(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 456 240**

(51) Int. CI.:

G06F 19/28 (2011.01)
G06F 19/22 (2011.01)
G06F 19/24 (2011.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **29.11.2007 E 07816282 (3)**

(97) Fecha y número de publicación de la concesión europea: **26.03.2014 EP 2215578**

(54) Título: **Método y sistema informático para evaluar anotaciones de clasificación asignadas a secuencias de ADN**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
21.04.2014

(73) Titular/es:

SMARTGENE GMBH (100.0%)
INDUSTRIESTRASSE 16
6300 ZUG, CH

(72) Inventor/es:

EMLER, STEFAN y
MICHEL, PIERRE-ANDRÉ

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín europeo de patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre concesión de Patentes Europeas).

DESCRIPCIÓN

Método y sistema informático para evaluar anotaciones de clasificación asignadas a secuencias de ADN

Campo de la invención

La presente invención se refiere a un método implementado por ordenador y a un sistema informático para la evaluación de anotaciones de clasificación asignadas a secuencias de ADN. Específicamente, la presente invención se refiere a un método implementado por ordenador y a un sistema informático para evaluar anotaciones de clasificación asignadas a secuencias de ADN almacenadas en una base de datos.

Antecedentes de la invención

La identificación de formas de vida, basada en secuencias se utiliza cada vez más para fines de diagnóstico. Al ser independiente del crecimiento y del metabolismo, este método ofrece ventajas significativas en términos de velocidad y precisión sobre técnicas convencionales basadas en el cultivo. Genes conservados presentes en todas las bacterias u hongos se amplifican y posteriormente se secuencian utilizando técnicas de secuenciación automatizada. Las secuencias obtenidas se comparan después con referencias en una base de datos. De este modo, incluso materiales aislados raros, inesperados o inusuales se pueden identificar y clasificar rápidamente. El análisis de las secuencias se puede aplicar a todos los genes conservados de todas las formas de vida, particularmente a microorganismos tales como bacterias y hongos. La identificación de microorganismos basada en las secuencias, depende de la comparación de la secuencia característica de la muestra con una base de datos que contiene secuencias de referencia que representan todos los géneros y especies pertinentes. Por tanto, es importante que una base de datos de referencia cumpla los siguientes requisitos:

1) Secuencia exacta: la base de datos contiene secuencias correctas de la diana solicitada, no tiene errores de secuenciación, ni fallos de lectura, no hay huecos artificiales, inserciones, no hay secuencias de vectores.

2) Anotación de clasificación correcta (es decir, denominación de los registros): las secuencias se anotan correctamente (por ejemplo, los nombres de las especies) y esta información se actualiza con respecto a cambios en la taxonomía.

3) Representativa: la base de datos representa todas las formas de vida pertinentes, por ejemplo, género y especie, incluyendo sus variantes genéticas (intraespecíficas, intragenómicas).

4) Actualización: las referencias se actualizan con respecto a especies descritas recientemente y a posibles cambios en la taxonomía (véase también 2).

Actualmente no existe una base de datos de referencia única que cumpla todos estos requisitos. Sin embargo, debido a que la calidad de los resultados de las comparaciones de secuencias depende en gran medida de las referencias disponibles, es crucial que estas bases de datos sean lo más fidedignas posible. En general, los científicos añaden registros a repositorios públicos que tienen una calidad aceptable en términos de contenido de la secuencia y de la anotación (por ejemplo, nombre de la especie). Sin embargo, hay muchos errores de secuenciación o anotaciones incorrectas en relación con la taxonomía actual. Se producen errores de anotación, por ejemplo, cuando las secuencias se presentan junto con una información incorrecta sobre el organismo o el gen a partir del cual se ha obtenido la secuencia, o con nombres de especies que no están actualizados (por ejemplo, cuando las especies se han clasificado de nuevo taxonómicamente, como es frecuentemente en el caso de bacterias). Cuando una secuencia de una muestra se busca en una base de datos de referencia, la lista resultante suele mostrar coincidencias correctas e incorrectas indistintamente, dejando en manos de la experiencia del usuario la determinación de qué referencias se identificaron de forma correcta o incorrecta. Por lo tanto, una secuencia correcta con una anotación incorrecta podría aparecer en la parte superior de la lista de coincidencias y, por lo tanto, indicar una identificación errónea de una bacteria, por ejemplo. Debido a que la identificación de agentes patógenos basada en la secuencia se está convirtiendo hoy en día en una parte del trabajo rutinario en los laboratorios de diagnóstico médico, veterinarios e industriales, existe una necesidad de que las búsquedas y las comparaciones de secuencias en bases de datos sean fáciles y fiables, por ejemplo, para la identificación de una especie bacteriana o fúngica o un subtipo de virus, o para hacer cotejar cualquier organismo desconocido con una base de datos de organismos bien caracterizados. En particular, los resultados de la búsqueda y la comparación de la similitud de secuencias se deben proporcionar de forma adecuada con respecto a la experiencia de los técnicos de laboratorio común, que, en general, no tienen experiencia científica o una amplia formación en bioinformática o en taxonomía de organismos (microorganismos).

El documento de EE.UU. 2007/0083334 describe sistemas y métodos para anotar secuencias biomoleculares. Después de la alineación(es) de secuencias, las secuencias biomoleculares se agrupan con métodos informáticos de acuerdo con un campo de homología progresiva, usando uno o varios algoritmos de agrupamiento. Una secuencia biomolecular se considera que pertenece a una agrupación, si la secuencia comparte una homología de secuencia basada en la alineación, superior a un determinado valor umbral con uno de los miembros de la agrupación. Según el documento de EE.UU. 2007/0083334, la agrupación computacional se puede efectuar usando cualquier programa informático de alineación, disponible comercialmente incluyendo un algoritmo de homología local. Por ejemplo, un

grupo muestra un cierto grado de homología, si los ácidos nucleicos son idénticos entre sí en un 90%.

El documento de EE.UU. 2007/0134692 describe un método y un sistema basado en la alineación para actualizar datos de anotación de un conjunto de sondas. Se generan una o varias agrupaciones mediante la transcripción a través de conjuntos de datos recuperados a partir de una o varias fuentes. Una o varias secuencias de sondas se alinea con una secuencia representativa de una o varias de las agrupaciones. La secuencia representativa se alinea con una secuencia del genoma y la secuencia del genoma se anota con información sobre la ubicación de la sonda. Las secuencias de sondas alineadas se cartografían en la secuencia del genoma, usando la alineación de la secuencia representativa y la secuencia del genoma. Una puntuación se calcula utilizando un número asociado con las secuencias de sondas alineadas y un número asociado con la formación de una ubicación de la sonda asociada con una región de la secuencia del genoma que se corresponde con la secuencia representativa alineada. Registros redundantes se pueden eliminar mediante el método de agrupación. Por ejemplo, si la alineación de transcritos en una agrupación se solapa en >97% en su longitud completa, entonces se determina que son redundantes y solo la secuencia más larga se conserva en la agrupación.

Compendio de la invención

Es un objeto de esta invención proporcionar un método implementado por ordenador y un sistema informático para evaluar (y evaluar de nuevo) anotaciones de clasificación que incluyen anotaciones taxonómicas, sistemáticas y/o funcionales, asignadas a secuencias de ADN. En particular, un objeto de la presente invención es proporcionar un método implementado por ordenador y un sistema informático para evaluar cualitativamente las anotaciones de clasificación, de tal manera que anotaciones erróneas y/o dudosas sean patentes para los técnicos de laboratorio que no tienen una amplia experiencia o entrenamiento en bioinformática o en taxonomía de organismos (microorganismos).

De acuerdo con la presente invención, estos objetos se consiguen especialmente a través de las características de las reivindicaciones independientes. Además, otras realizaciones ventajosas se deducen de las reivindicaciones dependientes y de la descripción.

