(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6373977号
(P6373977)

(45)発行日　平成30年8月15日(2018.8.15)　(24)登録日　平成30年7月27日(2018.7.27)

(51)Int.Cl.　　F I
G06F 19/22 (2011.01)　G06F 19/22
C12N 15/00 (2006.01)　C12N 15/00
C12Q 1/68 (2018.01)　C12Q 1/68

請求項の数 15　(全 16 頁)

(21)出願番号　特願2016-514498 (P2016-514498)
(86)(22)出願日　平成26年4月30日 (2014.4.30)
(65)公表番号　特表2016-524749 (P2016-524749A)
(43)公表日　平成28年8月18日 (2016.8.18)
(86)国際出願番号　PCT/IB2014/061098
(87)国際公開番号　WO2014/188290
(87)国際公開日　平成26年11月27日 (2014.11.27)
審査請求日　平成29年4月24日 (2017.4.24)
(31)優先権主張番号　61/826,619
(32)優先日　平成25年5月23日 (2013.5.23)
(33)優先権主張国　米国(US)

(73)特許権者　590000248
コーニンクレッカ　フィリップス　エヌヴェ
KONINKLIJKE PHILIPS N. V.
オランダ国　５６５６　アーエー　アインドーフェン　ハイテック　キャンパス　５
High Tech Campus 5, NL-5656 AE Eindhoven
(74)代理人　100122769
弁理士　笛田　秀仙
(74)代理人　100163809
弁理士　五十嵐　貴裕

最終頁に続く

(54)【発明の名称】ＤＮＡ配列の高速かつ安全な検索

(57)【特許請求の範囲】
【請求項１】
データベースに記憶されたデオキシリボ核酸（ＤＮＡ）又はリボ核酸（ＲＮＡ）配列に対する配列モデルを有する配列指標を生成するステップであって、当該生成するステップは、有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータとして前記データベースに記憶された各ＤＮＡ又はＲＮＡ配列に対する前記配列モデルを計算するステップを含み、前記配列モデルが、文脈木重み付け（ＣＴＷ）を使用して計算される、ステップと、
クエリＤＮＡ又はＲＮＡ配列に前記配列モデルを適用すること、並びにどれだけ良好に各配列モデルが前記クエリＤＮＡ又はＲＮＡ配列にフィットするかを決定することに基づいて前記クエリＤＮＡ又はＲＮＡ配列に最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップと、
を含む方法を実行するように電子データ処理装置により実行可能な命令を記憶する非一時的記憶媒体。
【請求項２】
前記識別するステップが、
有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータとして前記クエリＤＮＡ又はＲＮＡ配列に対するクエリモデルを計算するステップであって、前記クエリモデルが、文脈木重み付け（ＣＴＷ）を使用して計算される、ステップと、
前記クエリモデルを使用して達成可能な前記クエリＤＮＡ又はＲＮＡ配列の圧縮の量を

測定する圧縮計量の基準値を計算するステップと、
を含み、
前記クエリＤＮＡ又はＲＮＡ配列に前記配列モデルを適用することが、前記圧縮計量の前記基準値と、前記配列モデルを使用して前記クエリＤＮＡ又はＲＮＡ配列の圧縮率を測定する前記圧縮計量の値との間の差に基づいて各配列モデルに対する情報利得を推定することを含む、
請求項１に記載の非一時的記憶媒体。
【請求項３】
前記識別するステップが、前記配列モデルを使用し、前記データベースに記憶された前記ＤＮＡ又はＲＮＡ配列を使用しない、請求項１乃至２のいずれか一項に記載の非一時的記憶媒体。
【請求項４】
前記クエリＤＮＡ又はＲＮＡ配列に前記配列モデルを適用することが、
各配列モデルに対して、前記配列モデルを使用して前記クエリＤＮＡ又はＲＮＡ配列に対する符号語長を計算する、
ことを含む、請求項１に記載の非一時的記憶媒体。
【請求項５】
前記識別するステップが、
ＣＴＷを使用して有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータとして前記クエリＤＮＡ又はＲＮＡ配列に対するクエリモデルを計算するステップと、
前記クエリモデルを使用して前記クエリＤＮＡ又はＲＮＡ配列に対する基準符号語長を計算するステップと、
を含み、
前記クエリＤＮＡ又はＲＮＡ配列に前記配列モデルを適用することが、前記基準符号語長と、前記配列モデルを使用して前記クエリＤＮＡ又はＲＮＡ配列に対して計算された符号語長との間の差に基づいて各配列モデルに対する情報利得を推定することを含む、
請求項１に記載の非一時的記憶媒体。
【請求項６】
前記データベースに記憶された前記ＤＮＡ又はＲＮＡ配列が、ＤＮＡ染色体配列であり、
前記クエリＤＮＡ又はＲＮＡ配列が、染色体より小さいクエリＤＮＡ配列フラグメントである、
請求項１乃至５のいずれか一項に記載の非一時的記憶媒体。
【請求項７】
データベースに記憶されたデオキシリボ核酸（ＤＮＡ）又はリボ核酸（ＲＮＡ）配列に対する文脈木重み付け（ＣＴＷ）モデル$\{S_x, \Theta_{Sx}\}$を有する配列指標を生成するステップであって、S_xが前記ＤＮＡ又はＲＮＡ配列ｘに対する前記文脈木重み付けモデルを示し、Θ_{Sx}が文脈木モデルS_xのパラメータを示す、当該生成するステップと、
クエリＤＮＡ又はＲＮＡ配列ｙに前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用すること、並びにどれだけ良好に各ＣＴＷモデルが前記クエリＤＮＡ又はＲＮＡ配列ｙにフィットするかを決定することに基づいて前記クエリＤＮＡ又はＲＮＡ配列ｙに最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップと、
を有し、
前記生成するステップ及び前記識別するステップが、電子データ処理装置により実行される、方法。
【請求項８】
前記識別するステップが、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用し、前記データベースに記憶された前記ＤＮＡ又はＲＮＡ配列ｘを使用しない、請求項７に記載の方法。
【請求項９】

