applied microbiology

MDPI

Artículo

Predecir la dinámica de crecimiento del microbioma en Perturbaciones ambientales

jorge sol [1] y Yi-Hui Zhou 2,*

[1] Centro de Investigación en Bioinformática, Universidad Estatal de Carolina del Norte, Raleigh, NC 27695, EE. UU.;
[2] 3litus@gmail.com Departamentos de Ciencias Biológicas y Estadística, Universidad Estatal de Carolina del Norte, Raleigh, NC 27695, EE. UU
* Correspondencia: yihui_zhou@ncsu.edu

Resumen: MicroGrowthPredictor es un modelo que aprovecha las redes de memoria a corto plazo (LSTM) para predecir cambios dinámicos en el crecimiento del microbioma en respuesta a diversas perturbaciones ambientales. En este artículo, presentamos las capacidades innovadoras de MicroGrowthPredictor, que incluyen la integración del modelado LSTM con una novedosa técnica de estimación de intervalos de confianza. La red LSTM captura la compleja dinámica temporal de los sistemas de microbiomas, mientras que los novedosos intervalos de confianza proporcionan una medida sólida de la incertidumbre de la predicción. Incluimos dos ejemplos : uno que ilustra la composición y diversidad de la microbiota intestinal humana debido al tratamiento recurrente con antibióticos y el otro que demuestra la aplicación de MicroGrowthPredictor en un conjunto de datos intestinales artificiales. Los resultados demuestran la mayor precisión y confiabilidad de las predicciones basadas en LSTM facilitadas por MicroGrowthPredictor. La inclusión de métricas específicas, como el error cuadrático medio, valida el rendimiento predictivo del modelo. Nuestro modelo tiene un inmenso potencial para aplicaciones en ciencias ambientales, atención médica y biotecnología, fomentando avances en la investigación y el análisis de microbiomas. Además, cabe destacar que MicroGrowthPredictor es aplicable a datos reales con tamaños de muestra pequeños y observaciones temporales bajo perturbaciones ambientales, lo que garantiza su utilidad práctica en varios dominios.

Palabras clave: dinámica del microbioma; incertidumbre de predicción; aplicaciones medioambientales

Cita: Sun, G.; Zhou, Y.-H.
Predecir la dinámica de crecimiento del microbioma bajo perturbaciones ambientales. Aplica. Microbiol. 2024, 4, 948–958. https://doi.org/10.3390/applmicrobiol4020064

Editor académico: Bong-Soo Kim

Recibido: 7 de mayo de 2024
Revisado: 4 de junio de 2024
Aceptado: 7 de junio de 2024
Publicado: 10 de junio de 2024

1. Introducción

El microbioma humano, un intrincado ecosistema de billones de microorganismos que residen dentro y sobre el cuerpo humano, desempeña un papel crucial en el mantenimiento de la homeostasis fisiológica, las funciones metabólicas y las respuestas inmunitarias [1]. Las alteraciones en el microbioma se han relacionado con una gran cantidad de afecciones, que van desde trastornos gastrointestinales hasta enfermedades más sistémicas como diabetes, obesidad e incluso trastornos neurológicos [2]. Esta interacción simbiótica entre huésped y microbio subraya la necesidad de comprender la naturaleza dinámica del microbioma humano [3], en particular cómo cambia con el tiempo y en respuesta a diversos estímulos ambientales [4,5].

En condiciones normales, el microbioma intestinal está compuesto por una comunidad diversa de bacterias, siendo Firmicutes y Bacteroidetes los filos predominantes. Las perturbaciones ambientales , como los cambios en la dieta, el uso de antibióticos y la exposición a contaminantes, pueden alterar significativamente la composición y función del microbioma, lo que puede tener implicaciones potenciales para la salud. Por ejemplo, el tratamiento con antibióticos puede reducir drásticamente la diversidad microbiana, lo que a menudo resulta en un crecimiento excesivo de bacterias resistentes y una disminución de microbios beneficiosos, lo que puede alterar los procesos metabólicos y las funciones inmunes [6]. Comprender estas dinámicas poblacionales es crucial para desarrollar estrategias para mitigar los efectos adversos de tales perturbaciones en la salud humana.

Las tecnologías de secuenciación de alto rendimiento, en particular la secuenciación de ARNr 16S, han marcado el comienzo de una nueva era en los estudios de microbiomas, permitiendo evaluaciones detalladas de la diversidad microbiana y la abundancia relativa en diferentes poblaciones y condiciones humanas [7]. Sin embargo , la gran cantidad de datos generados por estas tecnologías presenta oportunidades y desaf

Uno de los principales desafíos es descifrar los patrones temporales y predecir estados futuros del microbioma, esencial para aplicaciones sanitarias preventivas y terapéuticas.

Históricamente, el modelado predictivo en biometría ha empleado varios métodos estadísticos , pero estos enfoques tradicionales a menudo no logran manejar la alta dimensionalidad y la no linealidad de los datos del microbioma. La llegada del aprendizaje automático, y más específicamente del aprendizaje profundo, ofrece nuevas vías prometedoras para datos tan complejos [8]. Las redes neuronales recurrentes (RNN) [9] y su variante avanzada, las redes de memoria a corto plazo (LSTM) [10], destacan en el análisis y predicción de secuencias temporales, proporcionando un marco excelente para modelar la dinámica del microbioma.

En este estudio, presentamos el modelo MicroGrowthPredictor, cuyo objetivo es aprovechar el poder de las redes LSTM para predecir cambios en el microbioma humano en respuesta a perturbaciones ambientales, un paso fundamental hacia la medicina personalizada y las intervenciones terapéuticas dirigidas.

