(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 632 602**

(51) Int. CI.:

G06F 19/18 (2011.01)
G06F 19/26 (2011.01)
C12Q 1/68 (2006.01)
C12Q 1/04 (2006.01)
C12N 1/20 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **28.06.2013 PCT/US2013/048719**

(87) Fecha y número de publicación internacional: **03.01.2014 WO14005094**

(96) Fecha de presentación y número de la solicitud europea: **28.06.2013 E 13759967 (6)**

(97) Fecha y número de publicación de la concesión europea: **17.05.2017 EP 2694669**

(54) Título: **Métodos de fabricación o creación de un consorcio microbiano sintético identificado mediante análisis computacional de secuencias de amplicones**

(30) Prioridad:

28.06.2012 US 201261665656 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.09.2017

(73) Titular/es:

TAXON BIOSCIENCES, INC. (100.0%)
3150 Paradise Drive
Tiburon, CA 94920, US

(72) Inventor/es:

KUNIN,VICTOR;
ASHBY, MATTHEW;
SCHERER, STEWART y
PATIN, NASSTASIA

(74) Agente/Representante:

PONS ARIÑO, Ángel

DESCRIPCIÓN

Métodos de fabricación o creación de un consorcio microbiano sintético identificado mediante análisis computacional de secuencias de amplicones

Campo de la invención

La presente invención se refiere, en general, a la ecología microbiana, a la bioinformática, a la biología computacional y a la microbiología. En realizaciones alternativas, la invención proporciona algoritmos computacionales, programas informáticos, software y otros métodos, sistemas y productos de fabricación (por ejemplo, ordenadores, dispositivos o aparatos) para identificar miembros de comunidades microbianas, su abundancia y distribución a partir de datos de secuencia de amplicones y comparar comunidades microbianas y consorcios microbianos. En realizaciones alternativas, la invención proporciona métodos de identificación de consorcios, seguidos opcionalmente de la construcción de consorcios microbianos artificiales a partir de cepas puras o de cultivos de enriquecimiento.

Antecedentes

Una de las tareas más importantes en la ecología microbiana es la identificación de los miembros de una comunidad y su abundancia relativa. Actualmente, las comunidades microbianas se caracterizan por secuenciación de moléculas de ARNr 16S amplificadas por PCR y el análisis informático de las secuencias. Existen varias líneas informáticas para efectuar dichos análisis, y consisten, en líneas generales, en la eliminación de las lecturas de baja calidad, el agrupamiento (la tarea de asignar un conjunto de objetos o secuencias a grupos o agrupaciones) y la clasificación de los representantes del grupo.

La etapa de agrupamiento se considera esencial. Esto se debe a que cada secuencia se considera como un representante de una célula, y, sin embargo, el proceso de secuenciación es inherentemente erróneo, y las secuencias con errores se pueden interpretar como nuevos organismos. En forma acumulativa, se sabe que estos errores inflan las estimaciones de la riqueza de la comunidad. Por lo tanto, el agrupamiento al 97 % de identidad es actualmente la práctica común en el campo.

Sin embargo, el agrupamiento tiene muchos inconvenientes intrínsecos. Además de los errores de secuenciación, el agrupamiento concentra la diversidad microbiana genuina. La secuencia más común del grupo se usa como representativa, mientras que las otras secuencias del agrupamiento se pierden. Además, los agrupamientos son intrínsecamente sensibles a los datos de entrada, no son estables con el tiempo y cambian cada vez que se agregan nuevos datos. Por lo tanto, un análisis realizado con N muestras debe reagruparse y, por lo tanto, volverse a realizar cuando se añade la muestra N + 1. Dado que la adición de muestras es una operación frecuente, evitar el agrupamiento podría ahorrar tiempo tanto al investigador como a los ordenadores.

El análisis de la composición de la comunidad solo es una etapa para interrogar a una comunidad microbiana. El aislamiento de cultivos es otra técnica valiosa. Los aislados pueden secuenciarse para identificar la secuencia correspondiente a la región indicadora (amplicón). Sin embargo, la naturaleza fluida de los grupos significa que el agrupamiento al que se asigna el aislamiento puede cambiar con respecto al número de miembros y la distribución entre las muestras, incluso para los conjuntos de datos previamente analizados.

Otro inconveniente es que la práctica actual de asignar la clasificación taxonómica a secuencias representativas del grupo requiere la reclasificación después de cada agrupamiento. Se trata de un procedimiento posiblemente costoso desde el punto de vista informático. Además, dado que las secuencias representativas de los grupos cambian, la taxonomía puede no coincidir, haciendo que el análisis de los datos sea aún más desconcertante.

La tecnología de secuenciación cambia cada pocos meses. Los cambios reducen principalmente el coste de la secuencia por base o aumentan la longitud de la lectura. A medida que la tecnología cambia, los nuevos datos de amplicones más largos no se pueden comparar directamente con los datos heredados. Las soluciones actuales incluyen bien el uso de datos antiguos de menor resolución o la resecuenciación de muestras antiguas. Cualquier solución tiene problemas: la primera solución descarta la mayor resolución que puede proporcionar la nueva tecnología, mientras que la segunda requiere un gran esfuerzo de recogida de muestras, que puede no estar disponible para muestras anteriores.

La identificación de los miembros de las comunidades microbianas es una etapa importante hacia la identificación de consorcios microbianos. Los consorcios microbianos realizan muchas tareas importantes en la naturaleza, en concreto, la biodegradación de compuestos complejos. Estos consorcios normalmente se estudian de una manera específica, cuando se selecciona una tarea en mano para la interrogación, se identifican los organismos de interés y se estudia la interacción. Esta estrategia de caso por caso permite una comprensión profunda de algunos consorcios, pero no presenta una visión general de la variedad de consorcios que hay en la naturaleza.

El documento WO 2011/159924 desvela composiciones microbianas, por ejemplo, consorcios, que se optimizan específicamente tanto para estimular la metanogénesis como para conversiones “metilotróficas” u otras conversiones. En realizaciones alternativas, el documento WO 2011/159924 desvela métodos de desarrollo de cambios de nutrientes y composiciones microbianas ambos optimizados específicamente para estimular la metanogénesis de un reservorio dado. El documento WO 2011/159924 también proporciona métodos de evaluación de la formación de biomasa potencialmente dañina y la precipitación a escala producida como consecuencia de la adición de cambios de nutrientes. En otras realizaciones, el documento WO 2011/159924 desvela métodos de simulación de biogás en condiciones subterráneas usando un modelo computacional.

Sumario

En realizaciones alternativas, la presente invención proporciona métodos de identificación de consorcios microbianos o de un grupo de microbios con distribuciones medioambientales correlacionadas de acuerdo con la reivindicación 1.

Los detalles de una o más realizaciones de la invención se exponen en los dibujos adjuntos y en la descripción que figura más adelante. Otras características, objetos y ventajas de la invención serán evidentes a la luz de la descripción y de los dibujos, y de las reivindicaciones.

Breve descripción de los dibujos

Los siguientes dibujos son ilustrativos de los aspectos de la invención, y no pretenden limitar el alcance de la invención englobado por las reivindicaciones.

La Figura 1 ilustra esquemáticamente un esquema de ejemplo del núcleo de una base de datos usado para la práctica de la invención. Los recuadros representan los principales tipos de datos, las líneas representan las conexiones entre ellos. Cada bloque representa un grupo de tablas, hojas de cálculo o archivos. Los datos para los marcadores, el procedimiento de secuenciación, las muestras y los metadatos, aislados, taxonomías y clasificación están interrelacionados.

La Figura 2 ilustra una visualización de una red concurrente microbiana, según lo descrito detalladamente más adelante.

La Figura 3 ilustra un método, o procedimiento, de ejemplo de la invención, que ilustra esquemáticamente un procedimiento de identificación de la composición de una comunidad de secuencias de amplicones o "lecturas", según lo descrito en el apartado de Procedimientos más adelante.

La Figura 4 ilustra un método, o procedimiento, de ejemplo de la invención, que ilustra esquemáticamente un procedimiento de identificación y creación de consorcios sintéticos. Las composiciones de las comunidades de múltiples muestras son la entrada. A partir de la distribución de los marcadores en las muestras se calcula una matriz de distancia para los marcadores. Los grupos de microbios con distribución relacionada se identificaron manualmente o por agrupamiento, y se denominaron consorcios. Los microbios correspondientes se identifican en colecciones de cultivos o cultivos de enriquecimiento y se combinan sintéticamente para formar un consorcio sintético.

Los símbolos de referencia similares de los diversos dibujos indican elementos similares.

