modelo estudiado disponía de 56 neuronas, el número total de neuronas del cerebro del gusano es de 302 para un total de 2230 conexiones, sin presencia de ciclos (Figura 11) (Durbin, 1987). El modelo utilizado, en cambio, requeriría de calcular los pesos de 302^2 conexiones.

Ser capaces de encontrar conexiones reales en los datos del gusano podría permitir utilizar modelos más sencillos y más parecidos al cerebro del gusano. Además, es también de interés comprobar las relaciones entre la topografía del modelo, y el comportamiento real del mismo y ver si existe alguna relación entre ambos.

Los datos originales no ofrecen a simple vista ninguna forma de detectar qué neuronas están conectadas entre sí. La topografía del modelo tampoco parece ofrecer ninguna manera fiable de detectarlas. Los pesos de las conexiones más elevados, independientemente del intervalo de tiempo escogido para entrenar el modelo de Ising, son los de las autoconexiones entre neuronas, que son precisamente los únicos que sabemos con seguridad que no existen (Figura 12). Además, si utilizamos $T=4$, donde estos se reducen considerablemente, la diagonal sigue teniendo algunos de los pesos más grandes.

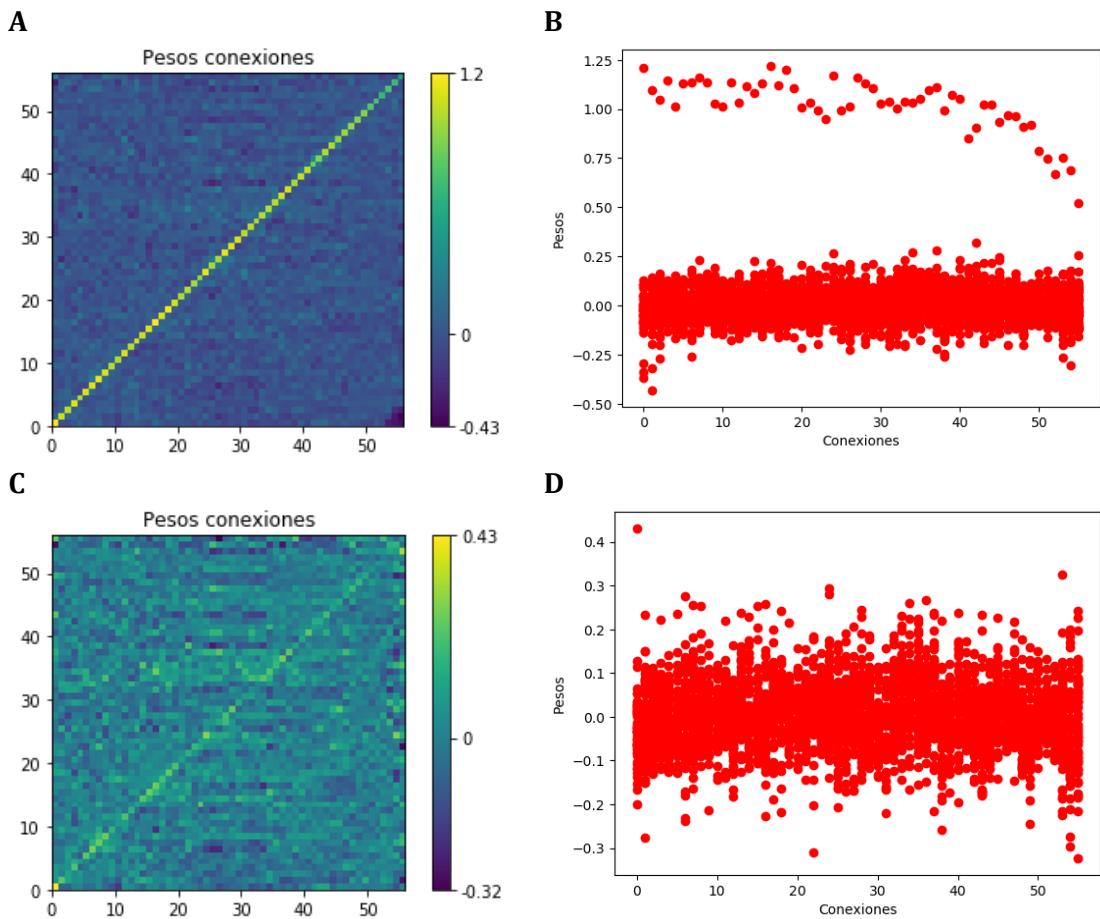


Figura 12. Pesos de las conexiones para el modelo con $T=1$ (arriba) y para $T=4$, (abajo). El modelo “ingenuo” con $T=1$, produce unos pesos desorbitadamente grandes en las autoconexiones de las neuronas, lo que hace sospechar que no reflejan bien la realidad. Al cambiar a $T=4$, estos pesos disminuyen y todos ellos se concentran en valores más acotados.

Siendo imposible conseguir las conexiones a partir de la topografía del modelo o de los datos originales, se procede a estudiar la entropía entre cada par de neuronas. Para reconstruir la red original, se parte de las siguientes suposiciones con respecto del conjunto de datos original:

1. **Las neuronas escogidas representan bien la red:** esto significa que las neuronas estudiadas están conectadas las unas a las otras en una proporción similar al resto de la red, su comportamiento es parecido, etc. Esto es importante porque para la reconstrucción del grafo de conexiones, se busca replicar algunas propiedades del cerebro completo del *C.elegans*. Así que si las neuronas a estudiar están conectadas de forma distinta que las del resto del sistema, el grafo resultante será incorrecto.
2. **Las neuronas conectadas están fuertemente correlacionadas:** la cantidad de información que se transfiere en una única conexión es siempre lo más pequeña posible. Por ejemplo: si una neurona se activa, se supone que el resto de neuronas responderá siempre de forma parecida o igual a la activación de dicha neurona. Esta suposición puede dar problemas si una neurona tiene muchas otras como entrada, y está correlacionada con el conjunto de sus activaciones, pero no con cada una por separado.

