Nuevas matemáticas para mapear redes ecológicas microbianas

New mathematics for mapping microbial ecological networks

Marco Tulio Angulo, CONACyT – Instituto de Matemáticas, Universidad Nacional Autónoma de México

RESUMEN. Los microbios alojados en nuestro cuerpo y planeta proveen muchos servicios esenciales para la salud humana. Mapear las redes ecológicas subyacentes a estas comunidades microbianas es un paso necesario para poder predecir su comportamiento, abriendo la puerta para crear nuevos tratamientos para enfermedades humanas y muchas otras condiciones. Sin embargo, los algoritmos existentes para mapear estas redes ecológicas microbianas no han sido muy exitosos. Esto es principalmente debido a que requieren asumir un modelo dinámico poblacional particular para la comunidad –que nunca es conocido a-priori– y siempre arrojan una sola red ecológica como resultado de la inferencia, a pesar de que puede haber varias redes ecológicas que son consistentes con los datos disponibles. Para superar estos retos, aquí desarrollamos un nuevo algoritmo de inferencia matemáticamente riguroso que no requiere asumir ningún modelo poblacional y que infiere todas las redes ecológicas que son consistentes con los datos. Ilustramos nuestro algoritmo con datos simulados y validamos su desempeño usando datos experimentales. Argumentamos como el algoritmo propuesto puede ser un paso clave para modelar ecológicamente comunidades microbianas complejas como el microbiota intestinal humano.

PALABRAS CLAVE: comunidades microbianas; inferencia; redes ecológicas; teoría de sistemas.

ABSTRACT. The microbes hosted in our body and on Earth provide many essential services for human well-being. Mapping the ecological networks underlying these microbial communities is a necessary step for predicting their behavior, opening the door to create new treatments for human diseases and many other conditions. However, the existing algorithms to map these networks have not been very successful. The limited success of these algorithms is because they require to assume a population dynamics model for the community —which is never a priori known—, and they always provide a single network as result of the inference, despite there might be several ecological networks that are consistent with the data. To overcome these two challenges, here we develop a new mathematically-rigorous inference algorithm that does not require assuming any population dynamics, and that infers all ecological networks that are consistent with the data. We illustrate our algorithm with simulated data and then validate its performance using experimental data. Our algorithm could be a pivotal step to ecologically model complex microbial communities such as the human gut microbiota.

KEYWORDS: Ecological networks, inference, microbial communities, systems theory.

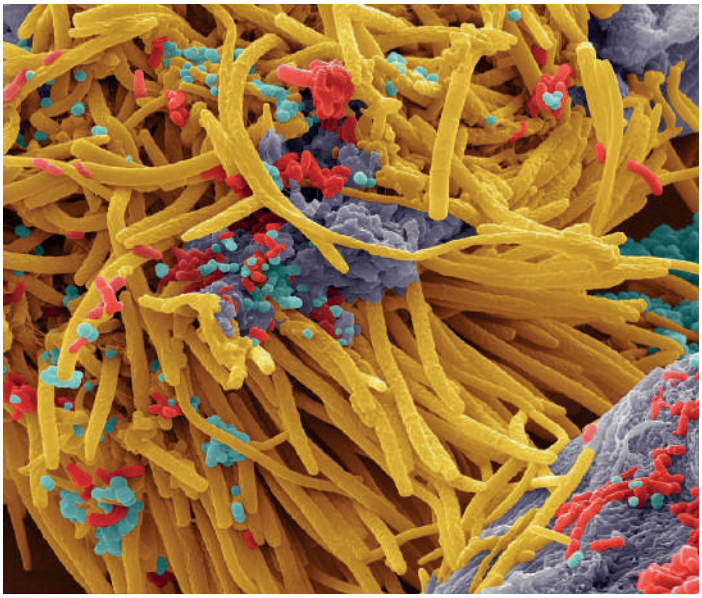
Introducción

Los microbios alojados en nuestro cuerpo y planeta proveen nutrientes y muchos otros servicios que son indispensables para mantener la salud humana (Mueller & Sachs, 2015; Shreiner, Kao, & Young, 2015). Avances muy recientes en tecnologías de secuenciación de ADN han hecho posible comenzar a entender la composición y funciones metabólicas de muchas de estas comunidades microbianas (Turnbaugh, et al., 2007). Esto ha revelado las profundas consecuencias de alterarlas. Por ejemplo, alteraciones en el *microbiota intestinal* humano —el agregado de microorganismos que residen en nuestros intestinos— se relacionan no solo a enfermedades gastrointestinales, sino a condiciones tan divergentes como autismo, obesidad, y el desarrollo de nuestro sistema inmune (Dietert, 2016). Alteraciones en otras comunidades microbianas pueden contribuir a disminuir la productividad de cultivos agrícolas (Mueller & Sachs, 2015), o perturbar el clima al cambiar la tasa de secuestro de carbono en los océanos (Guidi, et al., 2016). Modificar el balance de especies[1] en una comunidad microbiana podría entonces ofrecer nuevas formas de tratar o prevenir enfermedades, incrementar la productividad de cultivos, o incluso crear fármacos y biocombustibles (Alivisatos, et al., 2015; Dubilier, McFall-Ngai, & Zhao, 2015). Sin embargo, antes de que la medicina y la bioingeniería puedan alcanzar estas metas, primero es necesario entender cómo las especies de una comunidad microbiana interactúan entre sí.

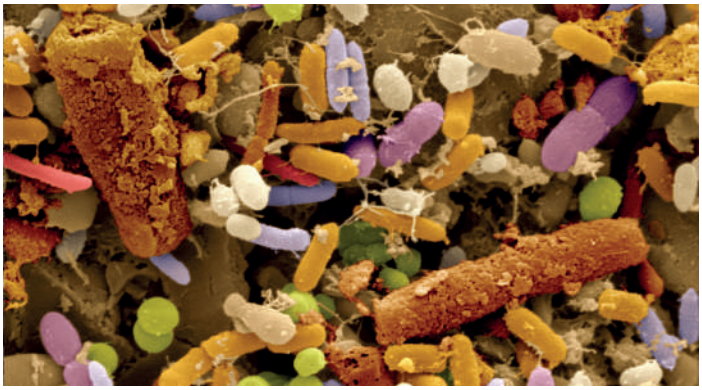
Mapear redes ecológicas de interacciones microbianas es un paso necesario para predecir el resultado de modificar a una comunidad (Widder, et al., 2016; Buffie, et al., 2015), o incluso para construir comunidades sintéticas que realicen funciones deseadas (Hudson, Anderson, Corbett, & Lamb, 2017). Una red ecológica es dirigida, ponderada y con signos, donde nodos representan especies y aristas interacciones ecológicas directas entre especies (e.g., predador-presa, parasitismo, comensalismo, mutualismo, amensalismo o competición). Enfatizamos que esta red ecológica es fundamentalmente distinta a redes construidas a partir de correlaciones o co-ocurrencia de especies, que no son dirigidas y por tanto no codifican ninguna relación causal que permita predecir el comportamiento dinámico del ecosistema subyacente (Friedman & Alm, 2012). Desafortunadamente, a pesar de que el conocimiento sobre la composición y funciones de muchas comunidades microbianas ha avanzado enormemente, el entendimiento de la estructura de sus redes ecológicas subyacentes se ha rezagado (Faust & Raes, 2012).