Descripción detallada

En aspectos alternativos, la divulgación proporciona métodos, sistemas y productos de fabricación (por ejemplo, ordenadores, dispositivos o aparatos) para identificar y comparar miembros de comunidades microbianas, consorcios microbianos o grupos de microbios con distribuciones medioambientales correlacionadas, a partir de datos de secuencia de amplicones, e identificar miembros de una comunicación microbiana, un consorcio microbiano o un grupo de microbios con distribuciones medioambientales correlacionadas, a partir del análisis de las secuencias de amplicones. En aspectos alternativos, la divulgación proporciona algoritmos computacionales, programas informáticos y otros métodos, sistemas y productos de fabricación (por ejemplo, ordenadores, dispositivos o aparatos) para identificar miembros de comunidades microbianas, su abundancia y distribución a partir de datos de secuencia de amplicones y comparar comunidades microbianas y consorcios microbianos.

En aspectos alternativos, la divulgación proporciona algoritmos computacionales, programas informáticos, software y otros métodos, sistemas y productos de fabricación (por ejemplo, ordenadores, dispositivos o aparatos) para identificar miembros de una comunidad microbiana, un consorcio microbiano o un grupo de microbios con distribuciones medioambientales correlacionadas, su abundancia y distribución a partir de datos de secuencia de amplicones y comparar comunidades microbianas y consorcios microbianos. En aspectos alternativos, la divulgación usa lecturas truncadas únicas (denominadas "marcadores") como representantes de organismos. Los marcadores únicos y sus apariciones en las muestras se almacenan en una base de datos. La base de datos también puede vincularse a otros tipos de datos, tales como la clasificación de marcadores, la presencia en colecciones de cultivos, etc.

En aspectos alternativos, los errores de secuenciación se evidencian mediante: (i) el recorte de las lecturas en una región predefinida (véase la Etapa 2 del Procedimiento); ii) la eliminación de las lecturas truncadas de baja calidad; e iii) la fijación de un umbral para que la abundancia mínima de los marcadores aparezca en el análisis.

Evitando el agrupamiento, la presente invención permite la consistencia del recuento de los miembros cuando se añaden o se eliminan muestras. Las diferencias en la distribución de la abundancia de los marcadores muy similares entre las muestras se pueden usar para identificar organismos ecológicamente distintos.

En aspectos alternativos, los métodos o métodos implementados por ordenador de la divulgación comprenden:

Base de datos:

Se mantiene una base de datos que contiene datos de experimentos previos, por ejemplo, como se ilustra en la Figura 1. La base de datos debe contener marcadores únicos. Los marcadores son secuencias de nucleótidos únicas recortadas hasta la región deseada como se describe en la etapa 1 del Procedimiento. La base de datos debe contener registros de la aparición de marcadores entre muestras analizadas previamente o conjuntos de datos. La base de datos puede contener clasificaciones taxonómicas de marcadores, enlaces a aislamientos o colecciones de cultivos que contienen secuencias de marcadores en sus genomas o cualquier otro dato asociado con marcadores.

Entrada:

1) Salida del secuenciador, que comprende o contiene lecturas y puntuaciones de calidad asociadas para un gen amplificado (por ejemplo, un gen amplificado por PCR) que, opcionalmente, comprende o es un ADNr 16S, 18S, 23S o 28S (ADN que codifica ARN ribosómico).
2) Una lista de muestras y “códigos de barras” asociados. Los códigos de barras son secuencias oligonucleotídicas cortas de identificación de muestras incluidas dentro de las secuencias de cebadores que permiten la multiplexación dentro de una sola serie.

Procedimiento

1) Se identifican las secuencias de nucleótidos que contienen los códigos de barras mencionados en el archivo de entrada y se registra su correspondencia con las muestras, y se eliminan los códigos de barras. Las secuencias que no contienen códigos de barras correctos se descartan.

2) Las lecturas se cortan para mantener solo regiones predefinidas. Las secuencias resultantes se denominan marcadores. En realizaciones alternativas, el corte se puede realizar reconociendo patrones dentro de la lectura o recortando lecturas a la longitud deseada. Por ejemplo, el recorte de longitud a la longitud 250 mantendrá las bases 1 a 250 y descartará otras bases. Los patrones se pueden reconocer, por ejemplo, para cualquier secuencia superior a 120 pares de bases (pb) que reconozca los patrones conservados GGTAGTC (SEQ ID NO: 1) en 5' de la secuencia y AATTGNCGGGG (SEQ ID NO: 2) en 3' de la secuencia, lo que permite menos de 2 desapareamientos, y el marcador resultante debe ser de entre 90 y 200 pares de bases de longitud. Las lecturas que son más cortas que la longitud definida o no coinciden con el patrón se descartan. Se pueden aplicar múltiples reglas o cualquier combinación de las reglas para recortar secuencias.

3) Se eliminan las lecturas truncadas de baja calidad. Las lecturas truncadas de alta calidad pueden identificarse como lecturas truncadas en las que al menos el X % de las bases tiene al menos puntuación Q de Y. Se pueden implementar múltiples reglas. Por ejemplo, solo se mantienen truncadas lecturas que tengan una puntuación Q de 20/25 para el 100/90 por ciento de las bases, respectivamente, y eliminarse las lecturas truncadas con nucleótidos ambiguos (tales como “N”). Las lecturas truncadas por debajo de este umbral se consideran de baja calidad y se eliminan. Pueden emplearse otros procedimientos para el control de la calidad, por ejemplo, los procedimientos descritos en Sogin (2006) *Proc. Natl. Acad. Sci*. EE.UU., vol. 103(32): 12115-12120; P. H. Victor Kunin, "PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data". *The Open Journal*, vol. 1, n.º 1.

4) Se introducen nuevos marcadores (que no están presentes en la versión anterior de la base de datos) y los recuentos de marcadores de cada muestra se introducen en la base de datos. Los recuentos de marcadores pueden ser recuentos reales (tales como 5) o fracciones del total de marcadores de la muestra (tal como el 0,5 % del total).

5) Se clasifican taxonómicamente nuevos marcadores. Existen múltiples procedimientos alternativos para la clasificación taxonómica que pueden usarse para poner en práctica la invención, como se describe, por ejemplo, en: Wu, *et al*. (2008) “An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP)”. PLoS ONE 3(7):e2566.doi:10.1371/journal.pone.0002566, y las referencias citadas en el mismo. Un procedimiento ilustrativo usado para poner en práctica la invención es una transferencia de la anotación de la secuencia más cercana de una base de datos pública, por ejemplo, tal como una base de datos

de Genbank.

6) En esta etapa, la base de datos contiene los recuentos de miembros de la comunidad reflejados por la abundancia de marcadores y la taxonomía de cada marcador. Esta información se debe presentar al ser humano en un formato de lectura o analizado además computacionalmente. Otros análisis pueden incluir la distribución en marcadores entre muestras; la identificación de consorcios microbianos entre muestras; la identificación de sesgos de procedimiento, etc. El formato de lectura para seres humanos puede comprender un gráfico, una representación o una tabla u otra forma de presentación. Por ejemplo, el formato de lectura para seres humanos puede ser una tabla en la que las muestras sean columnas, los marcadores sean filas y las celdas reflejen los recuentos del marcador en la muestra. Se establece un umbral para que la secuencia aparezca en el análisis. Por ejemplo, se puede requerir que las secuencias sean abundantes al menos un X % en al menos Y conjuntos de datos; por ejemplo, al menos un 1 % en al menos 1 conjunto de datos. Este umbral (denominado X anteriormente) puede ser variado por la solicitud, siendo aproximadamente el 1 % el más útil. Este procedimiento permite analizar los errores raros y los miembros irrelevantes de la comunidad, manteniéndose a la vez la mayoría de los miembros importantes de la comunidad.

En aspectos alternativos, los métodos usan la normalización para el tamaño de los datos y la representación de las abundancias microbianas como fracción del recuento microbiano total en una muestra. En realizaciones alternativas, la representación como fracciones puede ser necesaria a medida que los tamaños de la muestra varían y los conjuntos de datos no se pueden comparar fácilmente sin una normalización.

En aspectos alternativos, los datos de abundancia se transforman mediante una transformación logarítmica. Esta transformación puede ser necesaria, ya que existe un gran intervalo natural de los datos, y porque los recuentos de abundancia suelen ser reproducibles solo al nivel del orden de magnitud. Por lo tanto, la transformación logarítmica permite corregir las inexactitudes de las metodologías de secuenciación. Dado que la transformada logarítmica no es posible para valores 0, estos se pueden sustituir con valores arbitrarios muy pequeños, por ejemplo, al 0,01 %, o por debajo del 1 % de abundancia o un recuento absoluto de 1.

En aspectos alternativos, las etapas 1, 2 y 3 se pueden realizar en cualquier orden o en paralelo. En realizaciones alternativas, las etapas 1 y 5 no son esenciales, por ejemplo, en una realización, un protocolo de la invención comprende las etapas 2, 3, 4 y 6. Las etapas 4 y 5 se pueden realizar en paralelo o realizarse en cualquier orden.