De acuerdo con la presente invención, los objetos anteriormente mencionados se consiguen en particular, ya que para evaluar las anotaciones de clasificación (incluyendo anotaciones taxonómicas, sistemáticas y/o funcionales) asignadas a secuencias de ADN almacenadas en una base de datos, por ejemplo, una base de datos de referencia, las secuencias de ADN se agrupan, basándose en sus anotaciones de clasificación respectivas, por especies utilizando sistemas de clasificación establecidos para la clasificación taxonómica, sistemática y/o funcional. Posteriormente, para las parejas de secuencias de ADN, se determina en cada caso una medida de distancia entre las secuencias de ADN respectivas. La medida de distancia se determina alineando automáticamente las secuencias de ADN respectivas y definiendo la medida de distancia basándose en una puntuación de la similitud entre las secuencias de ADN alineadas. Por ejemplo, la medida de distancia entre dos secuencias de ADN se calcula como un valor complementario a la puntuación de la similitud, por ejemplo, restando una puntuación ponderada de la similitud de una. Por ejemplo, la puntuación ponderada de la similitud se calcula dividiendo la puntuación de la similitud entre las dos secuencias de ADN, entre la longitud menor de las dos secuencias de ADN. Posteriormente, se determina una secuencia centroide que tiene la medida global más corta de distancia a las secuencias de ADN. Preferiblemente, dentro de un grupo definido de secuencias de ADN, por ejemplo, secuencias de ADN relacionadas con una especie, la secuencia centroide es la secuencia de ADN de estas secuencias que tiene la medida acumulada más corta de distancia a las otras secuencias de ADN en el grupo. Alternativamente, la secuencia centroide es un objeto completamente virtual, calculado para tener la menor medida promedio de la distancia a todas las secuencias de ADN que se van a considerar. Cabe señalar que dentro del presente contexto, la expresión "secuencia centroide" se utiliza para incluir un objeto centroide, representativo de una secuencia de ADN real, así como un objeto centroide representativo de un objeto virtual. A cada una de las secuencias de ADN que se van a considerar, se asigna la medida de distancia entre la secuencia de ADN respectiva de las secuencias de ADN y la secuencia centroide, como un nivel de confianza cuantitativo para la anotación de clasificación de la secuencia de ADN respectiva de las secuencias de ADN. Preferiblemente, los niveles de confianza se almacenan en la base de datos asignada a la anotación respectiva y la secuencia de ADN que coincide con una especie conocida o un nombre de género. La evaluación y la valoración de las anotaciones de clasificación con estos niveles de confianza, hace que sea posible proporcionar a un usuario una indicación del grado de representatividad de una secuencia de ADN para una especie en particular. Por ejemplo, cuando un usuario realiza una consulta en la base de datos, con cada registro en la lista de secuencias de referencia que coinciden, se presenta un campo para el usuario que indica el nivel de confianza que representa la secuencia de ADN correspondiente para esa especie y/o género en particular. Dependiendo de la realización, el nivel de confianza cuantitativo, es decir, la medida de la distancia a una secuencia centroide, es un valor numérico o un valor cualitativamente descriptivo obtenido a partir del valor numérico. Para los niveles de confianza numéricos, una medida baja de la distancia indica una anotación de confianza, mientras que con una distancia mayor, el registro debe ser considerado con más cuidado en lo que respecta a proporcionar una identificación válida.

De acuerdo con la presente invención, la medida de la distancia se determina entre secuencias de ADN de una especie y las secuencias centroides se determinan para las secuencias de ADN de cada una de las especies. En una realización preferida, los valores atípicos se definen dentro de la especie, por lo que los valores atípicos son aquellas secuencias de ADN que tienen las mayores medidas de distancia a la secuencia centroide de la especie

respectiva. Por ejemplo, uno o varios valores atípicos se definen basándose en un valor umbral de la distancia máxima, una desviación definida a partir de una medida promedio de la distancia o un número o una cantidad definida de secuencias de ADN que tienen la mayor medida de distancia desde la secuencia centroide. Para los valores atípicos que tienen una medida menor de distancia a una secuencia centroide de otra especie, las anotaciones se marcan como incorrectas, por ejemplo, estableciendo un indicador respectivo en la base de datos.

En una realización, se genera un grafo de aristas ponderado a partir de las puntuaciones de similitud entre las secuencias de ADN. En este grafo, las secuencias de ADN son nodos en el grafo, y los nodos están conectados si la puntuación de similitud entre las secuencias de ADN respectivas es positiva (a las secuencias no alineables y diferentes se les asigna una similitud de cero). A la medida de la distancia entre las secuencias de ADN respectivas se asigna en cada caso un peso de arista. Para los nodos en el grafo, se calculan las densidades de conectividad local (número de conexiones con otros nodos). Las agrupaciones de nodos se definen a través de agregación progresiva hasta una densidad máxima de conectividad local, por lo que la medida de la distancia entre las secuencias de ADN asociadas con nodos en una agrupación (distancia dentro de la agrupación) es significativamente menor que una medida promedio de distancia entre las secuencias de ADN asociadas con los nodos del grafo (distancia promedio del grafo).

En una realización adicional, un valor umbral de agrupación se recibe en el ordenador del usuario, por ejemplo, como respuesta al usuario al ver el grafo que se muestra en una pantalla. Posteriormente, las agrupaciones de nodos se definen mediante la aplicación del valor umbral de agrupación como una distancia máxima dentro de la agrupación. Por lo tanto, los nodos asociados con secuencias de ADN que tienen una medida de distancia mayor que la distancia máxima dentro de la agrupación, no están incluidos en la agrupación. Después de la aplicación del valor umbral de agrupación, el grafo se muestra en la pantalla. Mediante la selección de diferentes valores umbrales de agrupación, el usuario puede seleccionar un nivel de granularidad del grafo en el sentido de que, con un valor relativamente elevado del valor umbral de agrupación, el grafo es típicamente una estructura coherente que conecta todos los nodos, mientras que para los valores umbrales menores de agrupación, el grafo típicamente se desintegra en múltiples agrupaciones.

Preferiblemente, en el enfoque basado en el grafo, la secuencia de ADN asociada con el nodo que tiene la mayor densidad de conectividad en una agrupación, es decir, el mayor número de conexiones con otros nodos, se define como la secuencia centroide de esa agrupación.

En una realización, la anotación de clasificación asociada con una secuencia centroide se asigna a secuencias de ADN asociadas con esa secuencia centroide. Específicamente, la anotación del centroide de una agrupación determinada se asigna a secuencias de ADN asociadas con los nodos de esa agrupación. Preferiblemente, esta anotación no sobrescribe la anotación de clasificación existente de una secuencia de ADN, sino que se añade como una recomendación que se puede mostrar a los usuarios.

Además de un método implementado por ordenador y un sistema informático para la evaluación de anotaciones de clasificación asignadas a secuencias de ADN, almacenadas en una base de datos, la presente invención también se refiere a un producto de programa informático que incluye medios de código de programa informático para controlar uno o varios procesadores de un ordenador, de tal manera que el ordenador ejecuta el método, en particular, un producto de programa informático que incluye un medio legible por ordenador que contiene el medio de código de programa informático.

Breve descripción de los Dibujos

La presente invención se explicará con más detalle, a modo de ejemplo, con referencia a los dibujos en los que:

La Figura 1 muestra un diagrama de bloques que ilustra esquemáticamente una configuración ejemplar de un sistema basado en ordenador para la puesta en práctica de realizaciones de la presente invención, comprendiendo dicha configuración un sistema informático con una base de datos, y estando conectada dicha configuración a un terminal de entrada de datos a través de una red de telecomunicaciones.

La Figura 2 muestra un diagrama de flujo que ilustra una secuencia ejemplar de las etapas para valorar las anotaciones de clasificación asignadas a secuencias de ADN.

La Figura 3 muestra un diagrama de flujo que ilustra una secuencia ejemplar de las etapas para la determinación de una o varias secuencias centroides.

La Figura 4 muestra un ejemplo de una agrupación de secuencias de ADN relacionadas con una secuencia centroide.

La Figura 5 muestra una alineación de 11 variaciones ejemplares de secuencias de ADN relacionadas con una especie.

La Figura 6 muestra un ejemplo de una interfaz de usuario que muestra a un usuario posibles coincidencias de una secuencia de una muestra, indicándose cada posible coincidencia con un nivel de confianza (dist).

Descripción detallada de las realizaciones preferidas

En la Figura 1, el número de referencia 3 se refiere a un terminal de entrada de datos. Como se ilustra en la Figura 1, el terminal de entrada de datos 1 incluye un ordenador personal 31 con un teclado 32 y un monitor de visualización 33, por ejemplo.