Este rezago es debido, principalmente, a dos limitaciones fundamentales de los algoritmos existentes para inferir redes ecológicas microbianas a partir de *datos metagenómicos* (i.e., datos con la abundancia de especies de la comunidad en cierto instante de tiempo). Primero, los algoritmos de inferencia existentes requieren elegir a-priori un modelo parametrizado de la dinámica poblacional subyacente a la comunidad (Steinway, Biggs, Loughran Jr, Papin, & Albert, 2015; Bucci, et al., 2016; Cao, Gibson, Bashan, & Liu, 2017). Cualquier elección es siempre extremadamente difícil de justificar, debido a que las especies en una comunidad interactúan a través de tantos mecanismos que producen dinámicas muy diversas incluso a la escala de dos especies (Jost & Ellner, 2000). Críticamente, cualquier error entre el modelo elegido y la dinámica poblacional "real" de la comunidad puede generar errores de inferencia arbitrariamente grandes (Angulo, Moreno, Lippner, Barabási, & Liu, 2017). La segunda limitación es la poca *informatividad* de los datos metagenómicos (Cao, Gibson, Bashan, & Liu, 2017), originada por la estabilidad de muchas comunidades microbianas como el microbiota intestinal humano (Lozupone, Stombaugh, Gordon, Jansson, & Knight, 2012). Dicha estabilidad hace que los datos metagenómicos tiendan a capturar pequeñas variaciones con respecto a los equilibrios de la comunidad microbiana, disminuyendo su informatividad. La falta de informatividad implica que existen varias redes ecológicas que explican los datos metagenómicos disponibles. Todas estas redes son igualmente útiles, pues cada una de ellas codifica un posible comportamiento de la comunidad ante modificaciones. A pesar de esto, todos los algoritmos de inferencia existentes siempre arrojan solo una red ecológica, sin importar la informatividad de los datos. Es importante resaltar que la mayoría de los métodos de inferencia para sistemas dinámicos en general comparten estas mismas dos limitaciones (Ljung, 1998).

1 Utilizamos el término "especie" en un contexto ecológico general: como un grupo de organismos adaptados a un conjunto particular de recursos en el ambiente. No representa necesariamente el rango taxonómico más bajo. De hecho, los microbios también podrían organizarse por cepas, géneros, o unidades taxonómicas operacionales.



FUENTE: http://extrastory.cz/images/2016/04-duben/04-4/bacteria-mouth.jpg



FUENTE: https://media.npr.org/assets/img/2015/03/12/feces_wide-0aac9a2e67eb0eb36c446ad79ed80cd3d987c746.jpg?s=1400

En este artículo desarrollamos un nuevo formalismo matemático para inferir redes ecológicas microbianas resolviendo las dos limitaciones fundamentales arriba mencionadas. En particular, construimos un algoritmo matemáticamente riguroso que permite: (I) inferir redes ecológicas sin necesidad de conocer la dinámica poblacional subyacente; (II) cuantificar la informatividad de los datos e inferir el conjunto de todas las redes ecológicas que son consistentes con los datos metagenómicos disponibles; y (III) utilizar datos metagenómicos tanto temporales como en estado estable para hacer la inferencia. El formalismo matemático que utilizamos generaliza nuestro trabajo reciente (Xiao, et al., 2017) en el punto (II), y también al utilizar datos temporales para realizar la inferencia. Ilustramos nuestro algoritmo usando datos en simulación y luego validamos su desempeño con datos experimentales. Terminamos argumentando como la adopción y aplicación del algoritmo propuesto podría ser un paso clave para modelar ecológicamente comunidades microbianas complejas como el microbiota intestinal humano.

Formulación del problema

La red ecológica $\mathcal{G}=(X,E)$ de una comunidad microbiana de N especies consta de nodos $X=\{1,\dots,N\}$ correspondientes a especies, y aristas $(j\to i)\in E$ representando una interacción ecológica directa de la j-ésima especie en la i-ésima especie. Las interacciones de una red ecológica pueden ser de dos clases: inhibición (negativas) o promoción (positivas) de crecimiento. La red ecológica de una comunidad microbiana esta codificada en su dinámica poblacional, que en general puede ser descrita por un conjunto de N ecuaciones diferenciales de la forma

$$\frac{dx_i(t)}{dt}=x_i(t)\,f_i\big(x(t)\big),\ i=1,\dots,N. \tag{1}$$

Aquí $x(t)=(x_1(t),\dots,x_N(t))\in\mathbb{R}^N$ es un vector donde $x_i(t)$ representa la abundancia absoluta de la i-ésima especie al tiempo t. Las funciones $f_i:\mathbb{R}^N\to\mathbb{R}$, $i=1,\dots,N$, determinan la dinámica poblacional de la comunidad, modelando las interacciones intra- y entre-especies. Estas interacciones pueden ocurrir a través de muy diversos mecanismos incluyendo alimentación cruzada, bacteriocinas, o intercambio de electrones, haciendo muy difícil saber la forma funcional adecuada para f_i. Por tanto, consideramos que las f_i's son funciones meromórficas desconocidas (i.e., cocientes de funciones analíticas). De esta forma, nuestra única suposición sobre la estructura de la dinámica poblacional de la comunidad es la variable x_i que aparece factorizada en la Ec. (1). Esta suposición modela que una especie que se extingue no puede reaparecer, y se satisface cuando la comunidad microbiana no es sujeta a perturbaciones externas como migración o invasión.

Matemáticamente, las interacciones ecológicas directas entre las especies de la comunidad están descritas por la matriz Jacobiana $J(x)\in\mathbb{R}^{N\times N}$ de la Ec. (1), con elementos $J_{ij}=\partial f_i/\partial x_j$. En particular, la clase de interacción ecológica (inhibición o promoción) esta codificada por el *patrón de signos* $S(x)=\text{sign}J(x)\in\{-1,0,+1\}^{N\times N}$ con elementos $s_{ij}=\text{sign}(J_{ij})$. Esto ocurre debido a que la j-ésima especie promueve, inhibe, o no tiene efecto directo sobre el crecimiento de la i-ésima especie si y solo si $J_{ij}(x)>0$, $J_{ij}(x)<0$, o $J_{ij}(x)\equiv 0$, respectivamente.

Asumiremos que el patrón de signos S es constante, lo que implica que la clase de las interacciones ecológicas no cambia con el tiempo. Note, sin embargo, que la magnitud de $J_{ij}(x)$ puede variar arbitrariamente e incluso depender de la abundancia de especies distintas a i y j. Esta suposición es muy débil en el sentido de que es satisfecha por la mayoría de las llamadas "respuestas funcionales" que caracterizan los mecanismos de interacción entre especies en modelos dinámicos poblacionales, como por ejemplo las respuestas funcionales Holling Tipo I y II, Crawley-Martin, entre otras (Xiao, et al., 2017).