前記識別するステップが、
前記クエリＤＮＡ又はＲＮＡ配列ｙに対するＣＴＷモデル$\{S_y, \Theta_{Sy}\}$を計算するステップであって、S_yが前記クエリＤＮＡ又はＲＮＡ配列ｙに対する文脈木モデルを示し、Θ_{Sy}が前記文脈木モデルS_yのパラメータを示す、当該計算するステップと、
前記クエリＤＮＡ又はＲＮＡ配列ｙに対する前記ＣＴＷモデル$\{S_y, \Theta_{Sy}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙの圧縮率を測定する圧縮計量の基準値を計算するステップと、
を含み、
前記クエリＤＮＡ又はＲＮＡ配列ｙに前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用することが、前記圧縮計量の前記基準値と、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙの圧縮率を測定する前記圧縮計量の値との間の差に基づいて各ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$に対する情報利得を推定することを含む、
請求項７乃至８のいずれか一項に記載の方法。
【請求項１０】
前記識別するステップが、
前記クエリＤＮＡ又はＲＮＡ配列ｙに対するＣＴＷモデル$\{S_y, \Theta_{Sy}\}$を計算するステップであって、S_yが前記クエリＤＮＡ又はＲＮＡ配列ｙに対する文脈木モデルを示し、Θ_{Sy}が文脈木モデルS_yのパラメータを示す、当該計算するステップと、
前記クエリＤＮＡ又はＲＮＡ配列ｙに対するＣＴＷモデル$\{S_y, \Theta_{Sy}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙに対する基準符号語長を計算するステップと、
を含み、
前記クエリＤＮＡ又はＲＮＡ配列ｙに前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用することが、前記基準符号語長と、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙに対して計算される符号語長との間の差に基づいて各ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$に対する情報利得を推定することを含む、
請求項７乃至８のいずれか一項に記載の方法。
【請求項１１】
前記クエリＤＮＡ又はＲＮＡ配列ｙに前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用することが、
各ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$に対して、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙに対する符号語長を計算する、
ことを含み、前記識別するステップが好適には、
前記クエリＤＮＡ又はＲＮＡ配列ｙに最も類似しているものとして、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して、前記クエリＤＮＡ又はＲＮＡ配列ｙに対する最も短い符号語長を持つ前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップ、
を含む、
請求項７乃至８のいずれか一項に記載の方法。
【請求項１２】
データベースに記憶されたデオキシリボ核酸（ＤＮＡ）又はリボ核酸（ＲＮＡ）配列をモデル化する配列指標から文脈木重み付け（ＣＴＷ）モデル$\{S_x, \Theta_{Sx}\}$を検索するステップであって、S_xが前記ＤＮＡ又はＲＮＡ配列ｘに対する文脈木モデルを示し、Θ_{Sx}が前記文脈木モデルS_xのパラメータを示す、当該検索するステップと、
クエリＤＮＡ又はＲＮＡ配列に前記検索されたＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用すること、並びにどれだけ良好に各ＣＴＷモデルが前記クエリＤＮＡ又はＲＮＡ配列ｙにフィットするかを決定することに基づいて前記クエリＤＮＡ又はＲＮＡ配列に最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップと、
を含む方法を実行するようにプログラムされた電子データ処理装置、
を有する装置。
【請求項１３】
前記識別するステップが、前記データベースに記憶された前記ＤＮＡ又はＲＮＡ配列を

使用しない、請求項１２に記載の装置。
【請求項１４】
前記クエリＤＮＡ又はＲＮＡ配列ｙに前記検索されたＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を適用することが、
各ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$に対して、前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙに対する符号語長を計算する、
ことを含む、請求項１２に記載の装置。
【請求項１５】
前記識別するステップが、前記識別された１以上のＤＮＡ又はＲＮＡ配列をモデル化する前記ＣＴＷモデル$\{S_x, \Theta_{Sx}\}$を使用して前記クエリＤＮＡ又はＲＮＡ配列ｙに対して計算された最も短い符号語長を持つことに基づいて、前記ＤＮＡ又はＲＮＡ配列ｙに最も類似しているものとして、前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップを含む、請求項１４に記載の装置。
【発明の詳細な説明】
【技術分野】
【０００１】
以下は、ゲノム配列指標付け（indexing）、記憶、検索（retrieval）、処理、ラベル付け、及び関連するタスク、並びに患者プライバシ及び医療データセキュリティのような態様並びに医療診断及び医療スクリーニング等のような応用に関する。例示的にデオキシリボ核酸（ＤＮＡ）配列を参照して記載されているが、以下は、ＤＮＡ配列、及びリボ核酸（ＲＮＡ）配列等のようなゲノム配列と連動した応用を見つける。
【背景技術】
【０００２】
ＤＮＡシークエンシングは、がん及び他の病気の診断、遺伝性疾患に対する医療スクリーニング、個人用医療、個人用薬物設計、遺伝人類学及び進化研究、系譜的研究、及び法医学人物同定等のような、多くの既存の及び期待される商業的、医療的及び科学的応用を持つ。医療分野において、臨床試験及びゲノムワイド関連研究は、特定の治療、薬物の有効性を評価し、ＤＮＡパターンと疾病との間の従属関係等を決定する典型的なツールである。臨床試験において、試験に含める適格性基準は、同様の表現型（例えば人種）及び機能性（例えば遺伝子がオン又はオフである）を持つＤＮＡ配列を持つ患者を含むことができる。ゲノムワイド関連研究において、試験を行うために、症例群（例えば突然変異を含む配列）及び対照群（突然変位を含まない配列）に分割されることができるＤＮＡ配列が、選択される。遺伝人類学において、ゴールは、一般に、人口移動を追跡する、又は経時的な遺伝的多様性を研究する等のために基準ＤＮＡサンプル（又は基準ＤＮＡサンプルプール）と強い類似性を持つＤＮＡサンプルを識別することである。これらは、ＤＮＡ配列比較を使用する応用の単なる例示的な例である。
【０００３】
人間のＤＮＡゲノムは、約３００００の遺伝子を集合的に暗号化するおおよそ３．２×10^9のヌクレオチドからなる。動物、植物及び他の生命体に対するゲノムは、幅広く異なることができるが、典型的には、同等の桁である。臨床試験に対して適格な患者、又は研究目的に対するＤＮＡ配列等を見つけるために、巨大なデータベースが、処理される必要がありうる。したがって、同様なＤＮＡ配列を位置特定する迅速な手順は、有利である。このような検索は、ＤＮＡゲノムの純粋なサイズ並びにギャップ、アライメントエラー、合計配列長の差、及び様々なタイプのノイズを含むことができる実験的に取得されたＤＮＡ配列の時々断片的な性質のような多くの問題により複雑にされる。
【０００４】
人間のＤＮＡに対処する場合、他の検討事項は、対象のプライバシである。ＤＮＡ配列は、遺伝的記録全体を暗号化しており、特定の疾患に対するリスク素因及び祖先情報等のような医療的に又は個人的にセンシティブな情報を明らかにすることができる。ＤＮＡ配列は、（一卵性の双生児を例外として）人間のユニーク識別子でもある。同様の検討事項

は、競走馬及び作物等のような商業的に価値のある生命体の非人間ゲノム配列データを処理する際にも生じることができる。このような情報の制御に関する関心は、米国における医療保険会社及び雇用主による個人のＤＮＡから得られた健康情報に基づく差別を禁止することを意図される、２００８年の遺伝情報差別禁止法（ＧＩＮＡ）により示される。しかしながら、ＧＩＮＡは、生命保険、身体障害保険及び長期ケア保険をカバーしていない。また、ＤＮＡ配列は、他のタイプの個人医療データと比較してユニークな検討事項を関与させる。人間のゲノムは、全体的に理解されるには程遠く、したがって、ＤＮＡから新しい個人的にセンシティブな情報を抽出する新しい技術に対する進行中の可能性が存在する。また、他の医療情報とは異なって、ＤＮＡ配列は、これら自体が識別子であるので、匿名化されることができない。したがって、ＤＮＡマッチングは、好ましくは、データセキュリティを強化する形で行われるべきである。

