applied microbiology

MDPI

Статья

Прогнозирование динамики роста микробиома в условиях Экологические возмущения

Джордж Сан [1] и И-Хуэй Чжоу 2,*

[1] Исследовательский центр биоинформатики, Университет штата Северная Каролина, Роли, Северная Каролина 27695, США;
[2] 3litus@gmail.com Факультет биологических наук и статистики, Университет штата Северная Каролина, Роли, Северная Каролина 27695, США
* Переписка: yihui_zhou@ncsu.edu

Аннотация: MicroGrowthPredictor — это модель, которая использует сети долговременной краткосрочной памяти (LSTM) для прогнозирования динамических изменений роста микробиома в ответ на различные возмущения окружающей среды. В этой статье мы представляем инновационные возможности MicroGrowthPredictor, которые включают интеграцию моделирования LSTM с новым методом оценки доверительного и Сеть LSTM фиксирует сложную временную динамику систем микробиома, а новые доверительные интервалы обеспечивают надежную меру неопределенности прогноза. Мы включаем два примера : один иллюстрирует состав и разнообразие микробиоты кишечника человека в результате повторного лечения антибиотиками, а другой демонстрирует применение MicroGrowthPredictor на наборе данных искусственного кишечника . Результаты демонстрируют повышенную точность и надежность прогнозов на основе LSTM, обеспечиваемых MicroGrowthPredictor. Включение конкретных показателей, таких как среднеквадратическая ошибка, подтверждает прогнозирующую эффективность модели. Наша модель имеет огромный потенциал для применения в науках об окружающей среде, здравоохранении и биотехнологиях, способствуя прогрессу в исследованиях и анализе микробиома. Более того, примечательно, что MicroGrowthPredictor применим к реальным данным с небольшими размерами выборки и временными наблюдениями в условиях возмущений окружающей среды, что обеспечивает его практическую полезность в различных областях.

Ключевые слова: динамика микробиома; неопределенность прогноза; экологические приложения

Цитирование: Сунь, Г.; Чжоу, Ю.-Х. Прогнозирование динамики роста микробиома при возмущениях окружающей среды. Прил. Микробиол. 2024, 4, 948–958. https://doi.org/10.3390/applmicrobiol4020064

Академический редактор: Бонг Су Ким

Поступила: 7 мая 2024 г.
Пересмотрено: 4 июня 2024 г.
Принято: 7 июня 2024 г.
Опубликовано: 10 июня 2024 г.

1. Введение

Микробиом человека, сложная экосистема из триллионов микроорганизмов, живущих в организме человека и на его поверхности, играет решающую роль в поддержании физиологического гомеостаза, метаболических функций и иммунных реакций [1]. Нарушения в микробиоме связаны с множеством состояний, начиная от желудочно-кишечных расстройств и заканчивая более системными заболеваниями, такими как диабет, ожирение и даже неврологические расстройства [2]. Такое симбиотическое взаимодействие хозяина и микроба подчеркивает необходимость понимания динамической природы микробиома человека [3], особенно того, как он меняется с течением времени и в ответ на различные стимулы окружающей среды [4,5].

В нормальных условиях микробиом кишечника состоит из разнообразного сообщества бактерий, преобладающими из которых являются Firmicutes и Bacteroidetes. Возмущения окружающей среды , такие как изменения в рационе питания, использование антибиотиков и воздействие загрязняющих веществ, могут значительно изменить состав и функцию микробиома, что может привести к потенциальным последствиям для здоровья. Например, лечение антибиотиками может резко сократить микробное разнообразие, что часто приводит к чрезмерному росту резистентных бактерий и уменьшению количества полезных микробов, что может нарушить метаболические процессы и иммунные функции [6]. Понимание этой демографической динамики имеет решающее значение для разработки стратегий по смягчению неблагоприятного воздействия таких возмущений на здоровье человека.

Технологии высокопроизводительного секвенирования, в частности секвенирование 16S рРНК, открыли новую эру в исследованиях микробиома, позволяя проводить детальную оценку микробного разнообразия и относительной численности в различных человеческих популяциях и состояниях [7]. Однако огромные данные, генерируемые этими технологиями, представляют как возможности, так и проблемы.

Одной из основных задач является расшифровка временных закономерностей и прогнозирование будущих состояний микробиома, что необходимо для профилактических и терапевтических применений в здравоохранении.

Для прогнозного моделирования в биометрии исторически использовались различные статистические методы , но эти традиционные подходы часто не справляются с многомерностью и нелинейностью данных микробиома. Появление машинного обучения, а точнее глубокого обучения, открывает новые многообещающие возможности для обработки таких сложных данных [8]. Рекуррентные нейронные сети (RNN) [9] и их усовершенствованный вариант, сети долгосрочной краткосрочной памяти (LSTM) [10], превосходны в анализе и прогнозировании временных последовательностей, обеспечивая отличную основу для моделирования динамики микробиома.

В этом исследовании мы представляем модель MicroGrowthPredictor, целью которой является использование возможностей сетей LSTM для прогнозирования изменений в микробиоме человека в ответ на возмущения окружающей среды, что является важным шагом на пути к персонализированной медицине и целенаправленным терапевтическим вмешательствам.

2. Материалы и методы 2.1.

Модель долговременной краткосрочной памяти (LSTM)

Сеть долгосрочной краткосрочной памяти (LSTM), специализированная форма архитектуры рекуррентной нейронной сети (RNN), специально разработана для решения проблем обучения на последовательных данных, особенно на долгосрочных зависимостях. Традиционные RNN, теоретически способные обрабатывать такие зависимости, на практике часто терпят неудачу из-за проблемы исчезновения градиента, когда информация теряется на каждом временном шаге во время обучения. Сети LSTM предназначены для преодоления этого ограничения, что делает их особенно подходящими для приложений в различных областях, таких как анализ временных рядов, обработка естественного языка и, что имеет отношение к нашей работе, анализ данных микробиома.