2. Materiales y Métodos 2.1.

Modelo de memoria larga a corto plazo (LSTM)

La red de memoria a largo plazo (LSTM), una forma especializada de la arquitectura de red neuronal recurrente (RNN), está diseñada explícitamente para abordar los desafíos de aprender a partir de datos secuenciales, en particular las dependencias a largo plazo. Los RNN tradicionales, aunque teóricamente son capaces de manejar tales dependencias, a menudo se quedan cortos en la práctica debido al problema del gradiente de desaparición, en el que la información se pierde en cada paso de tiempo durante el entrenamiento. Las redes LSTM están diseñadas para superar esta limitación, lo que las hace particularmente adecuadas para aplicaciones en diversos campos, como el análisis de series temporales, el procesamiento del lenguaje natural y, lo que es pertinente para nuestro trabajo, el análisis de datos de microbioma

Las redes LSTM introducen una estructura celular más sofisticada que las RNN tradicionales [11]. Cada célula LSTM contiene mecanismos llamados puertas que regulan el flujo de información dentro y fuera de la célula. Hay tres tipos de puertas dentro de una celda LSTM (Figura 1A):

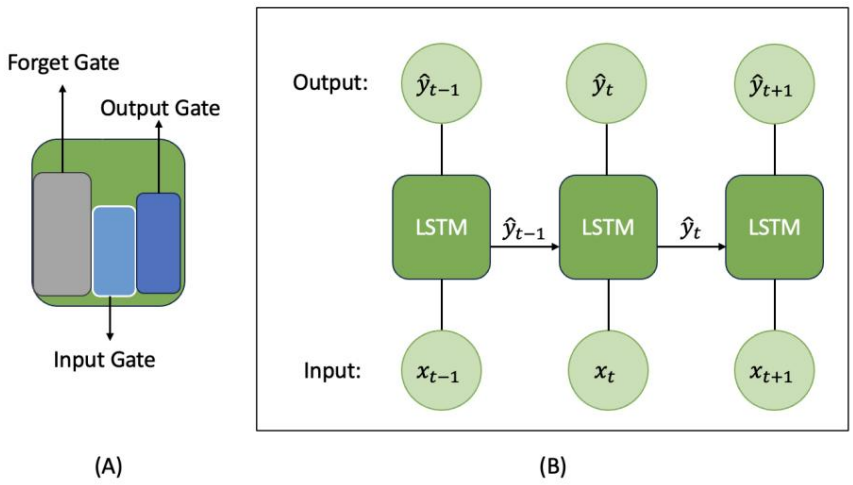

Figura 1. Arquitectura de memoria larga a corto plazo (LSTM): (A) Un acercamiento a una celda LSTM, que muestra sus tres puertas: la puerta de entrada, la puerta de olvido y la puerta de salida. (B) El flujo de datos de entrada y salida en una red LSTM desde el paso de tiempo t - 1 hasta el paso de tiempo t.

- Puerta de entrada: Modula la cantidad de nueva información que se agregará al estado de la celda.
- Puerta de olvido: determina la cantidad de información que se descartará del estado de la celda. La puerta de olvido ayuda a eliminar información microbiana irrelevante u obsoleta, manteniendo así solo los datos más pertinentes para un modelado preciso.
- Puerta de salida: Controla la cantidad de información que se enviará desde la celda. Para los datos del microbioma, la puerta de salida ayuda a decidir qué información microbiana procesada debería influir en las predicciones o análisis de la red en cada paso de tiempo.

Estas puertas trabajan juntas para actualizar el estado de la celda y permiten que el LSTM recuerde y olvide información en secuencias largas (Figura 1B), lo cual es crucial para aprender dependencias a largo plazo. La Figura 1B ilustra la transición de datos a través de la red LSTM de un paso de tiempo al siguiente. Muestra los datos de entrada y los datos de salida a medida que fluyen desde el paso de tiempo t − 1 hasta el paso de tiempo t. En cada paso de tiempo, la celda LSTM procesa los datos de entrada, junto con el estado de la celda del paso de tiempo anterior. Este procesamiento da como resultado un estado de celda actualizado y una salida, que luego se pasan al siguiente paso de tiempo. Este mecanismo secuencial permite que la red LSTM maneje de manera efectiva las dependencias temporales, asegurando que la información se transmita y utilice en diferentes pasos de tiempo para mejorar la predicción y el análisis en tareas de series de tiempo.

En el ámbito del análisis del microbioma, comprender la dinámica temporal y los patrones secuenciales es esencial, dada la naturaleza de la evolución y la interacción de las comunidades microbianas a lo largo del tiempo. Aquí, adoptamos notación específica para dilucidar la mecánica del modelo LS
Considere un conjunto de datos de entrenamiento D $\overset{t}{=}_1\{(x_t , y_t)\}$ donde xt denota el vector de relativa abundancias [12] de todos los taxones microbianos en el t-ésimo paso de tiempo e yt significa el resultado deseado correspondiente. El LSTM toma estas secuencias de entrada y las procesa a través de su intrincada estructura celular, capturando valiosas dependencias temporales presentes en los datos que son fundamentales para predicciones y análisis precisos en estudios de microbioma.

2.2. Estructura del modelo para la predicción del crecimiento del

microbioma El modelo LSTM empleado en este estudio está diseñado para ofrecer simplicidad y potencia. La capa de entrada está diseñada para procesar los niveles de abundancia relativa de taxones, acomodando una amplia gama de taxones microbianos denominados x_t . Esta capa, que comprende nodos n_{taxa} , cada uno de los cuales representa la abundancia relativa de un taxón particular, corresponde al recuento total de taxones únicos identificados en el conjunto de datos del microbioma.

Pasando a la arquitectura, nuestro modelo consta de dos capas ocultas ubicadas entre las etapas de entrada y salida. La capa oculta principal incorpora un LSTM con nh estados ocultos, que funcionan dentro de una sola capa. Esta configuración es crucial, ya que permite que el modelo capture e interprete la dinámica temporal inherente a la secuencia de entrada, cortesía de las células de memoria características del LSTM.

Para abordar el sobreajuste y mejorar la solidez del modelo, se implementa una estrategia de abandono después de la capa LSTM. Esta estrategia, regida por una probabilidad de abandono p preespecificada , implica la desactivación arbitraria de nodos, fortaleciendo la capacidad de generalización del modelo . Los nodos que no se ven afectados por la deserción pasan luego a la capa siguiente: un estrato completamente conectado que contiene nodos n_{fc} .

