明　細　書

発明の名称 ： 微生物叢のメタゲノム解析

技術分野

[0001] 本発明は微生物叢のメタゲノム解析を含み、特に細菌叢のメタ16SリボソームRNA（16S rRNA）解析、以下メタ16S解析という場合がある、を含む。

背景技術

[0002] 非特許文献１は子宮内の微生物叢、特に細菌叢（flora, microbiota, microbiome）が体外受精の成功率に影響を与えることを示している。特許文献１は、羊水に含まれる細菌から抽出した全ゲノムDNAに対して、各細菌共通の16Sリボソーム RNA遺伝子のユニバーサルプライマーを利用して16S rRNA遺伝子を特異的にPCR増幅する。さらに増幅産物に対してOperational Taxonomic Unit、以下OTUという、を用いた解析と他の解析とによって絨毛膜羊膜炎関連微生物を同定することを開示している。特許文献２は妊婦又は新生児のリスクを細菌分類群に基づいて決定することを開示している。

先行技術文献

特許文献

[0003] 特許文献1：特開２０１８－０８８８４３号公報

特許文献2：特表２０１７－５１９５１４号公報

非特許文献

[0004] 非特許文献1：Inmaculada Moreno, Francisco M. Codoner, Felipe Vilella, Diana Valbuena, Juan F. Martinez-Blanch, Jorge Jimenez-Almazan, Roberto Alonso, Pilar Alama, Jose Remohi, Antonio Pellicer, Daniel Ramona and Carlos Simon, “Evidence that the endometrial microbiota has an effect on implantation success or failure”, American Journal of Obstetrics and Gynecology, December 01, 2016, Volume 215, Issue 6, p.684-703

発明の概要

[0005] 発明者らは子宮内微生物叢及びその他一般の微生物叢をメタゲノム解析する方法を考案した。係る解析法では配列の相同性の閾値に基づくクラスタリングを利用する。

[0006] 発明者らの考案した解析法の一例ではいわゆる探索的クラスタリングを用いる。探索的クラスタリングではデータを割り当てるべきクラスターの変更とクラスターの重心（セントロイド）の変更を繰り返す。一例としてクラスターの割り当てに変化が無くなれば、そこで処理を終了する。その時点のクラスターをOTUとして、その重心を代表配列（Representative Sequence）とする。この代表配列を元にデータベースに照合して、OTUの生物種を判定する。

[0007] このような代表配列はしばしば疑似配列、すなわち実際には微生物叢の中の微生物が有してしない配列である可能性が高い。したがってOTUに対する判定にエラーを生じる可能性がある。本発明は微生物叢のメタゲノム解析において配列の相同性の閾値に基づくクラスタリングを行う場合に、疑似配列の影響を減らすのに適した手段を提供することを目的とする。

[0008] ［１］ 細菌叢のメタゲノム解析の方法であって、

細菌叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングを行い：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、前記代表配列をデータベースと照合することで、その二次OTUの系統推定を行う方法。

[0009] ［２］ 前記二次OTUの生成に際して、前記サンプル配列に対して前記一次OT

Uとは異なるクラスターを初めに割り当て、さらに前記クラスターの重心の配列に対して所定の閾値以上の相同性を有するサンプル配列に前記クラスターを割り当て直すことを、収束するまで繰り返す、

［１］の方法。

[0010] ［３］　前記サンプル配列に対してクラスターを初めに割り当てる際、前記サンプル配列からランダムにシード配列を選択するとともに、前記シード配列と前記シード配列に対して前記所定の閾値以上の相同性を有する配列とに最初のクラスターを割り当てる、

［２］の方法。

[0011] ［４］　前記一次OTUを作成した後に、前記二次OTUを生成するところ、

前記クラスターは前記一次OTUを包含するクラスターであり、

前記重心の配列は前記一次OTUの配列に前記一次OTUの配列数で重みづけをした平均である、

［２］又は［３］の方法。

[0012] ［５］　前記クラスターを割り当て直した際、全てのクラスターの割り当てが変化しなかった場合に、これを収束とする、

［２］～［４］のいずれかの方法。

[0013] ［６］　前記サンプルは被験者の子宮内腔液又は膣スワブであり、

前記サンプル配列は前記サンプルより抽出したゲノムDNAの16S rRNA遺伝子の部分配列をDNAシーケンシングして生成したものであり、

前記データベースはラクトバチルス属（Lactobacillus）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記系統推定において、前記二次OTUの細菌の属を推定し、

前記被験者の識別子と前記ラクトバチルス属を含む細菌の属の細菌叢中における構成比とを含むデータを生成する、

［１］～［５］のいずれかの方法。

[0014] ［７］　前記データベースはさらに、ビフィドバクテリウム属（Bifidobacterium）、プロピオニバクテリウム属（Propionibacterium）、ガードネレラ属