Como se ilustra en la Figura 1, el terminal de entrada de datos 3 está conectado al sistema informático 1 a través de una red de telecomunicaciones 2. Preferiblemente, la red de telecomunicaciones 2 incluye Internet y/o Intranet, haciendo que el sistema informático 1 sea accesible como un servidor de red a través de la red mundial o dentro una red IP diferente, respectivamente. La red de telecomunicaciones 2 también puede incluir otra red fija, tal como una red de área local (LAN) o una red digital de servicios integrada (RDSI) y/o una red inalámbrica, tal como una red de radio móvil (por ejemplo, el Sistema Global de telecomunicaciones Móviles (GSM) o el Sistema Universal de Telefonía Móvil (UMTS)), o una red de área local inalámbrica (WLAN). En una variante, al menos un terminal de entrada de datos 3 está conectado directamente al sistema de ordenador 1.

El sistema informático 1 incluye uno o varios ordenadores teniendo cada uno, uno o varios procesadores. Por otra parte, el sistema informático 1 comprende una base de datos 11 (referencia) que incluye registros almacenados de secuencias de ADN de referencia 111. Como se ilustra esquemáticamente en la Figura 1, el sistema informático 1 incluye diferentes módulos funcionales, a saber, un módulo de comunicación 120, un módulo de aplicación 121, un módulo comparador 122, un detector de centroide 123, un módulo de valoración 124, un detector de errores 125 y un generador de grafos 126. La base de datos 11 se implementa en un ordenador compartido con los módulos funcionales o en un ordenador independiente. Como se ilustra esquemáticamente en la Figura 1, la base de datos de referencia 11 incluye anotaciones de clasificación 112, que incluyen anotaciones taxonómicas, sistemáticas y/o funcionales, asociadas con secuencias de ADN 111. Típicamente, el contenido de la base de datos de referencia 11 incluye registros relacionados con secuencias de ADN recuperadas y obtenidas a partir de diferentes bases de datos de secuencias de ADN (públicas o privadas). El módulo de comunicaciones 120 incluye elementos convencionales del equipo informático y de programas informáticos configurados para el intercambio de datos a través de la red de telecomunicaciones 2 con uno o varios terminales de entrada de datos 3. El módulo de aplicación 121 es un módulo del equipo informático programado configurado para proporcionar a los usuarios de la terminal de entrada de datos 3 una interfaz de usuario 1211. Preferiblemente, la interfaz de usuario de 1211 se proporciona a través de un navegador de Internet convencional, tal como Microsoft Explorer o Mozilla Firefox. El módulo comparador 122, el detector de centroide 123, el módulo de valoración 124, el detector de errores 125 y el generador de grafos 126, son preferiblemente módulos de programas informáticos programados que se ejecutan en un procesador del sistema informático 1.

El número de referencia 7 se refiere a una base de datos de sistemas de clasificación (interconectada) accesible para el sistema informático 1 a través de la red de telecomunicaciones 2. La base de datos de sistemas de clasificación incluye sistemas de clasificación establecidos actuales para la clasificación taxonómica, sistemática y/o funcional de secuencias de ADN de formas vivas. Los sistemas de clasificación no son estáticos y están sujetos a cambio y/o adición.

En los siguientes párrafos, se describe la funcionalidad de los módulos funcionales con referencia a las Figuras 2 y 3.

En la etapa S1, en función de sus anotaciones de clasificación respectivas 112, el módulo comparador 122 agrupa por especies las secuencias de ADN 111 almacenadas en la base de datos de referencia 11 utilizando sistemas de clasificación establecidos actuales, disponibles a partir de la base de datos de sistemas de clasificación 7. La agrupación de las secuencias de ADN se realiza en todas las secuencias de ADN 111 o en un grupo seleccionado de las secuencias de ADN 111. Por ejemplo, el módulo comparador 122 es activado por un comando operador de una petición de usuario. En una realización, el módulo comparador 122 se activa periódicamente o de forma automática cada vez que se produce un cambio, una adición o una actualización en el sistema de clasificación 7, o se introduce (agrega) un número definido de nuevas secuencias de ADN 111 en la base de datos de referencia 11 y/o se asocia con una especie. En consecuencia, las anotaciones de clasificación 112 asignadas a secuencias de ADN 111 son evaluadas y evaluadas de nuevo continuamente y de forma repetida, por ejemplo, en función de los cambios en la base de datos de referencia 11 y/o la base de datos del sistema de clasificación 7.

En la etapa S2, el módulo comparador 122 genera una matriz para comparar las secuencias de ADN 111 (seleccionadas). Dependiendo de las realizaciones, se genera una matriz común para todas las secuencias de ADN 111, o se generan diferentes matrices para cada especie.

En la etapa S3, el módulo comparador 122 compara las secuencias de ADN 111 (seleccionadas). En primer lugar las secuencias de ADN respectivas se alinean automáticamente en la etapa S31.

La Figura 5 muestra un ejemplo de alineación de once secuencias (por ejemplo, secuencias ribosómicas bacterianas, utilizadas comúnmente para la identificación y la taxonomía de especies basadas en secuencias bacterianas) que representan "Abiotrophia defectiva". Como se puede observar en la Figura 5, estas secuencias no son idénticas; tienen diferencias o mutaciones que pueden reflejar o bien errores de la secuenciación o reflejar variaciones intraes-

pecíficas o intragenómicas verdaderas. A través de la alineación de estas secuencias, es evidente que estas variaciones se agrupan frecuentemente y que es posible determinar una secuencia que representa de forma excelente la alineación (en esta memoria, AY879307) y, por lo tanto, también se consideran las especies bacterianas con la anotación "Abiotrophia defectiva", con respecto a todas las secuencias de ADNr 16S de "Abiotrophia defectiva" publicadas.

En la etapa S32, el módulo comparador 122 determina una puntuación de similitud entre las secuencias de ADN 111 alineadas, por ejemplo una puntuación expresada como un porcentaje de la correspondencia entre secuencias. Las puntuaciones de similitud entre las secuencias de ADN 111 (seleccionadas) se almacenan en la matriz. Hay que destacar que la puntuación de similitud se puede determinar utilizando varios algoritmos de alineación diferentes, por ejemplo, algoritmos de alineación por parejas, globales, locales, ponderados y/o basadas en el perfil, y teniendo en cuenta otros elementos de las anotaciones además de la información de la clasificación.

En la etapa S4, la secuencia(s) centroide(s) C se determina para las secuencias de ADN 111 (seleccionadas). En primer lugar, en la etapa S41, el módulo comparador 122 determina una medida de distancia entre las secuencias de ADN 111 (seleccionadas) respectivas. La medida de distancia se determina basándose en las puntuaciones de similitud entre las secuencias de ADN 111 alineadas. En una realización, la medida de distancia se determina entre las secuencias de ADN 111 dentro de una especie. Preferiblemente, las mediciones de distancia entre las secuencias de ADN 111 (seleccionadas) se almacenan en la matriz.

Por ejemplo, la medida de distancia *dist(x, y)* entre dos secuencias de ADN *x* e *y,* se calcula mediante la determinación de un valor complementario de la puntuación de similitud, por ejemplo, *dist(x, y)* = 1 - *puntuación(x, y)*. Preferiblemente, la medida de distancia *dist(x, y)* entre dos secuencias de ADN *x* e *y,* se calcula mediante la determinación de un valor complementario de una puntuación ponderada de similitud, por ejemplo, en donde restando la puntuación ponderada de similitud de una, en donde la puntuación ponderada de similitud se calcula dividiendo la puntuación de similitud entre las dos secuencias de ADN alineadas *x* e *y,* entre la longitud menor l_x, l_y de las dos secuencias de ADN *x* e *y*:

$$dist(x,y) = 1 - \frac{puntuación(x,y)}{min(l_x, l_y)}$$

En la etapa S42, basándose en las medidas de distancia, el detector de centroide 123 determina la secuencia(s) centroide C para las secuencias de ADN 111 (seleccionadas). Esencialmente, para cada una de las especies agrupadas, la secuencia centroide C es la secuencia de ADN en el grupo que tiene la medida global menor de distancia *D* a las otras secuencias de ADN en el grupo. Alternativamente, una secuencia centroide C se define como un objeto virtual que se determina para tener la medida de distancia más corta a todas las secuencias de ADN en el grupo. En otras palabras, *c* es la secuencia centroide de un conjunto de secuencias *S*, si para todas las *N* secuencias *s* en el conjunto *S* diferentes de *c*:

$$D(c) < D(s),$$

en donde

$$D(s_i) = \sum_{j=1}^{N} dist(s_i, s_j).$$

Puede haber más de una secuencia centroide C (congruente) para las secuencias de ADN que tienen las mismas medidas de distancia.

La Figura 4 muestra un ejemplo de diez secuencias de ADN 50-59, que representan "Abiotrophia defectiva" tal y como se muestra en la Figura 5, con sus respectivas medidas de distancia *dist(x, y)* a la secuencia centroide C ("AY879307").