En aspectos alternativos, se usan procedimientos similares cuando las lecturas son de extremo par (solapadas o no solapadas), usándose el ensamblaje de lecturas emparejadas cuando se requiera, tal como la secuenciación con una plataforma de secuenciación de ácido nucleico Illumina, por ejemplo, GENOME ANALYZER IIX™ o HISEQ SYSTEM™ (Illumina, San Diego, CA), o equivalente). En este procedimiento se usará un par de lectura (ensamblado o no) como una “lectura” en el procedimiento descrito anteriormente.

Resultado:

El resultado es una descripción de las comunidades microbianas como recuentos de la abundancia de los miembros únicos de cada comunidad. Se proporciona en una forma de visualización (tabla de abundancias), y se almacena como base de datos o recogida de archivos que describen marcadores únicos, su distribución entre las muestras como lo evidencia la secuenciación.

Tratamiento de los errores de secuenciación

Una característica distintiva de la invención es la representación de los miembros de la comunidad con marcadores únicos en lugar de grupos. Dicha representación permite el uso de la base de datos y no requiere el agrupamiento. En aspectos alternativos, se considera que la función más importante del agrupamiento es la absorción de los errores de secuenciación en una sola secuencia OTU representativa que representa todas las secuencias dentro de una métrica de distancia definida (por ejemplo, el 0,03 por ciento). La presente invención rechaza la visión actual de que, sin agrupamiento, el ruido oscurecerá la señal y las conclusiones del análisis de la comunidad serán incorrectas.

La presente invención comprende el uso de las siguientes salvaguardias para limitar la influencia de los errores de secuenciación:

1) Las lecturas se recortan. A medida que las lecturas tienden a acumular más errores cerca de (uno o ambos) de los extremos, esta etapa reduce los posibles errores (etapa 2 del Procedimiento).
2) Las lecturas truncadas con un alto número de bases de baja calidad se eliminan del conjunto de datos (etapa 3 del Procedimiento).
3) Los errores de secuenciación que pasan por el filtro descrito en la etapa (2) se pueden dividir en raros y sistemáticos. Los errores raros del proceso de secuenciación dan lugar a marcadores que tienen una aparición insignificante, y se eliminan mediante el método descrito en el Procedimiento (6). Se considera que los marcadores con errores no eliminados por el procedimiento (6) tienen errores sistemáticos. Se espera que los

marcadores con errores sistemáticos correlacionen la distribución con su secuencia “madre” o “correcta”, teniendo el marcador correcto mucha mayor abundancia y alta similitud (identidad de secuencia) con el marcador con error. Estos rasgos de errores sistemáticos permiten la corrección en el análisis de datos y el procesamiento posterior.

Consistencia de los datos

Antes de la presente invención, cada vez que se añadía una muestra a un grupo previamente analizado de muestras, se debía volver a realizar el análisis (incluyendo el agrupamiento y la clasificación taxonómica). Por otra parte, debido a la naturaleza del agrupamiento, los grupos son de naturaleza fluida, y pueden agregar o arrojar marcadores únicos. A medida que los marcadores se añaden o se eliminan de los grupos, los recuentos de las abundancias de los grupos por muestra pueden cambiar. En realizaciones alternativas, la presente invención, mediante el uso de marcadores únicos, garantiza la consistencia de los recuentos de marcadores cuando se añaden o se eliminan muestras.

Debido a la naturaleza fluida de los grupos, antes de la presente invención, la unión entre tipos de datos era un procedimiento difícil. En realizaciones alternativas, la presente invención permite construir una base de datos con enlaces entre tipos de datos. Por ejemplo, los marcadores pueden unirse a clasificaciones taxonómicas, su distribución entre las muestras, aislados disponibles en las colecciones de cepas y otros datos con mayor simplicidad. Esta capacidad potencia la capacidad del operador para rastrear los datos.

La tecnología de secuenciación cambia cada pocos meses. Los cambios principalmente reducen el coste de la secuencia por base y/o aumentan la longitud de lectura. A medida que la tecnología cambia, los nuevos datos de amplicones más largos no se pueden comparar directamente con los datos heredados. Antes de la invención, una solución era descartar las ventajas proporcionadas por una mayor resolución de la tecnología más reciente y usar las mismas regiones que con la tecnología anterior para la consistencia. Como alternativa, se requería la resecuenciación de muestras antiguas, lo que requería un gran esfuerzo de recogida de muestras que podía no estar disponible para muestras más antiguas.

En aspectos alternativos, la base de datos puede contener marcadores obtenidos con diversas tecnologías de secuenciación, cubriendo diferentes regiones de amplicón o diferentes zonas de la misma región de amplicón. Pueden vincularse diversos marcadores identificadores del mismo grupo de organismos, permitiendo la comparación de los marcadores obtenidos con diversas tecnologías de amplificación, secuenciación o procesamiento de datos. Por ejemplo, si la tecnología anterior permitía la secuenciación de la región v5 de la molécula de ARNr 16S, la nueva tecnología permitía la secuenciación de ambas regiones v5 y v6. Por lo tanto, los marcadores obtenidos con la tecnología solo v5 pueden identificarse como incluidos en los marcadores de las regiones v5 y v6. Si hay otra secuencia disponible con la región solo v6, las secuencias v5 y v6 pueden unirse a través de una secuencia que contenga ambas regiones. Para concluir, la invención tiene la capacidad de permitir el análisis de los datos obtenidos con tecnología diferente.

Identificación de la divergencia ecológica de cepas muy similares

Antes de la presente invención, los marcadores muy similares (por ejemplo, marcadores con alta similitud de secuencia) estaban representados por un solo grupo. Cuando estos marcadores representaban organismos ecológicamente diferentes, se perdía la distinción de su distribución ecológica, y solo se informaba de la distribución combinada de todos los miembros del grupo. En realizaciones alternativas, la presente invención, rastreando cada marcador, permite la comparación de distribuciones ecológicas de secuencias altamente relacionadas.

La capacidad de la invención para rastrear marcadores muy similares a través de entornos proporciona una herramienta para distinguir entre errores de secuenciación y organismos genuinamente distintos. Los marcadores altamente similares pueden representar cepas altamente relacionadas o ser variantes dentro del mismo genoma. Las distribuciones de marcadores similares entre muestras (alta correlación) pueden ser el resultado de las variantes dentro de un genoma, las cepas relacionadas con una distribución ecológica similar o error de secuenciación. Por el contrario, las distribuciones de marcadores altamente divergentes entre muestras (como lo indica la baja correlación), solo pueden derivarse de cepas relacionadas con una distribución ecológica distinta. Por lo tanto, en realizaciones alternativas, la distribución de secuencias entre muestras se usa para identificar variantes ecológicamente relevantes de organismos similares. La "correlación" es cualquier forma de cálculo que identifica similitud o distancia, y puede ser opcionalmente la distancia Euclidiana, la correlación de Pearson, las distancias vectoriales, el Chi cuadrado, la distancia de Manhattan, métodos de ordenación que comprenden opcionalmente el uso de PCA, la disimilitud de Bray-Curtis y el escalamiento multidimensional no métrico (NMS o NMDS).

Los marcadores que tienen alta similitud de secuencia pero no se correlacionan de manera significativa se pueden identificar como representantes de distintos organismos con una distribución medioambiental distinta. Sin embargo, en algunos casos, los marcadores que representan organismos distintos pueden tener distribuciones correlacionadas. Estos se pueden identificar usando un procedimiento ligeramente diferente. Uno de los marcadores se puede designar como una referencia, y usarse su distribución para predecir la distribución de un marcador

correlacionado. Se puede usar una desviación significativa de la aparición esperada del marcador correlacionado en una o más muestras como una indicación de que el marcador representa un organismo distinto. La significación de la desviación se establecerá dependiendo del método preciso seleccionado, y hay un gran volumen de literatura que describe la evaluación de la importancia de las predicciones en función de la metodología usada.

En realizaciones alternativas, la presente invención permite utilizar información adicional procedente de marcadores más largos. Antes de la invención, los marcadores se agruparon. Por ejemplo, los marcadores de 100 pares de bases (pb) de longitud que tenían 1 diferencia se incluyeron en el 99 % de los grupos, así como los marcadores de 200 pb que tenían 2 diferencias. En realizaciones alternativas, la presente invención usa marcadores únicos, permitiendo así diferenciar entre variantes más ecológicas a medida que aumenta la longitud de secuencia.