Сети LSTM представляют более сложную структуру ячеек, чем традиционные RNN [11]. Каждая ячейка LSTM содержит механизмы, называемые воротами, которые регулируют поток информации в ячейку и из нее. В ячейке LSTM есть три типа ворот (рис. 1A):

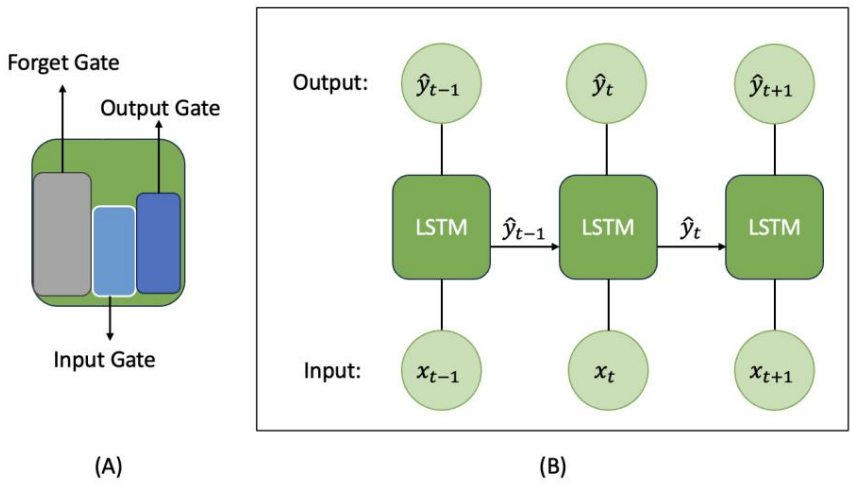

Рисунок 1. Архитектура длинной краткосрочной памяти (LSTM): (A) Увеличенное изображение ячейки LSTM, показывающее три ее вентиля: входной вентиль, вентиль забывания и выходной вентиль. (B) Поток входных и выходных данных в сети LSTM от временного шага t - 1 до временного шага t.

• Входной вентиль: модулирует количество новой информации, добавляемой к состоянию ячейки. • Ворота забывания: определяет объем информации, которая должна быть удалена из состояния ячейки.

«Ворота забывания» помогают исключить ненужную или устаревшую микробную информацию, сохраняя тем самым только наиболее важные данные для точного моделирования. • Выходной вентиль: контролирует количество информации, выводимой из ячейки.

Для данных о микробиоме выходной вентиль помогает решить, какая обработанная микробная информация должна влиять на прогнозы или анализ сети на каждом временном этапе.

Эти ворота работают вместе, обновляя состояние ячейки и позволяя LSTM запоминать и забывать информацию в длинных последовательностях (рис. 1B), что имеет решающее значение для изучения долгосрочных зависимостей. Рисунок 1B иллюстрирует переход данных через сеть LSTM от одного временного шага к другому. Он показывает входные и выходные данные по мере их прохождения от временного шага t - 1 к временному шагу t. На каждом временном шаге входные данные вместе с состоянием ячейки с предыдущего временного шага обрабатываются ячейкой LSTM. Результатом этой обработки является обновленное состояние ячейки и выходные данные, которые затем передаются на следующий временной шаг. Этот последовательный механизм позволяет сети LSTM эффективно обрабатывать временные зависимости, гарантируя, что информация переносится и используется на разных временных шагах для улучшения прогнозирования и анализа в задачах временных рядов.

В области анализа микробиома понимание временной динамики и последовательных закономерностей имеет важное значение, учитывая природу эволюции и взаимодействия микробных сообществ с течением времени. Здесь мы принимаем специальные обозначения для пояснения механики модели LS Рассмотрим набор обучающих данных $D = \{(x_t, y_t)\}_{t=1}^{T}$, где x_t обозначает вектор относительных численность [12] всех микробных таксонов на t-м временном шаге, а y_t означает соответствующий желаемый результат. LSTM принимает эти входные последовательности и обрабатывает их через свою сложную клеточную структуру, фиксируя ценные временные зависимости, присутствующие в данных, которые имеют решающее значение для точных прогнозов и анализа в исследованиях микробиома.

2.2. Структура модели для прогнозирования роста микробиома

Модель LSTM, использованная в этом исследовании, отличается простотой и эффективностью. Входной слой предназначен для обработки уровней относительной численности таксонов, вмещая обширный массив микробных таксонов, обозначенных как x_t . Состоящий из узлов n_{taxa} , каждый из которых представляет относительную численность определенного таксона, этот слой соответствует общему количеству уникальных таксонов, идентифицированных в наборе данных микробиома.

Переходя к архитектуре, наша модель состоит из двух скрытых слоев, расположенных между входным и выходным каскадами. Первичный скрытый уровень включает в себя LSTM с nh скрытыми состояниями, функционирующий в пределах одного слоя. Эта конфигурация имеет решающее значение, позволяя модели улавливать и интерпретировать временную динамику, присущую входной последовательности, благодаря характерным ячейкам памяти LSTM.

Чтобы устранить переобучение и повысить надежность модели, после уровня LSTM реализуется стратегия исключения. Эта стратегия, управляемая заранее заданной вероятностью отсева p, предполагает произвольную деактивацию узлов, усиливая способность модели к обобщению. Узлы, на которые не влияет отключение, затем передаются на следующий уровень — полностью связный слой, содержащий узлы NFC .

Вторичный скрытый слой использует функцию активации выпрямленной линейной единицы (ReLU) для точек данных, полученных из полностью связного слоя. Это придает существенную нелинейность , подготавливая модель к распознаванию сложных закономерностей в наборе данных. Прогнозы формулируются на основе выходных данных этого слоя.