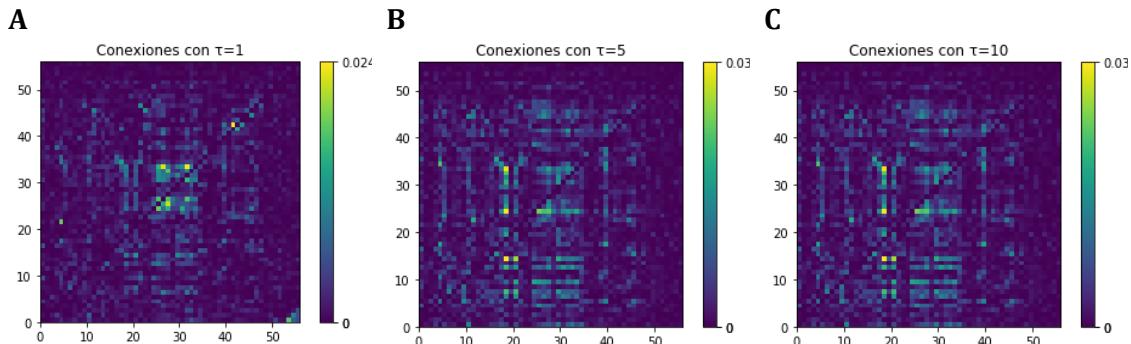


Figura 13. Matrices con la entropía de las conexiones entre pares de neuronas para $t+1$ (A), $t+5$ (B) y $t+10$ (C). A pesar de los cambios en el tiempo, en los tres casos parecen repetirse los picos altos y bajos, lo que nos permite deducir a golpe de vista cuáles son buenos candidatos a tener conexiones reales, y cuáles no.

3. **Cada neurona realiza siempre la misma función:** cada neurona tiene siempre la misma función de activación y obtiene información de cada neurona a la que está conectada de la misma manera. Esta suposición puede parecer trivial, pero existen evidencias en personas de que, ante daños cerebrales, algunas neuronas pueden cambiar su función (Stern, 2002).

7.2.1 Detección de conexiones a partir de actividad neuronal

Para calcular las conexiones reales, se ha utilizado la transferencia de información (Ecuación 8) entre pares de neuronas en distintos intervalos de tiempo (Figura 13), al igual que se ha hecho para escoger el mejor $t+T$ para simular el modelo. Se espera que las neuronas conectadas sean aquellas que menos información transmitan entre ellas. Es decir, que el comportamiento de una respecto de la otra sea siempre lo más determinista posible; una neurona conectada a otra tenderá siempre a reaccionar de la misma forma frente a la activación, o no activación, y por tanto, la entropía de la conexión entre ambas tenderá a ser muy baja.

Este método tiene como principal limitación su incapacidad para ver la transmisión de información entre conjuntos de neuronas, pues sólo analiza cada par de neuronas de forma individual. No obstante, añadir al algoritmo este mismo cálculo también para conjuntos de k -neuronas resultaría en un aumento del tiempo de ejecución inasumible.

Una vez se ha calculado la matriz con la entropía de las conexiones de las neuronas, se decide cuáles de ellas se consideran conexiones reales. Para hacer esto, se pueden utilizar dos criterios:

- Utilizar un ratio de conexión u otro rasgo del grafo original: el sistema cogerá siempre las mejores potenciales conexiones, que serán aquellas que tengan menor entropía, hasta que el grafo resultante de esas conexiones reproduzca la cualidad buscada. Recomendado si se dispone de datos reales, como el ratio de conexión del grafo.
- Utilizar un umbral: escoger un umbral en este caso no es trivial, ya que no sabemos a priori qué valores exactamente buscar, pero puede ser una buena opción si no se disponen de datos del grafo original a reconstruir. La opción más flexible es utilizar como umbral el valor de un percentil, escogido según el número de conexiones a detectar.

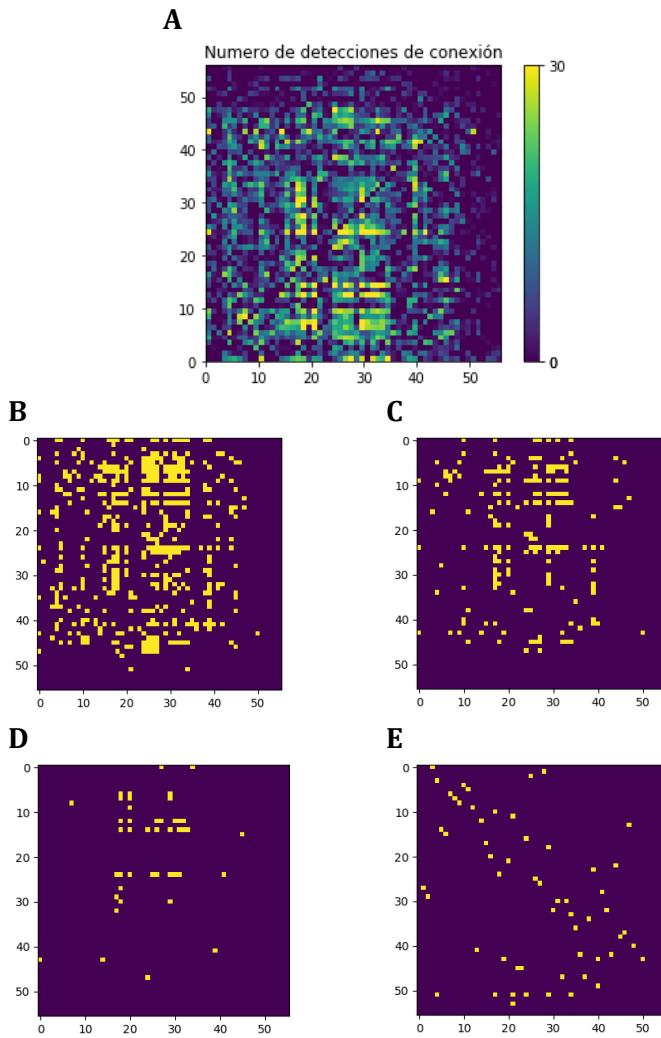


Figura 14: Número de veces que ha sido detectada una conexión, para todo $t+T$ con $0 < T < 30$ (A). Conexiones detectadas en al menos la mitad de las T estudiadas (B). Conexiones detectadas en un 75% de las T (C). Conexiones detectadas para el 100% de los T (D). Conexiones detectadas utilizando ratio real y eliminación de bucles (D).

El estudio de la transmisión de entropía se realiza para distintos intervalos de tiempo. De esta manera, comprobamos la influencia que tiene una neurona en otra en distintas escalas de tiempo. Con esto se comprueba que las conexiones se mantienen constantes a lo largo del tiempo, y se evitan falsos positivos de neuronas que parezcan estar muy correlacionadas en alguna escala temporal, debido a la mera casualidad.