La Figura 2 ilustra una visualización de la red de aparición simultánea de microbios. Esta figura se obtuvo de la siguiente manera: se obtuvieron recuentos de las abundancias microbianas a partir de una base de datos que contiene más de mil muestras únicas y más de 2,74 millones de marcadores únicos de ARNr 16S. Solo se seleccionaron muestras naturales no cultivadas para este análisis. Solo se seleccionaron marcadores con abundancia acumulada de más de 100 en todas las muestras. Se sometieron las abundancias a transformación logarítmica. Se compararon las abundancias de todos los microbios que aparecieron simultáneamente en más de 5 muestras, se descartaron los microbios que no aparecieron simultáneamente. Solo se consideraron como presencia en la muestra los recuentos de 10 o más, descartándose una abundancia menor. Se consideraron los marcadores que tenían una correlación de Pearson superior a 0,3. La matriz de correlación se agrupó con el programa MCL usando el valor de inflación de 1,1. Los resultados se representaron en una red en la que los nodos son marcadores y los bordes son correlaciones. La visualización se realiza con un programa BIOLAYOUT™ (Biotechnology and Biological Sciences Research Council (BBSRC), Swindon, RU). Debido a las limitaciones de espacio, solo se muestra una fracción de toda la red. Los consorcios microbianos se identifican fácilmente como grupos de nodos estrechamente conectados en el gráfico.

Identificación de los consorcios microbianos

En realizaciones alternativas, la invención proporciona métodos de identificación de consorcios microbianos o grupos de microbios con distribuciones medioambientales correlacionadas. Los consorcios microbianos realizan muchas tareas importantes en la naturaleza, en concreto, la biodegradación de compuestos complejos. Estos consorcios normalmente se estudian de una manera específica, cuando se selecciona una tarea en mano para la interrogación, se identifican los organismos de interés y se estudia la interacción. Esta estrategia de caso por caso permite una comprensión profunda de algunos consorcios, pero no presenta una visión general de la variedad de consorcios que hay en la naturaleza.

La identificación de consorcios es un problema inverso a la identificación de organismos con secuencias similares, pero con diversas distribuciones ecológicas. En cambio, los consorcios suelen tener organismos evolutivamente divergentes, con diversas secuencias que tienen una distribución medioambiental muy similar. Esos organismos estarían interactuando en la naturaleza, y esas interacciones pueden estar en forma de consorcios u otra forma de coexistencia.

Los microbios pueden tener diversos tipos de interacciones. Por ejemplo, dos microbios pueden depender enteramente uno del otro. Estos microbios también aparecerían siempre simultáneamente en las mismas muestras. Cabría esperar una fuerte correlación lineal entre las abundancias de estos dos microbios. Sin embargo, este tipo de interacción se informa con relativamente poca frecuencia en la literatura, y se espera que solo sea una fracción de todas las interacciones microbianas. El análisis previo de las secuencias presentadas simultáneamente al Proyecto de Bases de Datos Ribosómicas identificó un grupo limitado de organismos que se espera que se correlacionen de esta manera.

En la mayoría de los casos, los microbios no serían absolutamente dependientes entre sí, sino que formarían una interacción transitoria para participar en un consorcio que realizara alguna función. Estos microbios no aparecerían siempre juntos en la naturaleza. Sin embargo, si participan en algún proceso, para las muestras en las que tenga lugar el proceso, pueden estar correlacionados. Por lo tanto, se espera una correlación (más) débil para los formadores de consorcios transitorios.

Otra forma de correlación es la anticorrelación. Es decir, en el mismo tipo de muestra, la abundancia de un organismo se reduce cuando otro organismo está presente. Esta interacción se puede observar cuando los organismos se excluyen entre sí debido a la competitividad, o porque las condiciones medioambientales que favorecen a un organismo y suprimen al otro.

En realizaciones alternativas, la elección de las muestras para el análisis de correlación puede ser flexible. En una realización, una opción consiste en seleccionar todas las muestras de la base de datos. Otra opción es la de limitar las muestras por alguna característica. Esas características podrían ser el proyecto, el tipo de muestra o la fuente, seleccionando únicamente muestras sin cultivar o cultivadas, o cualquier combinación de las mismas. Opcionalmente, se pueden excluir las muestras que sean demasiado similares para ser informativas.

En realizaciones alternativas, para la identificación de consorcios, se requiere un conjunto de datos o una base de datos de abundancias microbianas en muestras. Esta base de datos puede construirse usando un método de la invención, por ejemplo, usando un ARNr, o mediante secuenciación génica, o mediante cualquier otro método conocido en la técnica. Los recuentos de la abundancia de cada secuencia de la base de datos se comparan con los recuentos de la abundancia de otras secuencias de la selección escogida de muestras. En realizaciones alternativas, las secuencias para las que se comparan los recuentos de la abundancia son todas las secuencias de la base de datos, o comprenden una subsección de secuencias, por ejemplo, incluyendo solo secuencias abundantes o secuencias solo significativamente diferentes o cualquier otro subconjunto de secuencias de la base de datos. Las abundancias microbianas de la base de datos pueden representarse como recuentos absolutos o como una fracción del total. Las abundancias pueden además someterse a transformación logarítmica, por ejemplo, para acomodar mejor el intervalo de datos y/o corregir las imprecisiones cuantitativas.

En realizaciones alternativas, se comparan las similitudes de las distribuciones de cada dos secuencias usando métricas de distancia. Estas métricas de distancia pueden incluir cualquiera o cualquier combinación de distancia Euclidiana, la correlación de Pearson, distancias vectoriales, Chi cuadrado, distancia de Manhattan o métodos de ordenación que comprenden opcionalmente el uso del Análisis de Componentes Principales (PCA), la ordenación de Bray-Curtis o la disimilitud de Bray-Curtis, y el escalamiento multidimensional no métrico (NMS o NMDS). La parte más importante es que estas métricas producen un valor numérico de distancia o similitud entre los dos microbios o marcadores. Se puede usar un umbral de distancia o similitud apropiado para designar marcadores similares, cuya similitud debe ser superior a 0.

En realizaciones alternativas, las similitudes entre marcadores se almacenan en una estructura de datos matricial en un ordenador, en forma de archivo, base de datos, en memoria de ordenador, o en un disco o una unidad.

En realizaciones alternativas, la matriz de similitud se visualiza como una red. En esta red, cada nodo puede ser un microbio o un marcador, y cada borde es una similitud entre ellos. En realizaciones alternativas, esta red ya presentaría consorcios en forma de componentes conectados en el gráfico. Las superficies de un gráfico que comparten más conexiones son microorganismos que ocurren simultáneamente, que se pueden identificar como consorcios.

En realizaciones alternativas, la red puede ser demasiado grande e incómoda de visualizar y analizar. Esto es así, en particular, en ausencia de agrupamiento basado en secuencias. Esta forma de agrupamiento absorbe secuencias similares, reduciendo el tamaño de los datos disponibles para el examen. En ausencia de agrupamiento basado en secuencias, la resolución es mayor, lo que coincide con el aumento de tamaño de los datos. El número de componentes puede llegar a ser demasiado grande para realizar el examen fácilmente, y el número de posibles pares de comparaciones crece como el cuadrado del número de microbios (o marcadores) examinados. Por lo tanto, en dicha realización, se desea la reducción de los datos hasta un tamaño manejable.

En realizaciones alternativas, para facilitar el análisis, se usa el agrupamiento basado en la distribución medioambiental (a diferencia del agrupamiento basado en secuencias). El agrupamiento es una técnica computacional de identificación del agrupamiento de un conjunto de objetos de manera que los objetos del mismo grupo (denominados grupo) son más similares (en algún sentido u otro) a los demás de los de otros grupos. En la presente solicitud, los objetos son microbios o marcadores, y la similitud se basa en la distribución medioambiental. Hay muchos algoritmos en la técnica que realizan el agrupamiento. En realizaciones alternativas, la invención comprende el uso del agrupamiento jerárquico, la identificación de componentes conectados, el agrupamiento basado en la conectividad, el agrupamiento basado en la distribución, el agrupamiento basado en la densidad, el agrupamiento de un solo enlace, el agrupamiento de Marcov (MCL, *Marcov clustering*) y/o el agrupamiento de centroides entre otros. En realizaciones alternativas, este agrupamiento identifica grupos de microbios con distribuciones ambientales similares. Los microbios que componen estos grupos pueden interpretarse como formadores de consorcios.

En realizaciones alternativas, los métodos de la invención permiten una reducción de la complejidad de las interacciones que un investigador examina desde todos los microbios de todas las muestras ensayadas hasta los microbios que son miembros de un solo grupo o un grupo de grupos basado en la distribución. En realizaciones alternativas, el papel de dichos microbios dentro de un supuesto consorcio se podría ensayar posteriormente en un laboratorio, en el que dichos microbios se pueden ensamblar artificialmente a partir de una colección de cultivos. Esta comunidad ensamblada se puede ensayar entonces para realizar una función del supuesto consorcio, e identificarse un grupo de organismos necesario para esta tarea.