【発明の概要】

【発明が解決しようとする課題】

【０００５】

以下は、前述の制限等を克服する改良された装置及び方法を検討する。

【課題を解決するための手段】

【０００６】

１つの例示的態様によると、不揮発性記憶媒体は、データベースに記憶されたＤＮＡ又はＲＮＡ配列に対する配列モデルを有する配列指標を生成するステップであって、有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータとして前記データベースに記憶される各ＤＮＡ又はＲＮＡ配列に対する前記配列モデルを計算するステップを含む当該生成するステップと、クエリＤＮＡ又はＲＮＡ配列に対する前記配列モデルのフィッティングの結果に基づいて前記クエリＤＮＡ又はＲＮＡ配列に最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップとを含む方法を実行するように電子データ処理装置により実行可能な命令を記憶する。

【０００７】

他の例示的態様によると、方法は、データベースに記憶されたＤＮＡ又はＲＮＡ配列に対する文脈木重み付け（ＣＴＷ、context tree weighting）モデル $\{S_x, \Theta_{Sx}\}$ を有する配列指標を生成するステップであって、S_xは、前記ＤＮＡ又はＲＮＡ配列ｘに対する文脈木モデルを示し、Θ_{Sx}は、文脈木モデルS_xのパラメータを示す、当該生成するステップと、クエリＤＮＡ又はＲＮＡ配列ｙに対するＣＴＷモデル $\{S_x, \Theta_{Sx}\}$ のフィッティングに基づいてクエリＤＮＡ又はＲＮＡ配列ｙに最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップとを有する。前記生成するステップ及び前記識別するステップは、電子データ処理装置により適切に実行される。

【０００８】

他の例示的態様によると、装置は、データベースに記憶されたＤＮＡ又はＲＮＡ配列をモデル化する配列モデルを配列指標から検索するステップであって、前記データベースに記憶された各ＤＮＡ又はＲＮＡ配列に対する前記検索された配列モデルが、有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータを有する、当該検索するステップと、クエリＤＮＡ又はＲＮＡ配列に対する前記検索された配列モデルのフィッティングに基づいて前記クエリＤＮＡ又はＲＮＡ配列に最も類似しているものとして前記データベースに記憶された１以上のＤＮＡ又はＲＮＡ配列を識別するステップとを含む方法を実行するようにプログラムされた電子データ処理装置を有する。

【０００９】

１つの利点は、ゲノム配列の高速比較を提供することにある。

【００１０】

他の利点は、匿名性を維持しながら高速比較を提供する形でゲノム配列に指標付けする指標付け方法を提供することにある。

【００１１】

他の利点は、指標記録とのクエリゲノム配列の高速比較を容易化するように計算済み有限記憶木ソースモデル及びモデルパラメータを含む前記指標記録を使用してゲノム配列に指標付けする指標付け方法を提供することにある。

【００１２】

多くの追加の利点及び利益は、以下の詳細な記載を読むと当業者に明らかになる。

【００１３】

本発明は、様々なコンポーネント及びコンポーネントの構成並びに様々な処理オペレーション及び処理オペレーションの構成の形を取り得る。図面は、好適な実施例を例示する目的のみであり、本発明を限定すると解釈されるべきではない。

【図面の簡単な説明】

【００１４】

【図１】ＤＮＡ配列を記憶及び指標付けするシステムを概略的に示す。

【図２】クエリＤＮＡ配列に類似したＤＮＡ配列を識別するように図１のシステムにより生成されるＤＮＡ配列指標を検索するシステムを概略的に示す。

【図３】囲みボックスにより示される各クエリ染色体に対する最大相互情報量を持つ、例示的な実際に実行されるＤＮＡ検索オペレーションからの相互情報量に対する推定値の表を示す。

【発明を実施するための形態】

【００１５】

ここに開示されるのは、（例えば固定又は可変次数）マルコフモデル又は文脈木重み付け（ＣＴＷ）モデル（ここで使用される例示的アプローチ）等のような有限記憶木ソースモデルを使用してＤＮＡ配列（又は、より一般的に、ゲノム配列、例えばＤＮＡ配列又はＲＮＡ配列等）を指標付けするアプローチである。前記ＤＮＡ配列に対する指標記録が、構築され、前記モデル及びパラメータを含む。この場合、ＣＴＷを使用してクエリＤＮＡ配列の直接的なモデル化により推定される符号語長と比較される、クエリＤＮＡ配列に対して同じ有限記憶木モデルを使用して得られる推定符号語長は、前記クエリ及び指標ＤＮＡ配列の類似性を定量的に評価する比較計量として機能する。前記符号長比較は、例えば、エントロピ又は情報利得（ＩＧ）又は同様の手段のような相互情報計量を使用して計算される。

【００１６】

このアプローチは、前記有限記憶木ソースモデル及びパラメータのみが、プレーンテキストで、すなわち暗号化されずに記憶されるので、ＤＮＡ配列がデータベースに記憶される患者のプライバシを保護する。有限長の部分配列の使用は、結果として生じるモデル及びパラメータが元のＤＮＡ配列より大幅に少ない情報を含むので、患者プライバシを保証し、前記有限記憶木ソースモデルの出力は、実際に本質的に統計的である。前記指標づけされたＤＮＡ配列（のセット）に対する前記モデル及びそのパラメータは、事前に計算されるので、検索は高速である。開示された類似性計量は、相互情報量が検索基準として使用されるので、編集又は設定距離のような他の軽量より柔軟かつ表現豊かである。ここに開示されるように、相互情報量は、ゲノム配列の時間的構造を探索する順次的なユニバーサル圧縮方法に基づいて適切に推定される。

【００１７】

図１を参照すると、ＤＮＡ配列を記憶及び指標付けする例示的システムが、記載される。（ここでx^Tとして示され、上付き文字ＴがＤＮＡ配列長を示す）指標付けされるべきＤＮＡ配列１０は、ＤＮＡ配列１０の代表的有限記憶木ソースモデルを生成するように処理される。この実例において、前記有限記憶木ソースモデルは、ＣＴＷ方法を使用して計算される文脈木重み付け（ＣＴＷ）モデルである。ＤＮＡ配列x^Tに適用されるモデル化モジュール１２の出力１４は、前記有限記憶木ソースモデル及びそのパラメータである。例示的なＣＴＷモデル化において、前記文脈木モデル（すなわち文脈又は部分配列）は、S_xとして（又はモデル化されたＤＮＡ配列x^Tのアイデンティティが明らかである場合に、より単純にＳとして）示され、前記パラメータは、ここでΘ_{S_x}として（又はモデル化さ

