applied microbiology

MDPI

Article

Prédire la dynamique de croissance du microbiome sous Perturbations environnementales

Georges Soleil [1] et Yi-Hui Zhou 2,*

[1] Centre de recherche en bioinformatique, Université d'État de Caroline du Nord, Raleigh, NC 27695, États-Unis ;
[2] 3litus@gmail.com Départements de sciences biologiques et de statistiques, North Carolina State University, Raleigh, NC 27695, États-Unis
* Correspondance : yihui_zhou@ncsu.edu

Résumé : MicroGrowthPredictor est un modèle qui exploite les réseaux de mémoire à long terme et à court terme (LSTM) pour prédire les changements dynamiques dans la croissance du microbiome en réponse à diverses perturbations environnementales. Dans cet article, nous présentons les capacités innovantes de MicroGrowthPredictor, qui incluent l'intégration de la modélisation LSTM avec une nouvelle technique d'estimation de l'intervalle de confiance. Le réseau LSTM capture la dynamique temporelle complexe des systèmes du microbiome, tandis que les nouveaux intervalles de confiance fournissent une mesure robuste de l'incertitude des prédictions. Nous incluons deux exemples : l'un illustrant la composition et la diversité du microbiote intestinal humain dues à un traitement antibiotique récurrent et l'autre démontrant l'application de MicroGrowthPredictor sur un ensemble de données intestinales artificielles . Les résultats démontrent la précision et la fiabilité améliorées des prédictions basées sur LSTM facilitées par MicroGrowthPredictor. L'inclusion de mesures spécifiques, telles que l'erreur quadratique moyenne, valide les performances prédictives du modèle. Notre modèle recèle un immense potentiel d'applications dans les sciences de l'environnement, les soins de santé et la biotechnologie, favorisant les progrès dans la recherche et l'analyse du microbiome. De plus, il convient de noter que MicroGrowthPredictor est applicable aux données réelles avec des échantillons de petite taille et aux observations temporelles sous perturbations environnementales, garantissant ainsi son utilité pratique dans divers domaines.

Mots-clés : dynamique du microbiome ; incertitude de prévision ; applications environnementales

Citation : Soleil, G. ; Zhou, Y.-H. Prédire la dynamique de croissance du microbiome sous perturbations environnementales. Appl. Microbiol. 2024, 4, 948-958. https://doi.org/10.3390/applmicrobiol4020064

Rédacteur académique : Bong-Soo Kim

Reçu : 7 mai 2024
Révisé : 4 juin 2024
Accepté : 7 juin 2024
Publié : 10 juin 2024

1. Introduction

Le microbiome humain, un écosystème complexe de milliards de micro-organismes résidant dans et sur le corps humain, joue un rôle crucial dans le maintien de l'homéostasie physiologique, des fonctions métaboliques et des réponses immunitaires [1]. Les perturbations du microbiome ont été liées à une pléthore de pathologies, allant des troubles gastro-intestinaux à des maladies plus systémiques telles que le diabète, l'obésité et même des troubles neurologiques [2]. Cette interaction symbiotique hôte-microbe souligne la nécessité de comprendre la nature dynamique du microbiome humain [3], en particulier comment il évolue au fil du temps et en réponse à divers stimuli environnementaux [4,5].

Dans des conditions normales, le microbiome intestinal est composé d'une communauté diversifiée de bactéries, les Firmicutes et les Bacteroidetes étant les phylums prédominants. Les perturbations environnementales , telles que les changements de régime alimentaire, l'utilisation d'antibiotiques et l'exposition à des polluants, peuvent modifier considérablement la composition et la fonction du microbiome, entraînant des implications potentielles sur la santé. Par exemple, le traitement antibiotique peut réduire considérablement la diversité microbienne, entraînant souvent une prolifération de bactéries résistantes et une diminution des microbes bénéfiques, ce qui peut perturber les processus métaboliques et les fonctions immunitaires [6]. Comprendre cette dynamique de population est crucial pour développer des stratégies visant à atténuer les effets néfastes de telles perturbations sur la santé humaine.

Les technologies de séquençage à haut débit, en particulier le séquençage de l'ARNr 16S, ont inauguré une nouvelle ère dans les études sur le microbiome, permettant des évaluations détaillées de la diversité microbienne et de l'abondance relative dans différentes populations et conditions humaines [7]. Cependant , les vastes données générées par ces technologies présentent à la fois des opportunités et des défis.

L'un des principaux défis consiste à déchiffrer les modèles temporels et à prédire les états futurs du microbiome, essentiels pour les applications de soins de santé préventives et thérapeutiques.

La modélisation prédictive en biométrie a historiquement utilisé diverses méthodes statistiques , mais ces approches traditionnelles ne parviennent souvent pas à gérer la haute dimensionnalité et la non-linéarité des données du microbiome. L'avènement du machine learning, et plus particulièrement du deep learning, offre de nouvelles perspectives prometteuses pour des données aussi complexes [8]. Les réseaux de neurones récurrents (RNN) [9] et leur variante avancée, les réseaux LSTM (Long Short-Term Memory) [10], excellent dans l'analyse et la prévision des séquences temporelles, fournissant un excellent cadre pour modéliser la dynamique du microbiome.

Dans cette étude, nous introduisons le modèle MicroGrowthPredictor qui vise à exploiter la puissance des réseaux LSTM pour prédire les changements dans le microbiome humain en réponse à des perturbations environnementales, une étape cruciale vers une médecine personnalisée et des interventions thérapeutiques ciblées.