Como resultado de esta parte, se obtiene una matriz donde se indica el número de escalas temporales en las que se ha detectado cada conexión (Figura 14, imagen A).

7.2.2 Construcción de la red

Para construir la red se utiliza una matriz de conexiones dirigida, ya calculada en el apartado anterior. A partir de ella seleccionamos las conexiones que mejor cumplan nuestros criterios. En este trabajo se han estudiado dos:

- Porcentaje de fiabilidad: escogemos aquellas conexiones que hayan sido por lo menos detectadas en un porcentaje de los intervalos de tiempo estudiados.

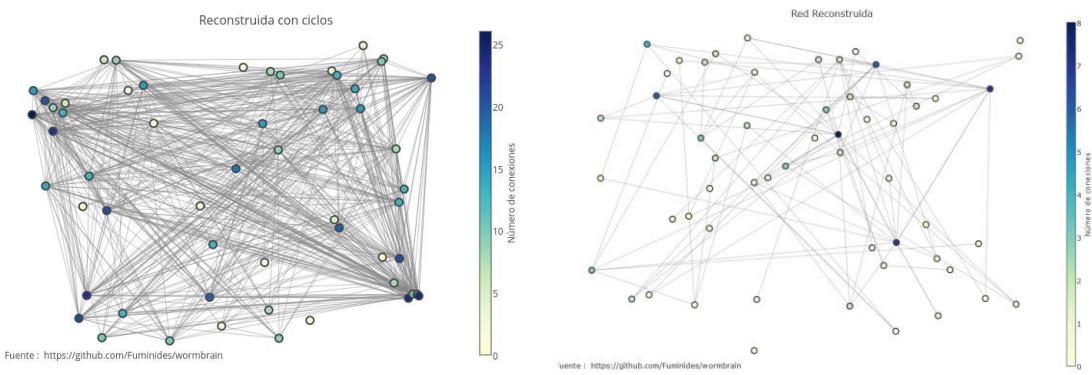


Figura 15. Red reconstruida con y sin ciclos. Hay un total de 646 conexiones en el grafo con ciclos, mientras que en segundo hay 57. A pesar de que el ratio de conexiones se mantiene con respecto al grafo original si se mantienen los ciclos, la densidad del mismo es 5 veces la original, sugiriendo que no representa de forma realista la red original.

- Reproducción de rasgos del grafo original: tomamos las mejores conexiones hasta que el grafo construido tenga alguna propiedad igual que el anterior, como por ejemplo, el ratio de conexión o la densidad.

Para estudiar la similitud entre el grafo original y el reconstruido, se utilizan las siguientes medidas:

- **Ratio de conexión del grafo:** número de conexiones por nodo.
- **Densidad del grafo:** número de conexiones presentes en relación al número de conexiones totales posibles.
- **Coeficiente de agrupamiento medio:** indica cómo de agrupado está cada nodo con sus vecinos en promedio.
- **Longitud media del camino más corto entre pares:** es la longitud media del camino más corto entre cada par de nodos.

La interpretación de las diferencias entre cada una de estas medidas en el grafo reconstruido y el original, depende de la forma de reconstrucción de la red. Si se utiliza en la reconstrucción el ratio de conexión, es obvio que el grafo reconstruido tendrá el mismo que el original. No obstante, grandes diferencias en alguna de estas magnitudes son indicativas de que el grafo reconstruido tiene diferencias importantes respecto del original.

Es también importante mencionar que el grafo original no tiene ciclos. Podría ser preferible respetar esta propiedad, debido a su importancia, antes que cualquiera de las otras medidas, independientemente de cómo se haya reconstruido el grafo (Figura 15). Para eliminar grafos de la matriz de conexiones se recorren cada una de las distintas conexiones de cada neurona; hasta que no es posible seguir o se ha encontrado la misma neurona dos veces, en ese caso, la última conexión recorrida se elimina.

| Grafo | Ratio de Conexión | Densidad | Coeficiente de agrupamiento | Longitud media de camino corto entre pares | Ciclos |
|--|-------------------|----------|-----------------------------|--|--------|
| Conexiones reales | 11.14 | 0.06 | 0.27 | 2.59 | No |
| Reconstruido con ratio de conexión | 11.13 | 0.20 | 0.58 | 1.66 | Sí |
| Reconstruido con 50% del ratio de conexión | 5.55 | 0.10 | 0.37 | 3.89 | Sí |
| Reconstruido con ratio de conexión y no ciclos | 0.96 | 0.018 | 0.0 | 8.94 | No |
| Reconstruido con percentil 50% | 1.63 | 0.03 | 0.1 | 2.98 | Sí |
| Reconstruido con percentil 50% y no ciclos | 1.10 | 0.02 | 0.01 | 3.86 | No |

Tabla 3. *Medidas para reconstrucciones distintas a partir de los datos de transmisión de entropía entre pares de neuronas, en comparación con el grafo de conexiones real. Como puede observarse, la forma de construir el grafo da lugar a resultados finales muy dispares.*

Como resultado final, se han creado distintos grafos de conexiones y se han comparado con el grafo original (Tabla 3). Escoger entre un método u otro es difícil, ya que ninguno consigue realmente parecerse al original, no obstante, reproducir el ratio de conexión parece ser una aproximación buena aunque algo conservadora.

Como nota final, es importante destacar la baja densidad de todos los grafos reconstruidos. Incluso en el que más conexiones posee, habría que calcular únicamente un 20% de las conexiones que se calculan actualmente para el modelo de Ising. También es destacable la falta de relación clara entre la topología del modelo y su comportamiento, que hace imposible reconstruir la red directamente del mismo.