れたＤＮＡ配列x^Tのアイデンティティが明らかである場合に、より単純にＳとして）示される、条件付き確率を有する。好ましくは、記述的注釈が、匿名アノテータ１６を介して提供される。患者プライバシが重要である応用において、前記注釈は、匿名であるべきであるが、ＤＮＡ配列１０のソースの関連する記述を構成すべきであり、例えばデモグラフィック情報、又は臨床情報等により前記ソースを記述する。前記応用が、匿名性を必要としない場合、アノテータ１６は、前記注釈に対象識別子を含めてもよい。指標記録フォーマッタ１８は、前記モデル及びパラメータ１４並びに前記注釈を含む指標記録を構築し、前記指標記録は、電子健康記録（ＥＨＲ）、又は学問上の目的で採用されるＤＮＡリポジトリ指標等のような、データベース２０に記憶される。

【００１８】

前記指標記録は、例えばＤＮＡ配列x^Tに対する（S_x, Θ_{Sx}）として表されるモデル及びパラメータ１４を含む。これは、ＤＮＡ配列x^Tを表すが、近似的表現であり、ＤＮＡ配列x^Tが導出された対象を識別するには不十分である。したがって、ＤＮＡ配列x^Tは、適切に安全なフォーマットで別に記憶される。このために、図１の例示的な実施例において、高度暗号規格（ＡＥＳ暗号）に適合する暗号化アルゴリズムを採用する暗号化モジュール２４は、ＤＮＡ配列１０を暗号化する。前記暗号化モジュールは、セキュリティ暗号化を実行し、オプションとして、結合された圧縮／暗号化アルゴリズムにより統合的に又は別のオペレーションのいずれかでロスレス圧縮を実行する。データベース記録フォーマッタ２６は、暗号化された（及びオプションとして圧縮された）ＤＮＡ配列をフォーマット化し、これを暗号化ＤＮＡ配列データベース２８に記憶する。

【００１９】

図１を参照し続けると、前記指標付けシステムは、以下のように適切に物理的に実現される。コンピュータ３０又は他の電子データ処理装置（例えばコンピュータ、又はセキュア暗号化伝送プロトコルによりリンクされたインターネットベースのサーバ等）は、データ処理モジュール１２、１８、２４、２６を実施するように適切にプログラムされる。匿名アノテータ１６は、例えば、ＥＨＲ又は他のデータベースからデモグラフィック又は他の関連情報を抽出する完全自動化システムとして、様々な形で実施されえ、当該情報の匿名化を適切に、又は人間のオペレータが前記関連情報を入力することを可能にするのにユーザインタフェース（例えば例示的なディスプレイ３２及びキーボード３４）を採用する半自動化システムとして、実行する。ＤＮＡ配列指標データベース２０は、磁気ディスク、個別ディスクの冗長アレイ（ＲＡＩＤ）、又は光ディスク等のような非一時的記憶媒体３６上で適切に実施される。同様に、暗号化ＤＮＡ配列データベース２８は、磁気ディスク、個別ディスクの冗長アレイ（ＲＡＩＤ）、又は光ディスク等のような非一時的記憶媒体３８上で適切に実施される。

【００２０】

例示的な図１において、同じコンピュータ３０が、指標付けモジュール１２、１８及びアノテータ１６又はその自動化された部分、並びに配列暗号化及び記憶モジュール２４、２６の両方を実施するのに対し、物理的に離れたデータ記憶媒体３６、３８が、指標２０及びデータベース２８をそれぞれ記憶する。このアプローチは、（単一のコンピュータ３０が適切に使用されるように）ワークフローブロックとして記憶及び指標付けされるべきＤＮＡ配列に対して典型的であり、指標２０及びデータベース２８を別の媒体上で保持することがセキュリティを強化することができるので、有利であることができる。このアプローチにおいて、ＤＮＡ配列１０に対する指標記録は、データベース２８に記憶された暗号化ＤＮＡ配列記録に対するリンクを記憶する（データベース記録フォーマッタ２６を指標記録フォーマッタ１８に接続し、前記指標記録における包含のために前記リンクを後者に伝えることを示す点線矢印により図１に概略的に示される）。

【００２１】

代替的な物理的実施が可能であると理解される。例えば、別々のコンピュータが、それぞれ、指標付けオペレーション１２、１６、１８及び暗号化／記憶オペレーション２４、２６を実施するのに使用されることができる。加えて又は代わりに、前記暗号化されたＤ

ＮＡ配列及び対応する指標記録は、同じ物理的非一時的記憶媒体に記憶されることができる。他の変形例として、前記指標記録の要素として前記暗号化されたＤＮＡ配列を含めることにより指標２０及び暗号化ＤＮＡ配列データベース２８を結合することが考えられる。これは、ＡＥＳ又は他の暗号化プロトコルが十分に安全であると見なされる場合に適切でありうる。（いかなる事象においても、復号鍵は、別々に、又は何らかの他の安全な形で記憶されるべきである）。

【００２２】

以下に、例示的なＣＴＷモデル化モジュール１２のオペレーションが、更に記載される。

【００２３】

前記文脈木重み付け（ＣＴＷ）方法（Willems et al., The Context Tree Weighting Method: Basic Properties, IEEE transactions on Information theory, 1995）は、深度が指定された最大深度Ｄを超過しない全ての木モデルに対応する符号化分布（coding distribution）を計算する。前記分布は、算術的符号化技術を使用して観測されたＤＮＡ配列１０を圧縮するのに使用されることができ、これは、結果として小さな冗長性を持つ符号語を生じる。実際に、実際の圧縮は、実行される必要がなく、むしろ、ここに開示された技術は、前記ＤＮＡ配列を圧縮するのに前記モデルを使用して得られる圧縮の量を示す符号語長を推定する。ソース配列の長さにより除算される符号語長は、エントロピの良好な推定値を与える。