2. Matériels et méthodes 2.1.

Modèle de mémoire à long terme (LSTM)

Le réseau LSTM (Long Short-Term Memory), une forme spécialisée de l' architecture de réseau neuronal récurrent (RNN), est explicitement conçu pour relever les défis de l'apprentissage à partir de données séquentielles, notamment les dépendances à long terme. Les RNN traditionnels, bien que théoriquement capables de gérer de telles dépendances, échouent souvent dans la pratique en raison du problème de gradient de disparition, dans lequel les informations sont perdues à chaque pas de temps pendant la formation. Les réseaux LSTM sont conçus pour surmonter cette limitation, les rendant ainsi particulièrement adaptés à des applications dans divers domaines tels que l'analyse de séries chronologiques, le traitement du langage naturel et, ce qui est pertinent pour notre travail, l'analyse des données du

Les réseaux LSTM introduisent une structure cellulaire plus sophistiquée que les RNN traditionnels [11]. Chaque cellule LSTM contient des mécanismes appelés portes qui régulent le flux d'informations entrant et sortant de la cellule. Il existe trois types de portes au sein d'une cellule LSTM (Figure 1A) :

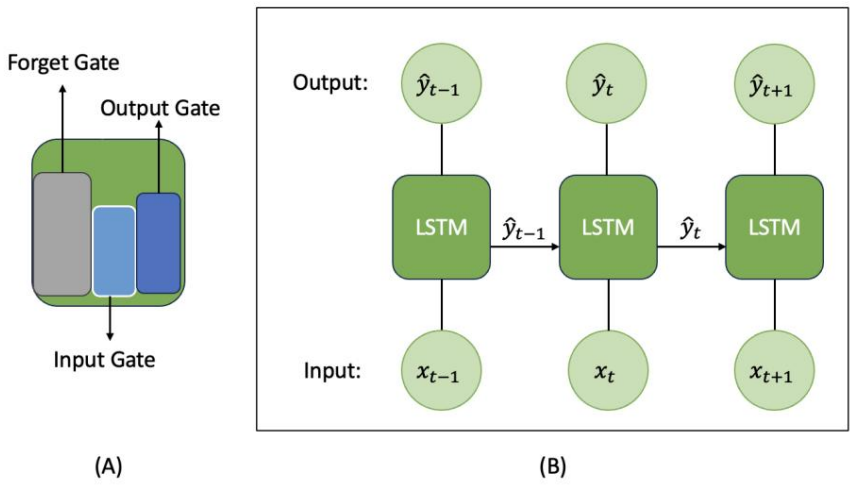

Figure 1. Architecture de mémoire à long terme (LSTM) : (A) Un zoom avant sur une cellule LSTM, montrant ses trois portes : la porte d'entrée, la porte d'oubli et la porte de sortie. (B) Le flux des données d'entrée et de sortie dans un réseau LSTM du pas de temps t - 1 au pas de temps t.

• Porte d'entrée : module la quantité de nouvelles informations à ajouter à l'état de la cellule. • Porte d'oubli : détermine l'étendue des informations à supprimer de l'état de la cellule.

La porte d'oubli aide à éliminer les informations microbiennes non pertinentes ou obsolètes, conservant ainsi uniquement les données les plus pertinentes pour une modélisation

précise. • Porte de sortie : contrôle la quantité d'informations à sortir de la cellule. Pour les données sur le microbiome, la porte de sortie aide à décider quelles informations microbiennes traitées doivent influencer les prédictions ou les analyses du réseau à chaque pas de temps.

Ces portes fonctionnent ensemble pour mettre à jour l'état de la cellule et permettent au LSTM de mémoriser et d'oublier des informations sur de longues séquences (Figure 1B), ce qui est crucial pour l'apprentissage des dépendances à long terme. La figure 1B illustre la transition des données via le réseau LSTM d'un pas de temps au suivant. Il montre les données d'entrée et les données de sortie telles qu'elles circulent du pas de temps t − 1 au pas de temps t. À chaque pas de temps, les données d'entrée, ainsi que l'état de la cellule du pas de temps précédent, sont traitées par la cellule LSTM. Ce traitement aboutit à un état de cellule mis à jour et à une sortie, qui sont ensuite transmis au pas de temps suivant. Ce mécanisme séquentiel permet au réseau LSTM de gérer efficacement les dépendances temporelles, garantissant que les informations sont transmises et utilisées sur différentes étapes de temps pour améliorer la prédiction et l'analyse des tâches de séries chronologiques.

Dans le domaine de l'analyse du microbiome, la compréhension de la dynamique temporelle et des modèles séquentiels est essentielle, compte tenu de la nature de l'évolution et de l'interaction des communautés microbiennes au fil du temps. Ici, nous adoptons une notation spécifique pour élucider la mécanique du modèle LSTM. Considérons un ensemble de données d'entraînement $D = \{(x_t, y_t)\}_{t=1}^{T}$ où x_t désigne le vecteur de relatif abondances [12] de tous les taxons microbiens au t-ème pas de temps et y_t signifie le résultat souhaité correspondant. Le LSTM prend ces séquences d'entrée et les traite à travers sa structure cellulaire complexe, capturant les dépendances temporelles précieuses présentes dans les données qui sont essentielles pour des prédictions et des analyses précises dans les études du microbiome.

2.2. Structure du modèle pour la prévision de la croissance

du microbiome Le modèle LSTM utilisé dans cette étude est conçu pour être à la fois simple et puissant. La couche d'entrée est conçue pour traiter les niveaux d'abondance relative des taxons, prenant en charge un large éventail de taxons microbiens notés x_t . Composée de nœuds n_{taxa} , chacun représentant l'abondance relative d'un taxon particulier, cette couche correspond au nombre total de taxons uniques identifiés dans l'ensemble de données du microbiome.

En ce qui concerne l'architecture, notre modèle se compose de deux couches cachées positionnées entre les étages d'entrée et de sortie. La couche cachée principale intègre un LSTM avec n_h états cachés, fonctionnant au sein d'une seule couche. Cette configuration est cruciale, car elle permet au modèle de capturer et d'interpréter la dynamique temporelle inhérente à la séquence d'entrée, grâce aux cellules mémoire caractéristiques du LSTM.

Pour remédier au surajustement et améliorer la robustesse du modèle, une stratégie d'abandon est mise en œuvre après la couche LSTM. Cette stratégie, régie par une probabilité d'abandon p prédéfinie , implique la désactivation arbitraire des nœuds, renforçant la capacité de généralisation du modèle . Les nœuds non affectés par l'abandon sont ensuite transmis à la couche suivante : une strate entièrement connectée contenant des nœuds NFC .

La couche cachée secondaire utilise la fonction d'activation de l'unité linéaire rectifiée (ReLU) sur les points de données dérivés de la couche entièrement connectée. Cela confère une non-linéarité essentielle , préparant le modèle à discerner des modèles complexes au sein de l'ensemble de données. Les prédictions sont formulées sur la base des résultats de cette couche.