（Gardnerella）、ストレプトコッカス属（Streptococcus）及びベイロネラ属（Veillonella）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記部分配列は、16S rRNA遺伝子の全体の配列の中で、これらの細菌の属の間で完全には保存されていない部分の配列である、

［６］の方法。

[0015] ［８］　前記データベースは、他の複数のデータベースをメタ解析して生成したデータベースである、

［６］又は［７］の方法。

[0016] ［９］　被験者より前記サンプルとして子宮内腔液又は膣スワブを採取し、

［１］～［５］のいずれかの方法で前記サンプルをメタゲノム解析し、ここで

前記サンプルより抽出したゲノムDNAの16S rRNA遺伝子の部分配列をDNAシーケンシングすることで前記サンプル配列を生成し、

前記データベースはラクトバチルス属（Lactobacillus）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記系統推定において、前記二次OTUの細菌の属を推定し、

ラクトバチルス属及びラクトバチルス属以外の細菌の属の細菌叢中における構成比に基づき、前記被験者の妊娠のしやすさを診断する方法。

[0017] ［１０］　真核生物の微生物叢のメタゲノム解析の方法であって、

微生物叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングを行い：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその

二次OTUの代表配列とし、前記代表配列をデータベースと照合することで、その二次OTUの系統推定を行う方法。

[0018] ［１１］　細菌又は真核生物の微生物叢のメタゲノム解析のプログラムであって、

前記微生物叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングをコンピューターに行わせ：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、前記コンピューターに前記代表配列をデータベースと照合させることで、その二次OTUの系統推定を前記コンピューターに行わせるプログラム。

[0019] 本発明により微生物叢のメタゲノム解析において配列の相同性の閾値に基づくクラスタリングを行う場合に、疑似配列の影響を減らすのに適した手段を提供することができる。

図面の簡単な説明

[0020] [図1]メタゲノム解析のフロー。

[図2]100%相同性によるグループ化。

[図3]最初のクラスターの割り当て。

[図4]クラスターの割り当て直し。

[図5]クラスターの割り当ての繰り返し。

[図6]代表配列の特定。

[図7]16S rRNA遺伝子。

[図8]実施例のフロー。

[図9]細菌の構成比。

[図10]参考例のクラスタリング。

発明を実施するための形態

[0021] 本実施形態では微生物叢、特に細菌叢のメタゲノム解析を行う。メタゲノム解析では非階層的クラスタリング、特に探索的クラスタリングを利用する。探索的クラスタリングを利用することでクラスタリングの精度と計算コストのバランスをとることができる。探索的クラスタリングの一例はk平均法（k-means）である。

[0022] 図１は、メタゲノム解析のフローである。メタゲノム解析の各処理はコンピューターにて機械的に行うが、別個のコンピューターにて各処理を分担してもよく、一個のコンピューターにて各処理を集中的に実行してもよい。ステップＳ１１で細菌叢を含むサンプルをDNAシーケンシングして塩基配列を得る。以下、かかる塩基配列をサンプル配列という。

[0023] 一態様において細菌叢は羊膜腔、子宮内腔、子宮頸管及び腟（膣）のいずれかから採取する。本明細書において「腟」は、英語及びドイツ語のVaginaを表すものとして用いる医学用語である。一態様においてサンプルとなるDNAは子宮頸管スワブ、腟スワブ（膣スワブ）、胎盤スワブ及び臍帯スワブ；尿、羊水、全血、血清、血漿、子宮内腔液及び子宮頸管粘液；並びにそれら以外の組織のいずれかから抽出する。

[0024] 一態様においてDNAシーケンサーがコンピューターにサンプル配列を送る。コンピューターがサンプル配列を自動的に処理する。他の態様においてDNAシーケンサーが、ネットワークを介して接続されたサーバーにサンプル配列を送る。サーバーがサンプル配列を記録する。コンピューターの求めに応じてサーバーがネットワークを介して接続されたコンピューターにサンプル配列を送る。コンピューターがサンプル配列を自動的に処理する。

[0025] 次に図１に示すステップＳ１２において、サンプル配列を100%相同性にてグループ化することで一次OTU（Primary OTU）を得る。一態様において、こ

のグループ化は探索的クラスタリングで行う。次にステップＳ１３でサンプル配列に対して100%未満の相同性閾値にて探索的クラスタリングを行うことで二次OTU（Secondary OTU）を得る。次にステップＳ１４にて二次OTU内の代表配列の系統推定を行うことで、その二次OTUの系統推定を行う。