Таким образом, наша модель MicroGrowthPredictor для прогнозирования динамики микробиома объединяет специально созданные слои, каждый из которых предназначен для интерпретации нюансов временной динамики в данных микробиома. Архитектура начинается с входного уровня, на котором размещаются узлы, представляющие таксоны n_{taxa} , и переходит в однослойный LSTM с nh скрытыми состояниями.

Хотя это и не подробно описано, мы предполагаем, что уровень LSTM сохраняет традиционный состав ячеек LSTM, включая входные, забывающие и выходные вентили для эффективной передачи информации. Эта структура позволяет модели изучать и сохранять долгосрочные зависимости, присущие последовательным данным.

После уровня LSTM применяется метод исключения с обозначенной вероятностью p , который служит механизмом регуляризации, снижая риски переобучения. Впоследствии вводится полностью связанный уровень с узлами NFC , кульминацией которого является плотный уровень, способный фиксировать нелинейные взаимозависимости в данных. Заключительный этап модели включает функцию активации ReLU, вносящую нелинейность и повышающую сложность модели для детальной интерпретации данных. Этот этап имеет решающее значение для формирования конечного результата, обеспечивая точные и плавные прогнозы на фоне динамически меняющегося ландшафта данных о микробиоме.

2.3. Обучение модели MicroGrowthPredictor

При работе с данными временных рядов и использовании LSTM для прогнозирования воздействия возмущений окружающей среды перекрестная проверка должна учитывать временные зависимости, присущие данным. Чтобы обеспечить надежные и точные прогнозы, мы применили метод перекрестной проверки временных рядов с использованием подхода скользящего окна. Набор данных был разделен на K последовательных складок без перемешивания. Для каждой складки k модель обучалась на первых k сгибах и тестировалась на k + 1 сгибе, повторяя до тех пор, пока каждая складка не стала тестовым набором. Этот метод обеспечивает соблюдение временных зависимостей и позволяет избежать утечки данных.

Оценочные показатели, такие как среднеквадратическая ошибка (MSE), собирались для каждого сгиба, и для оценки надежности модели рассчитывалась средняя производительность по всем сгибам.

2.4. Интервал прогнозирования

В то время как традиционные подходы к установлению доверительных или прогнозных интервалов в моделях глубокого обучения сталкиваются со значительными проблемами из-за нелинейности и сложной архитектуры этих моделей, недавние достижения начали прокладывать путь для более надежных решений. Одним из таких достижений является работа [13], в которой система отсева Монте-Карло (MC) была использована для внедрения метода, который, хотя и эффективен, оставляет место для дальнейшего совершенствования и применения в новых областях, таких как анализ данных

Наше исследование основано на этой фундаментальной работе, приняв принцип стохастических исключений после каждого скрытого слоя в архитектуре нейронной сети. Однако мы расширяем эту концепцию, адаптируя процесс отсева и последующую интерпретацию результатов модели специально к характеристикам и сложности данных микробиома.

Эта адаптация не только позволяет теоретическую интерпретацию результатов модели как случайной выборки из апостериорного прогнозируемого распределения, но также признает уникальное поведение данных в исследованиях микробиома.

Процесс построения эмпирического распределения прогнозируемых значений путем рассмотрения каждого прогноза во время исключения как выборки из основного распределения данных представляет собой детальный подход в нашем исследовании. Он отличается от классических методов, открывая окно в прогностические возможности и неопределенности модели, специально адаптированные к контексту микробиома, тем самым повышая надежность принятия решений на основе этих прогнозов.

В нашем подходе мы обозначаем тестовые данные, которые необходимо спрогнозировать, верхним индексом .

В основе интервала прогнозирования лежит условная вероятность p(y |x , D). Эту вероятность можно выразить как интеграл произведения p(y |x , θ) и p(θ |D) по вектору параметров θ, обозначенный следующим образом:

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

θ представляет вектор параметров модели глубокого обучения, а p(θ |D) соответствует апостериорному распределению. Однако получение аналитической формы для p(y |x , θ), как правило, невозможно. Чтобы преодолеть эту проблему, в ссылке предложен метод аппроксимации, использующий вариационное распределение, обозначенное как q (θ) . [14]. В результате получается следующее приближение:

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

где $\hat{\theta}_k$ q(θ). Это окончательное приближение, достигаемое путем выборки $\{\hat{\theta}_k\}_{k=1,\ldots,K}$ из вариационного распределения q(θ), использует метод интегрирования Монте-Карло.

Более того, это приближение эквивалентно реализации алгоритма исключения Монте-Карло, представленного в [13]. По сути, для данной точки данных тестирования (x) прогноз y со случайным эмпирическое оценивается несколько раз в точке x выпадением узлов и результирующее распределение выходных данных y служат оценкой p(y |x , D). Интервалы прогнозирования

улавливать изменчивость, происходящую из двух основных источников: неопределенности модели (η1) и собственного шума (η2).

Следующие шаги описывают процесс: Для каждой отдельной точки данных x тестового набора вычислите соответствующий выходной сигнал y˜ путем случайного исключения каждого узла с заданной вероятностью исключения p. Повторите этот процесс B раз, чтобы получить большое количество прогнозируемых значений y˜, каждое из которых варьируется из-за случайного исключения узлов. Затем вычислите неопределенность модели η1 путем вычисления среднего квадрата разницы между каждым прогнозируемым значением y˜ и средним значением всех прогнозируемых значений ¯y˜ . Это делается с помощью формула $\eta 1 = \frac{1}{B}\sum_{i=1}^{B}(\tilde{y}_i - \bar{\tilde{y}})^2$. Чтобы количественно оценить собственный шум в прогнозах, вычислите среднюю квадратичную разницу между каждым прогнозируемым значением y˜ и соответствующим ему истинным значением j y j , используя тестовый набор данных длиной V. Это дает нам собственный шум η2, рассчитанный как $\frac{1}{B}\sum_{j=1}^{B}(\tilde{y}_j - y_j)^2$. Объединив неопределенность модели и собственный шум, вычислите η2 = общую неопределенность η как квадратный корень из суммы η1 и η2, т.е. $\eta = \sqrt{\eta 1 + \eta 2}$. Наконец, определите верхнюю и нижнюю границы интервала прогнозирования путем добавления и вычитания zα/2, умноженного на η, из среднего прогнозируемого значения ¯y˜ . Здесь zα/2 представляет собой z-показатель, соответствующий желаемому уровню достоверности (1 α)100%. Формальный алгоритм указан в Алгоритме 1.