En résumé, notre modèle MicroGrowthPredictor pour prédire la dynamique du microbiome intègre des couches spécialement conçues, chacune conçue pour interpréter la dynamique temporelle nuancée des données du microbiome. L'architecture commence par une couche d'entrée hébergeant des nœuds représentatifs des taxons n_{taxa} , passant à un LSTM monocouche avec n_h états cachés.

Bien que cela ne soit pas explicitement détaillé, nous supposons que la couche LSTM conserve la composition conventionnelle des cellules LSTM, y compris les portes d'entrée, d'oubli et de sortie pour un transfert d'informations efficace. Cette structure joue un rôle déterminant pour permettre au modèle d'apprendre et de préserver les dépendances à long terme inhérentes aux données séquentielles.

Après la couche LSTM, une technique d'abandon avec une probabilité désignée p est appliquée pour servir de mécanisme de régularisation, atténuant les risques de surajustement. Par la suite, une couche entièrement connectée avec des nœuds NFC est introduite, aboutissant à une couche dense capable de capturer les interdépendances non linéaires dans les données. La phase finale du modèle intègre une fonction d'activation ReLU, introduisant la non-linéarité et améliorant la complexité du modèle pour une interprétation détaillée des données. Cette étape est cruciale pour façonner le résultat final, garantissant des prédictions précises et fluides dans le paysage dynamiquement changeant des données sur le microbiome.

2.3. Formation du modèle MicroGrowthPredictor

Lorsque l'on travaille avec des données de séries chronologiques et que l'on utilise le LSTM pour prédire l'impact des perturbations environnementales, la validation croisée doit tenir compte des dépendances temporelles inhérentes aux données. Pour garantir des prévisions robustes et précises, nous avons utilisé une méthode de validation croisée de séries chronologiques utilisant une approche à fenêtre glissante. L'ensemble de données a été divisé en K plis consécutifs sans brassage. Pour chaque pli k, le modèle a été entraîné sur les k premiers plis et testé sur le pli k + 1, en répétant jusqu'à ce que chaque pli serve d'ensemble de test. Cette méthode garantit le respect des dépendances temporelles et évite les fuites de données.

Des mesures d'évaluation, telles que l'erreur quadratique moyenne (MSE), ont été collectées pour chaque pli, et la performance moyenne pour tous les plis a été calculée pour évaluer la robustesse du modèle.

2.4. Intervalle de prédiction

Alors que les approches traditionnelles visant à établir des intervalles de confiance ou de prédiction dans les modèles d'apprentissage profond sont confrontées à des défis considérables en raison de la non-linéarité et de l'architecture complexe de ces modèles, des progrès récents ont commencé à ouvrir la voie à des solutions plus robustes . L'une de ces avancées est le travail de [13], dans lequel le cadre d'abandon de Monte Carlo (MC) a été exploité pour introduire une méthode qui, bien qu'efficace, laisse place à un raffinement et à une application supplémentaires dans de nouveaux domaines, tels que l'analyse des données du micro

Notre recherche s'appuie sur ces travaux fondamentaux, en adoptant le principe des abandons stochastiques après chaque couche cachée de l'architecture du réseau neuronal. Cependant, nous étendons ce concept en adaptant le processus d'abandon et l'interprétation ultérieure des résultats du modèle spécifiquement aux caractéristiques et à la complexité des données sur le microbiome.
Cette adaptation permet non seulement l'interprétation théorique du résultat du modèle en tant qu'échantillon aléatoire de la distribution prédictive postérieure, mais reconnaît également le comportement unique des données dans les études sur le microbiome.

Le processus de construction d'une distribution empirique de valeurs prédites en traitant chaque prédiction lors de l'abandon comme un échantillon de la distribution de données sous-jacente représente une approche nuancée dans notre étude. Elle s'écarte des techniques classiques en ouvrant une fenêtre sur les capacités prédictives et les incertitudes du modèle spécifiquement adaptées au contexte du microbiome, renforçant ainsi la robustesse de la prise de décision basée sur ces prédictions.

Dans notre approche, nous désignons les données de test que l'on cherche à prédire avec l'exposant . Le fondement de l'intervalle de prédiction réside dans la probabilité conditionnelle p(y |x , D). Cette probabilité peut être exprimée comme l'intégrale du produit de p(y |x , θ) et p(θ|D) sur le vecteur de paramètres θ, noté comme suit :

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

θ représente le vecteur de paramètres du modèle d'apprentissage profond et p(θ|D) correspond à la distribution a posteriori. Cependant, dériver une forme analytique pour p(y |x , θ) est généralement irréalisable. Pour surmonter ce défi, une technique d'approximation utilisant une distribution variationnelle notée q (θ) est proposée dans la réf. [14]. On obtient donc l'approximation suivante :

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

où ˆθk q(θ). Cette approximation finale, obtenue grâce à l'échantillonnage de { ˆθk}k=1,...,K à partir de la distribution variationnelle q(θ), utilise la technique d'intégration de Monte Carlo.
De plus, cette approximation équivaut à implémenter l' algorithme de dropout de Monte Carlo introduit dans [13]. Essentiellement, pour un point de données de test donné (x), la prédic-, y avec abandon aléatoire distribution est évalué plusieurs fois à x des nœuds et la sortie tive y résultante de la empirique servent d'estimation de p(y |x , D). Intervalles de prédiction

capturer la variabilité provenant de deux sources principales : l'incertitude du modèle (η1) et le bruit inhérent (η2).