La capa oculta secundaria emplea la función de activación de Unidad lineal rectificada (ReLU) en puntos de datos derivados de la capa completamente conectada. Esto imparte una no linealidad esencial, preparando el modelo para discernir patrones complejos dentro del conjunto de datos. Las predicciones se formulan en función del resultado de esta capa.

En resumen, nuestro modelo MicroGrowthPredictor para predecir la dinámica del microbioma integra capas especialmente diseñadas, cada una diseñada para interpretar la dinámica temporal matizada en los datos del microbioma. La arquitectura comienza con una capa de entrada que aloja nodos representativos de taxones n_{taxa} , pasando a un LSTM de una sola capa con nh estados ocultos.

Si bien no se detalla explícitamente, suponemos que la capa LSTM conserva la composición convencional de las celdas LSTM, incluidas las puertas de entrada, olvido y salida para una transferencia efectiva de información. Esta estructura es fundamental para permitir que el modelo aprenda y preserve las dependencias a largo plazo inherentes a los datos secuenciales.

Después de la capa LSTM, se aplica una técnica de abandono con una probabilidad p designada para que sirva como mecanismo de regularización y mitigue los riesgos de sobreajuste. Posteriormente, se introduce una capa completamente conectada con nodos N_{FC} , que culmina en una capa densa capaz de capturar interdependencias no lineales en los datos. La fase final del modelo incorpora una función de activación ReLU, que introduce no linealidad y mejora la complejidad del modelo para una interpretación detallada de los datos. Esta etapa es crucial para dar forma al resultado final, asegurando predicciones precisas y fluidas en medio del panorama dinámicamente cambiante de los datos del microbioma.

2.3. Entrenamiento del modelo MicroGrowthPredictor

Cuando se trabaja con datos de series temporales y se utiliza LSTM para predecir el impacto de las perturbaciones ambientales, la validación cruzada debe tener en cuenta las dependencias temporales inherentes a los datos. Para garantizar predicciones sólidas y precisas, empleamos un método de validación cruzada de series temporales utilizando un enfoque de ventana móvil. El conjunto de datos se dividió en K pliegues consecutivos sin barajar. Para cada pliegue k, el modelo se entrenó en los primeros k pliegues y se probó en el pliegue k + 1, repitiendo hasta que cada pliegue sirvió como conjunto de prueba. Este método garantiza que se respeten las dependencias temporales y evita la fuga de datos.

Se recopilaron métricas de evaluación, como el error cuadrático medio (MSE), para cada pliegue y se calculó el rendimiento promedio en todos los pliegues para evaluar la solidez del modelo.

2.4. Intervalo de predicción

Si bien los enfoques tradicionales para establecer intervalos de confianza o predicción en modelos de aprendizaje profundo enfrentan desafíos considerables debido a la no linealidad y la arquitectura compleja de estos modelos, los avances recientes han comenzado a allanar el camino para soluciones más sólidas. Uno de esos avances es el trabajo de [13], en el que se aprovechó el marco de abandono de Monte Carlo (MC) para introducir un método que, si bien es eficaz, deja espacio para un mayor refinamiento y aplicación en nuevos dominios, como el análisis de datos del microbior

Nuestra investigación se basa en este trabajo fundamental y adopta el principio de abandonos estocásticos después de cada capa oculta en la arquitectura de la red neuronal. Sin embargo, ampliamos este concepto adaptando el proceso de abandono y la interpretación posterior de los resultados del modelo específicamente a las características y complejidades de los datos del microbioma.
Esta adaptación no solo permite la interpretación teórica del resultado del modelo como una muestra aleatoria de la distribución predictiva posterior, sino que también reconoce el comportamiento único de los datos en los estudios de microbioma.

El proceso de construir una distribución empírica de los valores predichos tratando cada predicción durante el abandono como una muestra de la distribución de datos subyacente representa un enfoque matizado en nuestro estudio. Se diferencia de las técnicas clásicas al proporcionar una ventana a las capacidades e incertidumbres predictivas del modelo específicamente ajustadas al contexto del microbioma, reforzando así la solidez de la toma de decisiones basada en estas predicciones.

En nuestro enfoque, denotamos los datos de prueba que se pretende predecir con el superíndice . La base del intervalo de predicción reside en la probabilidad condicional p(y |x , D). Esta probabilidad se puede expresar como la integral del producto de p(y |x , θ) y p(θ|D) sobre el vector de parámetros θ, denotado de la siguiente manera:

$$p(y \mid x, D) = \int_{\theta} p(y \mid x, \theta) p(\theta \mid D) d\theta.$$

θ representa el vector de parámetros del modelo de aprendizaje profundo y p (θ | D) corresponde a la distribución posterior. Sin embargo, derivar una forma analítica para p(y |x , θ) generalmente no es factible. Para superar este desafío, en la ref. se propone una técnica de aproximación que utiliza una distribución variacional denominada q (θ) . [14]. En consecuencia, se obtiene la siguiente aproximación:

$$p(y \mid x, D) \approx \int_{\theta} p(y \mid x, \theta) q(\theta) d\theta \approx \frac{1}{K} \sum_{k=1}^{k} p(y \mid x, \hat{\theta}_k), \quad (1)$$

donde ˆθk q(θ). Esta aproximación final, lograda mediante el muestreo de { ˆθk}k=1,...,K de la distribución variacional q(θ), emplea la técnica de integración de Monte Carlo.
Además, esta aproximación es equivalente a implementar el algoritmo de abandono de Monte Carlo introducido en [13]. En esencia, para un punto de datos de prueba dado (x), la distribución empírica predictiva, y con resultante sirve se evalúa varias veces en x abandono aleatorio de nodos, y la salida tiva y como una estimación de p(y |x , D). Intervalos de predicción

capturar la variabilidad que se origina a partir de dos fuentes principales: la incertidumbre del modelo (η1) y el ruido inherente (η2).