Алгоритм 1: нейронная сеть LSTM и интервал прогнозирования. , p, t, nh , nf c

```
Требуется: x, y, x , да
Убедитесь, что θ, U, L
1 повторяют
2     z1 x из слоя LSTM c t и nh ; z2 z1 из-за случайного
3     исключения из p; z3 z2 из полносвязного слоя с nf c
4     узлами;
5. Примените ReLU к z3; yˆ z3
6     из выходного слоя;
7     Оцените yˆ с помощью y;
8. Обновление θ для модели mθ ; 9
 до последней эпохи; 10
для i = 1 до B do mθ
11    y˜ (x ) со случайным выпадением;
я 12 конец
13. Вычислить ¯y˜ и η;
14 U, L ¯y˜ ± zα/2 × η;
```

2.5. Настройка параметров

Чтобы оптимизировать производительность нашей модели MicroGrowthPredictor, мы используем двухэтапный процесс настройки.

На первом этапе мы предварительно выбираем количество скрытых блоков в слое LSTM (nh) и полносвязном слое (nf c) на основе предварительных экспериментов. Затем мы исследуем различные комбинации вероятности выпадения (p) и длины последовательности (T), которая представляет собой количество предыдущих точек данных, используемых в качестве признаков для прогнозирования. Производительность модели оценивается путем расчета среднеквадратической ошибки (MSE) на отдельном наборе тестовых данных, и мы выбираем комбинацию p и T, которая минимизирует эту ошибку.

Как только оптимальная вероятность выпадения и длина последовательности определены, мы переходим ко второму шагу, где мы точно настраиваем количество узлов как в LSTM, так и в полносвязных слоях. Для каждой архитектурной комбинации мы обучаем модель несколько раз с разными инициализациями, чтобы учесть изменения, возникающие в результате случайного исключения и начальных настроек веса. Мы рассчитываем MSE для каждого прогона обучения и выбираем архитектуру , которая приводит к наименьшей ошибке в наборе тестовых данных.

Этот строгий процесс настройки гарантирует, что наша модель MicroGrowthPredictor оптимально настроена для конкретного рассматриваемого набора данных, тем самым повышая ее прогнозную эффективность.

3. Результаты

В этом исследовании мы используем модель MicroGrowthPredictor и связанную с ней процедуру настройки для двух разных наборов данных: набора данных по антибиотику ципрофлоксацину (Cp) из [15] и набора данных по искусственному кишечнику, подробно описанному в [16]. Оба набора данных дают представление о временной динамике микробиома при различных возмущениях окружающей среды.

3.1. Ссылка на набор данных по

ципрофлоксацину [15] подчеркивает значительные изменения, вносимые в состав и разнообразие микробиоты кишечника человека из-за повторного лечения антибиотиками. Это исследование включало углубленное наблюдение за бактериальными сообществами в дистальном отделе кишечника у трех субъектов (D, E и F). Пробы стула собирались периодически в течение десяти месяцев, в общей сложности 52–56 образцов на человека. В течение этого периода времени каждому субъекту вводили два отдельных 5-дневных режима приема антибиотика ципрофлоксацина (ЦП) с интервалом в 6 месяцев. Интенсивный отбор проб — ежедневно в течение двух 19-дневных периодов, совпадающих с каждым курсом Cp, — позволил получить детальное представление о микробиоме во время воздействия антибиотиков. За пределами этих окон образцы отбирали еженедельно или ежемесячно, фиксируя микробный состав при отсутствии лечения.

В иллюстративных целях мы сосредоточимся на предмете D. Наш процесс оптимизации включает в себя создание контурного графика среднеквадратической ошибки (MSE) в зависимости от различных значений вероятности выпадения p и количества временных шагов. На рисунке 2 визуализирована эта взаимосвязь, которая помогает нам выбрать оптимальную комбинацию для уточнения модели MicroGrowthPredictor . Контурный график среднеквадратической ошибки построен с вероятностью выпадения p по оси x и количеством временных шагов по оси y. На контурном графике чем темнее штриховка , тем меньше ошибка. Мы используем функцию оптимизации для определения наилучшего сочетания вероятности выпадения и длины последовательности.