Les étapes suivantes décrivent le processus : Pour chaque point de données individuel x^* dans le ensemble de tests, calculez la sortie correspondante $\tilde{y}^*$ en supprimant de manière aléatoire chaque nœud avec une probabilité d'abandon p donnée. Répétez ce processus B fois pour obtenir un grand nombre de valeurs prédites $\tilde{y}^*$ dont chacune varie en raison de l'abandon aléatoire des nœuds. Ensuite, calculez l'incertitude du modèle η1 en calculant la différence quadratique moyenne entre chaque valeur prédite $\tilde{y}^*$ et la moyenne de toutes les valeurs prédites $\bar{\tilde{y}}^*$. Cela se fait en utilisant formule $\eta 1 = \frac{1}{B}\sum_{j=1}^{B}(\tilde{y}^*_j - \bar{\tilde{y}}^*)^2$. Pour quantifier le bruit inhérent dans les prédictions, calculez la différence quadratique moyenne entre chaque valeur prédite $\tilde{y}^*$ et sa vraie valeur j correspondante y j à l'aide de l'ensemble de données de test de longueur V. Cela nous donne le bruit inhérent $\eta 2^*$, calculé comme $\frac{1}{V}\sum_{j=1}^{V}(\tilde{y}^*_j - y^*_j)^2$. En combinant l'incertitude du modèle et le bruit inhérent, calculez η2 = incertitude globale η comme racine carrée de la somme de η1 et η2, c'est-à-dire $\eta = \sqrt{\eta 1 + \eta 2}$. Enfin, déterminez les limites supérieure inférieure de l'intervalle de prédiction en ajoutant et en soustrayant zα/2 fois η de la valeur moyenne prédite $\bar{\tilde{y}}^*$. Ici, zα/2 représente le score z associé au niveau de confiance souhaité (1 − α)100 %. Un algorithme formel est répertorié dans l'algorithme 1.

Algorithme 1 : réseau neuronal LSTM et intervalle de prédiction. , p, t, nh , nf c

Exiger : x, y, x^*, y^*
Assurer : θ, U, L 1
répéter z1
2 ← x de la couche LSTM avec t et nh ; z2 ← z1
3 par abandon aléatoire avec p ; z3 ← z2 de la couche
4 entièrement connectée avec des nœuds nf c ;
5 Appliquez ReLU à z3 ; yˆ
6 ← z3 de la couche de sortie ;
7 Évaluez yˆ avec y ;
8 Mettre à jour θ pour le modèle mθ ;
9 jusqu'à la dernière époque; 10 pour i = 1 à
11 $\tilde{y}^*$ B do ← mθ (x^*) avec abandon aléatoire ;
je 12 fin
13 Calculer $\bar{\tilde{y}}^*$ et η ;
14 U, L ← $\bar{\tilde{y}}^*$ ± zα/2 × η ;

2.5. Réglage des paramètres

Pour optimiser les performances de notre modèle MicroGrowthPredictor, nous utilisons un processus de réglage en deux étapes.

Dans la première étape, nous présélectionnons le nombre d'unités cachées dans la couche LSTM (nh) et la couche entièrement connectée (nf c) sur la base d'expériences préliminaires. Nous explorons ensuite différentes combinaisons de probabilité d'abandon (p) et de longueur de séquence (T), qui représente le nombre de points de données précédents utilisés comme caractéristiques de prédiction. Les performances du modèle sont évaluées en calculant l'erreur quadratique moyenne (MSE) sur un ensemble de données de test distinct, et nous sélectionnons la combinaison de p et T qui minimise cette erreur.

Une fois la probabilité d'abandon optimale et la longueur de séquence déterminées, nous passons à la deuxième étape, où nous affinons le nombre de nœuds dans les couches LSTM et entièrement connectées. Pour chaque combinaison architecturale, nous entraînons le modèle plusieurs fois avec différentes initialisations pour tenir compte des variations introduites par les abandons aléatoires et les paramètres de poids initiaux. Nous calculons le MSE pour chaque exécution de formation et sélectionnons l'architecture qui entraîne l'erreur la plus faible sur l'ensemble de données de test.

Ce processus de réglage rigoureux garantit que notre modèle MicroGrowthPredictor est configuré de manière optimale pour l'ensemble de données spécifique considéré, améliorant ainsi ses performances prédictives.

3. Résultats

Dans cette étude, nous utilisons le modèle MicroGrowthPredictor et la procédure de réglage associée pour deux ensembles de données distincts : l'ensemble de données sur l'antibiotique ciprofloxacine (Cp) de [15] et l'ensemble de données sur l'intestin artificiel détaillé dans [16]. Les deux ensembles de données offrent un aperçu de la dynamique temporelle du microbiome sous diverses perturbations environnementales.

3.1. Ciprofloxacin Dataset

Reference [15] souligne les modifications significatives imposées à la composition et à la diversité du microbiote intestinal humain en raison des traitements antibiotiques récurrents. Cette recherche impliquait une surveillance approfondie des communautés bactériennes dans l'intestin distal chez trois sujets (D, E et F). Des échantillons de selles ont été collectés périodiquement sur dix mois, pour un total de 52 à 56 échantillons par individu. Au cours de cette période, chaque sujet a reçu deux régimes distincts de 5 jours d'antibiotique ciprofloxacine (Cp), espacés de 6 mois. Un échantillonnage intense – quotidiennement sur deux périodes de 19 jours coïncidant avec chaque cours de Cp – a fourni une perspective détaillée du microbiome pendant l'exposition aux antibiotiques. En dehors de ces fenêtres, des échantillons ont été acquis de manière hebdomadaire ou mensuelle, capturant la composition microbienne en l'absence de traitement.

À des fins d'illustration, nous nous concentrons sur le sujet D. Notre processus d'optimisation consiste à générer un tracé de contour de l'erreur quadratique moyenne (MSE) par rapport à différentes valeurs de la probabilité d'abandon p et du nombre de pas de temps. La figure 2 visualise cette relation, guidant notre sélection d'une combinaison optimale pour affiner le modèle MicroGrowthPredictor . Le tracé de contour de l'erreur quadratique moyenne est tracé avec la probabilité d'abandon p sur l'axe des x et le nombre de pas de temps sur l'axe des y. Dans le tracé de contour, plus l' ombrage est foncé, plus l'erreur est faible. Nous utilisons une fonction d'optimisation pour identifier la meilleure combinaison de probabilité d'abandon et de longueur de séquence.