【００２４】

ＤＮＡ配列構造は、アミノ酸に対して及び後で順次的な形でタンパク質に対して符号化するようなものである。x^Tが観測されたＤＮＡ配列１０を示すとする。（より一般的には、x^Tは、同じ文脈木モデル及びパラメータにより一緒にモデル化される配列のセットを示すことができる）。この場合、ＣＴＷは、$P(x^T)$を推定するのに使用されることができ、ここでx^Tは、アルファベットＡ＝｛１，２，３，４｝からの値を持つベクトルとして適切に表される。（ＤＮＡアルファベットが、典型的には｛Ａ，Ｔ，Ｇ，Ｃ｝として表され、Ａがアデニンを示し、Ｔがチミンを示し、Ｇがグアニンを示し、Ｃがシトシンを示すのに対し、ＲＮＡアルファベットは、典型的には｛Ａ，Ｕ，Ｇ，Ｃ｝であり、チミンがウラシルを表すＵにより置き換えられることに注意する。アルファベットＡ＝｛１，２，３，４｝は、一般性を失うことなしにここで使用される。例えばメチル化のような情報を取得するように、４つより多いシンボルを持つアルファベットを採用することも考えられる。）x^Tで、観測された配列x^T内の位置ｔにおけるアルファベットＡからのシンボルを示す。前記ＤＮＡ配列に対する統計モデルは、前記文脈木を構築し、前記ＣＴＷアルゴリズムを使用して分布$P(x^T)$を、$P(x_t|\{x_{t-b}, b \in B\})$として推定することにより推定され、ここでＢは、適切な整数のセットである。「文脈」$\{x_{t-b}, b \in B\}$は、x^Tの$|B|$の異なる場所から得られたアルファベットＡからの値のセットからなる。典型的には、Ｂは、（最大深度Ｄまでの）x^Tに先行する値のセットとして記される。（前記観測されたＤＮＡ配列において実際に生じた）全ての可能な文脈は、確率分布$P(x_t|\{x_{t-b}, b \in B\})$と一緒に、それぞれ、文脈木（モデル）及びパラメータを構成する。

【００２５】

前記ＣＴＷアルゴリズムの出力は、前記文脈木モデル及び条件付き確率$\{S, \Theta_S\}$である。所定のＤＮＡ配列に対して、前記ＤＮＡ配列が$\{S, \Theta_S\}$を使用して圧縮された場合に得られる圧縮の量は、推定された符号語長Ｌにより特徴づけられることができる。ここに開示されるように、前記ＣＴＷ方法は、ツーパスアプローチで使用されることもでき、第１のステップにおいて、統計モデル$\{S, \Theta_S\}$が、観測されたＤＮＡ配列に対して算出され、第２のステップにおいて、前記モデルを使用して達成可能な前記ＤＮＡ配列の圧縮の量を示す前記符号語長が、推定される。前記推定は、第１のパスにおいて得られる$\{S, \Theta_S\}$により提供される固定の条件付き確率に基づき、比較すると、従来の（単一パス）ＣＴＷにおいて、前記符号語長は、各シンボルが処理されると常に更新されている確率に基づいて計算される。ここに更に開示されるように、このツーパスアプローチは、１つのＤＮＡ配

列（一般に一緒にモデル化された基準又は指標配列のセットでありうる、基準又は指標付けされた配列）に前記第１のステップを実行し、次いで、結果として生じるモデルを、第２の（クエリ）ＤＮＡ配列に対する符号語長を推定するのに使用することにより、２つの異なるＤＮＡ配列に対する類似性計量を規定するように拡張されることができる。前記モデルは、前記指標付けされたＤＮＡ配列から算出されたので、これは、前記指標付けされたＤＮＡ配列に対する最適に短い符号語長を生成すべきである。他方で、前記モデルが、前記クエリＤＮＡ配列に適用される場合、前記符号語長は、前記クエリＤＮＡ配列が前記指標付けされたＤＮＡ配列にどれだけ類似しているかに依存する。これらが類似している場合、前記モデルは、良好に「フィット」し、短い推定符号語長に対応する高い度合の圧縮を提供する。他方で、これらが類似していない場合、フィットが貧弱であり、前記クエリ配列に対する推定符号語長は、最適なモデルに対して得られるものより長い。前記クエリ配列から算出されたモデルに対して得られた符号語長は、適切な基準長さを提供する。例示的な定量的定式化は、以下のとおりである。

【００２６】

観測されたＤＮＡ配列x^Tを検討する。$\{S, \Theta_S\}$は、Ｄより大きくない深度の木ソースを記述するモデル（文脈）及びパラメータセット（条件付き確率）であると仮定する。この例において、$\{S, \Theta_S\}$が必ずしもx^Tから算出されないことに注意する。パラメータ$\{S, \Theta_S\}$を持つモデルが、ＤＮＡ配列x^Tを圧縮するのに使用される場合、圧縮された配列の長さは、

$$L(x^T|x_{-D}^1, S, \Theta_S) = -\sum_{t=1}^{T} log_2 P(x_t|x_{-D}^{t-1}, S, \Theta_S) = -\sum_{t=1}^{T} log_2 \theta_{\sigma_{\{x_{-D}^{t-1}\}}}^{x_t} \quad (1)$$

により与えられ、式（１）において、

$$\sigma_{\{x_{-D}^{t-1}\}}$$

は、Ｓから文脈への

$$x_{-D}^{t-1}$$

のマッピングであり、

$$P(x_t|x_{-D}^{t-1}, S, \Theta_S) = \theta_{\sigma_{\{x_{-D}^{t-1}\}}}^{x_t} \in \Theta$$

は、部分配列

$$\sigma_{\{x_{-D}^{t-1}\}}$$

がx^Tにおいて観測された後に生じるシンボルx^Tの確率である。$\{S, \Theta_S\}$が、x^Tを生成した実際のソースを記述する場合（例えば、上の例において、x^Tが前記指標付けされたＤＮＡ配列である場合）、$L(x^T|x^1_{-D}, S, \Theta_S)$は、最小の符号語長である理想的な符号

語長に対応する。しかしながら、$\{S,\Theta_S\}$が、何らかの他のソースを記述する場合（例えば、上の例において、x^Tが前記クエリ配列である場合）、$L(x^T|x^1_{-D},S,\Theta_S)$は、（少なくとも一般的には）前記モデルが他のＤＮＡ配列に対して算出され、観測されたＤＮＡ配列x^Tを効果的に記述しないので、前記理想的な符号語長より大幅に大きい。前記ＣＴＷ方法が、観測された（ＤＮＡ）配列のモデル及びパラメータを推定するのに使用される場合、結果として生じる符号語長は、前記理想的な符号語長から最小の距離（冗長性）を持つ。

【００２７】

類似性計量は、前記符号語長が、どれだけ良好に前記モデルが前記ＤＮＡ配列にフィットするかを示し、前記ＤＮＡ配列の符号語長が、式（１）の符号語長推定を使用して推定されるという、この概念を使用して規定されることができる。y^N及びx^Tが、必ずしも同じ長さではない２つの観測されたＤＮＡ配列であると仮定する。前の例に対する類推において、x^Tが長さＴの指標付けされたＤＮＡ配列であるとし、y^Nが長さＮのクエリＤＮＡ配列であるとする。$\{S_x,\Theta_{Sx}\}$が、前記ＣＴＷ方法を使用してx^Tに対して算出されたモデル及びパラメータセットであるとする。有利には、$\{S_x,\Theta_{Sx}\}$は、指標付けされたＤＮＡ配列x^T１０に対して事前に計算され、図１を参照して記載されるようにＤＮＡ指標２０に記憶されてもよい。更に、$L_{ctw}(y^N)$が、前記ＣＴＷ方法を使用して推定される（クエリ）ＤＮＡ配列y^Nに対する符号語長であるとする。換言すると、$L_{ctw}(y^N)$は、クエリＤＮＡ配列y^Nに対して算出されたモデル$\{S_y,\Theta_{Sy}\}$を使用して得られる符号語長である。したがって、$L_{ctw}(y^N)$は、前記ＣＴＷ方法を使用してy^Nに対して取得可能な最適な（すなわち最短の）符号語長である。この場合、差