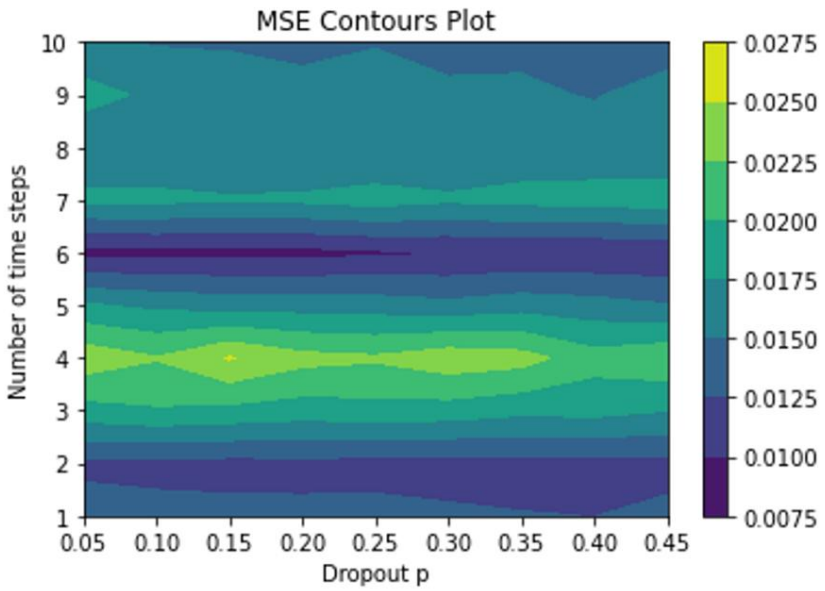

Рисунок 2. Контурный график среднеквадратической ошибки по p и t для субъекта D EU766613: Чем темнее контурный график, тем меньше ошибка. Мы можем определить наилучшую комбинацию вероятности выпадения и длины последовательности.

Впоследствии наш фокус смещается на определение оптимального количества узлов как для LSTM, так и для полностью связанных слоев, как показано на рисунке 3. Ось X представляет количество скрытых состояний в одном слое LSTM, а ось Y представляет число узлов в полносвязном слое. Различные комбинации приводят к изменению значения среднеквадратической ошибки . Контурный график обеспечивает прямое представление наименьшей MSE, обозначенной самой темной областью на рисунке.

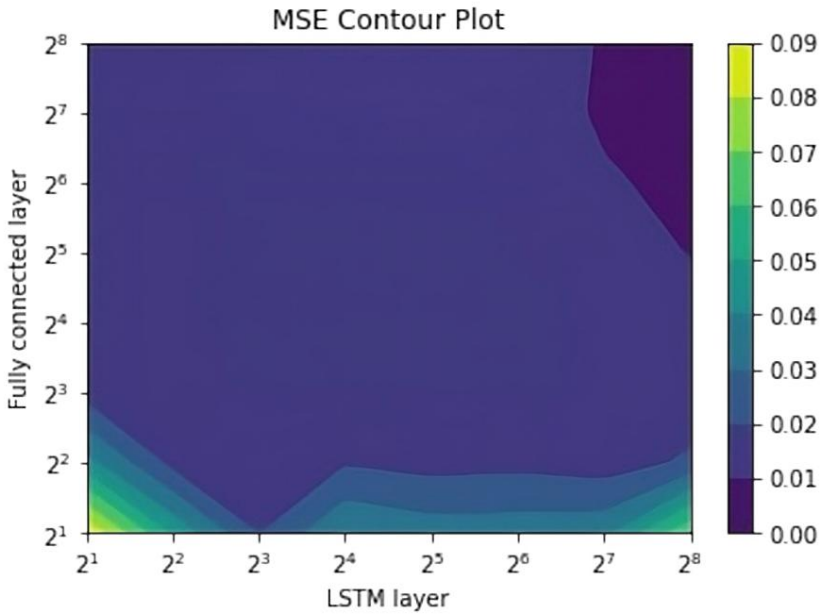

Рисунок 3. Контурный график среднеквадратической ошибки по nf с и nh для субъекта D EU766613: ось x представляет количество скрытых состояний в одном слое LSTM, а ось y представляет количество узлов в полностью связанном слое. слой. При различных сочетаниях среднеквадратичное значение меняется. Контурный график в основном дает нам прямое представление о наименьшем MSE, который представлен самой темной областью на рисунке.

Благодаря этому систематическому исследованию наша цель остается неизменной: определить конфигурацию, которая сводит к минимуму ошибку набора тестовых данных, тем самым повышая эффективность MicroGrowthPredictor.

Важно отметить, что в наш набор обучающих данных мы включили две трети наблюдаемых данных, стремясь обеспечить надежную основу для модели. Примечательно, что для каждого пациента были две точки данных, соответствующие назначению антибиотиков. Одна из этих точек была включена в обучающий набор, а другая была зарезервирована для набора прогнозов. По нашим наблюдениям, реакция на первый антибиотик была более отсроченной по сравнению со вторым. Это наблюдение объясняет, почему наши прогнозируемые данные демонстрируют задержку на рисунке 4.

Временное понимание, обеспечиваемое визуализацией траекторий относительной численности микробиома, имело решающее значение для понимания динамики изменений микробиома и их потенциальных последствий для здоровья хозяина. Для дальнейшего пояснения на рисунке 4 представлен анализ и прогноз относительного содержания бактероида EU766613 для субъекта D с использованием вышеупомянутых оптимальных параметров. Интервалы применения антибиотиков обозначены синей пунктирной вертикальной линией, а красная пунктирная линия разделяет периоды обучения и тестирования. В нашем исследовании повторного лечения антибиотиками мы уделяем приоритетное внимание включению обширных данных о применении антибиотиков, чтобы повысить прогностическую силу нашей модели. Этот подход, основанный на данных, повышает точность последующих прогнозов лечения, предлагая важнейший инструмент в борьбе с устойчивостью к антибиотикам посредством информированного стратегического применения методов лечения.

Визуализация подчеркивает способность модели MicroGrowthPredictor понимать динамику микробиома и формулировать прогнозы, основанные на этих выявленных закономерностях. Это достигается за счет обучения модели в течение 200 эпох с использованием скорости обучения 0,001. Кроме того, потеря среднеквадратической ошибки для данных обучения составляет 0,00081, а для данных тестирования — 0,01021.