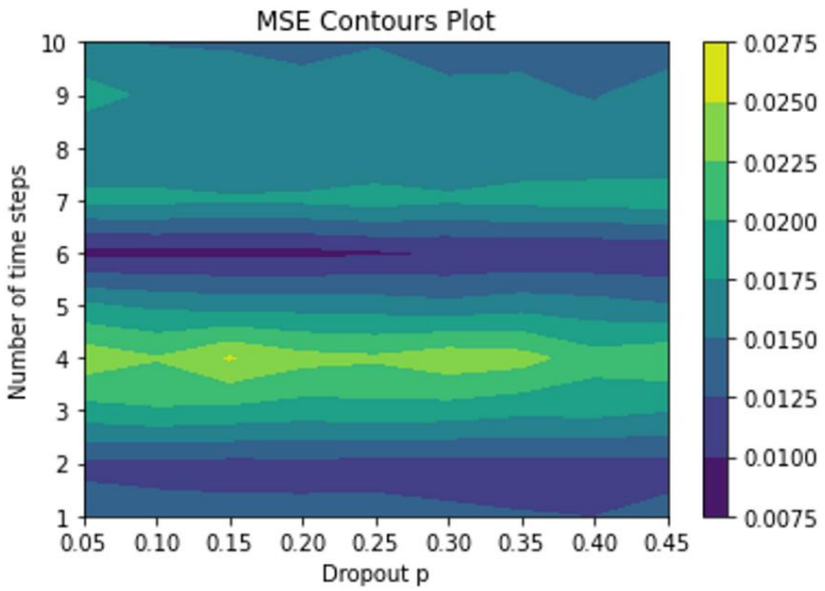

Figure 2. Tracé de contour de l'erreur quadratique moyenne sur p et t pour le sujet D EU766613 : plus le tracé de contour est sombre, plus l'erreur est petite. Nous pouvons identifier la meilleure combinaison de probabilité d'abandon et de longueur de séquence.

Par la suite, nous nous concentrons sur la détermination du nombre optimal de nœuds pour les couches LSTM et entièrement connectées, comme le montre la figure 3. L'axe des x représente le nombre d'états cachés dans la couche LSTM unique, et l'axe des y représente le nombre. de nœuds dans la couche entièrement connectée. Différentes combinaisons entraînent des modifications de la valeur de l'erreur quadratique moyenne. Le tracé de contour fournit une représentation directe du plus petit MSE, indiqué par la zone la plus sombre de la figure.

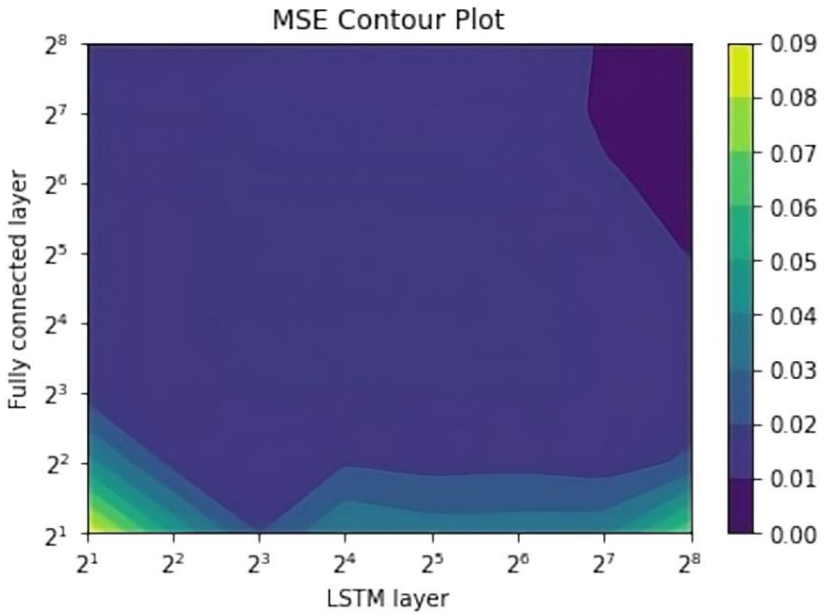

Figure 3. Tracé de contour de l'erreur quadratique moyenne sur nf c et nh pour le sujet D EU766613 : l'axe des x représente le nombre d'états cachés dans la couche LSTM unique et l'axe des y représente le nombre de nœuds dans la couche entièrement connectée. couche. Avec différentes combinaisons, la valeur quadratique moyenne change. Le tracé des contours nous donne essentiellement une impression directe du plus petit MSE, qui est représenté par la zone la plus sombre de la figure.

Grâce à cette exploration systématique, notre objectif reste cohérent : identifier une configuration qui minimise les erreurs de jeu de données de test, améliorant ainsi l'efficacité du MicroGrowthPredictor.

Il est essentiel de noter que dans notre ensemble de données de formation, nous avons inclus les deux tiers des données observées, dans le but de fournir une base solide au modèle. Notamment, il y avait deux points de données correspondant à l'administration d'antibiotiques pour chaque patient. L'un de ces points était inclus dans l'ensemble d'entraînement, tandis que l'autre était réservé à l'ensemble de prédiction. D'après nos observations, la réaction au premier antibiotique a présenté une réponse retardée par rapport au second. Cette observation explique pourquoi nos données prédites démontrent une tendance retardée dans la figure 4.

Les informations temporelles fournies par la visualisation des trajectoires d'abondance relative du microbiome étaient cruciales pour comprendre la dynamique des changements du microbiome et leurs implications potentielles pour la santé de l'hôte. Pour mieux comprendre, la figure 4 présente une analyse et une prédiction de l'abondance relative du bactérioïde EU766613 pour le sujet D, en utilisant les paramètres optimaux susmentionnés. Les intervalles d'administration des antibiotiques sont indiqués par une ligne verticale pointillée bleue, tandis que la démarcation pointillée rouge sépare les périodes de formation et de test. Dans notre étude sur les traitements antibiotiques répétés, nous accordons la priorité à l'inclusion de données détaillées sur les interventions antibiotiques afin de renforcer le pouvoir prédictif de notre modèle. Cette approche basée sur les données améliore la précision des prévisions de traitements ultérieurs, offrant ainsi un outil essentiel dans la lutte contre la résistance aux antibiotiques grâce à une application stratégique et éclairée des thérapies.

La visualisation souligne la capacité du modèle MicroGrowthPredictor à comprendre la dynamique du microbiome et à formuler des prédictions ancrées dans ces modèles identifiés. Ceci est réalisé avec le modèle formé sur 200 époques en utilisant un taux d'apprentissage de 0,001. De plus, la perte d'erreur quadratique moyenne pour les données d'entraînement est de 0,00081 et pour les données de test, elle est de 0,01021.