En la etapa S5, el módulo de valoración 124 asigna a las secuencias de ADN 111 (seleccionadas) la medida de distancia *dist(x, y)* entre la secuencia de ADN *i* respectiva y la secuencia centroide C, como un nivel de confianza cuantitativo para la anotación de clasificación asignada a la secuencia de ADN respectiva. Cuanto más pequeña sea la medida de la distancia asociada con una secuencia, mayor será la probabilidad de que esta secuencia particular esté próxima al centroide y que por lo tanto su anotación sea correcta. Por lo tanto, un valor pequeño de la medida de distancia *dist(x, y)* indica un nivel de confianza elevado, mientras que un valor alto de la medida de distancia *dist(x, y)* indica un nivel de confianza bajo. Un experto en la técnica entenderá, que el nivel de confianza asignado a las secuencias de ADN 111 (seleccionadas) se puede expresar alternativamente como un valor cuantitativo complementario de la medida de distancia *dist(x, y)* o como un valor de confianza cualitativo obtenido a partir de la medida de distancia *dist(x, y)*, por ejemplo, a partir de un conjunto de atributos verbales (por ejemplo, "muy alto", "alto", "medio", "bajo", "muy bajo") o un conjunto de colores.

En una etapa opcional S6, el detector de errores 125 identifica valores atípicos entre las secuencias de ADN de una especie. Los valores atípicos tienen la mayor medida de distancia a la secuencia centroide C de la especie respectiva. Por ejemplo, en la Figura 4, la secuencia de ADN 59 ("AJ496329") se detectaría como un valor atípico. En una realización, cualquier secuencia de ADN que tiene una medida de distancia a la secuencia centroide C superior a un valor umbral definido o a una desviación estándar, se determina como un valor atípico. En una realización, los valores atípicos se identifican y se retiran antes de determinar las secuencias centroides (de nuevo).

Posteriormente, en la etapa S7, el detector de errores 125 determina si un valor atípico detectado tiene o no una medida menor de distancia a una secuencia centroide de otra especie. Si ese es el caso, en la etapa S8, la anotación de clasificación del valor atípico se marca como incorrecta en la base de datos de referencia 11, por ejemplo, estableciendo un campo indicador. Además, en una realización, la anotación de clasificación de la secuencia centroide más próxima, se almacena asignada al valor atípico como una anotación de clasificación propuesta.

En una etapa S9 opcional adicional, aparte de los valores atípicos, el detector de centroide 123 asigna la anotación de clasificación asociada con una secuencia centroide C, a las secuencias de ADN 50-58 asociadas con esa secuencia centroide C

Si un usuario accede al sistema informático 1 para hacer una búsqueda en la base de datos de referencia 11 con una muestra de una secuencia de ADN transferida, por ejemplo, utilizando datos de secuencias de fragmentos de ADN procedentes de una muestra de ADN de un secuenciador 4 o de otra fuente, se muestra al usuario una interfaz de usuario con una lista de posibles coincidencias 6, como se muestra en la Figura 6, por ejemplo. Como se puede observar en la Figura 6, cada registro de la lista se proporciona con su respectiva medida de distancia (dist) al centroide C como un indicador del nivel de confianza. Típicamente, la lista se presenta con una valoración mediante similitud y el nivel de confianza es utilizado por un usuario como una medida de la fiabilidad de la anotación de clasificación respectiva. Por otra parte, los valores atípicos se pueden marcar visualmente en la lista, por ejemplo, resaltándolos o coloreándolos, mostrándolos u ocultándolos selectivamente de la lista, y anotaciones de clasificación alternativas que tienen un nivel de confianza mejor se puede mostrar, por ejemplo, como una propuesta de una clasificación más adecuada. El nivel de los valores de confianza se puede incluir adicionalmente y mostrar en cualquier agrupación, alineación o listas valoradas de secuencias de ADN, así como en los árboles filogenéticos, por ejemplo.

La Figura 3 muestra una secuencia ejemplar de etapas para un modo amplio de determinación de las secuencias centroides de las secuencias de ADN 111 (seleccionadas). Esencialmente, la etapa S40 es un enfoque alternativo o complementario de la detección del centroide realizada en la etapa S4. El procesamiento de la etapa S40 se puede activar después de la selección o detección a través del usuario de un nivel de complejidad mediante el detector de centroide 123. El nivel de complejidad se puede indicar, por ejemplo, mediante al menos un número definido de secuencias de ADN que tienen una medida de distancia entre ellas mismas, superior a un valor umbral de complejidad.

En la etapa S401, empleando las puntuaciones de similitud almacenadas en la matriz, el generador de grafos 126 genera un grafo de aristas ponderado 5. Los nodos en el grafo son representativos de las secuencias de ADN (seleccionadas) C, 50-59. Inicialmente, los nodos están conectados si la puntuación de similitud entre las secuencias de ADN respectivas es positiva, es decir, si no es cero. Un valor umbral de conectividad inicial se puede establecer para la puntuación de similitud para asegurar que los nodos forman un grafo coherente. Una medida de la distancia entre las secuencias de ADN respectivas se asigna en cada caso como un peso de la arista entre los respectivos nodos. La medida de la distancia se calcula, por ejemplo, tal y como se ha descrito anteriormente en el contexto de la etapa S41.

En la etapa S402, el generador de grafos 126 calcula las densidades de conectividad locales para los nodos en el grafo. La densidad de conectividad local de un nodo se define por el número de conexiones con otros nodos en el grafo.

En la etapa S403, el generador de grafos 126 define agrupaciones de nodos en el grafo. Las agrupaciones se definen a través de la agregación progresiva hasta máximos de densidad de conectividad local en el grafo. Esencialmente, la medida de distancia entre las secuencias de ADN asociadas con nodos dentro de una agrupación, es significativamente menor que una medida promedio de distancia entre las secuencias de ADN asociadas con los nodos del grafo. Un valor umbral de agrupación inicial (que permite una gran distancia dentro de la agrupación) se puede definir para la medida de distancia entre secuencias de ADN asociadas con nodos de una agrupación, de modo que todo el grafo forma solo una agrupación.

En la etapa S404, la agrupación se muestra a través de la interfaz de usuario 1211 a un usuario en la pantalla 33 del terminal de entrada de datos 3.

En la etapa S405, opcionalmente, se recibe un valor alternativo para el valor umbral de agrupación a través de la interfaz de usuario 1211, del usuario en el terminal de entrada de datos 3. Si en la etapa S406 se determina que se ha recibido un nuevo valor umbral de agrupación del usuario, el generador de grafos 126 define las agrupaciones en la etapa S403 utilizando el nuevo valor umbral de agrupación como una distancia máxima dentro de la agrupación.

Posteriormente, se muestra el grafo con la agrupación recién definida en la etapa S404. Si en la etapa S406 se determina que no se ha recibido ningún nuevo valor umbral de agrupación desde el usuario, el procesamiento continúa en la etapa S407.

En la etapa S407, el detector de centroide 123 determina la secuencia(s) centroide C para una o varias agrupaciones del grafo. Para cada agrupación, el detector de centroide 123 determina la secuencia de ADN asociada con el nodo que tiene la densidad de conectividad más elevada en la agrupación como la secuencia centroide C de esa agrupación. Posteriormente, el procesamiento continúa en la etapa S5 tal y como se ha descrito anteriormente con referencia a la Figura 2.

Cabe señalar que, en la descripción, el código del programa informático se ha asociado con módulos funcionales específicos y la secuencia de las etapas ha sido presentada en un orden específico, sin embargo, un experto en la técnica entenderá que el código del programa informático se puede estructurar de manera diferente y que el orden de al menos algunas de las etapas se puede alterar, sin desviarse del alcance de la invención. También hay que señalar que el método y el sistema propuestos no solo se pueden utilizar para la evaluación sin conexión a internet de anotaciones de clasificación en una base de datos, sino también en línea (en tiempo real o casi en tiempo real), por ejemplo, como un filtro para la introducción de la anotación de clasificación para un nueva secuencia de ADN que se va a añadir a una base de datos.