Para inferir la red ecológica de una comunidad microbiana, consideramos que tenemos un conjunto $\mathfrak{D}$ de datos metagenómicos. Cada dato metagenómico es un vector $y\in\mathbb{R}^N$ con la abundancia de cada especie en la comunidad en cierto instante de tiempo. A un dato metagenómico también le llamamos *muestra*. Los datos en $\mathfrak{D}$ pueden ser *temporales* o en *estado estable*. Los datos temporales son series de tiempo $\{y(t)\in\mathbb{R}^N, t\in\{t_0,\dots,t_L\}\}$ tales que $y(t)$ satisface la Ec. (1) para $t\in\{t_0,\dots,t_L\}$. Por ejemplo, series de tiempo con la respuesta temporal de la comunidad para distintas abundancias iniciales de especies (Fig. 1 A). Supondremos que estos datos temporales son muestreados con suficiente frecuencia para estimar razonablemente bien su derivada temporal $\{\dot{y}(t),\ t\in\{t_0,\dots,t_L\}\}$. Esto es posible, por ejemplo, si t_{k+1}-t_k es suficientemente pequeño, permitiendo ajustar una función continua $\{\hat{y}(t),\ t\in[t_0,t_f]\}$ a la serie de tiempo tal que $\dot{y}(t)\approx d\hat{y}(t)/dt$. Los datos en estado estable son constantes $\{y\in\mathbb{R}^N\}$ que corresponden a equilibrios no triviales de la Ec. (1), es decir, equilibrios donde al menos una especie está presente. Estos datos pueden representar la abundancia en estado estable de diferentes composiciones de especies de la misma comunidad microbiana (Fig. 2 A). Mapear la red ecológica de la comunidad consiste entonces en inferir el patrón de signos S a partir de datos metagenómicos $\mathfrak{D}$ dados, sin conocer cuál es la dinámica poblacional f_i de la comunidad.

Resultados

Mapeando redes ecológicas microbianas a partir de datos temporales

Para inferir el patrón de signos $S_i=\text{sign}(J_i)\in\{-1,0,1\}^N$ de la i-ésima especie, considere el subconjunto $\mathfrak{D}_i\subseteq\mathfrak{D}$ de todas las muestras en donde dicha especie está presente. Elija dos series de tiempo $y(\cdot),z(\cdot)\in\mathfrak{D}_i$ y sean $y(\tau),z(\tau')\in\mathfrak{D}_i$ dos muestras en instantes de tiempo τ, τ' (Fig. 1 A). Cuando se elige $y=z$, suponemos que $\tau\neq\tau'$ de tal forma que $y(\tau)$ y $z(\tau')$ corresponden a distintas muestras de la misma serie de tiempo. Como ambas series de tiempo son soluciones de la Ec. (1) y tienen presente a la i-ésima especie, note que satisfacen $\dot{y}_i(\tau)/y_i(\tau)-\dot{z}_i(\tau')/z_i(\tau')=f_i(y(\tau))-f_i(z(\tau'))$. Entonces, definiendo $\beta_i(y(\tau),z(\tau')):=\dot{y}_i(\tau)/y_i(\tau)-\dot{z}_i(\tau')/z_i(\tau')$ y aplicando el Teorema del Valor Intermedio para funciones multidimensionales, podemos usar esta última expresión para obtener:

$$\nu_i\big(y(\tau),z(\tau')\big)\cdot\big[y(\tau)-z(\tau')\big]=\beta_i\big(y(\tau),z(\tau')\big),\quad \forall y(\tau),z(\tau')\in\mathfrak{D}_i. \tag{2}$$

En esta ecuación, la notación "$\cdot$" indica el producto interno entre vectores, y $\nu_i(y,z)\in\mathbb{R}^N$ está definido como $\nu_i(y,z):=\int_0^1 J_i(y+\sigma(z-y))d\sigma$. Aquí, $J_i(x)=\partial f_i(x)/\partial x\in\mathbb{R}^N$ denota el i-ésimo renglón de la matriz $J(x)$. La Ec. (2) puede reescribirse de una manera más compacta como

$$\nu_i\big(y(\tau),z(\tau')\big)\cdot\delta_i\big(y(\tau),z(\tau')\big)=1,\qquad \forall y(\tau),z(\tau')\in\mathfrak{D}_i, \tag{3}$$

donde hemos definido el vector $\delta_i(y,z):=[y-z]/\beta_i(y,z)$. Note que $\delta_i(y,z)$ puede calcularse utilizando los datos $\mathfrak{D}_i$.

La observación crucial que permite inferir la red ecológica microbiana es la siguiente: como por hipótesis el patrón de signos S del Jacobiano $J(x)$ es constante, entonces el patrón de signos del vector $\nu_i(y,z)=\int_0^1 J_i(y+\sigma(z-y))d\sigma$ es el mismo para todo $y,z\in\mathfrak{D}_i$ y coincide con el patrón de signos $S_i=\text{sign}(J_i)$ que buscamos inferir. Entonces, para inferir el patrón de signos S_i, para cada par $y,z\in\mathfrak{D}_i$ y su correspondiente $\delta_i(y,z)$, calculamos el hiperplano $V_i(y,z)$ de todos los vectores $\{v_i|v_i\in\mathbb{R}^N\}$ que satisfacen la Ec. (3). Cada uno de estos hiperplanos $V_i(y,z)$ tiene asociado un conjunto de patrones de signo $\mathfrak{S}_i(y,z)\subseteq\{-1,0,1\}^N$ determinado por todos los ortantes de $\mathbb{R}^N$ a los que pertenece. Este conjunto puede ser calculado resolviendo 3^N programas lineales. Intersectando todos estos patrones de signo que han sido obtenidos, podemos calcular

$$\mathfrak{S}_i^*=\bigcap\nolimits_{y,z\in\mathfrak{D}_i}\mathfrak{S}_i(y,z) \tag{4}$$

El conjunto $\mathfrak{S}_i^*$ en esta ecuación constituye la base del algoritmo propuesto para inferir S_i, pues se calcula a partir de los datos $\mathfrak{D}_i$ y, debido a la observación hecha arriba, satisface que $S_i\in\mathfrak{S}_i^*$.

Note que si $\mathfrak{S}_i^*$ tiene un solo patrón de signos, entonces $S_i=\mathfrak{S}_i^*$. En tal caso, decimos que los datos $\mathfrak{D}_i$

son *totalmente informativos* para la i-ésima especie, pues ellos permiten inferir unívocamente su patrón de signos. Si el conjunto $\mathfrak{S}_i^*$ tiene más de un patrón de signos, el argumento presentado arriba prueba que todos ellos son consistentes con los datos $\mathfrak{D}_i$. Esto tiene dos implicaciones. Primero que, en este último caso, los datos $\mathfrak{D}_i$ no son suficientemente informativos como para inferir unívocamente el patrón de signos S_i de la i-ésima especie. A pesar de esto, utilizando la Ec. (4) podemos calcular *todos* los patrones de signos que son consistentes con los datos $\mathfrak{D}_i$. La segunda implicación es que la Ec. (4) infiere el patrón de signos bajo condiciones necesarias y suficientes de los datos $\mathfrak{D}_i$. Es decir, si utilizando esta ecuación no es posible inferir unívocamente el patrón de signos usando $\mathfrak{D}_i$, ningún otro algoritmo es capaz de hacerlo usando estos mismos datos.