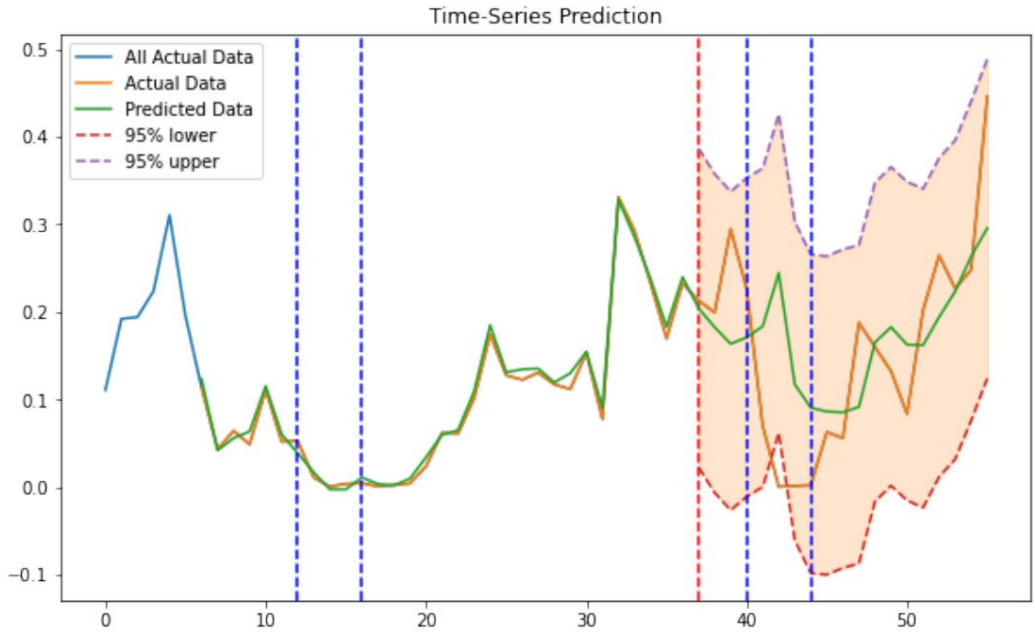

Figure 4. Trajectoires d'abondance relative du Bacteroid EU766613 pour le sujet D. Les paramètres choisis sont p = 0,05, t = 6, nh = 256 et nf = 256. Les bandes verticales bleues représentent les deux périodes de traitement antibiotique, et la la ligne pointillée rouge divise les données en formation et tests.

3.2. Ensemble de données

sur l'intestin artificiel L'ensemble de données fourni par [16] comprend des lectures résolues dans le temps du microbiote intestinal provenant d'un intestin humain artificiel. Ces données, capturées quotidiennement et toutes les heures, proviennent d'un intestin artificiel construit à l'aide de systèmes de bioréacteurs anaérobies à flux continu, garantissant une représentation précise de la dynamique du microbiote intestinal humain. Pendant un mois, quatre récipients ex vivo, chacun initialisé avec un inoculum fécal humain identique, ont été cultivés. Pour garantir la fidélité expérimentale, des paramètres clés tels que le pH, la température, le taux d'apport du milieu et la concentration en oxygène ont été rigoureusement respectés. Au jour 23, la dynamique microbienne a reçu un stimulus délibéré via l'introduction d'un bolus de Bacteroides ovatus, une souche isolée du donneur de selles. Cependant, des perturbations imprévues de l'approvisionnement en aliments dans deux navires entre les jours 11 et 13 ont introduit des variations microbiennes imprévues. Nous avons notamment observé des changements significatifs dans la population des Rikenellaceae, une famille de bactéries connue pour son rôle dans le microbiome intestinal humain. Les Rikenellaceae sont impliquées dans la dégradation des glucides complexes et jouent un rôle crucial dans le maintien de la santé intestinale et des fonctions métaboliques. Les changements dans cette population sont particulièrement intéressants car ils peuvent fournir des informations sur la manière dont les perturbations du régime alimentaire et les introductions microbiennes influencent la stabilité et le fonctionnement du microbiote intestinal.

Dans cet exemple, le premier navire sert d'ensemble de formation, tandis que le deuxième navire sert d'ensemble de test. Notre outil MicroGrowthPredictor, configuré avec une probabilité d'abandon optimale (p) de 0,25, a utilisé les cinq points temporels précédents pour identifier quatre paramètres et obtenir des prédictions optimales. La couche entièrement connectée était équipée de 256 nœuds et la couche LSTM comprenait 128 nœuds. Le modèle a subi un entraînement pendant 800 époques. L'erreur quadratique moyenne pour les données d'entraînement est de 0,00057 et pour les données de test, elle est de 0,01456. Sans utiliser notre modèle prédictif, un modèle additif généralisé (GAM) a un MSE de 0,0048 pour les données d'entraînement, soit environ 8,42 fois plus élevé. Les performances sur les données de test sont nettement moins bonnes, elles ne sont donc pas incluses à des fins de comparaison.

Les trajectoires de l'abondance relative du microbiome visualisées sur la figure 5 fournissent des informations essentielles sur la dynamique des changements du microbiome au fil du temps. La ligne bleue de la figure 5 représente toutes les données réelles, tandis que la ligne orange est mise en évidence simultanément avec la ligne prédite (verte). Dans notre algorithme d'apprentissage en profondeur, nous avons utilisé les cinq points temporels précédents pour prédire le suivant. Des variations notables, notamment pour les Rikenellaceae,

ont été observés en raison de la perturbation des deux premiers vaisseaux entre les jours 11 et 13. Ces visualisations révèlent des changements significatifs dans les populations microbiennes, soulignant la précision du modèle dans la capture des changements temporels. Les modèles observés correspondent à nos analyses statistiques, confirmant des changements substantiels dans la composition du microbiome lors des perturbations. Cet alignement renforce notre compréhension de la dynamique du microbiome et de ses réponses aux conditions expérimentales.