REIVINDICACIONES

1. Un método implementado por ordenador para la evaluación de anotaciones de clasificación (112) asignadas a secuencias de ADN (111) almacenadas en una base de datos (11), comprendiendo el método:

agrupar (S1) las secuencias de ADN (111) basándose en sus anotaciones de clasificación respectivas (112) por especies utilizando sistemas de clasificación establecidos;

determinar para parejas de secuencias de ADN (111) una medida de distancia entre las secuencias de ADN respectivas (111) mediante la alineación (S31) de forma automática de las secuencias de ADN respectivas (111) y determinar la medida de distancia (S41) basada en una puntuación de similitud entre las secuencias de ADN alineadas (111);

determinar una secuencia centroide (S4, S40), teniendo la secuencia centroide (C) una medida global más corta de distancia a las secuencias de ADN (111); y

asignar (S5) a las secuencias de ADN (111) la medida de distancia entre la secuencia de ADN respectiva y la secuencia centroide (C) como un nivel de confianza cuantitativo para la anotación de clasificación asignada a la secuencia de ADN respectiva;

en donde la medida de distancia se determina entre las secuencias de ADN (111) dentro de una especie; y las secuencias centroides se determinan para las secuencias de ADN (111) dentro de cada una de las especies.

2. El método según la reivindicación 1, que comprende además la identificación de valores atípicos dentro de la especie, teniendo los valores atípicos la mayor medida de distancia a la secuencia centroide (C) de la especie respectiva, y marcando las anotaciones (112) como incorrectas para valores atípicos que tienen una medida de distancia más pequeña a una secuencia centroide (C) de otra especie.

3. El método según una de las reivindicaciones 1 o 2, en el que el método comprende además generar (S401) a partir de las puntuaciones de similitud entre las secuencias de ADN (111) un grafo de aristas ponderado, siendo las secuencias de ADN (111) nodos en el grafo, estando conectados los nodos si la puntuación de similitud entre las secuencias de ADN respectivas (111) es positiva, y estando asignada la medida de distancia entre las secuencias de ADN respectivas (111) en cada caso como un peso de la arista; calcular (S402) las densidades de la conectividad local para los nodos en el grafo; y definir agrupaciones (S403) de nodos a través de la agregación progresiva hasta una densidad de conectividad local máxima, siendo la medida de distancia entre las secuencias de ADN (111) asociadas con nodos dentro de una agrupación significativamente más corta que una medida promedio de distancia entre las secuencias de ADN (111) asociadas con los nodos del grafo, en donde preferiblemente el método comprende adicionalmente recibir un valor umbral de agrupación (S405) de un usuario, que responde a mostrar el grafo en una pantalla (33); definir las agrupaciones (S403) de nodos aplicando el valor umbral de agrupación como una distancia máxima dentro de la agrupación; y mostrar el grafo (S404) en la pantalla (33) después de aplicar el valor umbral de agrupación.

4. El método según la reivindicación 3, en el que la secuencia de ADN asociada con el nodo que tiene la mayor densidad de conectividad en una agrupación se define como la secuencia centroide (C) de esa agrupación.

5. El método según una de las reivindicaciones 1 a 4, en el que la anotación de clasificación asociada con una secuencia centroide (C) se asigna a secuencias de ADN (111) asociadas con esa secuencia centroide (C).

6. El método según una de las reivindicaciones 1 a 5, en el que la determinación de la medida de distancia (S41) entre dos secuencias de ADN (111) incluye el cálculo de una puntuación ponderada de similitud dividiendo la puntuación de similitud entre las dos secuencias de ADN (111) entre la menor longitud de las dos secuencias de ADN (111), y restando la puntuación ponderada de similitud de una.

7. Un sistema informático (1) para la evaluación de las anotaciones de clasificación (112) asignadas a secuencias de ADN (111), comprendiendo el sistema (1):

una base de datos (11) que comprende una pluralidad de secuencias de ADN (111);

un módulo comparador (122) configurado para agrupar las secuencias de ADN (111) basándose en sus anotaciones de clasificación respectivas (112) por especie utilizando sistemas de clasificación establecidos (7), y para determinar para parejas de las secuencias de ADN (111) una medida de distancia entre las secuencias de ADN respectivas (111) mediante la alineación automática de las secuencias de ADN respectivas (111) y determinar la medida de distancia basada en una puntuación de similitud entre las secuencias de ADN alineadas (111);

un detector de centroide (123) configurado para determinar una secuencia centroide (C), teniendo la secuencia centroide (C) la medida global menor de distancia a las secuencias de ADN (111); y

un módulo de valoración (124) configurado para asignar a las secuencias de ADN (111) la medida de distancia entre la secuencia de ADN respectiva y la secuencia centroide (C) como un nivel de confianza cuantitativo para la anota-

ción de clasificación asignada a la secuencia de ADN respectiva;

en donde el módulo comparador (122) se configura adicionalmente para determinar la medida de distancia entre las secuencias de ADN (111) dentro de una especie; y el detector de centroide (123) se configura adicionalmente para determinar las secuencias centroides (C) para las secuencias de ADN (111) dentro de cada una de las especies

8. El sistema (1) según la reivindicación 7, que comprende además un detector de errores (125) configurado para identificar valores atípicos dentro de las especies, teniendo los valores atípicos la mayor medida de distancia a la secuencia centroide (C) de la especie respectiva, y para marcar anotaciones (112) como incorrectas para valores atípicos que tienen una medida de distancia más pequeña a una secuencia centroide (C) de otra especie.

9. El sistema (1) según una de las reivindicaciones 7 u 8, en el que el sistema (1) comprende además un generador de grafos (126) configurado para generar a partir de las puntuaciones de similitud entre las secuencias de ADN (111) un grafo de aristas ponderado, siendo las secuencias de ADN (111) nodos en el grafo, estando conectados los nodos si la puntuación de similitud entre las secuencias de ADN respectivas (111) es positiva, y estando asignada en cada caso la medida de la distancia entre las secuencias de ADN respectivas (111) como un peso de arista, para calcular densidades de conectividad locales para los nodos en el grafo, y para definir agrupaciones de nodos a través de la agregación progresiva hasta máximos de densidad de conectividad local, siendo la medida de distancia entre las secuencias de ADN (111) asociada con nodos dentro de una agrupación significativamente más corta que una medida promedio de distancia entre las secuencias de ADN (111) asociadas con los nodos del grafo.

10. El sistema (1) según la reivindicación 9, en el que el sistema (1) comprende además una interfaz de usuario (1211) configurada para recibir un valor umbral de agrupación de un usuario, que responde a mostrar el grafo en una pantalla (33); el generador de grafos (126) está configurado además para definir las agrupaciones de nodos mediante la aplicación del valor umbral de agrupación como una distancia máxima dentro de una agrupación, y para mostrar el grafo en la pantalla (33) después de aplicar el valor umbral de agrupación.

11. El sistema (1) según una de las reivindicaciones 9 o 10, en el que el detector de centroide (123) está configurado además para definir la secuencia de ADN asociada con el nodo que tiene la densidad de conectividad más alta en una agrupación, como la secuencia centroide (C) de esa agrupación.

12. El sistema (1) según una de las reivindicaciones 7 a 11, en el que el detector de centroide (123) está configurado además para asignar la anotación de clasificación asociada con una secuencia centroide (C) a secuencias de ADN (111) asociadas con esa secuencia centroide (C).

13. El sistema (1) según una de las reivindicaciones 7 a 12, en el que el módulo comparador (122) está configurado además para determinar la medida de distancia entre dos secuencias de ADN (111) restando una puntuación ponderada de similitud de una, la puntuación ponderada de similitud se calcula dividiendo la puntuación de similitud entre las dos secuencias de ADN (111) entre la longitud más corta de las dos secuencias de ADN (111).

14. Un producto de programa informático que comprende medios de código de programa informático para controlar uno o varios procesadores de un sistema informático (1), de tal manera que el sistema informático (1) realiza el método según una de las reivindicaciones 1 a 6.

15 El producto de programa informático según la reivindicación 14, que comprende además un medio legible por ordenador que contiene los medios de código de programa informático.