Construcción de consorcios sintéticos

Los microorganismos rara vez ocurren en la naturaleza como una sola especie. En la gran mayoría de los casos, interactúan con otros microorganismos que están presentes en el mismo ambiente, es decir, tienen la misma distribución medioambiental. Algunas de esas interacciones son de competitividad, mientras que otras interacciones implican cooperación; o una interacción puede implicar la competitividad en un aspecto y la cooperación en otro. La presente invención define organismos que tienden a coexistir como un consorcio y describe un procedimiento de

identificación de miembros de consorcios y construcción de consorcios sintéticos.

Los ejemplos de la función de interés de un consorcio incluyen, por ejemplo, sintetizar o degradar un compuesto de interés (por ejemplo, una metanogénesis utilizando metanol o una conversión “metilotrófica”); mantener la salud de un organismo hospedador, por ejemplo, un ser humano, o causar la enfermedad de un hospedador; formar una interacción mutuamente beneficiosa con una planta, un hongo o un animal; la prevención de enfermedades; la conservación y/o fermentación de productos alimenticios (por ejemplo, como ingrediente de un probiótico), la mejora de las cualidades del agua o del suelo; la biodegradación de contaminantes y la descontaminación; etc.

La identificación bioinformática de los consorcios anteriores puede usarse para ensamblar un consorcio sintético. La mayoría de los laboratorios de microbiología tienen acceso a colecciones de cepas en las que las cepas puras se mantienen o crecen aisladas de otros microbios. Esas cepas se pueden identificar por la secuenciación de ADNr 16S, la secuenciación del genoma u otros métodos (tales como métodos fenotípicos, recepción de cepas de una fuente de confianza, etc.). A veces, las cepas no se pueden purificar hasta un estado axénico deseado y se mantienen como cultivos de enriquecimiento.

En aspectos alternativos, los consorcios identificados mediante métodos bioinformáticos descritos en la presente invención pueden construirse sintéticamente. Las cepas correspondientes o los cultivos de enriquecimiento pueden referenciarse de forma cruzada. Esta referencia cruzada puede realizarse mediante secuenciación de ADNr 16S (cuando el cultivo tiene una secuencia idéntica al marcador) o mediante clasificación al mismo grupo taxonómico (tal como las especies). Cuando no se puede encontrar una coincidencia exacta entre un microbio identificado por los métodos bioinformáticos descritos en la presente invención dentro de la colección de cultivos, se puede sustituir con un organismo estrechamente relacionado. El organismo estrechamente relacionado puede ser un organismo de la misma especie o del mismo género, o tener un 95 % o más de identidad de secuencia de marcador o ARNr.

Los microbios identificados como presentes en el consorcio se pueden mezclar artificialmente en un laboratorio, creando un consorcio sintético. El consorcio sintético se puede ensayar entonces para realizar la tarea deseada de interés, tal como sintética o degradante de un compuesto de interés.

Distribuciones medioambientales

En aspectos alternativos, la divulgación comprende métodos de identificación y/o fabricación de un consorcio microbiano o un grupo de microbios que tenga una distribución medioambiental correlacionada, y consorcios microbianos o un grupo de microbios fabricados mediante estos métodos. En aspectos alternativos, la distribución medioambiental es una distribución de cualquier muestra medioambiental tal como, por ejemplo, un agua de producción, un agua de formación, una muestra núcleo, un corte de perforación, agua, un sedimento o un suelo. En aspectos alternativos, la distribución medioambiental es una distribución de cualquier medio ambiente que tenga un sustrato carbonoso, por ejemplo, incluyendo una formación rica en materia orgánica subterránea natural o artificial, tal como vertederos, biorreactores superficiales o subterráneos, o un depósito subterráneo artificial; o pizarra, carbón, arenas petrolíferas, betún, alquitrán, aceite, arenisca y caliza con desechos orgánicos u otros depósitos o formaciones ricos en hidrocarburos, por ejemplo, a través de la vía metilotrófica.

Referencias

[1] M. L. Sogin, *et al*., "Microbial diversity in the deep sea and the underexplored ’rare biosphere'", *Proc. Natl. Acad. Sci.* EE.UU., vol. 103, n.º 32, págs. 12115-12120, agosto de 2006.
[2] C. Quince, *et al.,* "Accurate determination of microbial diversity from 454 pyrosequencing data", *Nat. Methods*, vol. 6, n.º 9, págs. 639-641, septiembre de 2009.
[3] P. H. Victor Kunin, "PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data", *The Open Journal*, vol. 1, n.º 1.
[4] J. Kuczynski, *et al*., "Using QIIME to analyze 16S rRNA gene sequences from microbial communities", *Curr Protoc Bioinformatics*, vol. capítulo 10, p. Unidad 10.7., diciembre de 2011.
[5] V. Kunin, *et al.,* "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates", *Environ. Microbiol*, vol. 12, n.º 1, págs. 118-123, enero de 2010.

Algunas partes de la siguiente descripción detallada se presentan en términos de algoritmos y representaciones simbólicas de operaciones sobre bits de datos dentro de una memoria de ordenador. Estas descripciones y representaciones algorítmicas son los medios usados por los expertos en las técnicas de procesamiento de datos para transmitir más eficazmente el contenido de su labor a los expertos en la materia. Un algoritmo, en el presente documento y en general, se concibe como una secuencia autoconsistente de etapas que conducen a un resultado. Las etapas son aquellas que requieren manipulaciones físicas de cantidades físicas. Normalmente, aunque no necesariamente, estas cantidades adoptan la forma de señales eléctricas o magnéticas capaces de ser almacenadas, transferidas, combinadas, comparadas y manipuladas de otra manera. A veces, ha demostrado ser conveniente, principalmente por razones de uso común, para referirse a estas señales como bits, valores, elementos, símbolos, caracteres, términos, números o similares.

Se ha de tener en cuenta, sin embargo, que todos estos términos y términos similares deben asociarse con las cantidades físicas apropiadas y son meramente marcadores convenientes aplicados a estas cantidades. A menos que se indique específicamente lo contrario, como es evidente a partir de la presente descripción, se aprecia que a lo largo de la descripción, los análisis que utilizan términos tales como "procesamiento", "cómputo", "cálculo", "determinación", "visualización" o similares, se refieren a las acciones y a los procesos de un sistema informático o dispositivo electrónico similar que manipule y transforme datos representados como cantidades físicas (por ejemplo, electrónicos) dentro de los registros y de las memorias del sistema informático en otros datos representados de manera similar como cantidades físicas dentro de las memorias o de los registros del sistema informático u otros de dichos dispositivos de almacenamiento, transmisión o visualización de la información.

En aspectos alternativos, la divulgación proporciona productos de fabricación, o aparato, para realizar las operaciones de la invención. Estos productos de fabricación o aparatos pueden estar especialmente construidos para los fines requeridos o pueden comprender un ordenador de uso general activado selectivamente o reconfigurado por un programa informático almacenado en el ordenador. Dicho programa informático puede almacenarse en un soporte de almacenamiento informático de lectura tal como, pero sin limitación, cualquier tipo de disco incluyendo disquetes, discos ópticos, CD-ROM y discos magnéticos ópticos, memorias de solo lectura (ROM) , memorias de acceso aleatorio (RAM), EPROM, EEPROM, tarjetas magnéticas u ópticas, o cualquier tipo de medio adecuado para almacenar instrucciones electrónicas.

Los algoritmos y las presentaciones presentados en el presente documento no están inherentemente relacionados con ningún ordenador u otro aparato en particular. Se pueden usar diversos sistemas de uso general con programas de acuerdo con las enseñanzas de la presente memoria, o puede resultar conveniente construir un aparato más especializado para realizar las etapas del método. La estructura para una variedad de estos sistemas aparecerá en la siguiente descripción. Además, la presente invención no se describe con referencia a ningún lenguaje de programación en particular. En realizaciones alternativas, se usa una variedad de lenguajes de programación para implementar las realizaciones de la invención como se describe en el presente documento.

En aspectos alternativos, un medio de lectura por máquina incluye cualquier mecanismo para almacenar o transmitir información en una forma legible por una máquina (por ejemplo, un ordenador). Por ejemplo, un medio de lectura por máquina incluye un medio de almacenamiento de lectura por máquina (por ejemplo, memoria de solo lectura (ROM), memoria de acceso aleatorio (RAM), medios de almacenamiento en disco magnético, medios de almacenamiento ópticos, dispositivos de memoria flash, etc.), un medio de transmisión de lectura por máquina (señales eléctricas, ópticas, acústicas u otras formas de señales propagadas (por ejemplo, ondas portadoras, señales infrarrojas, señales digitales, etc.)), etc.

En la presente descripción, se exponen numerosos detalles. Será evidente, sin embargo, para un experto en la materia que la presente invención se puede poner en práctica sin estos detalles específicos.