Contrairement à l'idée selon laquelle plus de données conduisent à de meilleures prédictions, notre expérience impliquant des navires-écoles supplémentaires (dont 1, 3 et 4) pour prédire le deuxième navire a donné une erreur quadratique moyenne pour les tests de 0,0265, soit presque le double de l'erreur de test initiale. Il est intéressant de noter que la corrélation entre la valeur prédite et la valeur réelle pour tester le récipient 2 est de 0,70, soit 18 % de plus que le cas lorsque nous incluons les récipients 1, 3 et 4.
Cela suggère qu'un équilibre minutieux dans la sélection des données d'entraînement est crucial pour obtenir des prédictions précises.

Dans le domaine des études scientifiques, la croyance prévaut souvent selon laquelle l'incorporation de davantage d'ensembles de données ou d'informations pour la formation conduit à une précision accrue. Cependant, une considération critique se pose lorsque l'environnement dans lequel le modèle est formé diffère considérablement de l'environnement dans lequel il sera appliqué à des fins de test. Cette disjonction des conditions environnementales peut introduire des perturbations et des défis imprévus.

Dans notre expérience, l'hypothèse initiale selon laquelle davantage de données d'entraînement (y compris les navires 1, 3 et 4) amélioreraient intrinsèquement les prévisions a été remise en question par les résultats observés. Les perturbations de l'approvisionnement en aliments dans les deux premiers navires entre les jours 11 et 13 ont créé des variations dans la dynamique microbienne qui n'ont pas été correctement capturées par les données de formation supplémentaires. Les perturbations imprévues soulignent l'importance d'aligner les données de formation sur les conditions et perturbations attendues dans l'environnement de test.

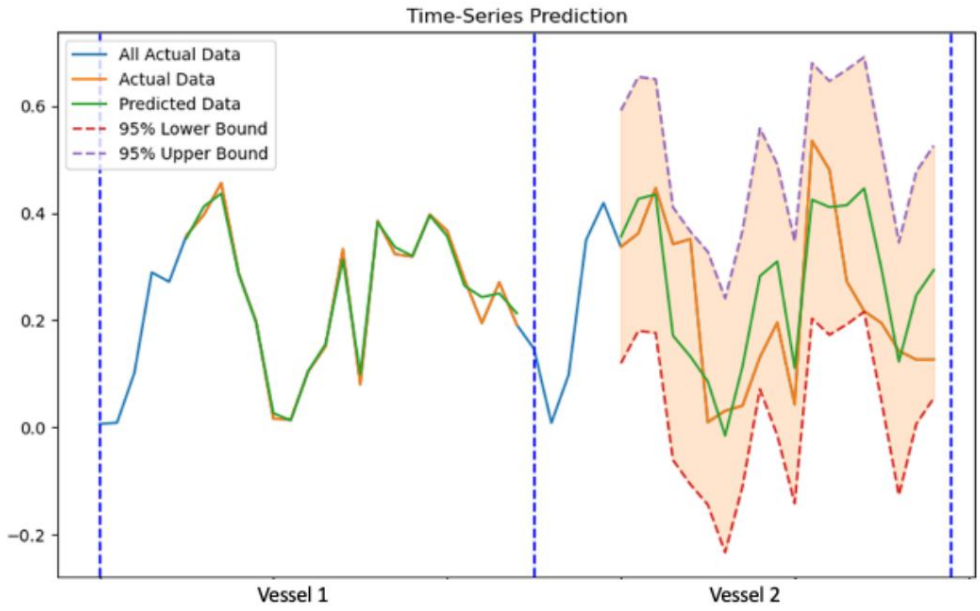

Figure 5. Trajectoires de l'abondance relative des Rikenellaceae dans les navires 1 et 2. La trajectoire entière du navire 2 est prédite par le modèle MicroGrowthPredictor formé sur les données du navire 1. Des intervalles de confiance sont fournis pour les données de test du navire 2. Dans cette expérience, la probabilité d'abandon optimale p de 0,25 a été utilisée. Le modèle a utilisé les cinq points temporels précédents pour identifier quatre paramètres et réaliser des prédictions optimales. La couche entièrement connectée était équipée de 256 nœuds et la couche LSTM comprenait 128 nœuds. Le modèle a subi un entraînement pendant 800 époque

Bien qu'il soit tentant de supposer qu'un échantillon plus grand conduira intrinsèquement à de meilleures prédictions, la clé réside dans la pertinence des données d'entraînement par rapport aux conditions de test. Dans les cas où les données de test impliquent différentes interruptions ou perturbations environnementales,

L'inclusion aveugle de divers ensembles de données peut conduire à des prédictions sous-optimales. L'équilibre délicat entre la quantité et la pertinence des données de formation devient crucial pour garantir l' adaptabilité du modèle aux scénarios du monde réel.

4. Discussion

Le modèle MicroGrowthPredictor exploite les connaissances issues des observations selon lesquelles un traitement antibiotique répété perturbe la communauté microbienne intestinale, affectant la diversité et l'abondance de groupes bactériens spécifiques. En analysant les données, le modèle prédit avec précision comment le microbiome évoluera au fil du temps en réponse à une perturbation antibiotique.

Cela permet de mieux comprendre l'impact des antibiotiques sur le microbiote intestinal et les implications potentielles pour la santé humaine.

De plus, la polyvalence du modèle est démontrée par son application à un ensemble de données intestinales artificielles. Les informations tirées de cet environnement contrôlé montrent l'adaptabilité de MicroGrowth-Predictor à divers systèmes de microbiome. L'ensemble de données sur l'intestin artificiel valide les capacités prédictives du modèle dans des conditions spécifiques, soulignant sa capacité à capturer des dynamiques temporelles complexes. Cela positionne le modèle comme étant précieux pour comprendre les effets des antibiotiques et des applications plus larges dans les sciences de l'environnement, les soins de santé et la biotechnologie.

Notre méthode aborde des problèmes du monde réel où la taille limitée des échantillons constitue une contrainte en raison de défis logistiques, éthiques ou financiers. En développant et en validant des méthodes qui fonctionnent bien avec des données limitées, nous proposons des solutions pratiques à de telles situations.