[0026] 用語「OTU」、Operational Taxonomic Unitは、生物の塩基配列をコンピューター上でその類似度を指標に分類したときに得られる、生物の分類単位をいう。生物が細菌である場合、例えば16S rRNA遺伝子の塩基配列が利用される。OTUの重要な用法の一つは、OTUがリンネの分類学や進化分類学上の分類単位と一致しているかどうかは無関係に、OTUに基づいて実在の生物、例えば細菌叢を構成する細菌、を分類することである。一方でOTUの系統推定を行うことでリンネの分類学や進化分類学上の分類単位とOTUとを関連付けることができる。このように系統推定されたOTUの情報は臨床的に利用しやすい。

[0027] 図１のステップＳ１１でサンプル配列が得られる。サンプル配列は細菌叢に含まれる各種、各系統の細菌の有する塩基配列の集合として得られる。サンプル配列の集合は細菌叢中の各属の細菌の存在比を含んでいる。また個々のサンプル配列はいずれか一つの属や種に由来する。一態様において、この時点では細菌の系統推定はされていない。一態様において、この時点では細菌叢中の各属の細菌の存在比は推定されていない。

[0028] サンプル配列は既知のゲノム配列を考慮して、所定の長さとなるように設定する。ただしDNAシーケンシングのエラーによるものも含めて、欠失や重複の発生が考慮されるため、その長さが完全に一致するわけではない。サンプル配列は種間での保存性が低く、種内では保存性が高いゲノム配列とすることができる。一態様においてサンプル配列を有するゲノムDNA断片は例えばプライマーやアダプターを用いてクローニングされる。

[0029] 図２は、サンプル配列において、100%の相同性を有する塩基配列同士をグループ化することを例示的に表す。上段に示すようにOTU aは100コピー、OTU bは10コピー、OTU cは5コピーの配列をそれぞれ有する。OTU bの配列はOTU aの配列に対して、二塩基の置換（ATTからAGC）が生じている。OTU cの配列

はOTU aに対して二塩基の置換（ATTからTGT）が生じている。このグループ化によって一次OTU（Operational Taxonomic Unit）を生成する。このグループ化は配列相同性の閾値100%のクラスタリングを行うことに等しい。100%の相同性とはサンプル配列の所定の長さの中における配列の完全一致を表す。

[0030] 図２の下段の各柱は各クラスターを表す。すなわち一次OTUを表す。各柱の高さは各一次OTUのコピー数を表している。各一次OTU間の距離が短いほど配列の相同性が高い。なお配列の相同性を二次元平面上に正確に表すことはできないため、図は模式的なものである。

[0031] 図１のステップＳ１２において、一次OTUごとの配列と配列のコピー数（Copy Number）との関係が判明する。配列のコピー数のことを以下配列数という場合がある。一態様において、一次OTU間の配列数の比は、細菌叢内における、これらの配列を有する細菌の存在比を表す。一次OTU以外のその他のOTUにおいて同様である。ただし本実施形態において「コピー数」という用語は、一次OTU及びその他のOTU内の個々のサンプル配列が同一のゲノムを有する細菌から得られた均質なものであることを必ずしも表していない。ゲノム配列上のサンプル配列以外の配列は細菌間で異なる可能性がある。したがって、一次OTU内の細菌が分類単位、種や属、に包含されることは推定である。

[0032] 一態様において、一次OTUを作成した後に二次OTUを作成する。そこで図３から図６に示す手順に従って探索的クラスタリングを行う。

[0033] 図３は、最初のクラスターの割り当てを模式的に表す。ここでは各サンプル配列に対して一次OTUとは異なるクラスターを初めに割り当てる。クラスター割り当ての手順として、まずサンプル配列全体の中からシードとなるサンプル配列、以下シード配列という、をランダムに選択する。実質的には一次OTU全体の中から一次OTUをランダムに選択する。

[0034] 図３に示す一態様では、一次OTU（Ｐｒ１）をシード配列として選択している。破線の円は一次OTU（Ｐｒ１）から見て類似している範囲を表す。すなわち一次OTU（Ｐｒ１）の配列に対して所定の閾値以上の相同性を有する範囲を表す。この範囲に最初のクラスターを割り当てる。

[0035] 図３に示すように一次OTU（Ｐｒ１）と異なる一次OTU（Ｐｒ２）が類似の範囲に含まれている。そこで一次OTU（Ｐｒ１）と一次OTU（Ｐｒ２）とに最初のクラスターを割り当てる。柱にハッチングの付されていることが、同一のクラスターの割り当てられた一次OTUであることを表す。これらの一次OTUが一つのクラスターを構成している。他の図においてハッチングは同様のことを表す。

[0036] 図３には図示していないが、一次OTU（Ｐｒ１）に類似しない他の一次OTUに対しても同様に他のクラスターを割り当てる。以上のようにして、一次OTUを包含するクラスターを生成する。