Los siguientes pasos describen el proceso: Para cada punto de datos individual x en el conjunto de pruebas, calcule la salida correspondiente y˜ eliminando aleatoriamente cada nodo con una probabilidad de abandono determinada p. Repita este proceso B veces para obtener una gran cantidad de valores predichos y˜, cada uno de los cuales varía debido a la eliminación aleatoria de nodos. A continuación, calcule la incertidumbre del modelo η1 calculando la diferencia cuadrática promedio entre cada valor predicho y˜ y la media de todos los valores predichos $\bar{y}$˜ . Esto se hace usando la fórmula $\eta 1 = \frac{1}{B}\sum^{B}(\tilde{y}_i - \bar{\tilde{y}})^2$. Para cuantificar el ruido inherente en las predicciones, calcule la diferencia cuadrática promedio entre cada valor predicho y˜ y su correspondiente valor j verdadero y j utilizando el conjunto de datos de prueba de longitud V. Esto nos da el ruido inherente η2, calculado como $\frac{1}{V}\sum_{j=1}^{V}(\tilde{y}_j - y_j)^2$. Combinando la incertidumbre del modelo y el ruido inherente, calcule η2 = incertidumbre general η como la raíz cuadrada de la suma de η1 y η2, es decir, $\eta = \sqrt{\eta 1 + \eta 2}$. Finalmente, determine los límites superior e inferior del intervalo de predicción sumando y restando zα/2 veces η del valor medio predicho $\bar{y}$˜ . Aquí, zα/2 representa la puntuación z asociada con el nivel de confianza deseado (1 − α)100%. Un algoritmo formal se enumera en el Algoritmo 1.

Algoritmo 1: Red neuronal LSTM e intervalo de predicción. , p, t, nh , nf c

```
Requerir: x, y, x , y
Asegúrese de que:
θ, U, L 1
    repita z1 ← x de la capa LSTM con t y nh ; z2 ←
    z1 mediante abandono aleatorio con p; z3 ← z2 de
    la capa completamente conectada con nf c nodos;
Aplique ReLU a z3; yˆ ←
    z3 de la capa de salida;
    Evalúe yˆ con y;
Actualización θ para el modelo
 mθ ; hasta la última
época; para i = 1 a
    y˜ B hacer ← mθ (x ) con abandono aleatorio;
i fin
Calcular ¯y˜ y η;
U, L ← ¯y˜ ± zα/2 × η;
```

2.5. Ajuste de parámetros

Para optimizar el rendimiento de nuestro modelo MicroGrowthPredictor, empleamos un proceso de ajuste de dos pasos.

En el primer paso, preseleccionamos el número de unidades ocultas en la capa LSTM (nh) y la capa completamente conectada (nf c) según experimentos preliminares. Luego exploramos diferentes combinaciones de la probabilidad de abandono (p) y la longitud de la secuencia (T), que representa el número de puntos de datos anteriores utilizados como características para la predicción. El rendimiento del modelo se evalúa calculando el error cuadrático medio (MSE) en un conjunto de datos de prueba separado, y seleccionamos la combinación de p y T que minimiza este error.

Una vez que se determinan la probabilidad de abandono óptima y la longitud de la secuencia, procedemos al segundo paso, donde ajustamos la cantidad de nodos tanto en el LSTM como en las capas completamente conectadas. Para cada combinación de arquitectura, entrenamos el modelo varias veces con diferentes inicializaciones para tener en cuenta las variaciones introducidas por la abandono aleatoria y la configuración de peso inicial. Calculamos el MSE para cada ejecución de entrenamiento y seleccionamos la arquitectura que genera el error más bajo en el conjunto de datos de prueba.

Este riguroso proceso de ajuste garantiza que nuestro modelo MicroGrowthPredictor esté configurado de manera óptima para el conjunto de datos específico bajo consideración, mejorando así su rendimiento predictivo.

3. Resultados

En este estudio, empleamos el modelo MicroGrowthPredictor y el procedimiento de ajuste asociado para dos conjuntos de datos distintos: el conjunto de datos del antibiótico ciprofloxacina (Cp) de [15] y el conjunto de datos de intestino artificial detallado en [16]. Ambos conjuntos de datos ofrecen información sobre la dinámica temporal del microbioma bajo diversas perturbaciones ambientales.

3.1. La referencia del conjunto

de datos de ciprofloxacina [15] subraya las importantes alteraciones impuestas a la composición y diversidad de la microbiota intestinal humana debido a los tratamientos antibióticos recurrentes. Esta investigación implicó una vigilancia en profundidad de las comunidades bacterianas en el intestino distal en tres sujetos (D, E y F). Se recogieron muestras de heces periódicamente durante diez meses, sumando entre 52 y 56 muestras por individuo. Dentro de este período de tiempo, a cada sujeto se le administraron dos regímenes separados de 5 días del antibiótico ciprofloxacina (Cp), con un intervalo de 6 meses. El muestreo intenso (diario durante dos períodos de 19 días coincidiendo con cada ciclo de Cp) proporcionó una perspectiva detallada del microbioma durante la exposición a los antibióticos. Fuera de estas ventanas, se adquirieron muestras semanal o mensualmente, capturando la composición microbiana en ausencia de tratamiento.

Con fines ilustrativos, nos centramos en el tema D. Nuestro proceso de optimización implica generar un gráfico de contorno del error cuadrático medio (MSE) frente a valores variables de la probabilidad de abandono p y el número de pasos de tiempo. La Figura 2 visualiza esta relación, guiando nuestra selección de una combinación óptima para refinar el modelo MicroGrowthPredictor . El gráfico de contorno del error cuadrático medio se traza con la probabilidad de abandono p en el eje x y el número de pasos de tiempo en el eje y. En el gráfico de contorno, cuanto más oscuro es el sombreado, menor es el error. Utilizamos una función de optimización para identificar la mejor combinación de probabilidad de abandono y longitud de secuencia.