8 CONCLUSIONES

A partir de la actividad neuronal de un gusano *C.elegans* en movimiento se ha inferido un modelo de Ising cinético, a partir del cual se han realizado una serie de análisis para comprobar su régimen de actividad, y se ha comparado la topología del propio modelo con el comportamiento de las muestras generadas a partir del mismo. Como conclusiones estos análisis, se han obtenido principalmente tres resultados:

- Se ha encontrado el mejor intervalo de tiempo para modelar las transiciones con Ising cinético y se han hecho varias optimizaciones en su entrenamiento.
- Se han encontrado signos de criticalidad en el cerebro del gusano: el modelo generado actúa en punto de criticalidad, según el estudio en función del índice de magnetismo, de la ley de Zipf y de la entropía del sistema. No obstante, no se ha podido establecer si la capacidad calorífica máxima del modelo diverge en el límite termodinámico.
- Se ha reconstruido la red original a partir de la simulación del modelo, y se ha comprobado la falta de una relación explícita entre la topología de este y la relación entre las neuronas cuando este está siendo simulado.
- Se ha construido un clasificador para reconocer el comportamiento de un gusano a partir de su actividad neuronal.

Con ello, se arroja algo más de luz al debate sobre si los tejidos neuronales funcionan en punto crítico, ya que los signos de criticalidad obtenidos son robustos. La confirmación de este fenómeno tendría multitud de consecuencias, entre las que se destacan:

1. En neurociencia siempre se ha trabajado con la hipótesis de que era posible entender cada parte del cerebro por separado. Que este operase en punto crítico indicaría que el sistema no puede explicarse únicamente desde este ángulo, y que es necesario tener en cuenta las interacciones entre distintas partes.
2. Los sistemas en punto crítico poseen sensibilidad máxima a estímulos externos. Esto tendría implicaciones en cómo el cerebro percibe la realidad.
3. Se ha especulado ya con la posibilidad de que la criticalidad podría ser el punto de partida para otros fenómenos más complejos que ocurren en nuestro cerebro, como por ejemplo la conciencia (Ouellette, 2014).

Además, la reconstrucción de la red permitiría la utilización de modelos más sencillos y más ajustados a la realidad para simular el tejido nervioso del gusano. Esto podría ser muy importante para el estudio de sistemas de tamaño mayor al estudiado (56 neuronas), como el propio sistema nervioso del *C.elegans* completo, que dispone de 302.

9 BIBLIOGRAFÍA

- Aguilera, M., Alquézar, C., & Izquierdo, E. J. (2017). Signatures of criticality in a maximum entropy model of the *C.elegans* brain during free behaviour.
- Beggs, J. M. (2008). The criticality hypothesis: How local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1864), 329-343.
- Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., & Walczak, A. M. (2012). Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13), 4786-4791.
- Botcharova, M., Farmer, S. F., & Berthouze, L. (2014). Markers of criticality in phase synchronization. *Frontiers in Systems Neuroscience*, 8
- Durbin, R. M. (1987). Studies on the development and organisation of the nervous system of *caenorhabditis elegans*. *Wormatlas*,
- Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4), 500.
- Mora, T., Walczak, A. M., Bialek, W., & Callan, C. G. (2010). Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12), 5405-5410.
- Nguyen, J. P., Shipley, F. B., Linder, A. N., Plummer, G. S., Liu, M., Setru, S. U., . . . Leifer, A. M. (2016). Whole-brain calcium imaging with cellular resolution in freely behaving *caenorhabditis elegans*. *Proceedings of the National Academy of Sciences*, 113(8), E1081.
- Ouellette, J. (2014). A fundamental theory to model the mind. *Quanta Magazine*.
- Stern, Y. (2002). What is cognitive reserve? theory and research application of the reserve concept. *Journal of the International Neuropsychological Society*, 8(03), 448-460.

Estudio y análisis de la red neuronal del gusano *Caenorhabditis elegans*

- Tetzlaff, C., Okujeni, S., Egert, U., Wrgtter, F., & Butz, M. (2010). Self-organized criticality in developing neuronal networks. *PLoS Comput Biol*, 6(12), e1001013.
- Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry, M. J., & Bialek, W. (2015). Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, 112(37), 11508-11513.
- Tyrcha, J., Roudi, Y., Marsili, M., & Hertz, J. (2013). The effect of nonstationarity on models inferred from neural data. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03), P03005
- Mora, T., & Bialek, W. (2011). Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2), 268-302.

10 ANEXOS

9.1 PICOS DE ENTROPIA SEGÚN UMBRAL

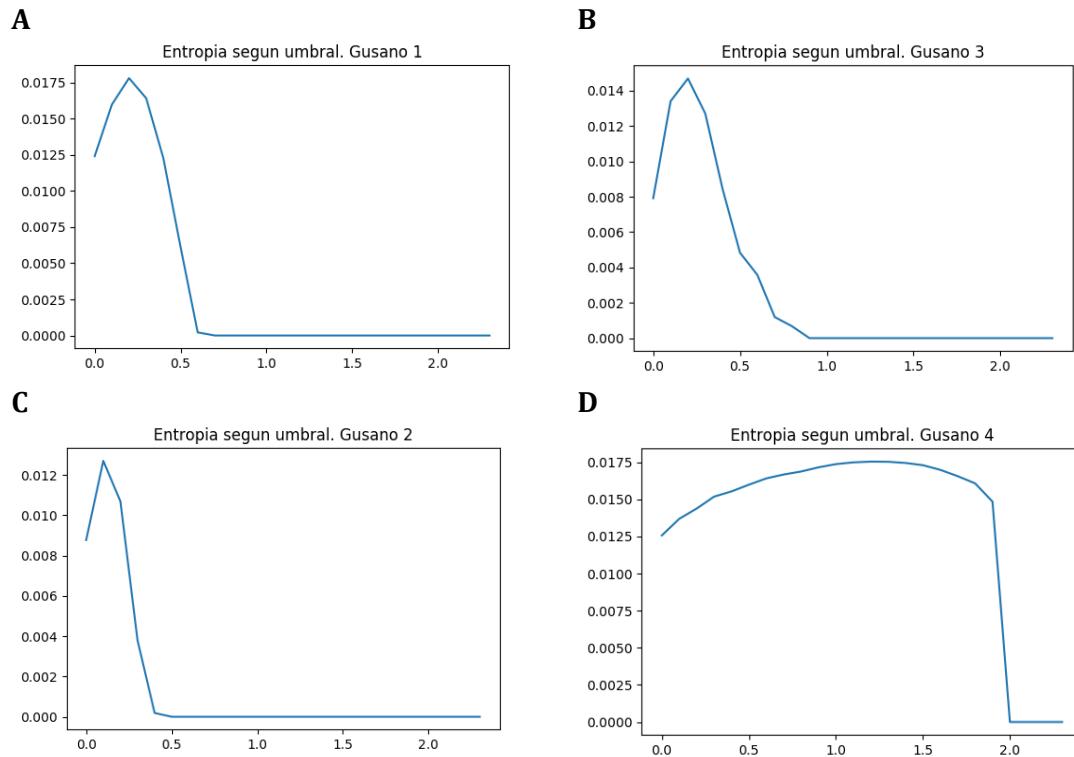


Figura 16. Picos de entropía para los cuatro primeros gusanos, sumando la entropía obtenida de cada neurona por separado. Esta medida es claramente insuficiente y muy sensible al ruido (como puede verse en el gusano 4) pero nos permite hacernos una idea de los resultados a encontrar al utilizar combinaciones de varias neuronas siguiendo el mismo método.