$$\begin{aligned}
&\frac{1}{N}L_{ctw}(y^N)-\frac{1}{N}L\left(y^N|\,S_x,\Theta_{S_x}\right)\\
&\quad=-\frac{1}{N}\sum_{t=1}^{N}\log_2 P_{ctw}(y_t|y_{-D}^{t-1})+\frac{1}{N}\sum_{t=1}^{N}\log_2 P\left(y_t|y_{-D}^{t-1},S_x,\Theta_{S_x}\right)\\
&\quad=-\frac{1}{N}\sum_{t=1}^{N}\log_2\frac{P_{ctw}(y_t|y_{-D}^{t-1})}{P\left(y_t|y_{-D}^{t-1},S_x,\Theta_{S_x}\right)}\\
&\quad=-\frac{1}{N}\sum_{t=1}^{N}\log_2\frac{P_{ctw}(y_t|y_{-D}^{t-1})}{\theta^{y_t}_{S_x,\sigma_{\{y_{-D}^{t-1}\}}}}
\end{aligned}\qquad(2)$$

が、計算されることができる。式（２）の差は、x^Tの分布がy^Nを記述（圧縮）するためにy^Nの代わりに使用される場合に、どれだけが得られることができるかを示すことが見られる。利得が高い場合、$\{S_x,\Theta_{Sx}\}$は、y^Nに良好にフィットするソースを記述し、したがって、我々は、y^N及びx^Tの両方が同じソースにより生成されることを仮定し、これらが類似していると見なすことができる。利得が低い場合、$\{S_x,\Theta_{Sx}\}$を使用して推定されるy^Nに対する符号語長は、非常に高い冗長性を持ち、$\{S_x,\Theta_{Sx}\}$は、y^Nを圧縮する助けにならず、これは、他のタイプの（ＤＮＡ）配列を生成する他のソースに対応することを意味する。したがって、我々は、y^N及びx^Tが異なるソースにより生成され、これが類似していないと言うことができる。一般に、利得が高いほど、モデル及びパラメータセット$\{S_x,\Theta_{Sx}\}$が、配列y^Nを、より良好に記述する。したがって、$\{S_x,\Theta_{Sx}\}$を持つソースがy^Nを生成したことは、更にもっともらしい。

【００２８】

前記ＣＴＷ方法を使用して推定されたソースシンボルごとの符号語長は、前記ＤＮＡソース配列のエントロピの推定値を与える。したがって、式（２）の類似性計量は、ＤＮＡ

配列y^NとＤＮＡ配列x^Tを生成したＤＮＡソースとの間の相互情報量の推定値でもある。式（２）により提供される相互情報量の推定値は、過小評価である。これは、相互情報量が真に非負であるので、見られることができる。対照的に、式（２）は、最適な（最小の）符号語長である$L_{ctw}(y^N)$と、非最適な（したがってより大きい）符号語長である$L(y^N|S_x,\Theta_{Sx})$との間の（１／Ｎによりスケーリングされた）差を取る。後に続くのは、式（２）が、一般的に、厳密に非負の真の相互情報値より一般的に小さい、負の値を取り上げることができる。式（２）により与えられる相互情報量の過小評価は、部分的に、第２項の符号化冗長性の結果として生じる。前記過小評価は、類似性計量としての式（２）の有用性を否定しないが、しかしながら、より高い類似性（すなわちより大きな情報利得）が、式（２）の類似性計量により出される「より小さい負」値により示される。

【００２９】

先行する記載の観点から、クエリＤＮＡ配列y^Nと、モデル及びパラメータセット$\{S_x,\Theta_{Sx}\}$が事前に計算され、指標データベース２０に記憶される、指標付けされたＤＮＡ配列x^Tとの間の類似性を測定する類似性計量Ｉは、式（２）を使用して適切に計算される、又は換言すると$I(y^N;x^T,\{S_x,\Theta_{Sx}\})$は、式（２）を使用して適切に推定される。

【００３０】

一例として、クエリＤＮＡ配列y^Nに最も類似しているＤＮＡ配列指標２０内の指標付けされたＤＮＡ配列x^Tを見つける問題を検討する。これは、

$$\max_{P(x^T)} I(Y^N;X^T)$$

を見つけることになる。$\{S_x,\Theta_{Sx}\}$が、x^Tの関数である場合、データ処理不等式、

$$\max_{P(x^T)} I(y^N;x^T) = \max_{P(x^T)} I\left(y^N;x^T,\{S_x,\Theta_{S_x}\}\right)$$

$$= \max_{P(x^T)} \left(I\left(y^N;\{S_x,\Theta_{S_x}\}\right) + I\left(y^N;x^T\middle|\{S_x,\Theta_{S_x}\}\right)\right) \qquad (3)$$

$$\geq \max_{P(x^T)} I\left(y^N;\{S_x,\Theta_{S_x}\}\right),$$

による。$\{S_x,\Theta_{Sx}\}$が、y^Nを生成したソースにマッチする場合、前記不等式は、等式になる。最も類似している指標付けされたＤＮＡ配列は、$I(Y^N;\{S_x,\Theta_{Sx}\})$を最大化するものである。

【００３１】

ここで図２を参照すると、クエリＤＮＡ配列y^Nに類似しているＤＮＡ配列を識別するように図１のシステムにより生成されたＤＮＡ配列指標２０を検索するシステムが、記載される。クエリＤＮＡ配列y^N４０が、受け取られる。文脈木重み付け（ＣＴＷ）モジュール１２（図１の指標付けシステムと併せて既に記載されている）は、クエリＤＮＡ配列y^Nに対するモデル及びパラメータ$\{S_y,\Theta_{Sy}\}$を算出するのに使用され（これはツーパスバージョンのＣＴＷの第１のパスである）、符号語長推定器モジュール４２は、$\{S_y,\Theta_{Sy}\}$を使用して得られた最適な（最小の）符号語長$L_{ctw}(y^N)$を推定するのに式（１）を使用する。

【００３２】

各指標付けされたＤＮＡ配列x^Tは、次いで、現在試験下の指標付けされたＤＮＡ配列x^Tに対する指標エントリを検索する検索モジュール５２を起動することにより開始する、試験ループ５０の反復により試験される。この指標エントリは、ＣＴＷを使用して（すなわち、図１を参照して記載されたＣＴＷモジュール１２により）x^Tに対して算出され