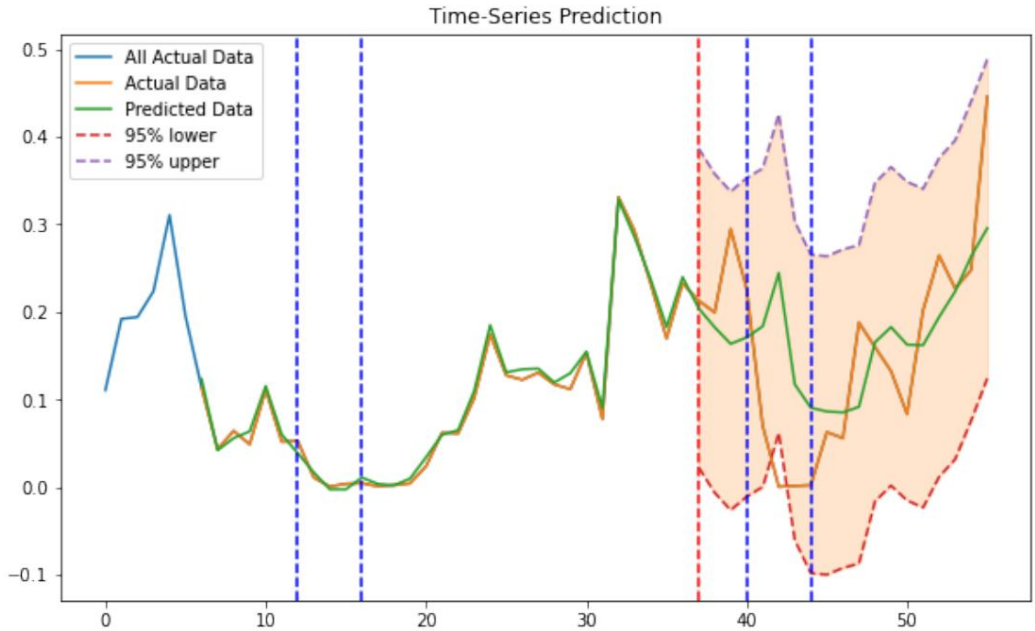

Рисунок 4. Траектории относительной численности бактероида EU766613 для субъекта D. Выбранные параметры : p = 0,05, t = 6, nh = 256 и nf = 256. Синие вертикальные полосы представляют два периода лечения антибиотиками, а красная пунктирная линия разделяет данные на обучение и тестирование.

3.2. Набор данных искусственного

кишечника. Набор данных, предоставленный [16], включает в себя показания микробиоты кишечника с временным разрешением, полученные из искусственного кишечника человека. Эти данные, собираемые как ежедневно, так и ежечасно, получены из искусственного кишечника, построенного с использованием анаэробных биореакторных систем непрерывного действия, что обеспечивает точное представление динамики микробиоты кишечника человека. В течение месяца культивировали четыре сосуда ex vivo, каждый из которых был инициализирован идентичным фекальным инокулятом человека. Для обеспечения точности эксперимента строго поддерживались такие ключевые параметры, как pH, температура, скорость ввода среды и концентрация кислорода. На 23-й день микробная динамика получила целенаправленный стимул путем введения болюса Bacteroides ovatus, штамма, выделенного из стула донора. Однако непредвиденные перебои в подаче корма в двух сосудах между 11 и 13 днями привели к незапланированным микробным изменениям. Примечательно, что мы наблюдали значительные изменения в популяции Rikenellaceae, семейства бактерий, известного своей ролью в микробиоме кишечника человека. Rikenellaceae участвуют в расщеплении сложных углеводов и играют решающую роль в поддержании здоровья кишечника и метаболических функций. Изменения в этой популяции особенно интересны, поскольку они могут дать представление о том, как нарушения в питании и введение микробов влияют на стабильность и функцию микробиоты кишечника.

В этом примере первое судно служит нашим обучающим набором, а второе — нашим тестовым набором. Наш инструмент MicroGrowthPredictor, настроенный с оптимальной вероятностью исключения (p) 0,25, использовал предыдущие пять временных точек для определения четырех параметров и достижения оптимальных прогнозов. Полносвязный уровень имел 256 узлов, а уровень LSTM — 128 узлов. Модель прошла обучение на протяжении 800 эпох. Среднеквадратическая ошибка для обучающих данных составляет 0,00057, а для тестовых данных — 0,01456. Без использования нашей прогнозной модели обобщенная аддитивная модель (GAM) имеет MSE 0,0048 для обучающих данных, что примерно в 8,42 раза выше. Производительность на данных тестирования значительно хуже, поэтому она не включена для сравнения.

Траектории относительной численности микробиома, представленные на рисунке 5, дают критическое представление о динамике изменений микробиома с течением времени. Синяя линия на рисунке 5 представляет все фактические данные, а оранжевая линия выделяется одновременно с прогнозируемой линией (зеленой). В нашем алгоритме глубокого обучения мы использовали предыдущие пять временных точек для прогнозирования следующей. Заметные вариации, особенно для Riken

наблюдались из-за разрушения первых двух сосудов между 11 и 13 днями. Эти визуализации показывают значительные сдвиги в микробных популяциях, подчеркивая точность модели в фиксации временных изменений. Наблюдаемые закономерности согласуются с нашим статистическим анализом, подтверждая существенные изменения в составе микробиома во время возму
Такое согласование укрепляет наше понимание динамики микробиомов и их реакции на экспериментальные условия.

Вопреки представлению о том, что больше данных приводит к лучшим прогнозам, наш эксперимент с участием дополнительных тренировочных судов (включая 1, 3 и 4) для прогнозирования второго судна дал среднеквадратическую ошибку для тестирования 0,0265, что почти вдвое превышает исходную ошибку тестирования. Интересно, что корреляция между прогнозируемым значением и фактическим значением для тестирования сосуда 2 составляет 0,70, что на 18% выше, чем в случае, когда мы включаем сосуды 1, 3 и 4.
Это говорит о том, что тщательный баланс при выборе обучающих данных имеет решающее значение для достижения точных прогнозов.