El algoritmo de inferencia propuesto en este artículo se basa en representar la Ec. (4) a través de un histograma de la frecuencia $\omega_{i,k}$ con la que cada patrón de signo $k\in\{-1,0,1\}^N$ ocurre en el conjunto $\{\mathfrak{S}_i(y,z)|\forall y,z\in\mathfrak{D}_i\}$. Definiendo $\omega_i^*:=\max_k\omega_{i,k}$ y una *precisión* $\varepsilon\in[0,1]$, el algoritmo propuesto construye el conjunto de patrones de signo $\hat{\mathfrak{S}}_i^\varepsilon$ para inferir S_i como

$$\hat{\mathfrak{S}}_i^\varepsilon=\left\{k\in\{-1,0,1\}^N \mid \omega_{i,k}\in\left[\omega_i^*-\varepsilon,\omega_i^*\right]\right\} \tag{5}$$

Note que si $\hat{\mathfrak{S}}_i^*$ tiene al menos un elemento entonces $\hat{\mathfrak{S}}_i^0=\hat{\mathfrak{S}}_i^*$. Esto prueba que el algoritmo de la Ec. (5) también infiere todos los patrones de signos consistentes con $\mathfrak{D}_i$ bajo condiciones necesarias y suficientes.

En la práctica, sin embargo, es posible que ruidos en la medición de los datos metagenómicos $\mathfrak{D}_i$ ocasionen que la intersección en la Ec. (4) sea vacía. En tal caso, el algoritmo de la Ec. (5) entrega como inferencia el conjunto de todos los patrones de signo con máxima frecuencia de ocurrencia. Esta formulación también permite cuantificar la *informatividad* $I_{ij}^\varepsilon(\mathfrak{D}_i)\in[0,1]$ de los datos $\mathfrak{D}_i$ para inferir el signo s_{ij} de la interacción entre la especie j y la especie i. Definimos $I_{ij}^\varepsilon(\mathfrak{D}_i):=\omega_k^*/\max(1,\sigma_{ij}^\varepsilon)$, donde σ_{ij}^ε es el número de cambios de signo en la j-ésima entrada entre todos los vectores en $\hat{\mathfrak{S}}_i^\varepsilon$. Si los datos $\mathfrak{D}_i$ son totalmente informativos para la i-ésima especie, entonces $\omega_i^*=1$ y $\hat{\mathfrak{S}}_i^0$ contiene un solo elemento. En este caso, $I_{ij}^0(\mathfrak{D}_i)=1$ para $j=1,\ldots,N$ y la informatividad es máxima, coincidiendo con el hecho de que es posible inferir unívocamente s_{ij} para todo j. En otro caso, se satisface $I_{ij}^\varepsilon(\mathfrak{D}_i)\in[0,1]$ con $I_{ij}^\varepsilon(\mathfrak{D}_i)<\omega_i^*$. Entonces, al utilizar el algoritmo de la Ec. (5), la informatividad $I_{ij}^\varepsilon(\mathfrak{D}_i)$ es una medida de la confidencia en el patrón de signos inferido para la interacción j→i.

Para ilustrar el funcionamiento del algoritmo de la Ec. (5), considere la comunidad microbiana con $N=2$ especies de la Fig. 1. Como datos disponibles $\mathfrak{D}$, utilizamos dos series de tiempo $y(\cdot),z(\cdot)$ correspondientes a la respuesta de la comunidad a dos abundancias iniciales de especies (Fig. 1 A). Eligiendo una muestra en cada serie de tiempo $y(\tau),z(\tau')$, el vector $y(\tau)$-$z(\tau')$ de la Ec. (2) es el vector que une estas dos muestras. El vector $\delta_i(y(\tau),z(\tau'))$ de la Ec. (3) tiene esta misma orientación. Para $N=2$, el hiperplano $V_i(y(\tau),z(\tau'))$ de todos los vectores que satisfacen la Ec. (3) es simplemente una línea (Fig. 1 B). Para cada par de muestras $y(\tau),z(\tau')$ la línea $V_i(y(\tau),z(\tau'))$ correspondiente tiene asociados un conjunto de patrones de signos $S_i(y(\tau),z(\tau'))$, que pueden ser contados en un histograma (Fig. 1 C). En este histograma, observamos que solo un patrón de signos tiene frecuencia de ocurrencia máxima $\omega_{i,k}=1$ para cada especie (marcas en Fig. 1 C), determinando correcta y unívocamente el patrón de signos S_i para las dos especies de la comunidad. De hecho, $I_{ij}^0(\mathfrak{D}_i)=1$ para $i,j=1,2$. Por tanto, los datos $\mathfrak{D}$ son totalmente informativos para ambas especies en este ejemplo. De esta forma, el algoritmo de la Ec. (5) infiere unívocamente la red ecológica de la comunidad solo a partir de datos metagenómicos (panel derecho de la Fig. 1 C), sin conocer su dinámica poblacional.

Mapeando redes ecológicas microbianas a partir de datos en estado estable

El algoritmo de la Ec. (5) aplica para cualquier conjunto de muestras, en particular, si son muestras en estado estable. De hecho, en este caso el algoritmo tiene una interpretación geométrica más simple (Xiao, et al., 2017). Para ver esto, note que para cualquiera par de muestras en estado estable $y,z\in\mathfrak{D}_i$, se tiene $\beta_i(y,z)=0$. Por lo tanto, la Ec. (2) implica que el conjunto de $V_i(y,z)$'s son simplemente hiperplanos ortogonales a los vectores y-z, y todos ellos intersectan al origen de $\mathbb{R}^N$. Como consecuencia, la intersección en la Ec. (4) puede resultar a lo menos en una línea. Esto implica que $\hat{\mathfrak{S}}_i^0$ en la Ec. (5) tiene a lo menos tres patrones de signo, i.e., $\hat{\mathfrak{S}}_i^0=\{-\hat{s}_i,0,\hat{s}_i\}$ para algún $\hat{s}_i\in\{-1,0,1\}^N$. Como consecuencia, $\mathfrak{D}_i$ nunca puede ser totalmente informativo cuando solo contiene datos en estado estable. Para aumentar la informatividad de $\mathfrak{D}_i$ sería necesario agregar datos