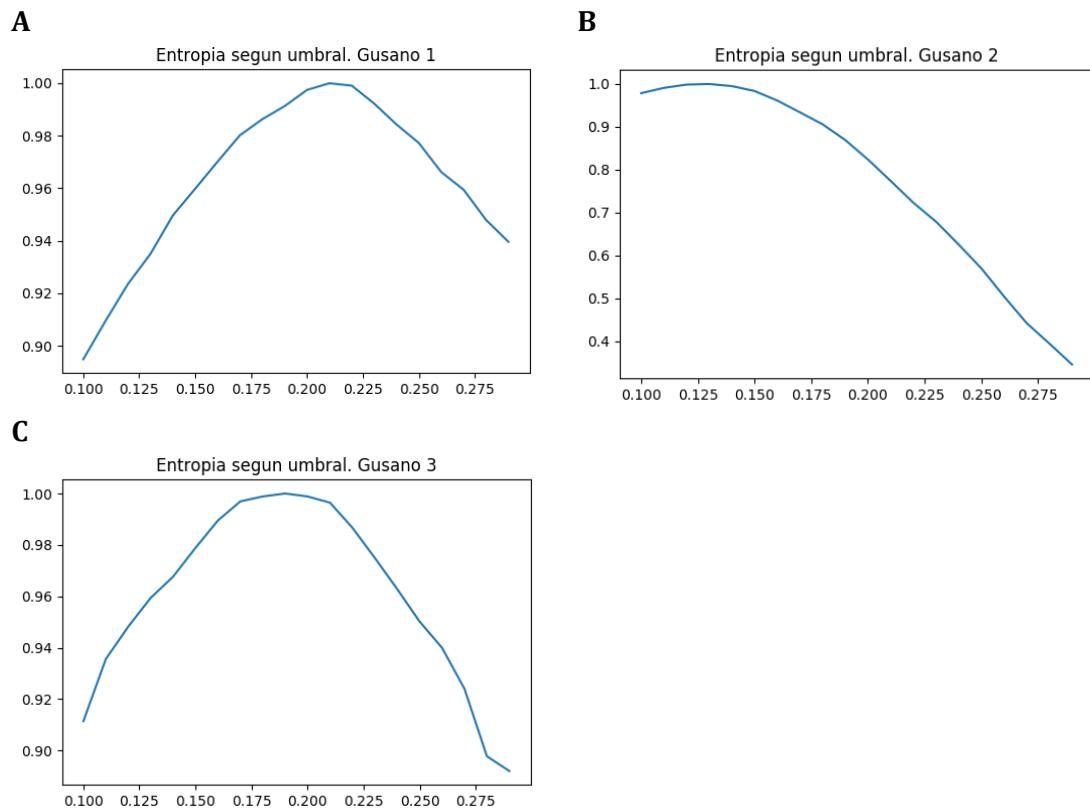


Figura 17. Picos de entropía de la figura 1 a más resolución. Como puede verse, el máximo de cada gusano es similar, pero no son iguales, lo que sugiere que utilizar un valor fijo precalculado para todos los gusanos podría no ser la forma óptima de discretizar la actividad neuronal.

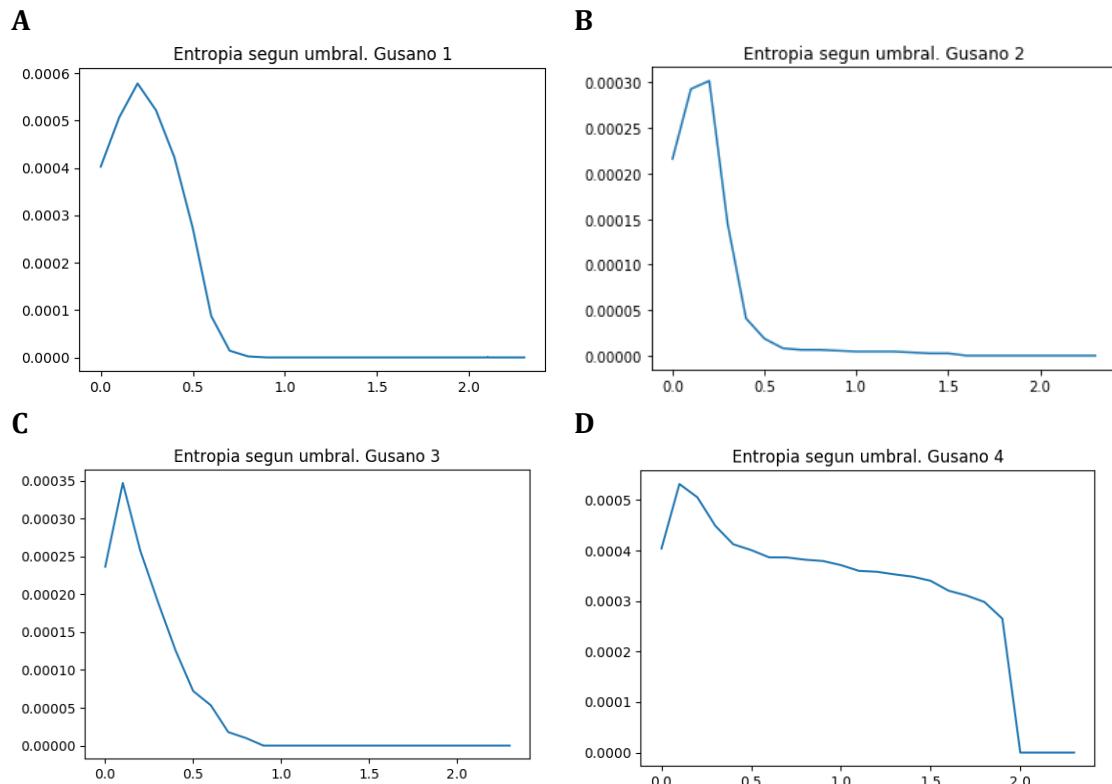


Figura 18. Picos de entropía análogos a los de la figura 1, pero calculados a partir de la suma de la entropía de todas las combinaciones de 2 neuronas de los gusanos. En este caso, los picos comienzan a ser más pronunciados, y la resistencia al ruido (especialmente notable en el gusano 4).