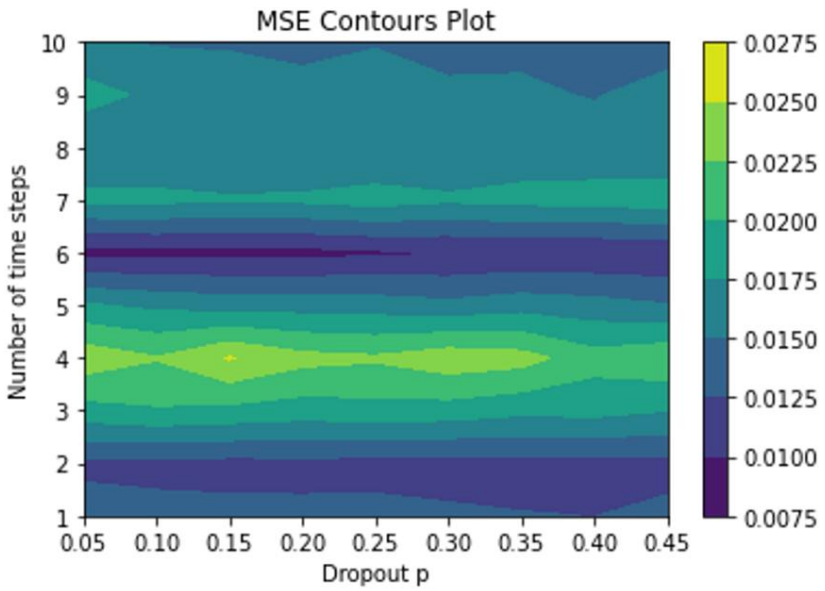

Figura 2. Gráfico de contorno del error cuadrático medio sobre p y t para el sujeto D EU766613: cuanto más oscuro es el gráfico de contorno, menor es el error. Podemos identificar la mejor combinación de probabilidad de abandono y duración de la secuencia.

Posteriormente, nuestro enfoque cambia a determinar el recuento óptimo de nodos tanto para LSTM como para las capas completamente conectadas, como se muestra en la Figura 3. El eje x representa el número de estados ocultos en una sola capa de LSTM, y el eje y representa el número de nodos en la capa completamente conectada. Diferentes combinaciones dan como resultado cambios en el valor del error cuadrático medio. El gráfico de contorno proporciona una representación directa del MSE más pequeño, indicado por el área más oscura de la figura.

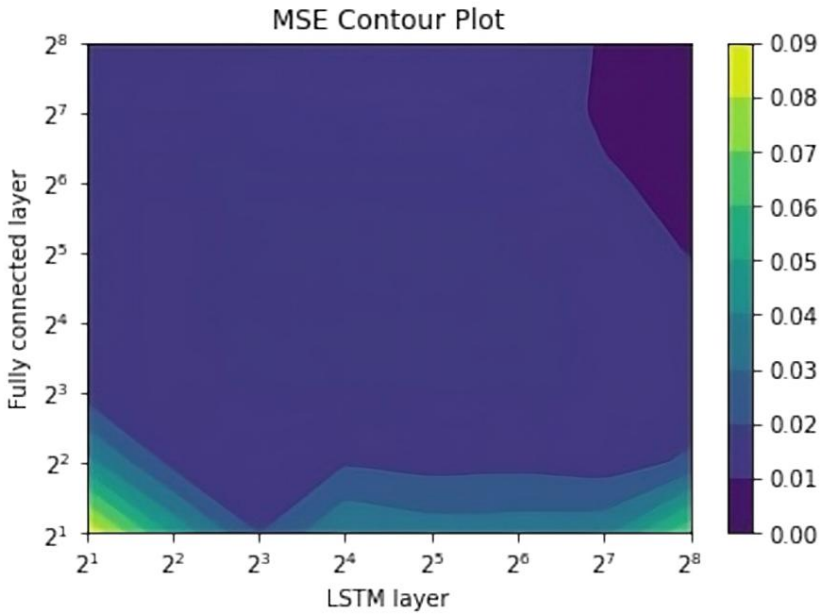

Figura 3. Gráfico de contorno del error cuadrático medio sobre nf c y nh para el sujeto D EU766613: el eje x representa el número de estados ocultos en la capa LSTM única y el eje y representa el número de nodos en la capa completamente conectada capa. Con diferentes combinaciones, el valor cuadrático medio cambia. Básicamente, el gráfico de contorno nos da una impresión directa del MSE más pequeño, que está representado por el área más oscura de la figura.

A través de esta exploración sistemática, nuestro objetivo sigue siendo consistente: identificar una configuración que minimice el error del conjunto de datos de prueba, mejorando así la eficacia de MicroGrowthPredictor.

Es esencial tener en cuenta que en nuestro conjunto de datos de entrenamiento, incluimos dos tercios de los datos observados, con el objetivo de proporcionar una base sólida para el modelo. En particular, hubo dos puntos de datos correspondientes a la administración de antibióticos para cada paciente. Uno de estos puntos se incluyó en el conjunto de entrenamiento, mientras que el otro se reservó para el conjunto de predicción. Según nuestras observaciones, la reacción al primer antibiótico mostró una respuesta tardía en comparación con el segundo. Esta observación explica por qué nuestros datos pronosticados demuestran un patrón retrasado en la Figura 4.

La información temporal proporcionada por la visualización de las trayectorias de la abundancia relativa del microbioma fue crucial para comprender la dinámica de los cambios del microbioma y sus posibles implicaciones para la salud del huésped. Para aclarar más, la Figura 4 presenta un análisis y predicción de la abundancia relativa del bacteroide EU766613 para el sujeto D, utilizando los parámetros óptimos antes mencionados. Los intervalos de administración de antibióticos se indican con una línea vertical de puntos azules, mientras que la demarcación de puntos rojos segrega los períodos de entrenamiento y prueba. En nuestro estudio sobre tratamientos antibióticos repetidos, priorizamos la inclusión de datos extensos sobre intervenciones con antibióticos para reforzar el poder predictivo de nuestro modelo. Este enfoque basado en datos mejora la precisión de las predicciones de tratamientos posteriores, ofreciendo una herramienta fundamental para combatir la resistencia a los antibióticos mediante la aplicación estratégica e informada de terapias.

La visualización subraya la capacidad del modelo MicroGrowthPredictor para comprender la dinámica del microbioma y formular predicciones basadas en estos patrones identificados. Esto se logra con el modelo entrenado durante 200 épocas utilizando una tasa de aprendizaje de 0,001. Además, la pérdida por error cuadrático medio para los datos de entrenamiento es 0,00081 y para los datos de prueba, es 0,01021.