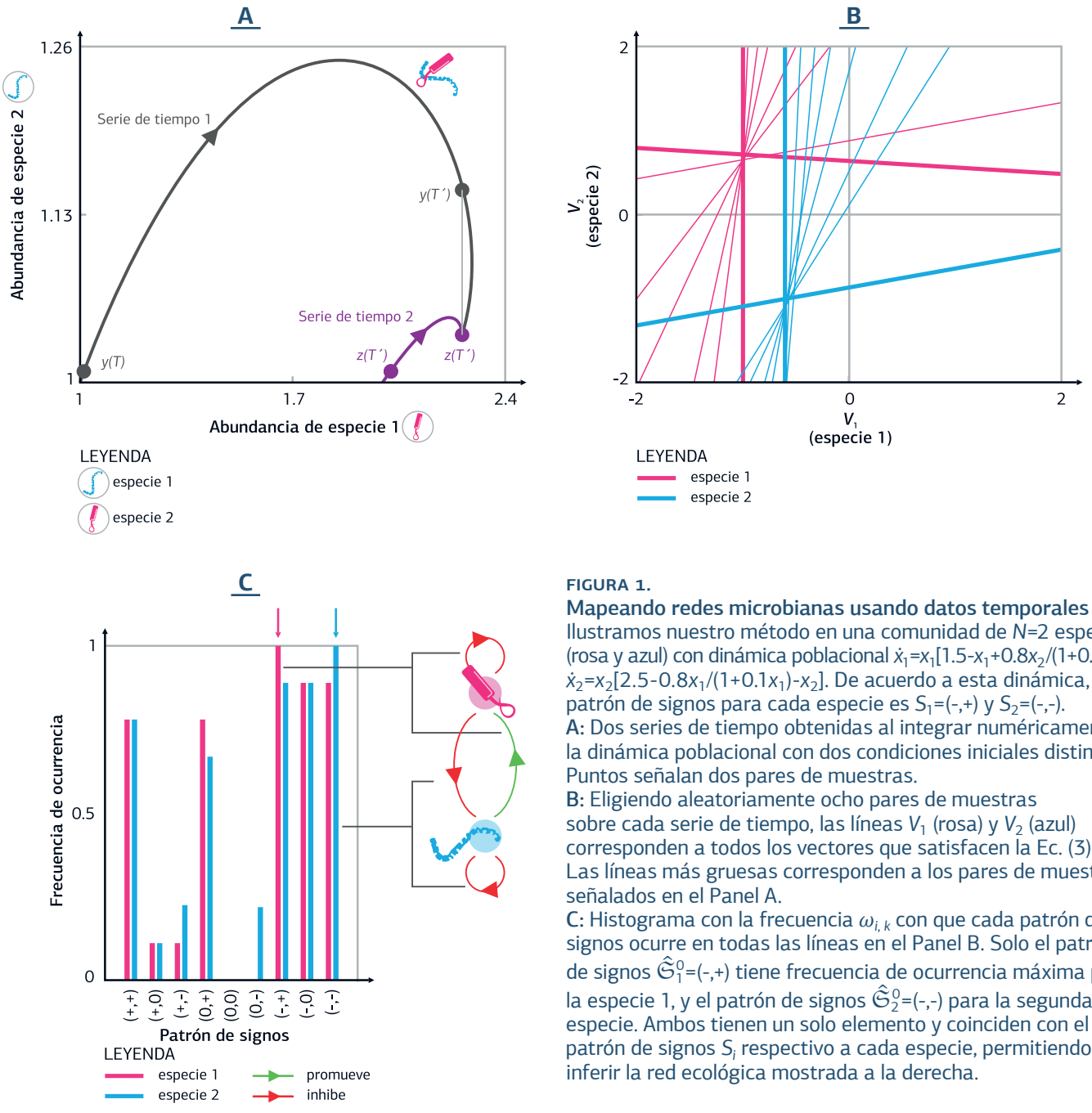


FIGURA 1.
Mapeando redes microbianas usando datos temporales
Ilustramos nuestro método en una comunidad de N=2 especies (rosa y azul) con dinámica poblacional $\dot{x}_1=x_1[1.5-x_1+0.8x_2/(1+0.1x_2)]$, $\dot{x}_2=x_2[2.5-0.8x_1/(1+0.1x_1)-x_2]$. De acuerdo a esta dinámica, el patrón de signos para cada especie es S_1=(-,+) y S_2=(-,-).
A: Dos series de tiempo obtenidas al integrar numéricamente la dinámica poblacional con dos condiciones iniciales distintas. Puntos señalan dos pares de muestras.
B: Eligiendo aleatoriamente ocho pares de muestras sobre cada serie de tiempo, las líneas V_1 (rosa) y V_2 (azul) corresponden a todos los vectores que satisfacen la Ec. (3). Las líneas más gruesas corresponden a los pares de muestras señalados en el Panel A.
C: Histograma con la frecuencia $\omega_{i,k}$ con que cada patrón de signos ocurre en todas las líneas en el Panel B. Solo el patrón de signos $\widehat{\mathfrak{S}}_1^0$=(-,+) tiene frecuencia de ocurrencia máxima para la especie 1, y el patrón de signos $\widehat{\mathfrak{S}}_2^0$=(-,-) para la segunda especie. Ambos tienen un solo elemento y coinciden con el patrón de signos S_i respectivo a cada especie, permitiendo inferir la red ecológica mostrada a la derecha.

temporales, como mostramos en la sección anterior. Otra opción, sin embargo, es usar *información previa*. En particular, es razonable suponer que todas las interacciones inter-especie son inhibitorias s_{ii}=-1, pues esta condición frecuentemente garantiza la estabilidad de la comunidad (May, 1973). Usando esta información previa, es posible decidir cuál de los tres patrones de signo en $\widehat{\mathfrak{S}}_i^0$ es el correcto, infiriendo una única red ecológica microbiana. Utilizando ε>0 se puede ajustar la sensibilidad a errores de medición en los datos metagenómicos disponibles.

En la Fig. 2, ilustramos el funcionamiento del algoritmo de la Ec. (5) en una comunidad de dos especies con datos en estado estable. En este caso, los datos corresponden a la abundancia en estado estable de tres composiciones posibles de la comunidad: cuando las especies 1 o 2 crecen en monocultivo, y cuando ambas especies crecen en co-cultivo