Contrairement à de nombreux modèles de boîte noire, notre approche offre des informations claires sur la façon dont les perturbations environnementales influencent les populations microbiennes au fil du temps, ce qui est crucial pour comprendre les processus biologiques et concevoir des interventions ciblées. Plus précisément, nous discutons de son potentiel à contribuer aux plans de traitement personnalisés en prédisant les réponses individuelles aux changements alimentaires, aux traitements antibiotiques et aux interventions probiotiques.

En résumé, MicroGrowthPredictor apparaît comme un outil puissant dépassant les approches de modélisation traditionnelles. Le modèle, piloté par des informations dérivées des données plutôt que par l'intégration directe des connaissances, intègre des réseaux LSTM avec une estimation de l'intervalle de confiance pour contribuer à une compréhension holistique de la dynamique du microbiome. Les applications réussies du modèle au microbiote intestinal humain réel et aux ensembles de données intestinales artificielles soulignent son efficacité et son impact potentiel. Nous prévoyons que MicroGrowthPredictor jouera un rôle central dans l'avancement de la recherche sur le microbiome, en offrant des informations précieuses et en contribuant à une prise de décision éclairée dans divers domaines.

Contributions des auteurs : Conceptualisation, Y.-HZ ; méthodologie, GS et Y.-HZ ; validation, GS et Y.-HZ ; rédaction – ébauche originale, GS et Y.-HZ ; rédaction-révision et édition, GS et Y.- HZ ; visualisation, GS et Y.-HZ ; surveillance, Y.-HZ; administration de projet, Y.-HZ; acquisition de financement , Y.-HZ Tous les auteurs ont lu et accepté la version publiée du manuscrit.

Financement : Cette recherche a été financée par l'Environmental Protection Agency des États-Unis, numéro de subvention 84045001, le National Institute of Health P30ES025128 et le programme des centres de recherche en ingénierie de la National Science Foundation dans le cadre de l'accord de coopération NSF n° EEC-2133504.

Déclaration de disponibilité des données : les données sont contenues dans l'article.

Conflits d'intérêts : Les auteurs déclarent que la recherche a été menée en l'absence de toute relation commerciale ou financière pouvant être considérée comme un conflit d'intérêts potentiel.

Les références

1. Altve¸s, S. ; Yildiz, Hong Kong ; Vural, HC Interaction du microbiote avec le corps humain dans la santé et les maladies. Biosci. Microbiote Alimentaire Santé 2020, 39, 23-32. [Référence croisée] [Pub Med]
2. Smith, J. ; Johnson, M. Dynamique du microbiome sous perturbations environnementales. J. Microbiome Res. 2022, 10, 123-145.
3. Brun, EM; Sadarangani, M. ; Finlay, BB Le rôle du système immunitaire dans la régulation des interactions hôte-microbe dans l'intestin. Nat. Immunol. 2013, 14, 660-667. [Référence croisée] [Pub Med]
4. Candela, M. ; Biagi, E. ; Maccaferri, S. ; Turroni, S. ; Brigidi, P. Le microbiote intestinal est un facteur plastique répondant aux changements environnementaux. Tendances Microbiol. 2012, 20, 385-391. [Référence croisée]

5. Heure, GT ; Dohnalová, L. ; Thaiss, CA La dimension du temps dans les interactions hôte-microbiome. mSystems 2019, 4, e00216-18. [Référence croisée] [Pub Med]
6. Volontaire, BP ; Russell, SL ; Finlay, BB Changer l'équilibre : effets des antibiotiques sur le mutualisme hôte-microbiote. Nat. Révérend Microbiol. 2011, 9, 233-243. [Référence croisée] [Pub Med]
7. Brun, E. ; Williams, D. Modélisation prédictive de la croissance du microbiome à l'aide des réseaux LSTM. J. Informatique. Biol. 2021, 45, 321-335.
8. Ching, T. ; Himmelstein, DS; Beaulieu-Jones, BK; Kalinine, AA; Fais, BT ; Bien, médecin généraliste ; Ferrero, E. ; Agapow, PM ; Zietz, M. ; Hoffman, MM; et coll. Opportunités et obstacles à l'apprentissage profond en biologie et en médecine. JR Soc. Interface 2018, 15, 20170387. [Référence croisée] [Pub Med]
9. Medsker, LR ; Jain, L. Réseaux de neurones récurrents. Des. Appl. 2001, 5, 2.
10. Graves, A. ; Graves, A. Mémoire longue à court terme. Dans l'étiquetage de séquence supervisé avec des réseaux de neurones récurrents ; Springer : Berlin/Heidelberg, Allemagne, 2012 ; p. 37-45.
11. Yu, Y. ; Six.; Hu, C. ; Zhang, J. Une revue des réseaux de neurones récurrents : cellules LSTM et architectures de réseau. Calcul neuronal. 2019, 31, 1235-1270. [Référence croisée] [Pub Med]
12. Zhou, YH ; Gallins, P. Une revue et un didacticiel des méthodes d'apprentissage automatique pour la prédiction des traits de l'hôte du microbiome. Devant. Genet. 2019, 10, 579. [Réf. croisée] [Pub Med]
13. Zhu, L. ; Laptev, N. Prédiction approfondie et confiante pour les séries chronologiques chez Uber. Dans Actes de la Conférence internationale de l'IEEE 2017 sur les ateliers d'exploration de données (ICDMW), Orléans, LA, États-Unis, 18-21 novembre 2017 ; IEEE : Piscataway, New Jersey, États-Unis, 2017 ; pp. 103-110.
14. Gal, Y. ; Ghahramani, Z. Dropout comme approximation bayésienne : représentation de l'incertitude du modèle dans l'apprentissage profond. Dans Actes de la Conférence internationale sur l'apprentissage automatique, PMLR, New York, NY, États-Unis, 20-22 juin 2016 ; pp. 1050-1059.
15. Dethlefsen, L. ; Relman, DA Récupération incomplète et réponses individualisées du microbiote intestinal distal humain à des perturbations répétées des antibiotiques. Proc. Natl. Acad. Sci. États-Unis 2011, 108, 4554-4561. [Référence croisée] [Pub Med]
16. Silverman, JD; Durand, Hong Kong ; Bloom, RJ ; Mukherjee, S. ; David, LA Les modèles linéaires dynamiques guident la conception et l'analyse des études sur le microbiote dans les intestins humains artificiels. Microbiome 2018, 6, 202.