[0037] 図４は、クラスターの割り当て直しを模式的に表す。重心Ｃｔ１はクラスターの重心すなわち、一次OTU（Ｐｒ１）と一次OTU（Ｐｒ２）との重心である。破線の円は重心Ｃｔ１から見て類似している範囲を表す。すなわち重心Ｃｔ１の配列に対して所定の閾値以上の相同性を有する範囲を表す。新たに生じた重心Ｃｔ１に類似する範囲内の一次OTU（Ｐｒ３）にクラスターを割り当て直す。

[0038] 図５は、クラスターの割り当ての繰り返しを模式的に表す。重心Ｃｔ２は一次OTU（Ｐｒ３）を有するクラスターの重心すなわち、一次OTU（Ｐｒ１）と一次OTU（Ｐｒ２）と一次OTU（Ｐｒ３）の重心である。破線の円は重心Ｃｔ２から見て類似している範囲を表す。すなわち重心Ｃｔ２の配列に対して所定の閾値以上の相同性を有する範囲を表す。新たに生じた重心Ｃｔ２に類似する範囲内の一次OTU（Ｐｒ４）にクラスターを割り当て直す。

[0039] 本実施形態ではクラスターの割り当て直しを、クラスターが収束するまで繰り返す。これにより二次OTUを生成する。

[0040] 図６は、二次OTU（Ｓｃ）及び代表配列の特定を模式的に表す。重心Ｃｔ３は一次OTU（Ｐｒ４）を有するクラスターの重心すなわち、一次OTU（Ｐｒ１）と一次OTU（Ｐｒ２）と一次OTU（Ｐｒ３）と一次OTU（Ｐｒ４）の重心である。破線の円は重心Ｃｔ３から見て類似している範囲を表す。すなわち重心Ｃｔ３の配列に対して所定の閾値以上の相同性を有する範囲を表す。新たに

生じた重心Ｃｔ３に類似する範囲内に、新たにクラスターを割り当てたとしても、クラスターを構成する一次OTUの組み合わせは変化しない。したがって重心Ｃｔ３を得る前にクラスターは収束していたことが分かる。

[0041] 以上のように探索的クラスタリングを行う。探索的クラスタリングにおいて相同性の所定の閾値は100%未満である。一態様において閾値は90%以上の任意の値である。一態様において閾値は91，92，93，94，95，96，97，97.5，98，98.5，99及び99.5%のいずれかである。閾値を低くすることで解析精度を高めることができる。閾値を高くすることで解像度を高めることができる。また計算時間を少なくできる。

[0042] 探索的クラスタリングの収束の結果として得られるクラスターはOTUである。本実施形態において二次OTU（Ｓｃ）と称する。ここで探索的クラスタリングの途中で得られるクラスターもOTUであることに留意する。最初にランダムに選んだシードとこれに類似するサンプル配列とからなるクラスターはOTUである。

[0043] 収束までに行うクラスターの割り当て直しの回数は1回以上である。上記実施形態では2回行っている。この回数は3、4、5、6、7、8及び9回となる場合がある。一態様において割り当て直しの回数に上限を設ける。一態様において割り当て直しの回数が上限に達した時に、クラスターが収束したものと推定する。

[0044] 図４から図６に示す一態様において、クラスターの重心の配列は各一次OTUの配列の単純な平均である。他の態様においてクラスターの重心の配列は各一次OTUの配列に、各一次OTUの配列数で重みづけをした平均でもある。

[0045] 収束時において二次OTUに含まれる一次OTUが１個であることが許容される。また全ての一次OTUと全ての二次OTUとが一致することが許容される。

[0046] 図６に示す一態様において、重心に基づくクラスターの割り当て直しで、不図示のクラスターを含めて全てのクラスターが変化しなかったときに、クラスターが収束したものと推定する。他の態様において、全てのクラスター内の一次OTUの組み合わせの変化量が所定値を下回った場合に、これを収束と

する。

[0047] 図３において一次OTU（Ｐｒ１）をシードとして用いた。そしてシードから最初のクラスターを決定した。これに対して、配列の相同性を考慮せずにランダムに選んだ複数の一次OUTからなるグループを最初のクラスターとしてもよい。

[0048] 図１に示すステップＳ１４において系統推定を行うために、まず各二次OTUの代表配列を選定する。一態様において各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とする。図６に示す一態様において、一次OTU（Ｐｒ３）の配列数が二次OTU（Ｓｃ）の中で最も多い。したがって、二次OTU（Ｓｃ）の代表配列を一次OTU（Ｐｒ３）の配列とする。