В сфере научных исследований часто преобладает мнение, что включение большего количества наборов данных или информации для обучения приводит к повышению точности. Однако возникает критическое соображение, когда среда, в которой обучается модель, существенно отличается от среды, в которой она будет применяться для тестирования. Такое несоответствие условий окружающей среды может привести к непредвиденным сбоям и проблемам.

В нашем эксперименте первоначальное предположение о том, что большее количество обучающих данных (включая сосуды 1, 3 и 4) по своей сути улучшит прогнозы, было поставлено под сомнение наблюдаемыми результатами. Нарушения подачи корма в первых двух сосудах между 11 и 13 днями привели к изменениям в динамике микробов, которые не были должным образом отражены в дополнительных данных обучения. Непредвиденные сбои подчеркивают важность согласования данных обучения с условиями и нарушениями, ожидаемыми в среде тестирования.

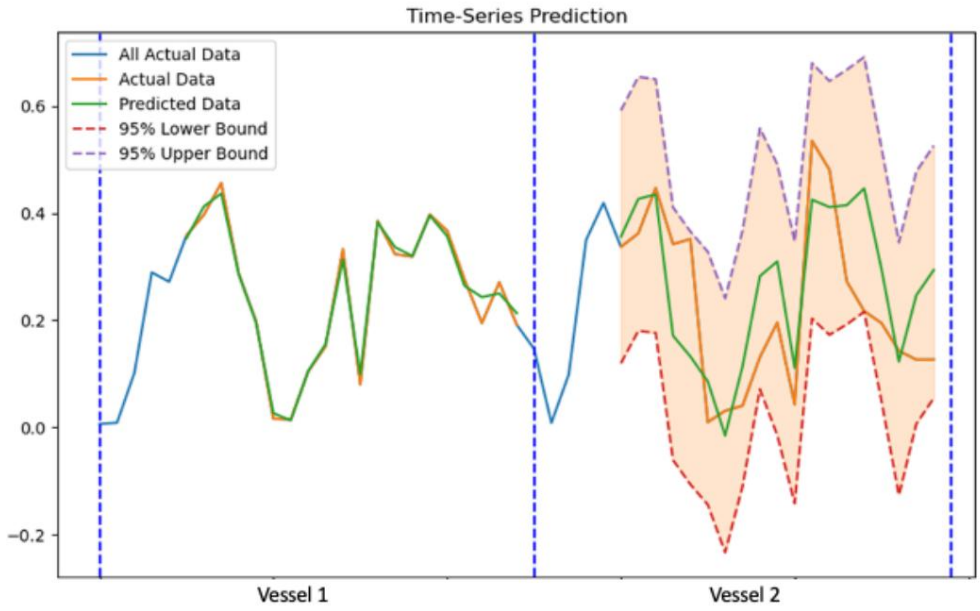

Рисунок 5. Траектории относительной численности Rikenellaceae в сосудах 1 и 2. Вся траектория движения сосуда 2 прогнозируется с помощью модели MicroGrowthPredictor, обученной на данных из сосуда 1. Доверительные интервалы указаны для данных испытаний сосуда 2. В этом эксперименте использовалась оптимальная вероятность выпадения p , равная 0,25. Модель использовала предыдущие пять временных точек для определения четырех параметров и достижения оптимальных прогнозов. Полносвязный уровень имел 256 узлов, а уровень LSTM — 128 узлов. Модель прошла обучение на протяжении 800 эпох.

Хотя заманчиво предположить, что больший размер выборки по своей сути приведет к лучшим прогнозам, ключевой момент заключается в релевантности обучающих данных условиям тестирования. В случаях, когда данные испытаний включают различные нарушения или возмущения окружающей среды,

слепое включение различных наборов данных может привести к неоптимальным прогнозам. Хрупкий баланс между количеством и актуальностью обучающих данных становится решающим в обеспечении адаптивности модели к реальным сценариям.

4. Дискуссия

Модель MicroGrowthPredictor использует знания, полученные из наблюдений о том, что повторное лечение антибиотиками разрушает микробное сообщество кишечника, влияя на разнообразие и численность конкретных групп бактерий. Анализируя данные, модель точно предсказывает, как микробиом будет меняться с течением времени в ответ на воздействие антибиотиков.

Это обеспечивает более глубокое понимание влияния антибиотиков на микробиоту кишечника и потенциальных последствий для здоровья человека.

Кроме того, универсальность модели демонстрируется ее применением к набору данных об искусственном кишечнике. Результаты, полученные в этой контролируемой среде, показывают способность MicroGrowth- Predictor адаптироваться к различным системам микробиома. Набор данных искусственного кишечника подтверждает прогностические возможности модели в конкретных условиях, подчеркивая ее способность улавливать сложную временную динамику. Это делает модель ценной для понимания воздействия антибиотиков и более широкого применения в науках об окружающей среде, здравоохранении и биотехнологиях.

Наш метод решает реальные проблемы, когда ограниченный размер выборки является ограничением из-за логистических, этических или финансовых проблем. Разрабатывая и проверяя методы, которые хорошо работают с ограниченными данными, мы предлагаем практические решения для таких си

В отличие от многих моделей «черного ящика», наш подход предлагает четкое представление о том, как возмущения окружающей среды влияют на микробные популяции с течением времени, что имеет решающее значение для понимания биологических процессов и разработки целевых вмешательств. В частности, мы обсуждаем его потенциал для внесения вклада в персонализированные планы лечения путем прогнозирования индивидуальной реакции на изменения в питании, лечение антибиотиками и пробиотические вмешательства.