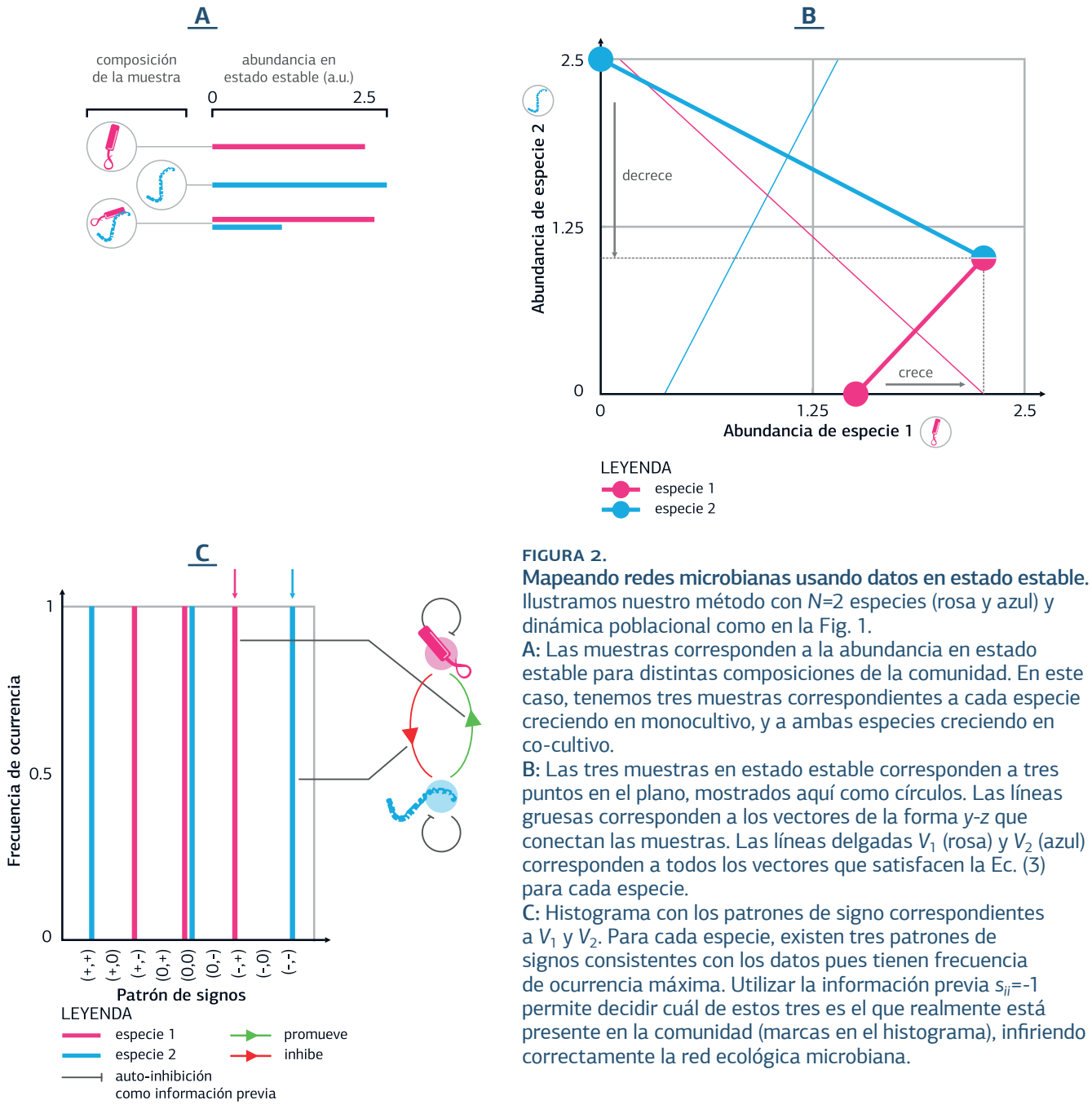


FIGURA 2.
Mapeando redes microbianas usando datos en estado estable. Ilustramos nuestro método con N=2 especies (rosa y azul) y dinámica poblacional como en la Fig. 1.
A: Las muestras corresponden a la abundancia en estado estable para distintas composiciones de la comunidad. En este caso, tenemos tres muestras correspondientes a cada especie creciendo en monocultivo, y a ambas especies creciendo en co-cultivo.
B: Las tres muestras en estado estable corresponden a tres puntos en el plano, mostrados aquí como círculos. Las líneas gruesas corresponden a los vectores de la forma y-z que conectan las muestras. Las líneas delgadas V_1 (rosa) y V_2 (azul) corresponden a todos los vectores que satisfacen la Ec. (3) para cada especie.
C: Histograma con los patrones de signo correspondientes a V_1 y V_2. Para cada especie, existen tres patrones de signos consistentes con los datos pues tienen frecuencia de ocurrencia máxima. Utilizar la información previa s_{ii}=-1 permite decidir cuál de estos tres es el que realmente está presente en la comunidad (marcas en el histograma), infiriendo correctamente la red ecológica microbiana.

(Fig. 2 A). Estos tres datos en estado estable pueden graficarse en el plano, en donde los vectores de la forma y-z son rectas que conectan estos tres estados estables (Fig. 2 B). Como N=2, el hiperplano V_i (y,z) ortogonal a y-z es simplemente una línea, y por tanto a lo más podemos obtener una línea ortogonal V_1 para la especie 1, y otra línea ortogonal V_2 para la especie 2 (Fig. 2 B). Cada una de estas líneas V_i provee tres patrones de signo de la forma $\hat{\mathfrak{S}}_i^0$={-$\hat{s}_i$,0,$\hat{s}_i$}. Con base a la información previa s_{ii}=-1 inferimos correcta y unívocamente el patrón de signos para ambas especies (Fig. 2 C). Nuestro algoritmo, a pesar de ser construido a partir de un formalismo puramente matemático, produce el mismo resultado que la siguiente observación empírica: como la abundancia en estado estable de la especie 1 es mayor en co-cultivo que en monocultivo, entonces la especie 2 debe promover el crecimiento de la especie 1 (Fig. 2 B).

Esto permite inferir la red ecológica cuando las muestras en estado estable difieren en solo una especie. En este sentido, nuestro formalismo permite extender esta observación empírica al caso general en donde las muestras en estado estable difieren en un número arbitrario de especies.

Aplicación usando datos experimentales.

Validamos el desempeño del algoritmo de la Ec. (5) para inferir la red ecológica de una comunidad microbiana de dos especies, *paramecium* y *didinium*, usando datos temporales de su abundancia. Los datos temporales son el resultado de un experimento en co-cultivo utilizando una solución con 0.5 g/l de Cerophyl (Jost & Ellner, 2000). Utilizamos una serie de tiempo con aproximadamente 70 muestras temporales tomadas con una frecuencia de 0.5 días, cada muestra con la cantidad de individuos de cada especie por mililitro de solución (Fig. 3 A). Para calcular la derivada temporal de esta serie de tiempo utilizamos interpolación por splines de tercer orden. Elegimos esta comunidad microbiana porque conocemos de antemano su red ecológica: *didinium* es predador de *paramecium*. Este hecho nos permite cuantificar el éxito de nuestro algoritmo para mapear redes ecológicas experimentales. Aplicando el algoritmo de la Ec. (5) con 70 pares de muestras elegidas al azar, obtenemos el histograma de la Fig. 3 B con la frecuencia de ocurrencia de cada patrón de signo. Con base a este histograma y eligiendo ε=0.1, obtenemos $\widehat{\mathfrak{S}}_1^{\varepsilon}$={(+,-),(-,-)} y $\widehat{\mathfrak{S}}_2^{\varepsilon}$={(+,-)} (marcas rosa en Fig. 2 B). La informatividad de los datos es I_{11}^{ε}=0.88/2=0.44 y I_{12}^{ε}=0.88 para la primera especie, y I_{21}^{ε}=I_{22}^{ε}=0.87. Esto implica que el patrón de signos para la segunda especie es S_2=(+,-) con alta confiabilidad (0.87), identificando correctamente a *didinium* como predador (i.e., "+" indica que la especie 1 promueve su crecimiento, y "-" indica que la especie 2 no puede sobrevivir sin la especie 1). Para la primera especie, inferimos s_{12}=-1 con alta confiabilidad (0.88), indicando que la especie 2 inhibe el crecimiento de la especie 1 (e.g., la especie 2 consume a la especie 1). Esto identifica a *paramecium* como la presa. Sin embargo, debido a que I_{11}^{ε}=0.44, los datos no son suficientemente informativos para inferir unívocamente el signo de s_{11}, pues $\widehat{\mathfrak{S}}_{11}^{\varepsilon}$={+,-}. Esta falta de informatividad puede ocurrir porque el signo de esta interacción no es constante, debido a que esta especie exhibe crecimiento logístico. En todo caso, el algoritmo propuesto infiere correctamente la estructura