[0049] 図１に示すステップＳ１４において、さらに代表配列をデータベースと照合する。以上により二次OTUの系統推定を行う。一態様において、系統推定は、二次OTUの細菌の門、綱、目、科、属、種、及び亜種のいずれかの分類群を推定することである。

[0050] 一態様においてデータベースは公共データベースである。他の態様においてデータベースは、一つの又は複数の他のデータベースをメタ解析して生成したデータベースである。一態様においてデータベースはコンピューターに対してネットワークを介して接続されている。

[0051] メタ解析する際は、例えば、他のデータベースの中の配列を、正確なアノテーションが行われている配列とそうでない配列とで区別する。一態様において正確なアノテーションが行われている配列は基準株（Type Strain）の配列である。一態様において、メタ解析して生成したデータベース中に同一の配列が複数含まれないようにする。このようなデータベースの最適化により代表配列をデータベースと照合する際の計算コストを大きく減らすことができる。一態様においてトップヒットのみの算出で行う。

[0052] ＜実施例＞

[0053] 図７は約1500bpの塩基長を有する16S rRNA遺伝子を模式的に表す。本実施

例ではサンプルとして16S rRNA遺伝子の部分配列を有するDNAを用いる。16S rRNA遺伝子は9個の超可変領域、以下可変領域というと、これらの間にある領域、以下保存領域という、とを有する。部分配列は5’側の可変領域2つを含む300ｂｐの領域からなる。

[0054] 本実施例では上記部分配列を有するDNAを被験者の子宮内腔液又は膣スワブからクローニングする。これらのスワブには特有の細菌叢が含まれる。得られたDNAをサンプルとしてDNAシーケンシングを行うことでサンプル配列を得る。DNAシーケンシングは次世代シーケンサーを用いて行う。

[0055] 次世代シーケンサーから出力される配列の数は非常に多い。上記クラスタリングを行うことで、データベースでの相同性検索を全ての配列に対してする必要がなくなる。一つの観点において、これは代表配列を相同性検索することで二次OTU内の全てのサンプル配列に対して系統推定ができることによる。

[0056] 図８は配列決定後のメタゲノム解析のフローを示す。まずステップＳ８１にて、配列の生データ（Raw Data）からアダプター配列を除去することで、クリーンなFASTQ形式のデータ（Cleaned fastq）が得られる。プログラム名Trimmomaticが利用できる。本データ及び以下の各データは被験者の識別子を含む。

[0057] 図８に示すステップＳ８２にて、FASTQ形式のデータ中のリードを結合することで、結合されたFASTQ形式のデータ（merged fastq）が得られる。プログラム名fastq-joinが利用できる。

[0058] 図８に示すステップＳ８３にて、FASTQ形式のデータ中のプライマー配列を除去する。さらにステップＳ８４にて、FASTQ形式のデータをFASTA形式のデータに変換する。これらではプログラム名PRINSEQが利用できる。

[0059] 図８に示すステップＳ８５にて、FASTA形式のデータに含まれるサンプル配列から一次OTUを作成する。プログラム名Qiime pick otus.pyのUCLUSTが利用できる。相同性（similarity）の閾値は1.00（100%）とする。さらにステップＳ８６にて、各一次OTUから配列の情報を抽出する。プログラム名Qiime pi

ck rep set.py が利用できる。

[0060] 図８に示すステップＳ８７にて、サンプル配列から二次OTUを作成する。プログラム名Qiime pick otus.pyが利用できる。相同性（similarity）の閾値は0.995（99.5%）とする。さらに二次OTUに含まれる一次OTUから最も配列数の多い一次OTUを選択し、その配列を代表配列とする。

[0061] 本例では探索的クラスタリングのためにQiime pick otus.pyのUCLUSTを利用する。他の態様においてこれ以外のプラットフォームで探索的クラスタリングをする。係るプラットフォームの例示としてはusearch、CD-HIT及びvsearchが挙げられる。

[0062] 図８に示すステップＳ８８にて、代表配列とデータベース（ＤＢ）を照らし合わせることで各二次OTUに系統情報を割り当てたデータ（Assigned txt）を生成する。プログラム名BLASTが利用できる。さらにステップＳ８９にてデータから低品質のリードを除去する。

[0063] 本実施例で用いるデータベース（ＤＢ）はラクトバチルス属（Lactobacillus）の16S rRNA遺伝子の配列及びアノテーションを含む。他の態様においてデータベースはさらに、ビフィドバクテリウム属（Bifidobacterium）、プロピオニバクテリウム属（Propionibacterium）、ガードネレラ属（Gardnerella）、ストレプトコッカス属（Streptococcus）及びベイロネラ属（Veillonella）の16S rRNA遺伝子の配列及びアノテーションを含む。係る態様において上記部分配列は、これらの細菌の属の間で完全には保存されていない。一態様において上記部分配列は、16S rRNAの全体の配列の中で保存性の低い部分の配列である。一態様において上記部分配列は、これらの細菌の属を区別できる程度に保存性が低い。