たモデル及びパラメータセット{S_x,Θ_{Sx}}を提供する。オペレーション５４において、式（１）は、x^Tに対して算出されたモデル及びパラメータセット{S_x,Θ_{Sx}}を使用してモデル化されたクエリ配列y^Nに対して（非最適、及び一般的により大きい）符号語長$L(y^N|S_x,\Theta_{Sx})$を推定するのに再び使用される。換言すると、オペレーション５４は、ツーパスＣＴＷアルゴリズムの第２のパスを実行するが、x^Tに対して算出されたモデル及びパラメータセット{S_x,Θ_{Sx}}を使用する。試験ループ５０は、相互情報量の推定値$(1/N)L_{ctw}(y^N)-(1/N)L(y^N|S_x,\Theta_{Sx})$を計算することにより終了する。

【００３３】

代案として、オペレーション５４は、省略されることができ、式（２）の最後の表現が、$(1/N)L_{ctw}(y^N)-(1/N)L(y^N|S_x,\Theta_{Sx})$を直接的に計算するのに、代わりに使用されることができる。

【００３４】

試験ループ５０は、試験下の各指標付けされたＤＮＡ配列x^Tに対して繰り返される。（これは、ＤＮＡ指標２０において指標付けされたあらゆるＤＮＡ配列であってもよく、又は代わりに、匿名化された注釈に基づいてフィルタリングすることにより生成される前記指標のサブセットであってもよい）。セレクタモジュール６０は、次いで、クエリＤＮＡ配列y^Nに最も類似している１つ（又はそれ以上）の指標付けされたＤＮＡ配列を選択する。これは、例えば式（３）により、単一の最も類似している指標付けされたＤＮＡ配列を選択してもよく、又は「上位Ｋ」の最も類似している指標付けされたＤＮＡ配列が、選択されてもよく（すなわち、最も高い相互情報量を持つＫの指標付けされたＤＮＡ配列）、「上位Ｋ」の最も類似している指標付けされたＤＮＡ配列は、相互情報計量により測定される類似性によりランク付けされ、又は閾値が使用されてもよく、例えば相互情報計量が閾値を超過する全ての指標付けされたＤＮＡ配列が、選択される、又はその他である。出力モジュール６２は、次いで、セレクタモジュール６０により選択された前記１以上の最も類似している指標付けされたＤＮＡ配列を表示する又は他の形で人間知覚可能形式で提示する。

【００３５】

図２の説明的な例において、処理コンポーネント１２、４２、５０、６０、６２は、処理コンポーネント１２、４２、５０、６０、６２の機能を実施する適切なソフトウェアにより、指標付けモジュール１２、１８、２４、２６を実施する同じコンピュータ３０又は他の電子データ処理装置により実施される。代わりに、異なるコンピュータが、それぞれ図１及び２のシステムにより実行される指標付け及び検索オペレーションに対して使用されてもよい。出力モジュール６２は、前記選択された指標付けされたＤＮＡ配列に関する情報をディスプレイ３２上に表示してもよく、又はこの情報を他のコンピュータ（例えば暗号化ＤＮＡ配列データベース２８に対するアクセスを制御するリポジトリコンピュータ）に送信してもよく、又は（プリンタ又は他のマーキングエンジンと連動して）印刷されたレポートを生成してもよく、又はその他であってもよい。これが、データセキュリティ及び対象プライバシを危険にさらすので、出力モジュール６２が、典型的には、実際の指標付けされたＤＮＡ配列を実際に符号及び提供しないと理解されるべきである。むしろ、前記出力モジュールは、（クエリＤＮＡ配列y^Nに対する類似性に基づいて）関心配列を識別子、実際の配列は、適切なセキュリティ検査処理が実行された後に復号され、認可された個人に提供される。

【００３６】

ＤＮＡ配列指標付けモジュール１２、１８、２４、２６及び／又はＤＮＡ配列検索モジュール１２、４２、５０、６０、６２が、指標付けモジュール１２、１８、２４、２６及び／又は検索モジュール１２、４２、５０、６０、６２の機能を実行するようにコンピュータ３０により実行可能な命令（すなわちソフトウェア）を符号化する非一時的記憶媒体として実施されうるとも理解されるべきである。前記非一時的記憶媒体は、例えば、ハードディスクドライブ又は他の磁気記憶媒体、ランダムアクセスメモリ（ＲＡＭ）、読取専用メモリ（ＲＯＭ）、フラッシュメモリ又は他の電子記憶媒体、光ディスク又は他の光記

憶媒体、又はこれらの様々な組み合わせ等の１以上を有してもよい。

【００３７】

簡潔な総括のために、図１の例示的な指標付けシステムの実施例は、ＤＮＡ配列（のセット）x_i^{Ti}, ｉ＝１，２，...，ｎのＤＮＡデータベース２８及び対応する匿名化されたＤＮＡ配列指標２０を作成することを含む指標付けを実行する。これを行うために、モデル及びパラメータ$\{S_{xi}, \Theta_{Sxi}\}$は、前記ＣＴＷ方法を適用することにより各ＤＮＡ配列（のセット）x_i^{Ti}, ｉ＝１，２，...，ｎに対して推定され、$\{S_{xi}, \Theta_{Sxi}\}$セットは、他の関連情報（すなわち、注釈、オプションとして匿名化される）と一緒に指標データベース２０に記憶される。

【００３８】

図２の検索プロセスは、クエリ（例）ＤＮＡ配列y^N４０を与えられる。前記ＣＴＷアルゴリズムが、適用され、ソースシンボルごとの符号語長$(1/N)L_{ctw}(y^N)$が、モジュール１２、４２を使用してy^Nに対して推定される。指標データベース２０内の各ＤＮＡ指標記録ｉ，ｉ＝１，２，...，ｎに対して、前記符号語長は、$\{S_{xi}, \Theta_{Sxi}\}$を仮定して、y^N内の部分配列をS_{xi}からの文脈にマッピングし、対応するパラメータを使用して

$$\frac{1}{N}L\left(y^N \mid S_{x_i}, \Theta_{S_{x_i}}\right) = -\sum_{t=1}^{N} \log_2 \theta_{S_{x_i}, \sigma_{\{y_{-D}^{t-1}\}}}^{y_t}$$

を計算する（ＣＴＷ第２パスモジュール５４）ことによりy^Nに対して推定される。（y^Nからのある部分配列に対するS_{xi}内に文脈が存在しない場合、対応するパラメータは、１／２のような何らかの適切な値に適切にセットされる。）情報利得推定値$(1/N)L_{ctw}(y^N) - (1/N)L(y^N \mid S_{xi}, \Theta_{Sxi})$を最大化するＤＮＡ配列を指標付けする記録

$$\hat{\imath}$$

.