En realizaciones alternativas, la "complementariedad" puede definirse como un porcentaje de identidad o un porcentaje de identidad de secuencia, por ejemplo, en realizaciones alternativas, dos cadenas de ácido nucleico son un 80 %, 81 %, 82 %, 83 %, 84 %, 85 %, 86 %, 87 %, 88 %, 89 %, 90 %, 91 %, 92 %, 93 %, 94 %, 95 %, 96 %, 97 %, 98 %, 99 % o más, o completamente (100 %) idénticas o un 80 %, 81 %, 82 %, 83 %, 84 %, 85 %, 86 %, 87 %, 88 %, 89 %, 90 %, 91 %, 92 %, 93 %, 94 %, 95 % , 96 %, 97 %, 98 %, 99 % o más, o completamente complementarias. Cuanto más complementarias sean las dos cadenas, más probable es que el código resultante codifique una proteína en particular o, en el caso de la presente invención, cuanto más complementarias sean dos secuencias de amplicón, mayor grado de certeza habrá de que dos miembros (a partir de los cuales se derivan las secuencias de amplicón) pertenezcan a la misma comunidad microbiana.

Sistemas informáticos y dispositivos de almacenamiento de datos

En realizaciones alternativas, los métodos de la invención, en su totalidad o en parte, requieren necesariamente la implementación usando una máquina, un sistema informático o equivalente, dentro del cual se pueda ejecutar un conjunto de instrucciones para hacer que el ordenador o la máquina realice uno o más de los protocolos o metodologías de la invención. En realizaciones alternativas, la máquina puede conectarse (por ejemplo, conectarse en red) a otras máquinas, por ejemplo, en una red de área local (LAN), una intranet, una extranet o Internet, o cualquier equivalente de la misma. La máquina puede funcionar en la capacidad de un servidor o una máquina de cliente en un entorno de red cliente-servidor, o como una máquina igual en un entorno de red de igual a igual (o distribuido). La máquina puede ser un ordenador personal (PC), un PC Tablet, un decodificador (STB), un Asistente Personal Digital (PDA), un teléfono celular, un dispositivo web, un servidor, un enrutador de red, un conmutador o un puente, o cualquier máquina capaz de ejecutar un conjunto de instrucciones (secuenciales o de otro tipo) que especifiquen las acciones que vayan a ser realizadas por esa máquina. Se entenderá que el término "máquina" también incluirá cualquier colección de máquinas, ordenadores o productos de fabricación que ejecuten individual o conjuntamente un conjunto (o conjuntos múltiples) de instrucciones para realizar una o más de las metodologías de la invención.

En las realizaciones alternativas, un sistema informático ilustrativo de la invención comprende un dispositivo de procesamiento (procesador), una memoria principal (por ejemplo, memoria de solo lectura (ROM), memoria flash, memoria dinámica de acceso aleatorio (DRAM) tal como DRAM síncrona (SDRAM) o DRAM Rambus (RDRAM), etc.), una memoria estática (por ejemplo, memoria flash, memoria estática de acceso aleatorio (SRAM), etc.) y un dispositivo de almacenamiento de datos que se comunican entre sí a través de un bus.

En realizaciones alternativas, un procesador representa uno o más dispositivos de procesamiento de uso general tales como un microprocesador, una unidad de procesamiento central o similar. Más concretamente, el procesador puede ser un microprocesador de cálculo de conjunto de instrucciones complejas (CISC), microprocesador de cálculo de instrucciones reducidas (RISC), microprocesador de palabras de instrucción muy largas (VLIW) o un procesador que implemente otros conjuntos de instrucciones o procesadores que implementen una combinación de conjuntos de instrucciones. El procesador también puede ser uno o más dispositivos de procesamiento de uso especial, tales como un circuito integrado específico de la aplicación (ASIC), una matriz de puerta programable por campo (FPGA), un procesador de señales digitales (DSP), un procesador de red o similar. En realizaciones alternativas, el procesador está configurado para ejecutar las instrucciones (por ejemplo, el proceso lógico de procesamiento) para realizar las operaciones y las etapas descritas en el presente documento.

En realizaciones alternativas, el sistema informático comprende además un dispositivo de interfaz de red. El sistema informático también puede incluir una unidad de visualización de vídeo (por ejemplo, una pantalla de cristal líquido (LCD) o un tubo de rayos catódicos (CRT)), un dispositivo de entrada alfanumérico (por ejemplo, un teclado), un dispositivo de control del cursor y un dispositivo de generación de señales (por ejemplo, un altavoz).

En realizaciones alternativas, el dispositivo de almacenamiento de datos (por ejemplo, la unidad de disco) comprende un soporte informático de lectura de almacenamiento sobre el que se almacenan uno o más conjuntos de instrucciones (por ejemplo, software) que incorporan uno cualquiera o más de los protocolos, de las metodologías o de las funciones de la presente invención. Las instrucciones también pueden residir, total o al menos parcialmente, dentro de la memoria principal y/o dentro del procesador durante su ejecución por el sistema informático, también constituyendo la memoria principal y el procesador medios de almacenamiento accesibles por la máquina. Las instrucciones además se pueden transmitir o recibir por una red a través del dispositivo de interfaz de red.

En realizaciones alternativas, el soporte informático de lectura de almacenamiento se usa para almacenar conjuntos de estructuras de datos que definen estados de identificación de usuario y preferencias de usuario que definen perfiles de usuario. Los conjuntos de estructuras de datos y los perfiles de usuario también se pueden almacenar en otras secciones del sistema informático, tales como la memoria estática.

En realizaciones alternativas, mientras que el soporte informático de lectura de almacenamiento de un ejemplo de realización es un solo medio, la expresión "medio de almacenamiento accesible por la máquina" puede considerarse que incluye un solo medio o varios medios (por ejemplo, una base de datos centralizada o distribuida y/o cachés y servidores asociados) que almacenen uno o más conjuntos de instrucciones. En realizaciones alternativas, la expresión "medio de almacenamiento accesible por la máquina" también puede considerarse que incluye cualquier medio que sea capaz de almacenar, codificar o transportar un conjunto de instrucciones para su ejecución por la máquina y que haga que la máquina realice una o más de las metodologías de la presente invención. Por consiguiente, en realizaciones alternativas, la expresión "medio de almacenamiento accesible por la máquina" se considerará que incluye, pero sin limitación, memorias de estado sólido, y medios ópticos y magnéticos.

En realizaciones alternativas, la información y las señales se representan usando cualquier tecnología y/o técnica conocida en la materia. Por ejemplo, los datos, las instrucciones, los comandos, la información, las señales, los bits, los símbolos y los chips usados para poner en práctica las composiciones (dispositivos, ordenadores) y los métodos de la invención pueden representarse por tensiones, corrientes, ondas electromagnéticas, campos magnéticos o partículas, campos ópticos o partículas, o cualquier combinación de los mismos.

En realizaciones alternativas, los diversos bloques lógicos ilustrativos, módulos, circuitos y etapas algorítmicas usados para describir realizaciones ilustrativas de la invención se pueden implantar como hardware electrónico, software informático, o combinaciones de ambos. Para ilustrar claramente esta capacidad de intercambio de hardware y software, se han descrito anteriormente varios componentes ilustrativos, bloques, módulos, circuitos y etapas, en general, en términos de su funcionalidad. Que dicha funcionalidad se implante como hardware o como software depende de las restricciones de aplicación y diseño impuestas en particular al sistema en general. Los expertos en la materia pueden implantar la funcionalidad descrita de diversas maneras para cada aplicación particular, pero dichas decisiones de implantación no deben interpretarse como causantes de un alejamiento del alcance de la presente invención.

En vista de las presentes enseñanzas, a los expertos en la materia, se les ocurrirán fácilmente modificaciones de la presente invención. La descripción de la invención del presente documento es ilustrativa y no restrictiva. La presente invención solo estará limitada por las siguientes reivindicaciones, que incluyen la totalidad de dichas realizaciones y modificaciones observadas junto con la memoria descriptiva anterior y los dibujos adjuntos. Por lo tanto, el alcance de la invención debe determinarse no con referencia a la descripción anterior, sino que debe determinarse con

referencia a las reivindicaciones adjuntas junto con su alcance completo de equivalentes.