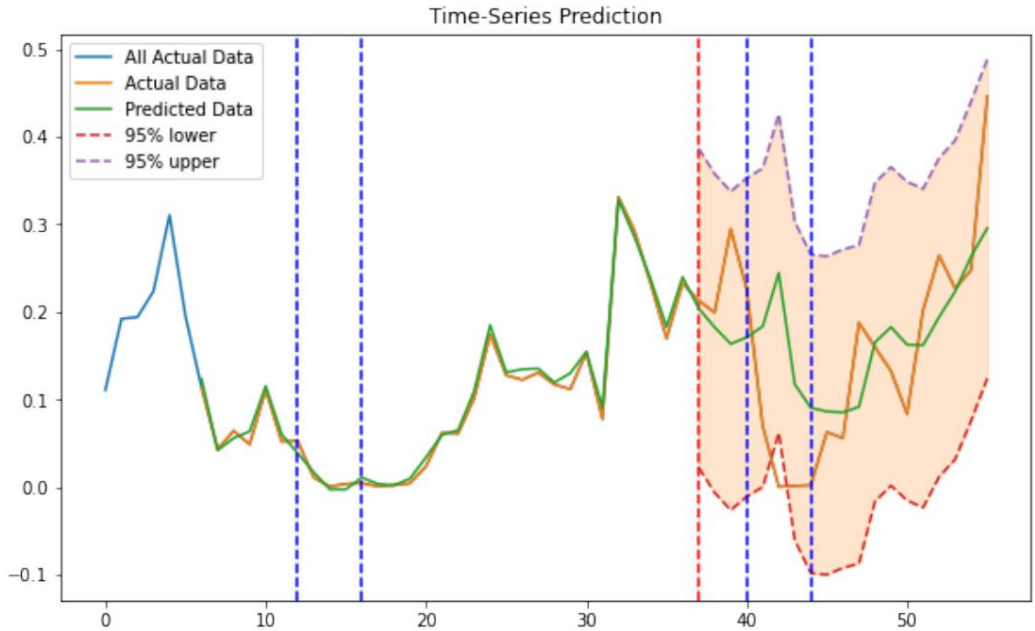

Figura 4. Trayectorias de abundancia relativa del bacteroide EU766613 para el sujeto D. Los parámetros elegidos son p = 0,05, t = 6, nh = 256 y nf = 256. Las bandas verticales azules representan los dos períodos de tratamiento con antibióticos y las La línea de puntos roja divide los datos en entrenamiento y prueba.

3.2. Conjunto de datos del

intestino artificial El conjunto de datos proporcionado por [16] comprende lecturas resueltas en el tiempo de la microbiota intestinal procedente de un intestino humano artificial. Estos datos, capturados tanto diariamente como cada hora, se originan a partir de un intestino artificial construido utilizando sistemas de biorreactores anaeróbicos de flujo continuo, lo que garantiza una representación precisa de la dinámica de la microbiota intestinal humana. Durante un mes, se cultivaron cuatro vasos ex vivo, cada uno inicializado con un inóculo fecal humano ic Para garantizar la fidelidad experimental, se mantuvieron estrictamente parámetros clave como el pH, la temperatura, la tasa de entrada del medio y la concentración de oxígeno. El día 23, la dinámica microbiana recibió un estímulo deliberado mediante la introducción de un bolo de Bacteroides ovatus, una cepa aislada del donante de heces. Sin embargo, interrupciones imprevistas en el suministro de alimento en dos embarcaciones entre los días 11 y 13 introdujeron variaciones microbianas no planificadas. En particular, observamos cambios significativos en la población de Rikenellaceae, una familia de bacterias conocida por su papel en el microbioma intestinal humano. Rikenellaceae participa en la descomposición de carbohidratos complejos y desempeña un papel crucial en el mantenimiento de la salud intestinal y las funciones metabólicas. Los cambios en esta población son particularmente interesantes porque pueden proporcionar información sobre cómo las alteraciones en la dieta y las introducciones microbianas influyen en la estabilidad y función de la microbiota intestinal.

En este ejemplo, el primer barco sirve como nuestro conjunto de entrenamiento, mientras que el segundo barco funciona como nuestro conjunto de prueba. Nuestra herramienta MicroGrowthPredictor, configurada con una probabilidad de abandono óptima (p) de 0,25, utilizó los cinco puntos de tiempo anteriores para identificar cuatro parámetros y lograr predicciones óptimas. La capa completamente conectada estaba equipada con 256 nodos y la capa LSTM estaba compuesta por 128 nodos. El modelo fue entrenado durante 800 épocas. El error cuadrático medio de los datos de entrenamiento es 0,00057 y el de los datos de prueba, 0,01456. Sin utilizar nuestro modelo predictivo, un modelo aditivo generalizado (GAM) tiene un MSE de 0,0048 para los datos de entrenamiento, que es aproximadamente 8,42 veces mayor. El rendimiento de los datos de prueba es significativamente peor, por lo que no se incluye para comparación

Las trayectorias de abundancia relativa de microbiomas visualizadas en la Figura 5 brindan información crítica sobre la dinámica de los cambios de microbiomas a lo largo del tiempo. La línea azul en la Figura 5 representa todos los datos reales, mientras que la línea naranja se resalta simultáneamente con la línea prevista (verde). En nuestro algoritmo de aprendizaje profundo, utilizamos los cinco puntos de tiempo anteriores para predecir el siguiente. Variaciones notables, especialmente para Rikenellaceae,

se observaron debido a la interrupción de los dos primeros vasos entre los días 11 y 13. Estas visualizaciones revelan cambios significativos en las poblaciones microbianas, lo que subraya la precisión del modelo para capturar cambios temporales. Los patrones observados se alinean con nuestros análisis estadísticos, lo que confirma cambios sustanciales en la composición del microbioma durante las perturbaciones. Esta alineación fortalece nuestra comprensión de la dinámica del microbioma y sus respuestas a las condiciones experimentales.