Подводя итог, MicroGrowthPredictor представляет собой мощный инструмент, превосходящий традиционные подходы к моделированию. Модель, основанная на знаниях, полученных на основе данных, а не на прямой интеграции знаний, включает сети LSTM с оценкой доверительного интервала, чтобы способствовать целостному пониманию динамики микробиома. Успешное применение модели как к реальной микробиоте кишечника человека, так и к наборам данных искусственного кишечника подчеркивает ее эффективность и потенциальное воздействие. Мы прогнозируем, что MicroGrowthPredictor сыграет ключевую роль в продвижении исследований микробиома, предлагая ценную информацию и способствуя принятию обоснованных решений в различных областях.

Вклад автора: Концептуализация, Ю.-Х3; методология, Г.С. и Ю.-Х.З.; валидация, ГС и Ю.-Х3; письменность — первоначальный черновик, Г.С. и Я.-Х3; написание-рецензирование и редактирование, Г.С. и Ю.- Х.З.; визуализация, ГС и Ю.-Х3; надзор, Ю.-Х.З.; администрация проекта, Ю.-Х3; приобретение финансирования , Ю.-Х.З. Все авторы прочитали и согласились с опубликованной верси

Финансирование: Это исследование финансировалось Агентством по охране окружающей среды США, номер гранта 84045001, Национальным институтом здравоохранения P30ES025128 и Программой центров инженерных исследований Национального научного фонда в соответствии с соглашением о сотрудничестве NSF № EEC-2133504.

Заявление о доступности данных: данные содержатся в статье.

Конфликты интересов: Авторы заявляют, что исследование проводилось при отсутствии каких-либо коммерческих или финансовых отношений, которые могли бы быть расценены как потенциальный конфликт интересов.

Рекомендации

1. Альтвёс, С.; Йылдыз, Гонконг; Вурал Х.К. Взаимодействие микробиоты с организмом человека в норме и при заболеваниях. Биология. Микробиота Food Health 2020, 39, 23–32. [Перекрестная ссылка] [ПабМед]

2. Смит Дж.; Джонсон, М. Динамика микробиома при возмущениях окружающей среды. J. Микробиом Рез. 2022, 10, 123–145.

3. Браун, Э.М.; Садарангани, М.; Финлей, Б.Б. Роль иммунной системы в регулировании взаимодействия хозяина и микроба в кишечнике. Нат. Иммунол. 2013, 14, 660–667. [Перекрестная ссылка] [ПабМед]

4. Кандела, М.; Бьяджи, Э.; Маккаферри, С.; Туррони, С.; Бриджиди, П. Кишечная микробиота является пластическим фактором, реагирующим на изменения окружающей среды. Тенденции Микробиол. 2012, 20, 385–391. [Перекрестная ссылка]

5. Ура, ГТ; Доналова, Л.; Таисс, Калифорния. Измерение времени во взаимодействии хозяина и микробиома. mSystems 2019, 4, e00216-18. [Перекрестная ссылка] [ПабМед]

6. Уиллинг, БП; Рассел, СЛ; Финли, Б.Б. Изменение баланса: влияние антибиотиков на мутуализм между хозяином и микробиотой. Нат. Преподобный Микробиол. 2011, 9, 233–243. [Перекрестная ссылка] [ПабМед]

7. Браун Э.; Уильямс, Д. Прогнозное моделирование роста микробиома с использованием сетей LSTM. Дж. Компьютер. Биол. 2021, 45, 321–335.

8. Чинг, Т.; Химмельштейн, Д.С.; Больe-Джонс, Британская Колумбия; Калинин А.А.; Делай, БТ; Уэй, терапевт; Ферреро, Э.; Агапов, премьер-министр; Зиц, М.; Хоффман, ММ; и другие. Возможности и препятствия для глубокого обучения в биологии и медицине. JR Soc. Интерфейс 2018, 15, 20170387. [Перекрестная ссылка] [ПабМед]

9. Медскер, Л.Р.; Джайн, Л. Рекуррентные нейронные сети. Дес. Прил. 2001, 5, 2.

10. Грейвс, А.; Грейвс, А. Длинная кратковременная память. В контролируемой маркировке последовательностей с помощью рекуррентных нейронных сетей; Спрингер: Берлин/Гейдельберг, Германия, 2012 г.; стр. 37–45.

11. Ю, Ю.; Шесть.; Ху, К.; Чжан Дж. Обзор рекуррентных нейронных сетей: ячейки LSTM и сетевые архитектуры. Нейронный компьютер. 2019, 31, 1235–1270. [Перекрестная ссылка] [ПабМед]

12. Чжоу, Ю. Х.; Галлинс, П. Обзор и руководство по методам машинного обучения для прогнозирования свойств микробиома хозяина. Передний. Жене. 2019, 10, 579. [CrossRef] [ПабМед]

13. Чжу, Л.; Лаптев Н. Глубокое и уверенное предсказание временных рядов в uber. В материалах Международной конференции IEEE по интеллектуальному анализу данных (ICDMW) 2017 г., Орлеан, Луизиана, США, 18–21 ноября 2017 г.; IEEE: Пискатауэй, Нью-Джерси, США, 2017 г.; стр. 103–110.

14. Гал, Ю.; Гахрамани, З. Отсев как байесовское приближение: представление неопределенности модели в глубоком обучении. В материалах Международной конференции по машинному обучению, PMLR, Нью-Йорк, США, 20–22 июня 2016 г.; стр. 1050–1059.

15. Детлефсен, Л.; Релман, Д.А. Неполное восстановление и индивидуализированные реакции микробиоты дистального отдела кишечника человека на неоднократное воздействие антибиотиков. Учеб. Натл. акад. наук. США 2011, 108, 4554–4561. [Перекрестная ссылка] [ПабМед]

16. Сильверман, доктор медицинских наук; Дюран, Гонконг; Блум, Р.Дж.; Мукерджи, С.; Дэвид, Л.А. Динамические линейные модели помогают проектировать и анализировать исследования микробиоты в искусственном кишечнике человека. Микробиом 2018, 6, 202.