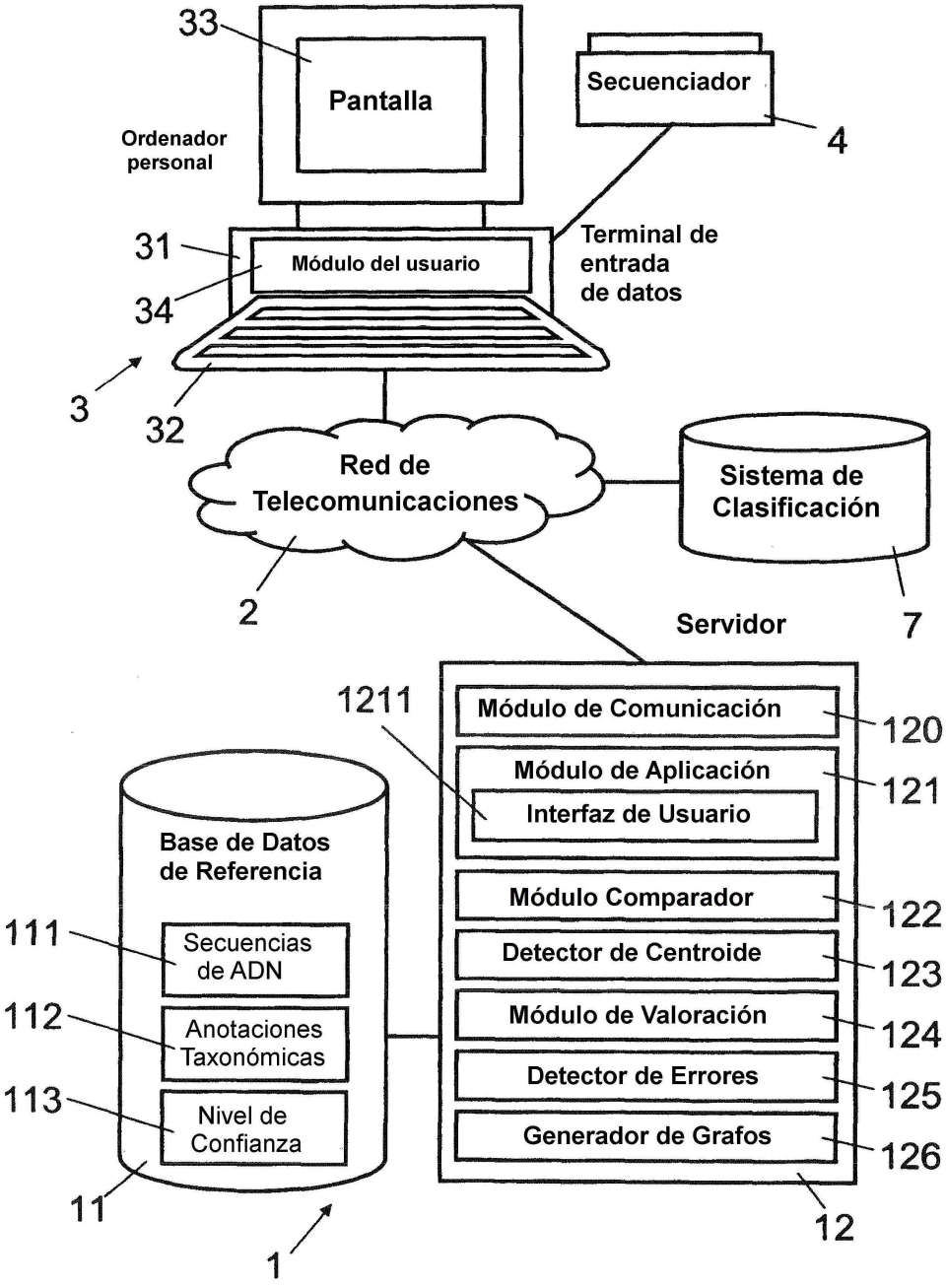
33
Pantalla
Ordenador personal
Secuenciador
4
31
Módulo del usuario
34
Terminal de entrada de datos
3
32
Red de Telecomunicaciones
Sistema de Clasificación
2
7
Servidor
1211
Módulo de Comunicación
120
Módulo de Aplicación
121
Interfaz de Usuario
Módulo Comparador
122
Detector de Centroide
123
Módulo de Valoración
124
Detector de Errores
125
Generador de Grafos
126
Base de Datos de Referencia
111
Secuencias de ADN
112
Anotaciones Taxonómicas
113
Nivel de Confianza
11
1
12

Fig. 1

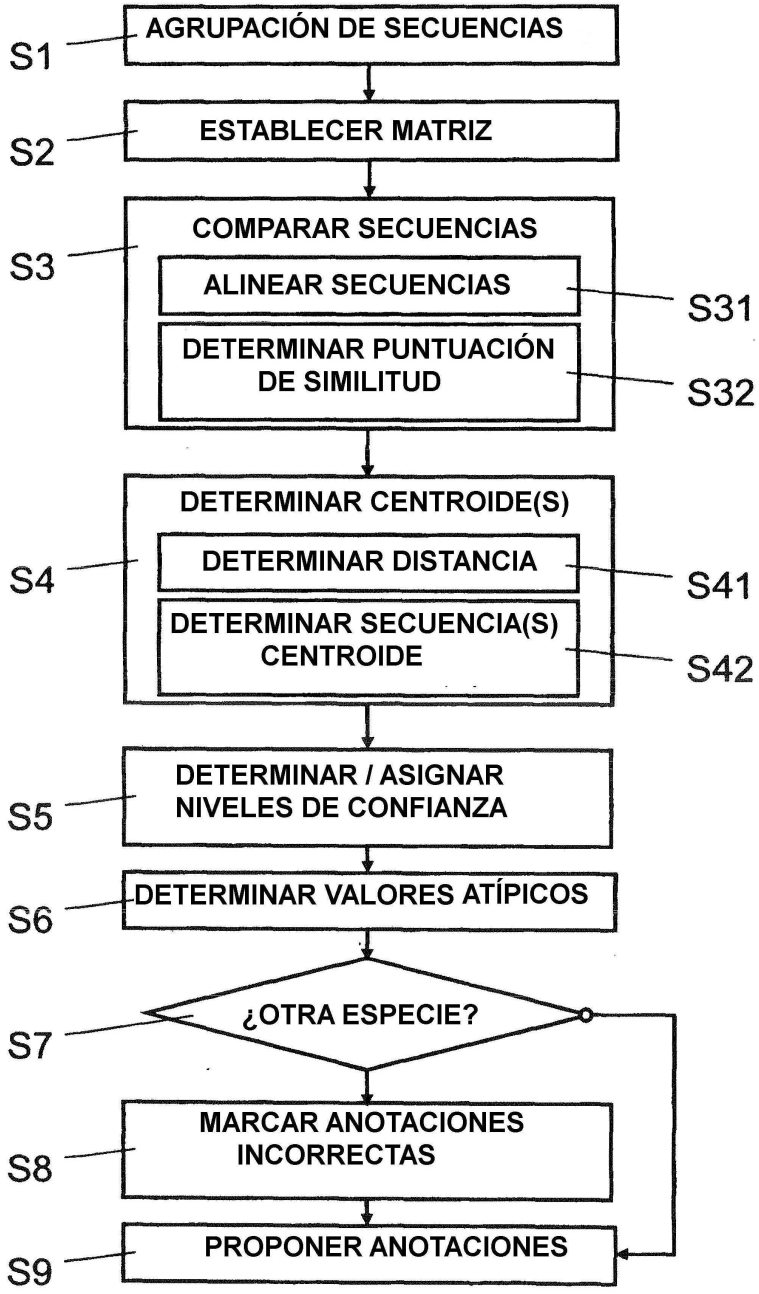
S1
AGRUPACIÓN DE SECUENCIAS
S2
ESTABLECER MATRIZ
S3
COMPARAR SECUENCIAS
ALINEAR SECUENCIAS
S31
DETERMINAR PUNTUACIÓN DE SIMILITUD
S32
S4
DETERMINAR CENTROIDE(S)
DETERMINAR DISTANCIA
S41
DETERMINAR SECUENCIA(S) CENTROIDE
S42
S5
DETERMINAR / ASIGNAR NIVELES DE CONFIANZA
S6
DETERMINAR VALORES ATÍPICOS
S7
¿OTRA ESPECIE?
S8
MARCAR ANOTACIONES INCORRECTAS
S9
PROPONER ANOTACIONES