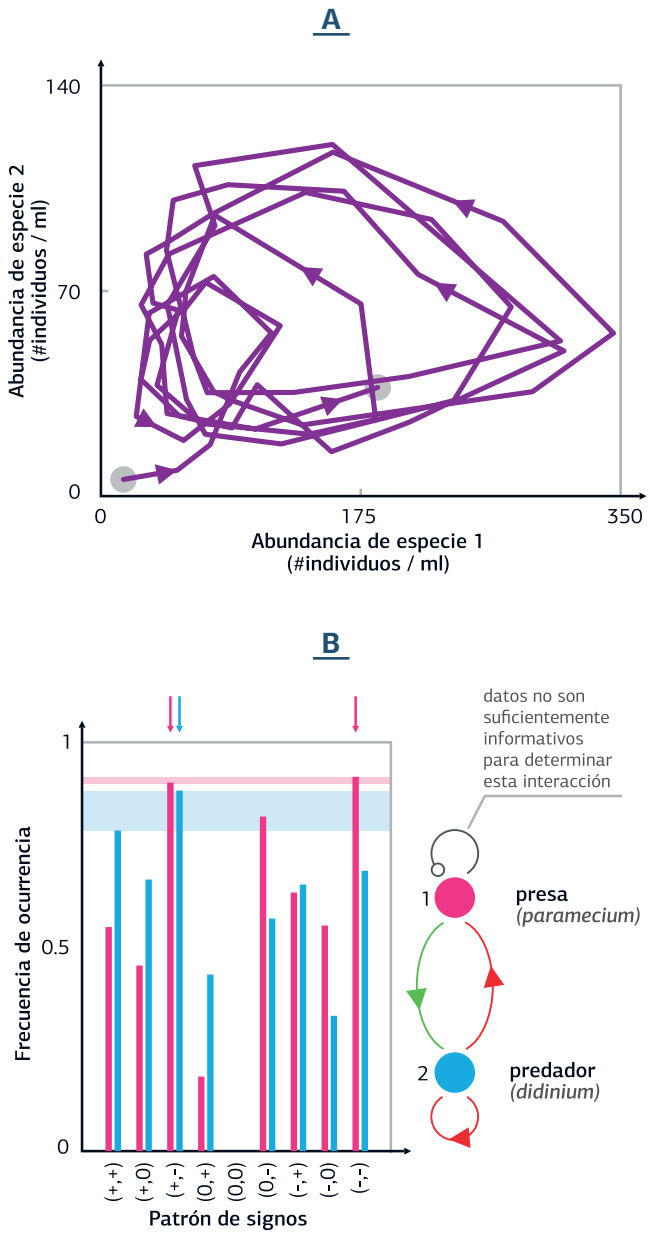


FIGURA 3.
Inferencia usando datos temporales experimentales.
A: Datos temporales experimentales de una comunidad con dos especies *paramecium* (especie 1) y *didinium* (especie 2) creciendo en co-cultivo.
B: Histograma con la frecuencia de ocurrencia de cada patrón de signos obtenido al utilizar el algoritmo de la Ec. (5). Usando ε=0.1 encontramos $\widehat{\mathfrak{S}}_1^{\varepsilon}$={(+,-),(-,-)} para la primera especie, pues dos patrones de signo tienen máxima frecuencia de ocurrencia con (marcas rosa). Para la segunda especie, encontramos $\widehat{\mathfrak{S}}_2^{\varepsilon}$={(+,-)}. Esto permite inferir la red ecológica de la derecha, identificando a *paramecium* como presa y a *didinium* como predador, únicamente a partir de los datos. Con respecto a la interacción intra-especie de la primera especie, note que su signo no puede ser determinado con base a los datos disponibles.

predador-presa de esta comunidad microbiana a partir de datos temporales experimentales.

También aplicamos nuestro método a una comunidad microbiana sintética con ocho especies bacterianas encontradas en el suelo (Friedman, Higgins, & Gore, 2017). Los datos experimentales consisten de muestras en estado estable de la comunidad con distintas composiciones (Fig. 4). Elegimos esta comunidad porque, de nuevo, podemos cuantificar el desempeño de nuestro algoritmo para mapear redes microbianas con datos experimentales. Para esto, primero consideramos 36 muestras que contienen 8 "solos" (abundancia en estado estable de las 8 especies creciendo en monocultivo), y 28 "duos" (abundancia en estado estable de todos los pares de 8 especies creciendo en co-cultivo). Como en este conjunto de datos siempre existen dos muestras que difieren solo en una especie para todas las especies, esto permite inferir fácilmente la red ecológica de la comunidad como discutimos con anterioridad (Fig. 4 A). Comparamos el desempeño de nuestro algoritmo al tratar de inferir esta misma red ecológica usando muestras con composiciones de especies más complejas. Enfatizamos que obtener muestras de solos y dúos es simplemente imposible en muchas comunidades microbianas como el microbiota humano, lo que motiva la construcción del algoritmo presentado en este trabajo. Aplicamos el algoritmo de la Ec. (5) a un conjunto de 65 muestras en estado estable con composición más complejas: tríos, septetos y octetos (Fig. 4B). El algoritmo propuesto infiere correctamente más del 78% de las interacciones ecológicas de la red. De hecho, las interacciones inferidas incorrectamente corresponden a interacciones "débiles" que son muy sensibles a ruido (Xiao, et al., 2017). El algoritmo propuesto tiene un desempeño similar en otras comunidades microbianas, y las redes inferidas pueden predecir la respuesta de comunidades experimentales a alteraciones de su composición (Xiao, et al., 2017).

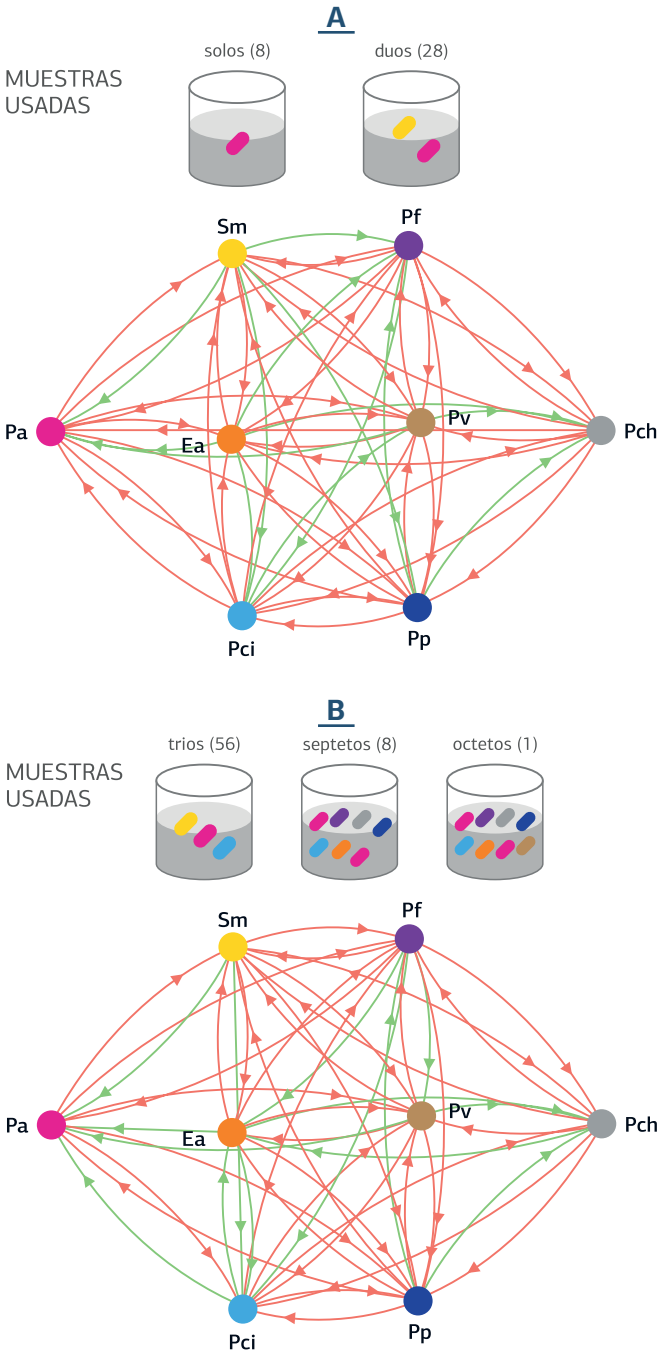


FIGURA 4.
Inferencia usando datos experimentales en estado estable. Consideramos una comunidad sintética con ocho especies bacterianas: Ea (*Enterobacter aerogenes*), Pa (*Pseudomonas aurantiaca*), Pch (*Pseudomonas chlororaphis*), Pci (*Pseudomonas citronellolis*), Pf (*Pseudomonas fluorescenc*), Pp (*Pseudomonas putida*), Pv (*Pseudomonas veronii*), Sm (*Serratia mercescens*).
A: Resultados de la inferencia utilizando muestras de "solos" (cada especie creciendo en monocultivo) y "duos" (todos los pares de especies creciendo en co-cultivo).
B: Inferencia de la red ecológica microbiana utilizando el algoritmo de la Ec. (5) con ε=0.1 y muestras de tríos, septetos y octetos, que tienen composiciones más complejas.