REIVINDICACIONES

1. Un método de identificación de un consorcio microbiano o de un grupo de microbios con distribuciones medioambientales correlacionadas, que comprende:

(a) proporcionar abundancias de marcadores en dos o más muestras, en las que cada marcador es representativo de un grupo de microbios con distribuciones medioambientales correlacionadas, y el marcador comprende una composición producida mediante:

(i) el suministro de una colección de muestras, en la que las muestras son secuencias de ácidos nucleicos de una o más comunidades microbianas o un grupo de microbios con distribuciones medioambientales correlacionadas, y el procesamiento de las muestras mediante:
(ii) la identificación de secuencias de nucleótidos que contienen códigos de barras identificadores de la muestra y el registro de su correspondencia con una determinada muestra y luego, la retirada de los códigos de barras y el descarte de las secuencias que no contienen los códigos de barras correctos o que contienen códigos de barras que no se corresponden con ninguna de las muestras;
(iii) el corte o truncamiento de las secuencias de nucleótidos o las “lecturas” de (a) y la designación de las secuencias de nucleótidos específicas de la región que quedan como “marcadores”, de modo que un marcador es una versión procesada o truncada de una lectura, y el resto de secuencias de nucleótidos identificadas o “lecturas” de la etapa (ii) y el mantenimiento de solo regiones previamente definidas;
(iv) la filtración cualitativa de las lecturas truncadas mediante la eliminación de las lecturas truncadas ambiguas y la eliminación de las lecturas truncadas de baja calidad, en la que una lectura de baja calidad está por debajo del umbral de calidad;
(v) la clasificación taxonómica de los marcadores restantes y, opcionalmente, la generación de una salida de datos que comprende una descripción de comunidades microbianas como recuentos de la abundancia de los miembros únicos de cada comunidad;
(vi) la importación de las secuencias e identificadores de nuevos marcadores a una base de datos; y
(viii) la importación de los marcadores de las abundancias de recuentos de las muestras a la base de datos;
(viii) la constricción o modificación de la base de datos que comprende los marcadores únicos de las etapas (ii) a (vii), en relación con sus abundancias en las muestras;
(ix) la exportación de datos de abundancia de marcadores de al menos dos muestras de la base de datos;
en la que se fija un umbral para los marcadores que aparecen en el análisis,
identificando de este modo la composición de comunidades microbianas o un grupo de microbios con distribuciones medioambientales correlacionadas;

(b) identificar similitudes de abundancias en muestras entre pares de microbios, mediante la comparación de las abundancias de un microbio con otro microbio de cada muestra usando métricas de distancia;
(c) repetir la etapa (b) para al menos un par más de microbios;
(d) almacenar las similitudes obtenidas en la etapa (b) y (c) en forma de una estructura de datos matricial en un formato digital;
(e) realizar bien un análisis de red, un análisis de grupos o un agrupamiento en la estructura matricial de datos de similitud obtenida en la etapa (d), implicando el análisis de red la representación de los datos en la que los microbios o los marcadores se designan como nodos de la red y las similitudes entre los marcadores o los microbios obtenidos en las etapas (b) y (c) se designan como bordes de la red; y
(f) designar los microbios que están conectados en la red o asignados al mismo grupo como un consorcio, de manera que las etapas del método de identificación de un consorcio microbiano o grupo de microbios con distribuciones medioambientales correlacionadas se implantan por ordenador.

2. El método de la reivindicación 1, que comprende además la etapa de combinar los correspondientes cultivos microbianos, en el que los cultivos microbianos se componen de cepas puras, cepas enriquecidas o cualquier combinación de las mismas.

3. El método de las reivindicaciones 1 o 2, en el que la distribución medioambiental es una distribución de cualquier muestra medioambiental, tal como, por ejemplo, un agua de producción, un agua de formación, una muestra núcleo, un corte de perforación, agua, un sedimento o un suelo; o la distribución medioambiental es una distribución desde cualquier medio ambiente que tenga un sustrato carbonoso, por ejemplo, incluyendo una formación rica en materia orgánica subterránea natural o artificial, tal como vertederos, biorreactores superficiales o subterráneos, o un depósito subterráneo artificial; o pizarra, carbón, arenas petrolíferas, betún, alquitrán, aceite, arenisca y caliza con desechos orgánicos u otros depósitos o formaciones ricos en hidrocarburos.

4. El método de cualquiera de las reivindicaciones 1 a 3, en el que las abundancias se deducen a partir del número de copias de secuencias génicas distintas detectadas en cada muestra.

5. El método de la reivindicación 4, en el que el gen es una secuencia de genes de ARNr 16S.

6. El método de la reivindicación 5, en el que las abundancias se representan como recuentos absolutos o fracción del total.

7. El método de la reivindicación 6, método que comprende la transformación logarítmica de los datos de abundancia.

8. El método de las reivindicaciones 1 a 7, en el que la métrica de distancia comprende: una distancia Euclidiana, un Chi cuadrado, una correlación, una distancia de Manhattan, un método de ordenación que comprende opcionalmente el uso de un análisis de componentes principales, una disimilitud de Bray-Curtis y/o un escalamiento multidimensional no métrico, opcionalmente, NMS o NMDS.

9. El método de la reivindicación 8, en el que cada muestra puede comprender todas las muestras disponibles o cualquier fracción de las muestras.

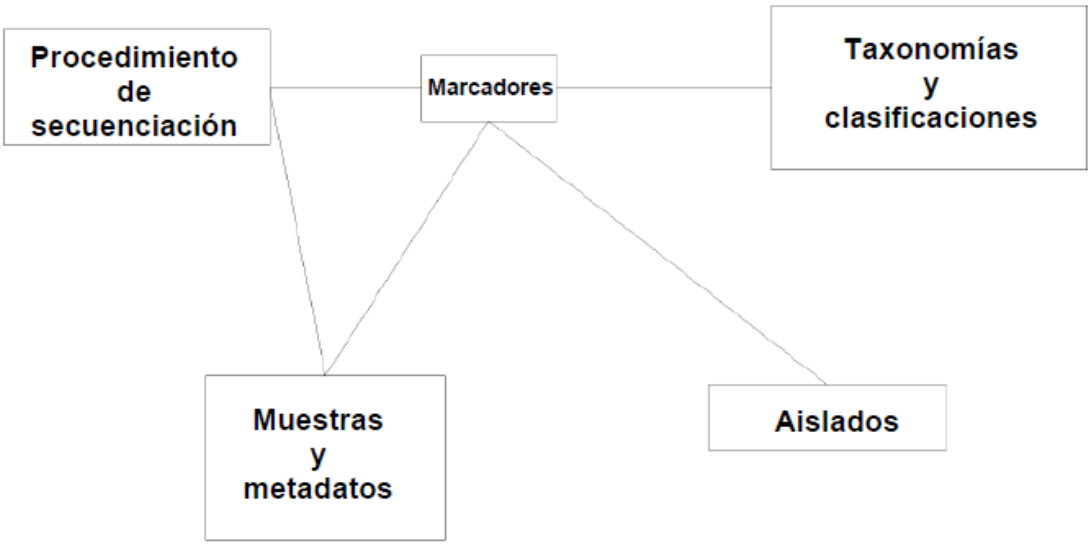
10. El método de una cualquiera de las reivindicaciones 1 a 9, en el que la estructura de datos matricial se almacena en una memoria de ordenador, en una unidad de disco, en un archivo, en una colección de archivos o una base de datos.

11. El método de una cualquiera de las reivindicaciones 1 a 10, en el que el análisis de los grupos o agrupamiento es un agrupamiento jerárquico, una identificación de componentes conectados, un agrupamiento basado en la conectividad, un agrupamiento basado en la distribución, un agrupamiento basado en la densidad, un agrupamiento de un solo enlace, un agrupamiento de Marcov (MCL) o un agrupamiento de centroides.

12. El método de una cualquiera de las reivindicaciones 1 a 11, en el que los microbios comprenden todos los microbios conectados en la red o todos los microbios del grupo o cualquier fracción de los mismos.

13. El método de una cualquiera de las reivindicaciones 1-12, en el que las etapas del método se implantan por ordenador usando un producto de programa informático para implantar estas etapas o un producto de programa informático para el procesamiento de datos, comprendiendo el producto de programa informático un proceso lógico ejecutable por ordenador contenido en un soporte informático de lectura para implantar estas etapas.