Fig. 2

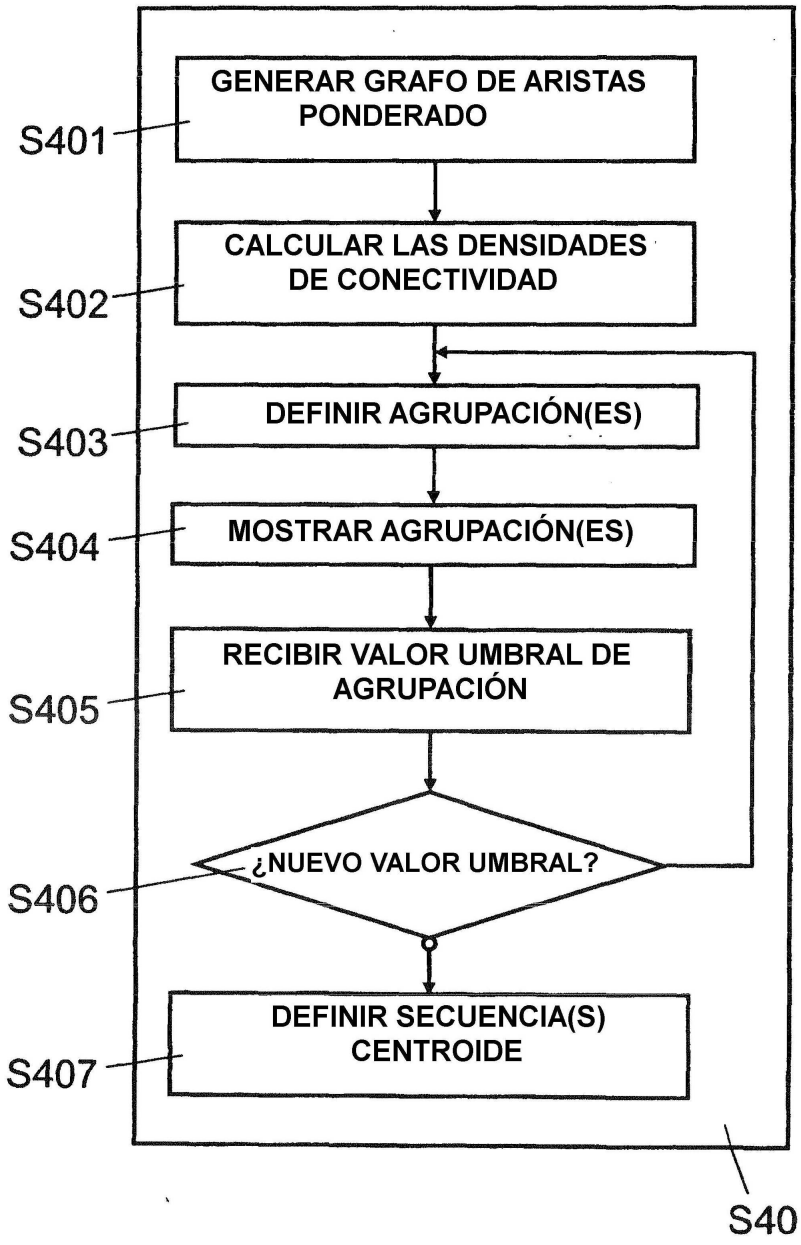
GENERAR GRAFO DE ARISTAS PONDERADO
S401
CALCULAR LAS DENSIDADES DE CONECTIVIDAD
S402
DEFINIR AGRUPACIÓN(ES)
S403
MOSTRAR AGRUPACIÓN(ES)
S404
RECIBIR VALOR UMBRAL DE AGRUPACIÓN
S405
¿NUEVO VALOR UMBRAL?
S406
DEFINIR SECUENCIA(S) CENTROIDE
S407
S40

Fig. 3

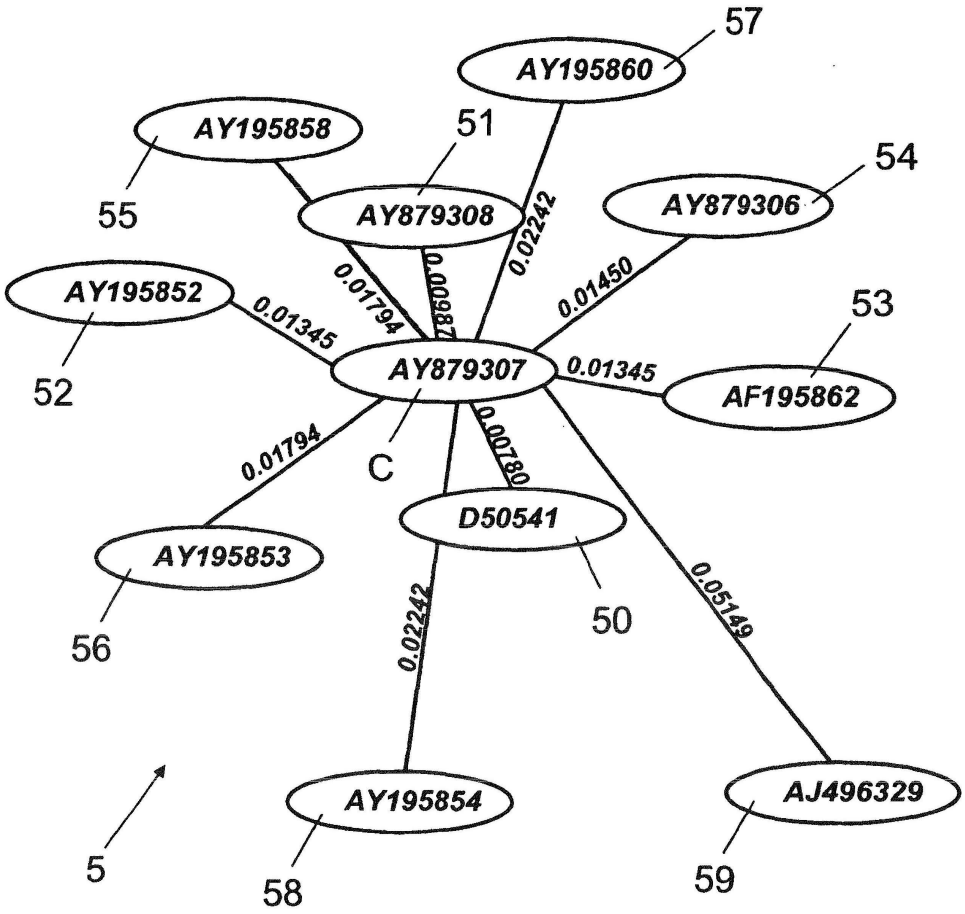
57
AY195860
AY195858
51
55
AY879308
54
AY879306
0.02242
0.01794
0.00887
0.01450
AY195852
0.01345
53
52
AY879307
0.01345
AF195862
0.01794
C
0.00780
D50541
AY195853
56
50
0.05149
0.02242
AY195854
AJ496329
5
58
59