Discusión

Inferir exitosamente la red ecológica de una comunidad microbiana permitirá predecir varios aspectos importantes de su comportamiento, incluyendo su estabilidad (Angulo & Slotine, 2017). También abrirá la puerta para diseñar sistemáticamente estrategias de control para restaurar comunidades microbianas alteradas (Angulo, Moog, & Liu, 2017). En particular,

note que la "universalidad dinámica" encontrada en el microbiota intestinal (Bashan, et al., 2016) hace posible combinar muestras de distintos humanos saludables. Por tanto, aplicando nuestro algoritmo a estas muestras permitiría inferir, por primera vez, la red ecológica "universal" del microbiota intestinal humano. Este hecho ilustra cómo la aplicación y adopción del algoritmo desarrollado en este trabajo podría contribuir significativamente a resolver algunos de los retos más difíciles que enfrenta nuestra sociedad, incluyendo enfermedades humanas complejas. Como aporte a las ciencias básicas y aplicadas, creamos un nuevo formalismo que muestra que matemáticas relativamente simples pueden ayudar a resolver problemas biológicos complejos relacionados a la salud humana.

El algoritmo propuesto puede extenderse al caso cuando distintas muestras corresponden a respuestas de la comunidad con distintos parámetros (e.g., distintos nutrientes), considerando estos parámetros como "nodos" adicionales sin dinámica. También es posible modificar el algoritmo para el caso cuando el patrón de signos no permanece constante y cuando el número de especies es muy grande, de manera completamente análoga a nuestros resultados recientes (Xiao, et al., 2017).

Existe una necesidad nacional de participar en el desarrollo de esta área emergente del control de comunidades microbianas, para poder cosechar sus potenciales frutos en salud y energías renovables. Cosechar estos beneficios requerirá una mayor interacción entre microbiólogos y matemáticos aplicados (teóricos de sistemas). Querétaro cuenta con una infraestructura científica y tecnológica sólida en estas dos disciplinas. Nuestro trabajo contribuye a convertir a este estado en líder nacional en esta área emergente de la ciencia.

RESUMEN CURRICULAR

Marco Tulio Angulo: Obtuvo el grado de Ingeniero en Automatizacion (con Honores) por la UAQ en 2017, y el grado de Doctor en Ingenieria (con Honores) por la UNAM en 2012. Posteriormente, realizo estudios de postdoctorado en el Center for Complex Networks Research (CCNR) en Northeastern University; y en Channing Division of Network Medicine, Harvard Medical School, en Boston, Estados Unidos. En 2016 regresó a Mexico para integrarse al Instituto de Matemáticas de la UNAM, como Cátedra CONACyT, en el Nodo Multidisciplinario de Matemáticas Aplicadas (NoMMA).

REFERENCIAS BIBLIOGRÁFICAS

Alivisatos, A. P., Blaser, M., Brodie, E. L., Chun, M., Dangl, J. L, Donohue, T. J, Jansson, J. K. (2015). A unified initiative to harness Earth's microbiomes. *Science, 350*(6260), 507--508.

Angulo, M. T., & Slotine, J.-J. (2017). Qualitative stability of nonlinear networked systems. *IEEE Transactions on Automatic Control, 62*(8), 4080--4085.

Angulo, M. T., Moog, C. H., & Liu, Y.-Y. (2017). Controlling microbial communities: a theoretical framework. *bioRxiv*, 149765.

Angulo, M. T., Moreno, J. A., Lippner, G., Barabási, A.-L., & Liu, Y.-Y. (2017). Fundamental limitations of network reconstruction from temporal data. *Journal of the Royal Society Interface, 14*(127), 20160966.

Bashan, A., Gibson, T. E., Friedman, J., Carey, V. J., Weiss, S. T., Hohmann, E. L., & Liu, Y.-Y. (2016). Universality of human microbial dynamics. *Nature, 534*(7606), 259.

Bucci, V., Tzen, B., Li, N., Simmons, M., Tanoue, T., Bogart, E., . . . Liu, Q. (2016). MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome biology, 17*(1), 121.

Buffie, C. G., Bucci, V., Stein, R. R., McKenney, P. T., Ling, L., Gobourne, A., ... Viale, A. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature, 517*(7533).

Cao, H.-T., Gibson, T. E., Bashan, A., & Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons. *BioEssays*, 27 (2).

Dietert, R. (2016). The Human Superorganism: How the Microbiome Is Revolutionizing the Pursuit of a Healthy Life.

Dubilier, N., McFall-Ngai, M., & Zhao, L. (2015). Create a global microbiome effort. *Nature, 525*(7575), 631--634.

Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews Microbiology, 10*(8), 538.

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS computational biology , 8*(9), e1002687.

Friedman, J., Higgins, L. M., & Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. *Nature ecology & evolution, 1*(5), 0109.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., ... R, J. (2016). Plankton networks driving carbon export in the oligotrophic ocean *Nature, 532*(7600), 365.

Hudson, L. E., Anderson, S. E., Corbett, A. H., & Lamb, T. J. (2017). Gleaning insights from fecal microbiota transplantation and probiotic studies for the rational design of combination microbial therapies. *Clinical microbiology reviews, 30*(1), 191--231.

Jost, C., & Ellner, S. P. (2000). Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proceedings of the Royal Society of London B: Biological Sciences, 267*(1453), 1611-1620.

Ljung, L. (1998). *System identification.* Springer.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature, 489*(7415), 220.

May, R. M. (1973). Qualitative stability in model ecosystems. *Ecology, 54*(3), 638--641.

Mueller, U., & Sachs, J. L. (2015). Engineering microbiomes to improve plant and animal health. *Trends in microbiology, 23*(10), 606--617.

Shreiner, A. B., Kao, J. Y., & Young, a. V. (2015). The gut microbiome in health and in disease. *Current opinion in gastroenterology, 31*(1), 69.

Steinway, S. N., Biggs, M. B., Loughran Jr, T. P., Papin, J. A., & Albert, R. (2015). Inference of network dynamics and metabolic interactions in the gut microbiome. *PLoS computational biology, 11*(6), e1004338.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature, 449*(7164), 804.

Widder, S., Allen, R. J., Pfeiffer, T., Curtis, T. P., Wiuf, C., Sloan, W. T., ... Shou, W. (2016). Challenges in microbial ecology: building predictive understanding of community function and dynamics. *The ISME journal, 10*(11), 2557.

Xiao, Y., Angulo, M. T., Friedman, J., Waldor, M. K., Weiss, S. T., & Liu, Y.-Y. (2017). Mapping the ecological networks of microbial communities. *Nature Communications, 8*(1), 2042.

@