Contrariamente a la idea de que más datos conducen a mejores predicciones, nuestro experimento que involucró embarcaciones de entrenamiento adicionales (incluidas 1, 3 y 4) para predecir la segunda embarcación arrojó un error cuadrático medio para la prueba de 0,0265, casi el doble del error de la prueba original. Curiosamente, la correlación entre el valor previsto y el valor real para probar el vaso 2 es 0,70, que es un 18% mayor que el caso cuando incluimos los vasos 1, 3 y 4.
Esto sugiere que un equilibrio cuidadoso en la selección de datos de entrenamiento es crucial para lograr predicciones precisas.

En el ámbito de los estudios científicos, a menudo prevalece la creencia de que incorporar más conjuntos de datos o información para la capacitación conduce a una mayor precisión. Sin embargo, surge una consideración crítica cuando el entorno en el que se entrena el modelo difiere significativamente del entorno en el que se aplicará para las pruebas. Esta disyunción en las condiciones ambientales puede introducir perturbaciones y desafíos imprevistos.

En nuestro experimento, los resultados observados cuestionaron la suposición inicial de que más datos de entrenamiento (incluidos los vasos 1, 3 y 4) mejorarían inherentemente las predicciones. Las interrupciones en el suministro de alimento en los dos primeros recipientes entre los días 11 y 13 crearon variaciones en la dinámica microbiana que no fueron capturadas adecuadamente por los datos de entrenamiento adicionales. Las interrupciones imprevistas subrayan la importancia de alinear los datos de entrenamiento con las condiciones y perturbaciones esperadas en el entorno de prueba.

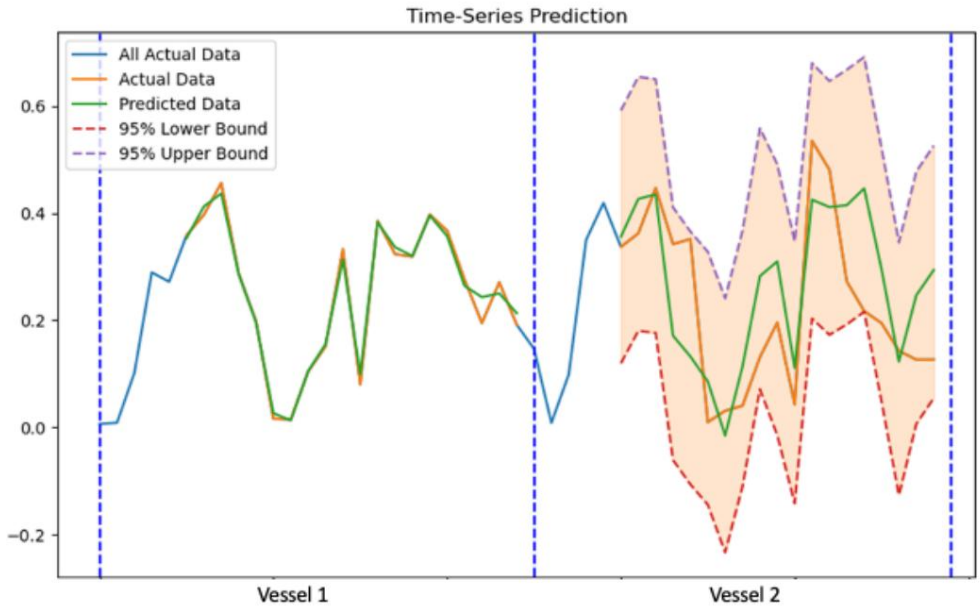

Figura 5. Trayectorias de la abundancia relativa de Rikenellaceae en los recipientes 1 y 2. La trayectoria completa del recipiente 2 se predice mediante el modelo MicroGrowthPredictor entrenado con datos del recipiente 1. Se proporcionan intervalos de confianza para los datos de prueba del recipiente 2. En este experimento, se utilizó la probabilidad óptima de abandono p de 0,25. El modelo utilizó los cinco puntos temporales anteriores para identificar cuatro parámetros y lograr predicciones óptimas. La capa completamente conectada estaba equipada con 256 nodos y la capa LSTM estaba compuesta por 128 nodos. El modelo fue entrenado durante 800 épocas.

Si bien es tentador suponer que un tamaño de muestra más grande conducirá inherentemente a mejores predicciones, la clave radica en la relevancia de los datos de entrenamiento para las condiciones de prueba. En los casos en que los datos de prueba impliquen diferentes interrupciones o perturbaciones ambientales,

La inclusión ciega de diversos conjuntos de datos puede conducir a predicciones subóptimas. El delicado equilibrio entre la cantidad y la relevancia de los datos de entrenamiento se vuelve crucial para garantizar la adaptabilidad del modelo a escenarios del mundo real.

4. Discusión

El modelo MicroGrowthPredictor aprovecha el conocimiento de las observaciones de que el tratamiento repetido con antibióticos altera la comunidad microbiana intestinal, afectando la diversidad y abundancia de grupos bacterianos específicos. Al analizar los datos, el modelo predice con precisión cómo cambiará el microbioma con el tiempo en respuesta a la perturbación de los antibióticos.

Esto proporciona una comprensión más profunda del impacto de los antibióticos en la microbiota intestinal y sus posibles implicaciones para la salud humana.

Además, la versatilidad del modelo se demuestra mediante su aplicación a un conjunto de datos de intestino artificial. Los conocimientos extraídos de este entorno controlado muestran la adaptabilidad de MicroGrowth-Predictor a diversos sistemas de microbiomas. El conjunto de datos del intestino artificial valida las capacidades predictivas del modelo en condiciones específicas, destacando su competencia para capturar dinámicas temporales intrincadas. Esto posiciona al modelo como valioso para comprender los efectos de los antibióticos y aplicaciones más amplias en ciencias ambientales, atención médica y biotecnología.

Nuestro método aborda problemas del mundo real donde los tamaños de muestra limitados son una limitación debido a desafíos logísticos, éticos o financieros. Al desarrollar y validar métodos que funcionan bien con datos limitados, brindamos soluciones prácticas para tales situaciones.