9.2 ANÁLISIS DE VALORES PROPIOS

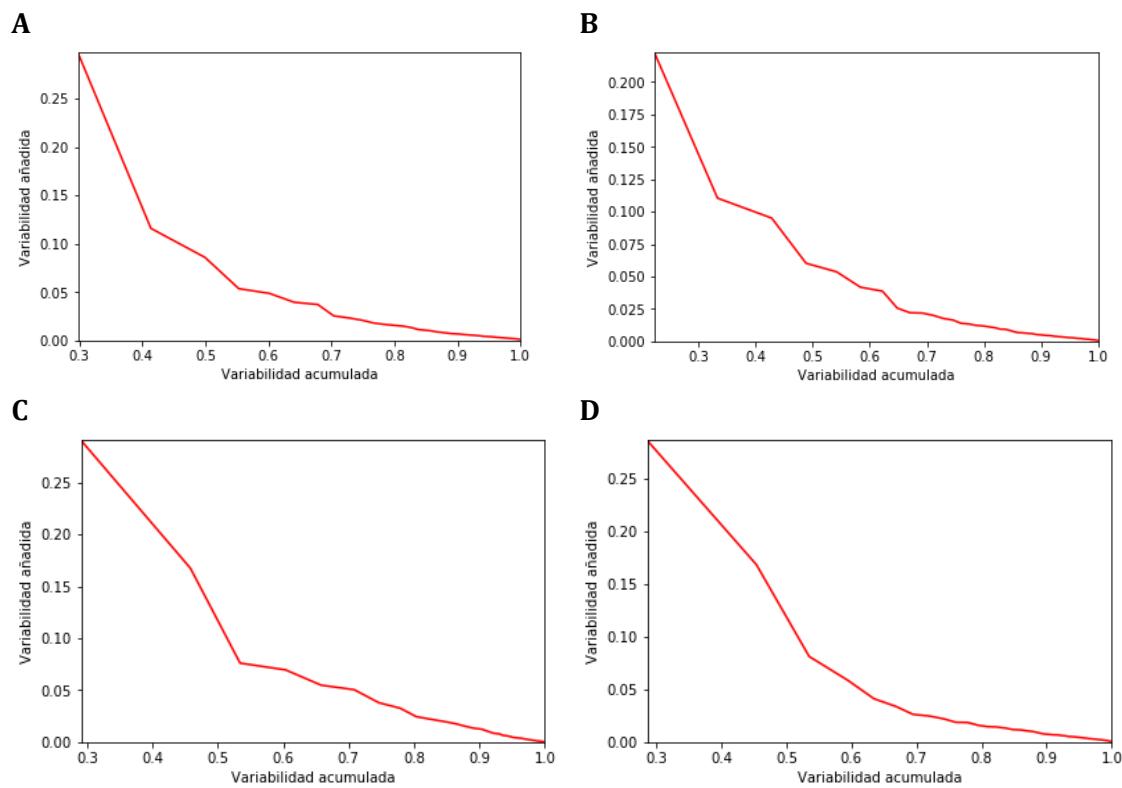


Figura 19. Estudio de variabilidad recogida para el sujeto 1 (A), sujeto 2 (B), sujeto 3(C) y sujeto 4(D). No se encuentran grandes diferencias en ninguno de los 4 sujetos, lo que indica que el comportamiento del cerebro de los mismos no se diferencia mucho de uno a otro.

9.1 ANÁLISIS DE ZIPF ORIENTADO A ESTADOS

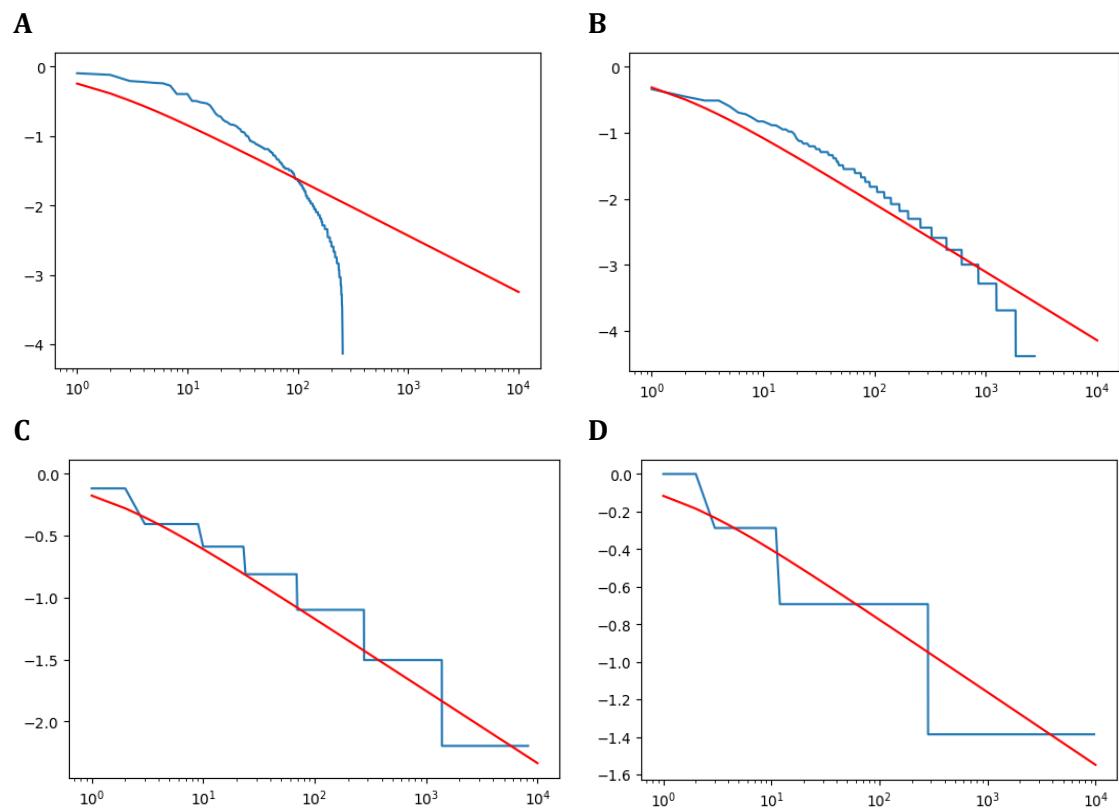


Figura 20. Probabilidad de los estados para una muestra de tamaño 10000 generada a partir de distintos sistemas de tamaño reducido de 8, 12, 20 y 36 neuronas. Ambos ejes en escala logarítmica. La probabilidad de cada estado es el eje azul, la distribución de Zipf ajustada a los datos es la línea roja..