Fig. 4

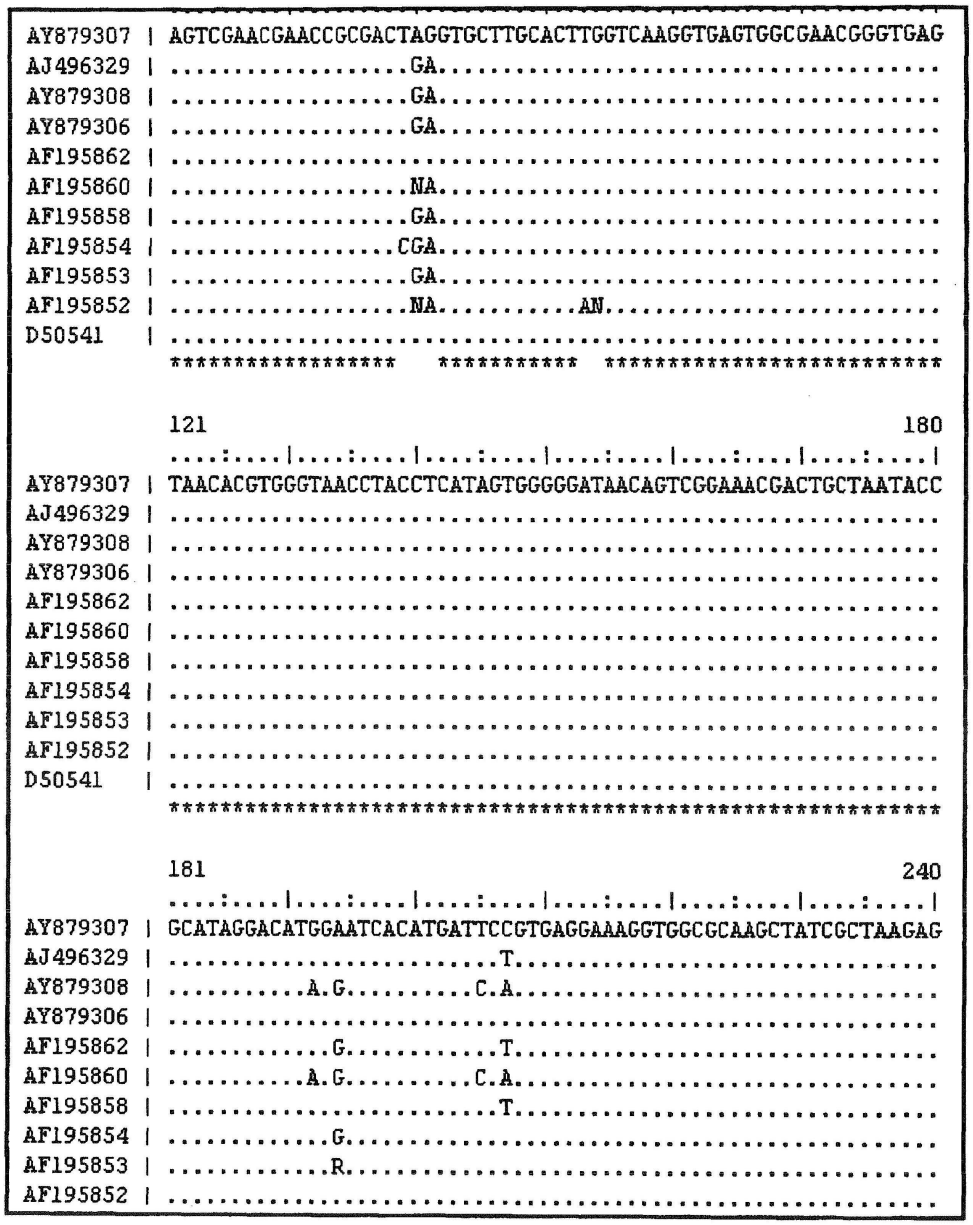
AY879307 | AGTCGAACGAACCGCGACTAGGTGCTTGCACTTGGTCAAGGTGAGTGGCGAACGGGTGAG
AJ496329 | .....GA.....
AY879308 | .....GA.....
AY879306 | .....GA.....
AF195862 | .....
AF195860 | .....NA.....
AF195858 | .....GA.....
AF195854 | .....CGA.....
AF195853 | .....GA.....
AF195852 | .....NA.....AN.....
D50541 | .....
121 180
AY879307 | TAACACGTGGGTAACCTACCTCATAGTGGGGGATAACAGTCGGAAACGACTGCTAATACC
AJ496329 | .....
AY879308 | .....
AY879306 | .....
AF195862 | .....
AF195860 | .....
AF195858 | .....
AF195854 | .....
AF195853 | .....
AF195852 | .....
D50541 | .....
181 240
AY879307 | GCATAGGACATGGAATCACATGATTCCGTGAGGAAAGGTGGCGCAAGCTATCGCTAAGAG
AJ496329 | .....T.....
AY879308 | .....A.G.....C.A.....
AY879306 | .....
AF195862 | .....G.....T.....
AF195860 | .....A.G.....C.A.....
AF195858 | .....T.....
AF195854 | .....G.....
AF195853 | .....R.....
AF195852 | .....

Fig. 5

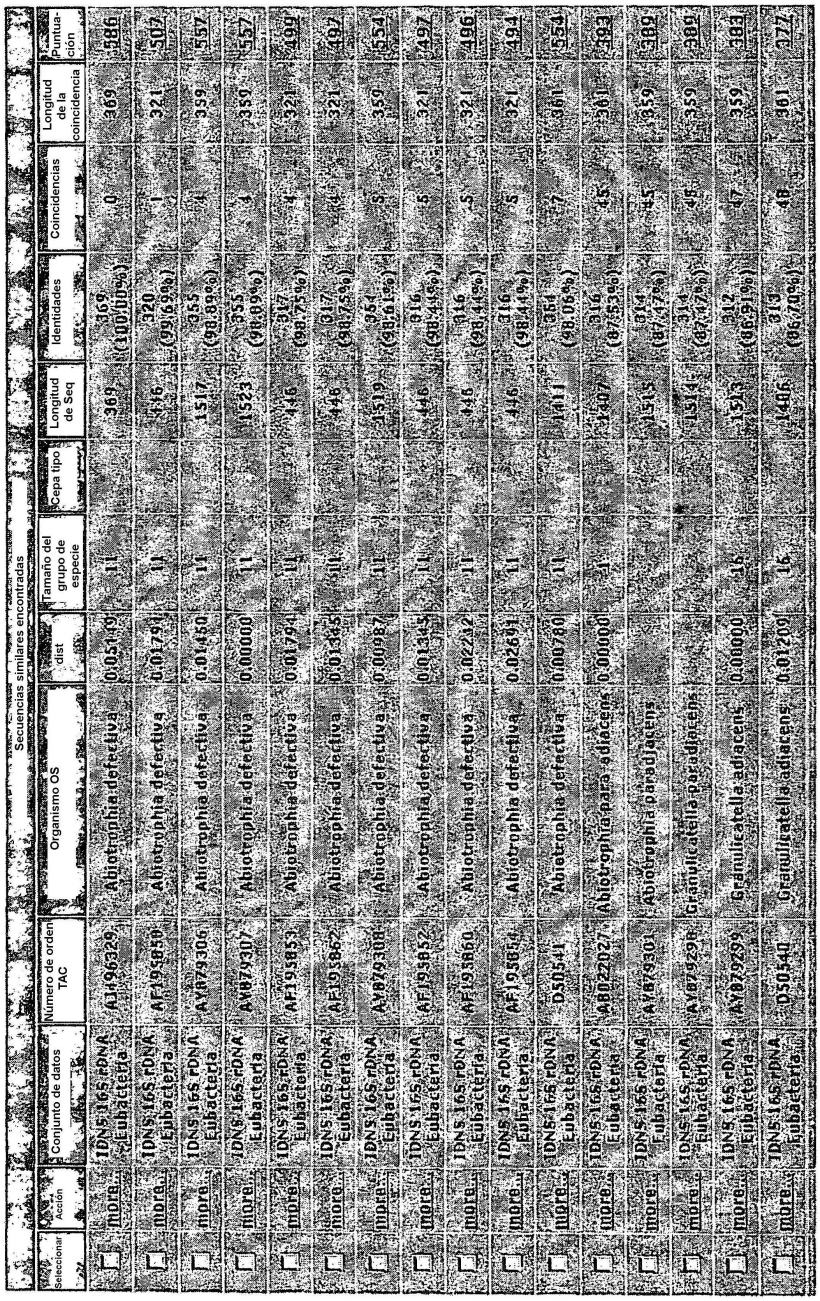
Secuencias similares encontradas
Seleccionar
Acción
Conjunto de datos
Número de orden TAC
Organismo OS
dist
Tamaño del grupo de especie
Cepa tipo
Longitud de Seq
Identidades
Coincidencias
Longitud de la coincidencia
Puntuación
more...
IDNS 16S rDNA Eubacteria
AJ496329
Abiotrophia defectiva
0.00519
11
369
369 (100.00%)
0
369
586
AF195850
Abiotrophia defectiva
0.01791
11
416
320 (99.69%)
1
321
502
AY879306
Abiotrophia defectiva
0.01450
11
1517
355 (98.89%)
4
359
557
AY879307
Abiotrophia defectiva
0.00000
11
1523
355 (98.89%)
4
359
557
AF195853
Abiotrophia defectiva
0.01794
11
446
317 (98.75%)
4
321
499
AF195867
Abiotrophia defectiva
0.01345
11
446
317 (98.75%)
4
321
497
AY879308
Abiotrophia defectiva
0.00987
11
1519
351 (98.61%)
5
359
554
AF195852
Abiotrophia defectiva
0.01345
11
446
316 (98.44%)
5
321
497
AF193860
Abiotrophia defectiva
0.02232
11
446
316 (98.44%)
5
321
496
AF195854
Abiotrophia defectiva
0.02681
11
446
316 (98.44%)
5
321
494
D50541
Abiotrophia defectiva
0.00760
11
1411
354 (98.06%)
7
361
554
AB022027
Abiotrophia para adiacens
0.00000
1
1407
316 (87.53%)
45
361
393
AY879301
Abiotrophia para adiacens
1515
314 (87.47%)
45
359
389
AY879298
Granulicatella para adiacens
1514
314 (87.47%)
45
359
389
AY879299
Granulicatella adiacens
0.00000
11
1513
312 (86.91%)
47
359
384
D50540
Granulicatella adiacens
0.01200
16
1406
313 (86.70%)
48
361
377
6

Fig. 6