A diferencia de muchos modelos de caja negra, nuestro enfoque ofrece información clara sobre cómo las perturbaciones ambientales influyen en las poblaciones microbianas a lo largo del tiempo, algo crucial para comprender los procesos biológicos y diseñar intervenciones específicas. Específicamente, discutimos su potencial para contribuir a planes de tratamiento personalizados al predecir respuestas individuales a cambios en la dieta, tratamientos con antibióticos e intervenciones con probióticos.

En resumen, MicroGrowthPredictor surge como una potente herramienta que supera los enfoques de modelado tradicionales. El modelo, impulsado por conocimientos derivados de datos en lugar de la integración directa de conocimientos, incorpora redes LSTM con estimación de intervalos de confianza para contribuir a una comprensión holística de la dinámica del microbioma. Las aplicaciones exitosas del modelo tanto a la microbiota intestinal humana del mundo real como a conjuntos de datos de intestino artificial subrayan su eficacia y su impacto potencial. Prevemos que MicroGrowthPredictor desempeñará un papel fundamental en el avance de la investigación del microbioma, ofreciendo información valiosa y contribuyendo a una toma de decisiones bien informada en diversos campos.

Contribuciones de los autores: Conceptualización, Y.-HZ; metodología, GS e Y.-HZ; validación, GS e Y.-HZ; redacción: borrador original, GS e Y.-HZ; redacción: revisión y edición, GS e Y.- HZ; visualización, GS e Y.-HZ; supervisión, Y.-HZ; administración de proyectos, Y.-HZ; adquisición de financiación , Y.-HZ Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación fue financiada por la Agencia de Protección Ambiental de EE. UU., subvención número 84045001, el Instituto Nacional de Salud P30ES025128 y el Programa de Centros de Investigación en Ingeniería de la Fundación Nacional de Ciencias bajo el acuerdo cooperativo NSF No. EEC-2133504.

Declaración de disponibilidad de datos: los datos están contenidos en el artículo.

Conflictos de intereses: Los autores declaran que la investigación se realizó en ausencia de relaciones comerciales o financieras que pudieran construirse como un potencial conflicto de intereses.

Referencias

1. Altvés, S.; Yildiz, HK; Vural, HC Interacción de la microbiota con el cuerpo humano en la salud y la enfermedad. Biosci. Salud alimentaria de la microbiota 2020, 39, 23–32. [Referencia cruzada] [PubMed]
2. Smith, J.; Johnson, M. Dinámica del microbioma bajo perturbaciones ambientales. J. Microbioma Res. 2022, 10, 123–145.
3. Marrón, EM; Sadarangani, M.; Finlay, BB El papel del sistema inmunológico en el control de las interacciones huésped-microbio en el intestino. Nat. Inmunol. 2013, 14, 660–667. [Referencia cruzada] [PubMed]
4. Candela, M.; Biagi, E.; Maccaferri, S.; Turroní, S.; Brigidi, P. La microbiota intestinal es un factor plástico que responde a los cambios ambientales. Tendencias Microbiol. 2012, 20, 385–391. [Referencia cruzada]

5. Uhr, GT; Dohnalová, L.; Thaiss, CA La dimensión del tiempo en las interacciones huésped-microbioma. mSystems 2019, 4, e00216-18. [Referencia cruzada] [PubMed]
6. Dispuesto, BP; Russell, SL; Finlay, BB Cambiando el equilibrio: efectos de los antibióticos en el mutualismo huésped-microbiota. Nat. Rev. Microbiol. 2011, 9, 233–243. [Referencia cruzada] [PubMed]
7. Marrón, E.; Williams, D. Modelado predictivo del crecimiento del microbioma utilizando redes LSTM. J. Computación. Biol. 2021, 45, 321–335.
8. Ching, T.; Himmelstein, DS; Beaulieu-Jones, BK; Kalinin, AA; Hazlo, BT; Camino, médico de cabecera; Ferrero, E.; Agapow, PM; Zietz, M.; Hoffman, MM; et al. Oportunidades y obstáculos para el aprendizaje profundo en biología y medicina. JR Soc. Interfaz 2018, 15, 20170387. [Referencia cruzada] [PubMed]
9. Medsker, LR; Jain, L. Redes neuronales recurrentes. Des. Aplica. 2001, 5, 2.
10. Tumbas, A.; Graves, A. Memoria larga a corto plazo. En Etiquetado de secuencias supervisadas con redes neuronales recurrentes; Saltador: Berlín/Heidelberg, Alemania, 2012; págs. 37–45.
11. Yu, Y.; Seis.; Hu, C.; Zhang, J. Una revisión de redes neuronales recurrentes: células LSTM y arquitecturas de red. Computación neuronal. 2019, 31, 1235–1270. [Referencia cruzada] [PubMed]
12. Zhou, YH; Gallins, P. Una revisión y un tutorial de métodos de aprendizaje automático para la predicción de rasgos del microbioma del huésped. Frente. Gineta. 2019, 10, 579. [Referencia cruzada] [PubMed]
13. Zhu, L.; Laptev, N. Predicción profunda y segura de series temporales en uber. En Actas de la Conferencia Internacional IEEE sobre Talleres de Minería de Datos (ICDMW) de 2017, Orleans, LA, EE. UU., 18 a 21 de noviembre de 2017; IEEE: Piscataway, Nueva Jersey, EE. UU., 2017; págs. 103-110.
14. Gal, Y.; Ghahramani, Z. La deserción como aproximación bayesiana: representación de la incertidumbre del modelo en el aprendizaje profundo. En Actas de la Conferencia Internacional sobre Aprendizaje Automático, PMLR, Nueva York, NY, EE. UU., 20 a 22 de junio de 2016; págs. 1050-1059.
15. Dethlefsen, L.; Relman, DA Recuperación incompleta y respuestas individualizadas de la microbiota intestinal distal humana a la perturbación repetida de los antibióticos. Proc. Nacional. Acad. Ciencia. EE.UU. 2011, 108, 4554–4561. [Referencia cruzada] [PubMed]
16. Silverman, JD; Durand, HK; Bloom, RJ; Mukherjee, S.; David, LA Los modelos lineales dinámicos guían el diseño y análisis de Estudios de microbiota dentro de intestinos humanos artificiales. Microbioma 2018, 6, 202.