[0064] 他のデータベースの例はNCBI、GreenGenes、SILVA、EzBioCloud及びRibosomal Database Projectである。他の態様において、データベースは他のデータベースをメタ解析して生成したデータベースである。他のデータベースの例は上記例示したデータベースである。上記例示したデータベースの例やこれらをメタ解析して生成したデータベースは、キメラやミスアノテーション

を含んでいてもよい。上述の通り16S rRNA遺伝子の部分配列を用いることで、全長配列を用いた時よりもキメラやミスアノテーションの影響を小さくすることができる。また全長配列を用いた時よりも計算コストを引き下げることができる。

[0065] 得られたデータ（Assigned txt）から、細菌叢中における細菌の構成比と被験者の識別子とを含むデータが得られる。一態様において細菌の構成比は、二次OTUの配列数に比例する。二次OTUには当該細菌の分類がアノテーションされる。一態様において細菌の構成比は、例えば図９に示すようなラクトバチルス属及びラクトバチルス属以外の細菌の属の細菌叢中における構成比である。

[0066] 一態様において、コンピューターは診断者に対して、細菌の構成比と被験者の識別子を提供する。一態様において情報の提供はネットワークを介して接続されたサーバー及び診断者の端末を介して行う。診断者は、細菌叢の構成比に基づき、被験者の体外受精による妊娠のしやすさや、自然妊娠の妊娠のしやすさを診断する。

[0067] ＜参考例１＞

[0068] 図１０は一次OTUを作成せず、サンプル配列に対して探索的クラスタリングを行った場合を表す。重心Ｃｔ４はクラスターを構成する配列Ｓｑ１、Ｓｑ２及びＳｑ３に類似する。ただし、重心Ｃｔ４は疑似配列である。すなわちサンプルには含まれない配列である。

[0069] 図１０に示す重心Ｃｔ４をデータベースと比較することでOTUの系統推定を行った場合、系統推定を誤る可能性がある。細菌分類の全体にわたって保存性の高い16S rRNA遺伝子のおいてこの傾向は特に顕著である。言い換えればミスアノテーションを起こしやすい。したがって閾値を97%前後とすることが好ましい。このような閾値は細菌を属又は属よりも大きな分類で系統推定するのに適する。

[0070] これに対して上記実施形態では、探索的クラスタリングを利用する点で上記実施形態と共通する。しかしながら代表配列を二次OTU内の一次OTUの配列

から選択するため、疑似配列を用いて系統推定することがない。このため疑似配列の影響を減らすことができる。また上記実施形態ではDNAシーケンシングでエラーが生じたサンプル配列も二次OTUに取り込むことができる。したがって、二次OTUに含まれる配列数はサンプルや細菌叢の実体をよく表す。上記実施形態のメタゲノム解析では、探索的クラスタリングのみを行ったメタゲノム解析よりもより細かな系統推定ができる。

[0071] なおシーケンシングにおけるエラーは互いに配列が類似する一次OTUが生じる背景でもある。例えば図２に示す場合において、OTU aが真の配列を有するサンプル配列からなり、OTU b及びOTU cがエラーを含むサンプル配列からなると考えてもよい。図３から図６ではエラーを含むサンプル配列をも取り込みながら二次OTUを作成する。しかしながら代表配列を最も配列数の多い一次OTUから作成するので、細菌の系統推定の正確性を高めることができる。

[0072] 本参考例の方法に従いOTUを生成し、そのOTUの中で100%相同性を有するサンプル配列でサブOTUを生成してもよい。かかるサブOTUの内、最も配列数の多いサブOTUの配列を代表配列としてもよい。すなわち上記実施形態における一次OTUの生成と二次OTUの生成との順序を逆にしてもよい。

[0073] 上記実施形態に従い、一次OTUを生成した後に、二次OTUを生成することが好ましい。これにより二次OTUの生成において、一次OTU内の完全一致する配列間で重複する計算を省略することができる。

[0074] ＜参考例２＞

[0075] 上記実施形態では探索的クラスタリングと呼ばれる非階層的クラスタリング手法を用いて一次OTU及び二次OTUを作成することで、クラスタリングの精度と計算コストのバランスをとる。これと異なる手法として階層的クラスタリングによってサンプル配列をクラスタリングすることで、一次OTU及び二次OTUを生成する段階的な手法を用いずにOTUを生成することが例示される。

[0076] 階層的クラスタリングの例としてSingle linkage、Complete linkage及びAverage linkageが挙げられる。一次OTU及び二次OTUを生成する段階的な手法を用いない場合、Single linkageではOTUが非常に大きくなるため精度が高め