Figura 1


Procedimiento de secuenciación
Marcadores
Taxonomías y clasificaciones
Muestras y metadatos
Aislados

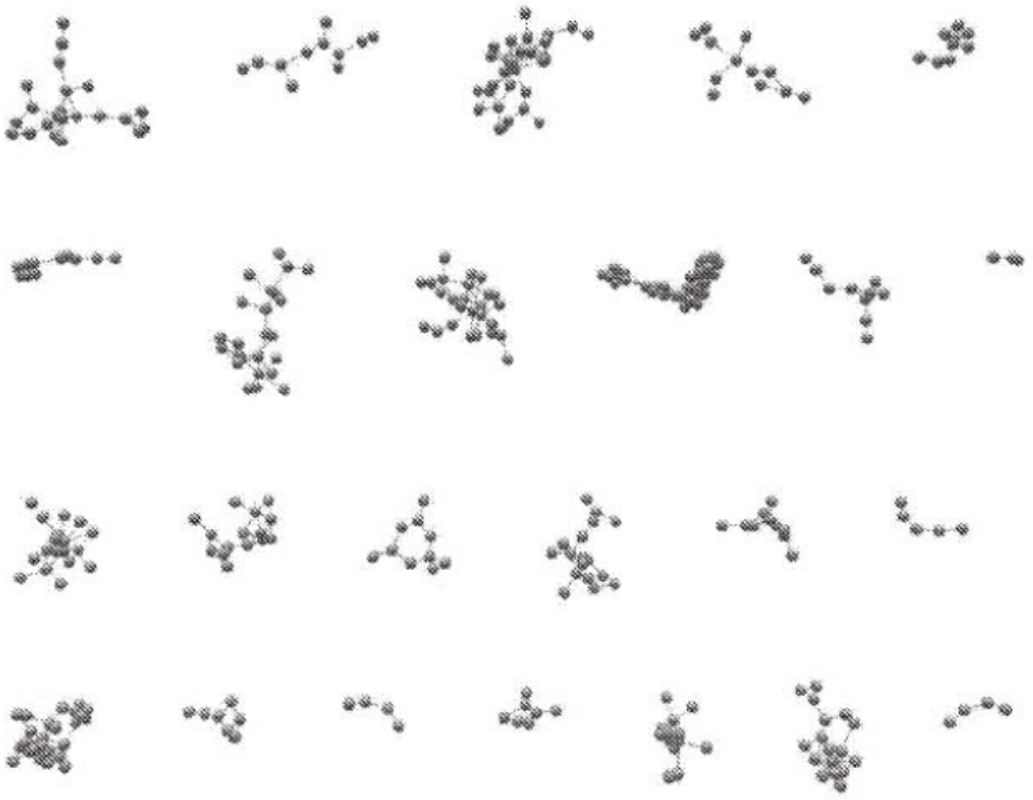


Figura 2

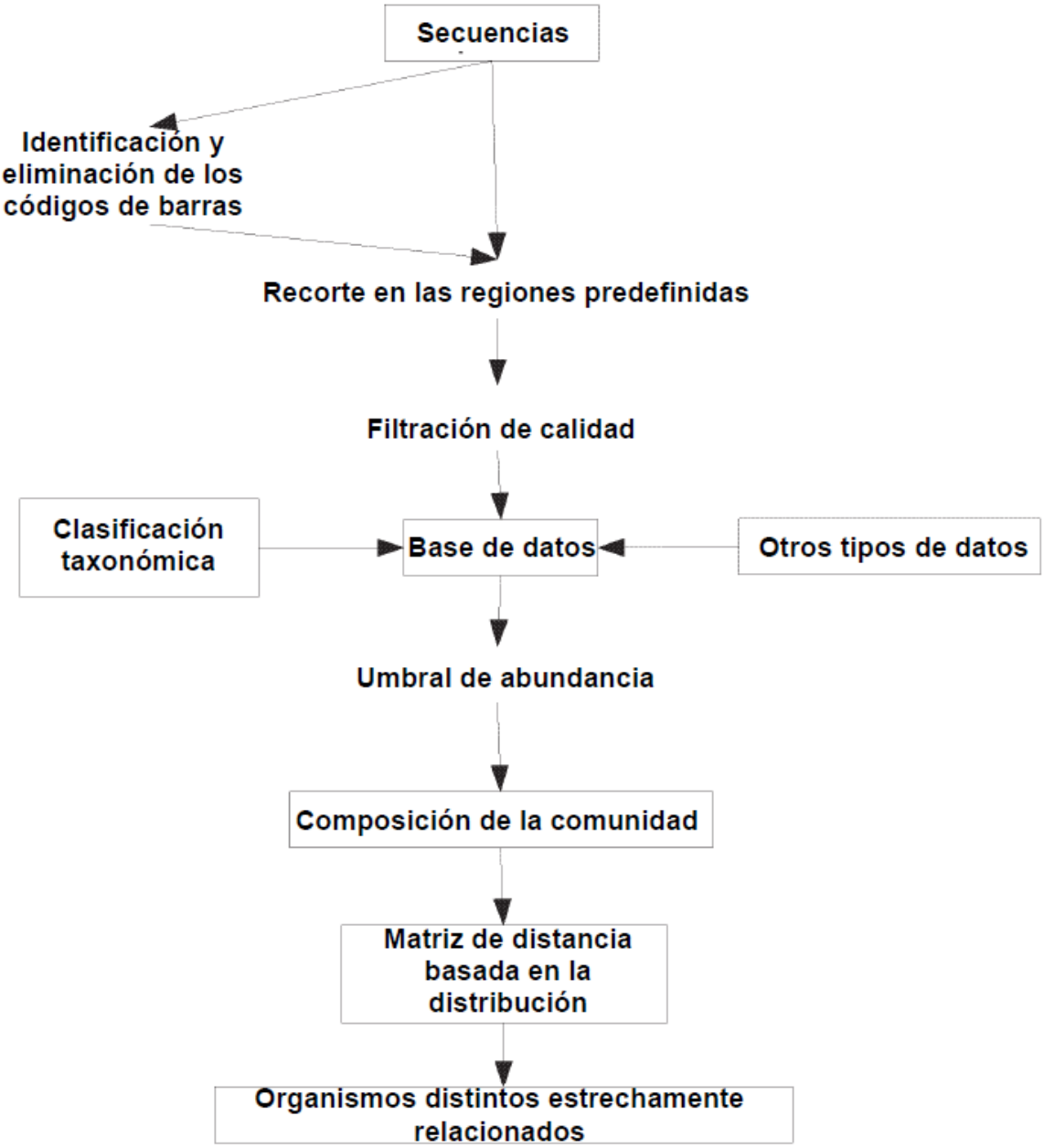

Secuencias
Identificación y eliminación de los códigos de barras
Recorte en las regiones predefinidas
Filtración de calidad
Clasificación taxonómica
Base de datos
Otros tipos de datos
Umbral de abundancia
Composición de la comunidad
Matriz de distancia basada en la distribución
Organismos distintos estrechamente relacionados

Figura 3

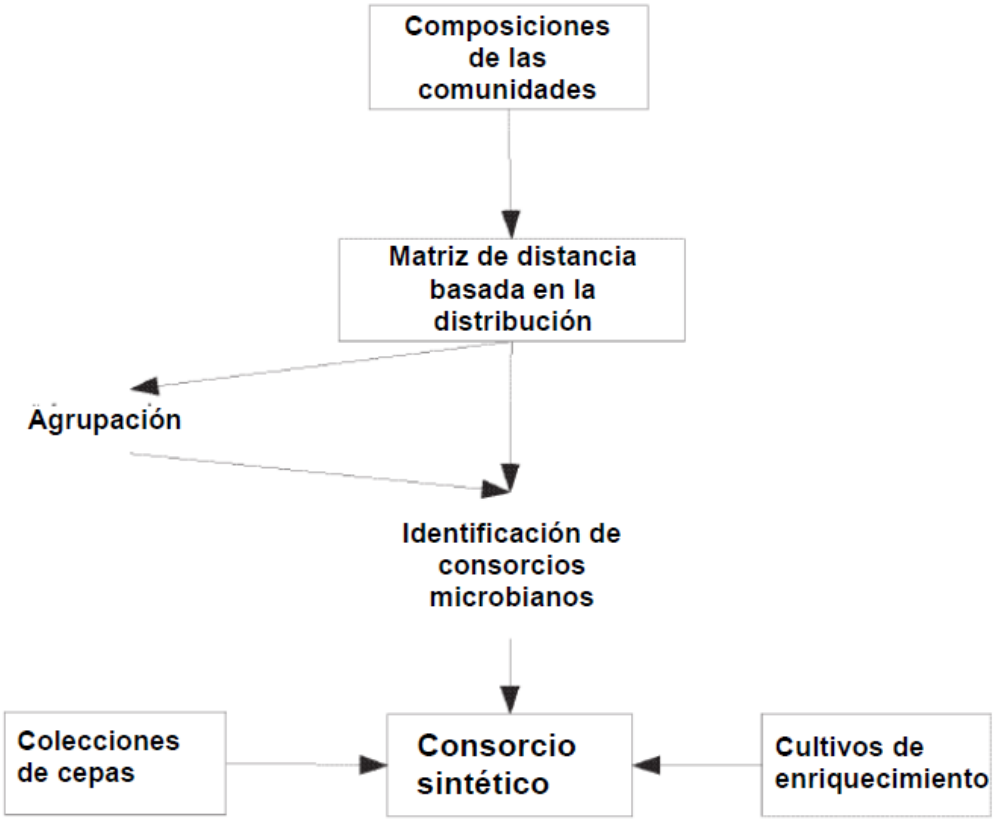

Composiciones de las comunidades
Matriz de distancia basada en la distribución
Agrupación
Identificación de consorcios microbianos
Colecciones de cepas
Consorcio sintético
Cultivos de enriquecimiento

Figura 4