9.2 ANÁLISIS DE CRITICALIDAD ORIENTADO A ESTADOS

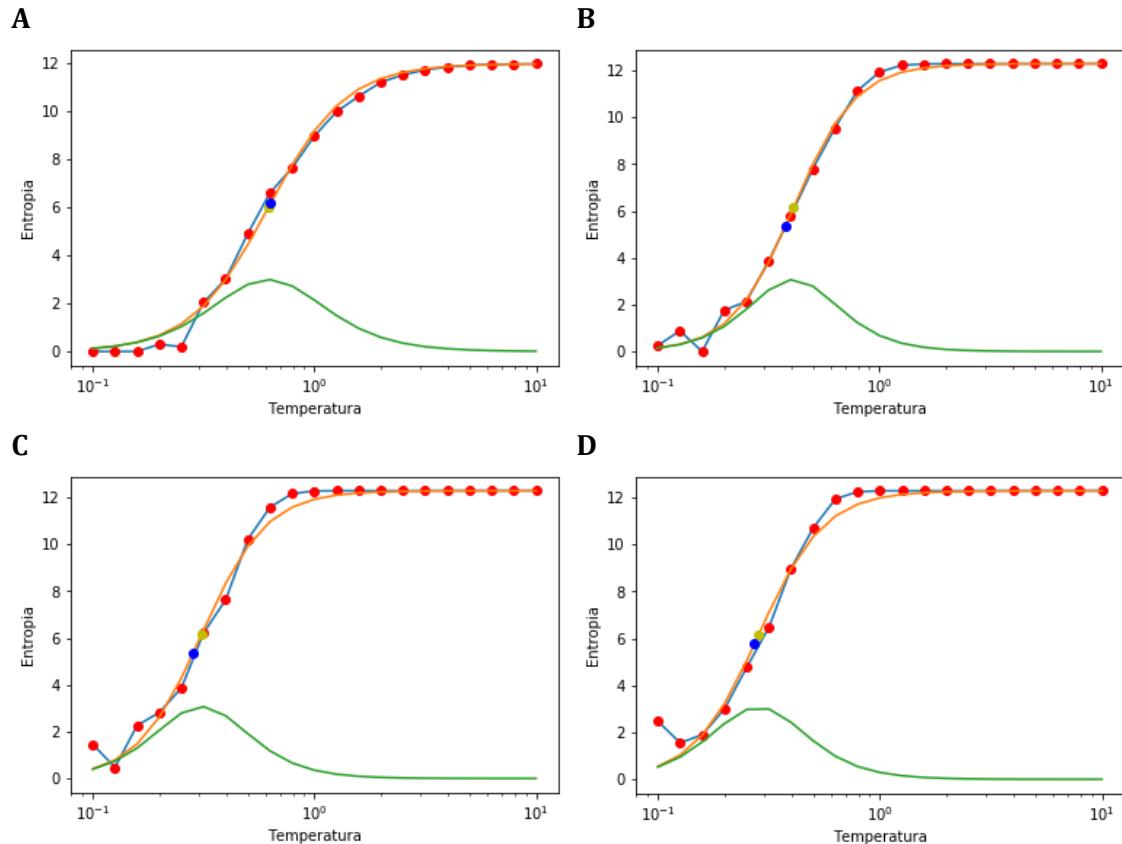


Figura 21. Análisis de entropía del sistema para 7 (A), 14 (B), 28 (C) y 56 (D) neuronas. Los valores reales de entropía del sistema están marcados con un punto rojo, la línea naranja representa la función sigmoidal que mejor se ajusta a estos, y la línea verde es la capacidad calorífica del sistema en cada uno de esos instantes (derivada de la sigmoidal). El punto amarillo indica dónde se encuentra el punto crítico en la función sigmoidal, y el punto azul donde se encuentra la muestra original. Como puede verse en los cuatro casos, la muestra original siempre está muy cerca del punto crítico, pero el sistema (temperatura 10^0) se encuentra bastante alejado del mismo.