にくい。段階的な手法を用いない場合、Complete linkageでは小さなOTUが大量に生成する。したがってデータベースとの照合に計算コストを要する。段階的な手法を用いない場合、Average linkageでは適度な大きさで、適度な数のOTUを生成できる。しかしながら、段階的な手法を用いないことで、その階層的クラスタリングにかかる計算コスト自体が非常に高くなる。またいずれの階層的クラスタリング手法においても、段階的な手法を用いない場合は、代表配列として重心を選んだ場合に、代表配列がしばしば疑似配列となる。

[0077] 上記実施形態で利用している探索的クラスタリングを用いて一次OTU及び二次OTUを生成する段階的な手法は、階層的クラスタリングを用いて一段階でOTUを生成する手法と比較して精度と計算コストのバランスに優れる。そして上記実施形態では、代表配列を二次OTU内の一次OTUから選択するため、疑似配列を用いて系統推定することがない。このため疑似配列の影響を減らすことができる。

[0078] なお、本発明は上記実施の形態に限られたものではなく、趣旨を逸脱しない範囲で適宜変更することが可能である。

[0079] 上記実施形態では各属にわたってメタ16S解析をした。本発明の他の観点において、16Sリボソーム遺伝子及び23Sリボソーム遺伝子の間のＩＴＳ（Internal Transcribed Spacer）を細菌類のクラスタリングに用いる。

[0080] 他の態様において特定の属内にメタゲノム解析の範囲を絞り込んでもよい。その一態様において16Sリボソーム遺伝子に換えて、recA遺伝子、rpoB遺伝子及びその他のハウスキーピング遺伝子を用いる。またこれらのハウスキーピング遺伝子を複数組み合わせてMLSA（Multilocus sequence typing）を行ってもよい。

[0081] 上記実施形態では、細菌叢のメタゲノム解析を行った。本発明の他の観点は、原生生物や真菌類といった真核生物を含む微生物叢のメタゲノム解析である。真核生物のクラスタリングには18Sリボソーム遺伝子や、真核生物のリボソームＤＮＡ中のＩＴＳ（Internal Transcribed Spacer）を用いることができる。

[0082] 上記実施形態では、配列の相同性の閾値に基づくクラスタリングとして、探索的クラスタリングを利用した。本発明の他の観点において、配列の相同性の閾値に基づくクラスタリングとして、階層的クラスタリングを用いる。一態様において階層的クラスタリングはSingle linkage、Complete linkage及びAverage linkageのいずれかである。他の態様において階層的クラスタリングでは解析パッケージmothurのclusterコマンド<https://mothur.org/wiki/cluster/>を利用する。一態様においてサンプル配列に対して、100%の相同性を有する塩基配列同士をmothurでグループ化することで、一次OTU（Operational Taxonomic Unit）を生成する。次にmothurの階層的クラスタリングによってクラスターの重心の配列に対して所定の閾値以上の相同性を有するサンプル配列にクラスターを割り当てる。ただし所定の閾値は100%未満である。次に各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、代表配列をデータベースと照合することで、その二次OTUの系統推定を行う。このように一次OTUと二次OTUと生成することで階層的クラスタリングを用いた場合でも疑似配列の影響を減らすことができる。

[0083] 一次OTUの生成に階層的クラスタリングを用いた上で、二次OTUの生成に探索的クラスタリングを用いてもよい。一次OTUの生成に探索的クラスタリングを用いた上で、二次OTUの生成に階層的を用いてもよい。

[0084] また上述のUCLUSTやmothur に限らず、様々な探索的クラスタリング及び階層的クラスタリングを配列の相同性の閾値に基づくクラスタリングとして用いることができる。他の態様において<http://qiime.org/scripts/pick_otus.html>に挙げられるblast、prefix/suffix、Trie、uclust_ref、usearch_ref、usearch61、sumaclust、sortmerna_v2及びswarmを使用する。

[0085] 本発明の他の観点は、上記メタゲノム解析に係る処理をコンピューターで実行するためのプログラムである。本発明の他の観点は、上記プログラムを格納するディスク、テープ、フラッシュメモリ、及びその他の不揮発性記憶媒体である。本発明の他の観点は、上記プログラムを、ネットワークを介し

てダウンロード提供することである。本発明の他の観点は、ネットワークを介して要求を受け取り、要求に応じて上記プログラムをコンピューター上で実行し、要求に応じて処理結果を、ネットワークを介して提供することである。一態様においてプログラムを実行するコンピューターの動作を、ネットワークで接続された複数の装置で実行する。

[0086] この出願は、２０２１年３月８日に出願された日本出願特願２０２１－３６０９３を基礎とする優先権を主張し、その開示の全てをここに取り込む。

符号の説明

[0087] Ｃｔ１－４　重心，　ＤＢ　データベース，　Ｐｒ１－４　一次OTU，　Ｓ１１－１４　ステップ，　Ｓ８１－８９　ステップ，　Ｓｃ　二次OTU，　Ｓｑ１－３　配列

請求の範囲

[請求項1] 細菌叢のメタゲノム解析の方法であって、

細菌叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングを行い：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、前記代表配列をデータベースと照合することで、その二次OTUの系統推定を行う方法。

[請求項2] 前記二次OTUの生成に際して、前記サンプル配列に対して前記一次OTUとは異なるクラスターを初めに割り当て、さらに前記クラスターの重心の配列に対して所定の閾値以上の相同性を有するサンプル配列に前記クラスターを割り当て直すことを、収束するまで繰り返す、

請求項１に記載の方法。

[請求項3] 前記サンプル配列に対してクラスターを初めに割り当てる際、前記サンプル配列からランダムにシード配列を選択するとともに、前記シード配列と前記シード配列に対して前記所定の閾値以上の相同性を有する配列とに最初のクラスターを割り当てる、

請求項２に記載の方法。

[請求項4] 前記一次OTUを作成した後に、前記二次OTUを生成するところ、

前記クラスターは前記一次OTUを包含するクラスターであり、

前記重心の配列は前記一次OTUの配列に前記一次OTUの配列数で重みづけをした平均である、

請求項２又は３に記載の方法。

[請求項5] 前記クラスターを割り当て直した際、全てのクラスターの割り当てが変化しなかった場合に、これを収束とする、

請求項２～４のいずれかに記載の方法。

[請求項6] 前記サンプルは被験者の子宮内腔液又は膣スワブであり、

前記サンプル配列は前記サンプルより抽出したゲノムDNAの16S rRNA遺伝子の部分配列をDNAシーケンシングして生成したものであり、

前記データベースはラクトバチルス属（Lactobacillus）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記系統推定において、前記二次OTUの細菌の属を推定し、

前記被験者の識別子と前記ラクトバチルス属を含む細菌の属の細菌叢中における構成比とを含むデータを生成する、

請求項１～５のいずれかに記載の方法。

[請求項7] 前記データベースはさらに、ビフィドバクテリウム属（Bifidobacterium）、プロピオニバクテリウム属（Propionibacterium）、ガードネレラ属（Gardnerella）、ストレプトコッカス属（Streptococcus）及びベイロネラ属（Veillonella）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記部分配列は、16S rRNA遺伝子の全体の配列の中で、これらの細菌の属の間で完全には保存されていない部分の配列である、

請求項６に記載の方法。

[請求項8] 前記データベースは、他の複数のデータベースをメタ解析して生成したデータベースである、

請求項６又は７に記載の方法。

[請求項9] 被験者より前記サンプルとして子宮内腔液又は膣スワブを採取し、

請求項１～５のいずれかに記載の方法で前記サンプルをメタゲノム解析し、ここで

前記サンプルより抽出したゲノムDNAの16S rRNA遺伝子の部分配列

をDNAシーケンシングすることで前記サンプル配列を生成し、

前記データベースはラクトバチルス属（Lactobacillus）の16S rRNA遺伝子の配列及びアノテーションを含み、

前記系統推定において、前記二次OTUの細菌の属を推定し、

ラクトバチルス属及びラクトバチルス属以外の細菌の属の細菌叢中における構成比に基づき、前記被験者の妊娠のしやすさを診断する方法。

[請求項10] 真核生物の微生物叢のメタゲノム解析の方法であって、

微生物叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングを行い：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、前記代表配列をデータベースと照合することで、その二次OTUの系統推定を行う方法。

[請求項11] 細菌又は真核生物の微生物叢のメタゲノム解析のプログラムであって、

前記微生物叢を含むサンプルをDNAシーケンシングして得た塩基配列、以下サンプル配列という、に対して以下を含むクラスタリングをコンピューターに行わせ：

前記サンプル配列に対して、100%の相同性を有する塩基配列同士をグループ化することで、一次OTU（Operational Taxonomic Unit）を

生成すること、及び

前記サンプル配列に対してクラスタリングを行うことで、所定の重心の配列と、これに対して所定の閾値以上の相同性を有するサンプル配列からなる二次OTUを生成すること、ただし前記所定の閾値は100%未満である、

各二次OTUに含まれる一次OTUの中で最も配列数の多い一次OTUの配列をその二次OTUの代表配列とし、前記コンピューターに前記代表配列をデータベースと照合させることで、その二次OTUの系統推定を前記コンピューターに行わせるプログラム。