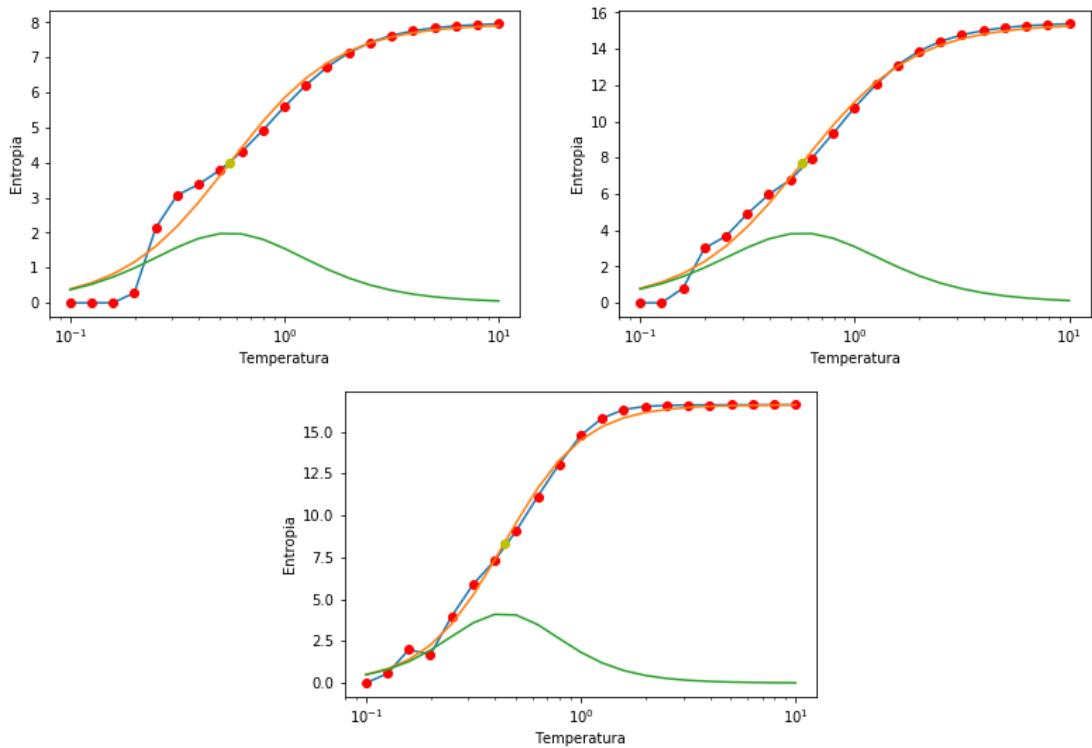


Figura 22. Mismo análisis que en la figura 21, pero para 4, 8 y 12 neuronas, con muestras de tamaño 10000. Los cambios obtenidos entre la figura anterior y esta denotan que el insuficiente tamaño de la muestra es la explicación para las diferencias que presentan el análisis orientado a estados y el orientado a transiciones.

Estos análisis son análogos a los realizados utilizando la fórmula de entropía de las transiciones entre estados, pero calculando la entropía con la fórmula de Shannon a partir de muestras sintéticas generadas con el modelo. Sin embargo, la incapacidad de generar una muestra sintética de tamaño aceptable, 2^{56} , hace estos resultados poco fiables.

Es cierto que parece que la muestra original y el punto de criticalidad tienen entropía muy similar, lo que indicaría que esta proviene de un sistema operando en dicho estado, pero nuestro modelo parece muy alejado del mismo (Figura 21). El cálculo de la entropía utilizando esta fórmula es muy sensible a la falta de muestras (Figura 22), y por tanto, no ofrece resultados muy fiables.

9.3 TRANSFERENCIA DE ENTROPÍA A LO LARGO DEL TIEMPO

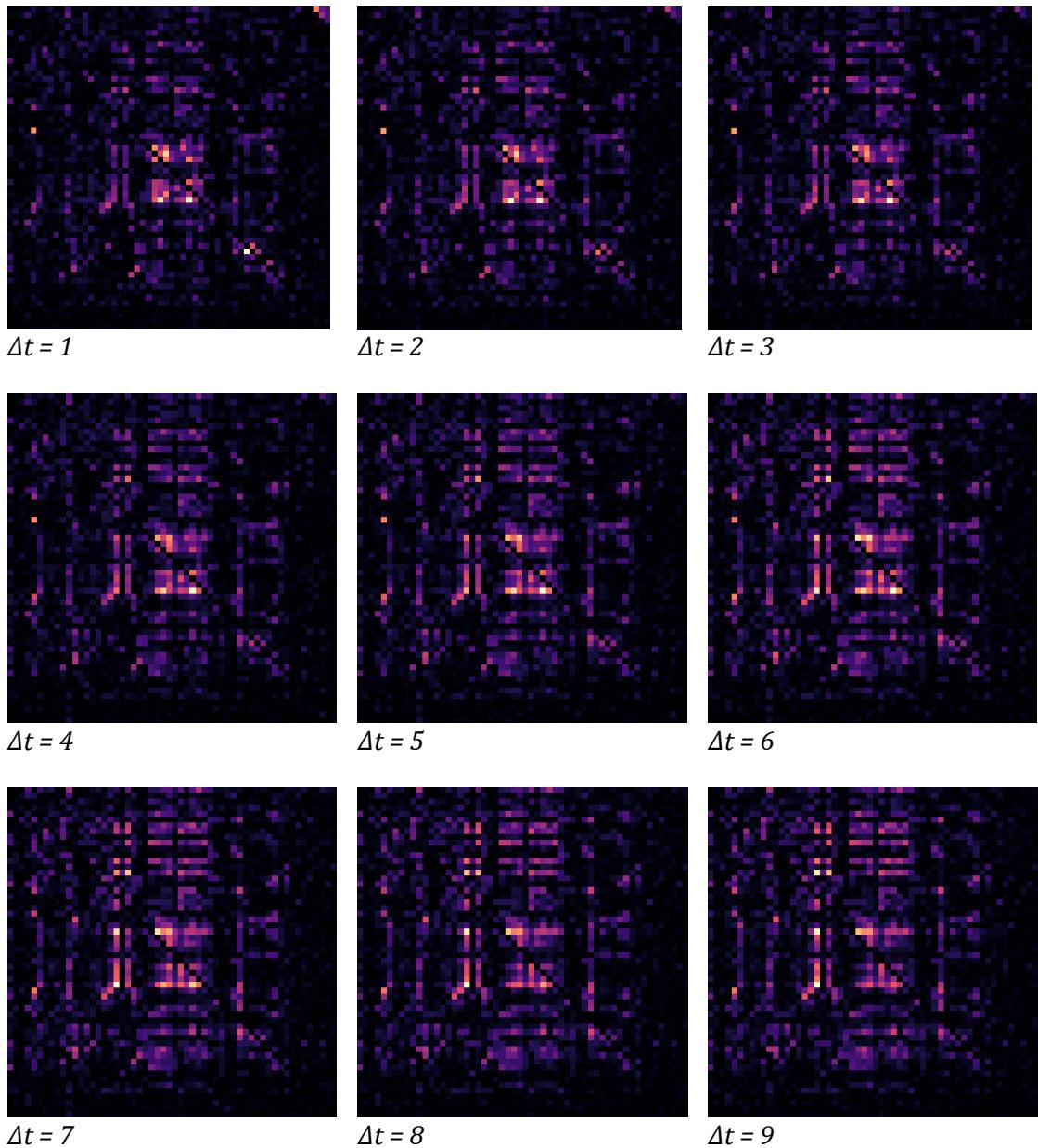


Figura 23. Transmisión de entropía entre pares de neuronas para distintos intervalos de tiempo. La relación entre pares de neuronas no parece cambiar mucho independientemente de la distancia a la que se escojan, por lo que se considera una buena medida para decidir cuales están conectadas realmente.