が、選択され（モジュール６０）、前記関連情報が、クエリを行っているパーティに返される（モジュール６２）。

【００３９】

指標データベース２０において、ＤＮＡ配列（のセット）に対応するモデル及びパラメータセット$\{S_{xi}, \Theta_{Sxi}\}$を記憶することのみを必要とすることが理解される。この情報は、実際の配列を生成したソースの確率的特徴のみを提供するので、単独では、前記ＤＮＡ配列を再構成するのに使用されることができない。

【００４０】

図３を参照すると、開示された検索プロセスの説明的な例が、記載される。この例は、GenBankからの１４のＤＮＡ配列を使用する。ゴールは、染色体ごとにデータベースを構成することである。この例において、前記ＣＴＷ方法は、各染色体、すなわち本例において染色体１，２，３，５，８，９，１０，１４に対して前記モデル及びパラメータセットを推定するのに深度Ｄ＝９（３つのコドンに対応する）を使用する。これらのモデル及びパラメータセットは、前記指標データベースに記憶される。前記クエリＤＮＡ配列は、人間のＤＮＡ配列フラグメントであり、ゴールは、これがいずれの染色体から来るのかを決定することである。染色体１，２，３，５，８，９，１０，１４に対応する前記指標付けされたＤＮＡ配列とともに図２の検索システムを使用して、前記クエリＤＮＡ配列フラグメントと異なる（指標付けされた）染色体に対応する前記モデル及びパラメータとの間の相互情報計量の推定値が、計算され、前記相互情報計量を最大化する染色体が、返される。図３は、複数のクエリ配列に対するこのような推定値の結果を提示する。図３において観測されるのは、提案された方法が、ＤＮＡのクエリピースがいずれの染色体からくるのかを正しく検出したことである。注意すべきは、前記クエリＤＮＡフラグメントが、完全

な染色体ではなく、むしろ、ＤＮＡ配列長Ｎのクエリフラグメントy^Nが、長さＴの指標付けされた（完全な染色体）ＤＮＡ配列x^Tの小さな一部であることである。

【００４１】

例示的な実施例は、例として意図され、多くの変形例が考えられる。例えば、ＣＴＷが、例示的実施例において採用されているが、様々な有限長マルコフ連鎖モデル又は可変次数マルコフモデルのような、他の有限記憶木ソースモデルが、採用されることができる。一般に、前記アプローチは、（好ましくは暗号化された）データベース２８に記憶されたＤＮＡ（又はＲＮＡ）配列に対する配列モデルを有する配列指標２０を生成する。データベース２８に記憶された各ＤＮＡ（又はＲＮＡ）配列に対する配列モデルは、有限記憶木ソースモデル及び前記有限記憶木ソースモデルに対するパラメータを有する。説明用の例において、各指標付けされたＤＮＡ配列x^Tに対する前記配列モデルは、ＣＴＷを使用してx^Tから算出されたモデル及びパラメータセット$\{S_{xi},\Theta_{Sxi}\}$である。

【００４２】

検索フェーズにおいて、データベース２８に記憶された１以上のＤＮＡ（又はＲＮＡ）配列は、クエリＤＮＡ（又はＲＮＡ）配列４０に対する前記配列モデルのフィッティングに基づいて前記クエリＤＮＡ（又はＲＮＡ）配列に最も類似しているとして識別される。例示的な実施例において、符号語長は、前記クエリＤＮＡ配列に対する前記配列モデルのフィッティングを評価するのに使用される。より一般的には、前記有限記憶木ソースモデルを使用して達成可能な前記クエリＤＮＡ配列の圧縮の量を測定するいかなる圧縮計量も、モデルフィットを評価するのに使用されることができる。前記圧縮計量が、より高いレベルの圧縮が前記クエリＤＮＡ（又はＲＮＡ）配列に前記モデルを適用することにより達成可能であることを示す場合に、前記配列モデルは、前記クエリＤＮＡ（又はＲＮＡ）配列に、より良好にフィットする。

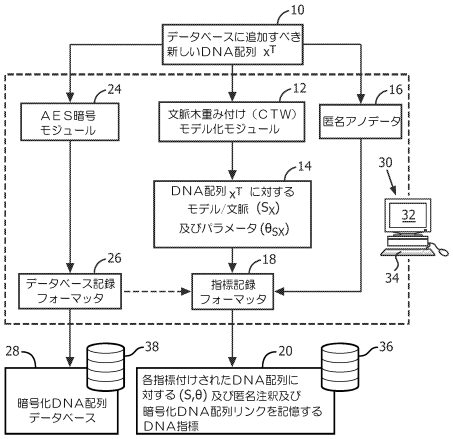
【００４３】

例示的な類似性（又は比較）計量は、（近似）情報利得（又は、同等に、相互情報量又はエントロピの変化）表現として定式化される。式（２）は、一例である。しかしながら、これらは、場合により単純化されることができる。例えば、Ｎによる正規化は、１つのクエリＤＮＡ配列のみが存在する（したがってＮが全ての場合において同じである）場合には、式（２）において省略されてもよい。実際に、１つのクエリＤＮＡ配列のみが、前記検索において採用されている場合、前記類似性計量は、$L_{ctw}(y^N)$項がこの場合に一定のオフセットであるので、$L(y^N|S_{xi},\Theta_{Sxi})$単独で与えられる推定符号語（すなわち圧縮計量）にされることができる。近似情報利得を得るために、前記類似性又は比較計量は、前記クエリＤＮＡ（又はＲＮＡ）配列から算出された有限記憶木ソースモデルを使用して前記クエリＤＮＡ（又はＲＮＡ）配列を圧縮するために得られた（ＣＴＷ符号語長推定値のような）圧縮計量の値（これは説明的な例において$(1/N)L_{ctw}(y^N)$である）を、前記データベースの前記ＤＮＡ（又はＲＮＡ）配列から算出された前記配列モデルを使用して前記クエリＤＮＡ（又はＲＮＡ）配列に対して得られた前記比較計量の値（これらは説明定な例において$(1/N)L(y^N|S_{xi},\Theta_{Sxi})$である）と適切に比較する。

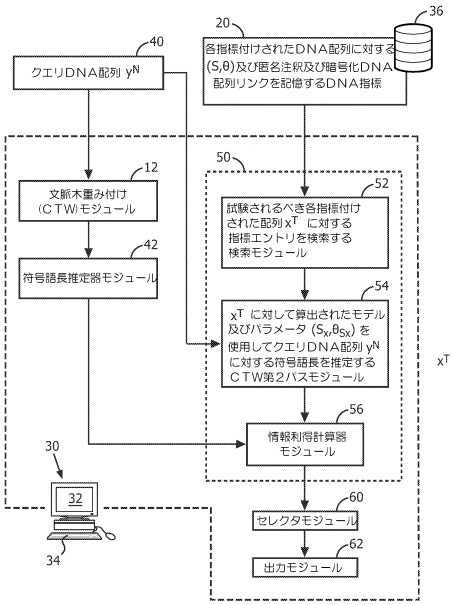
【００４４】

本発明は、好適な実施例を参照して記載されている。明らかに、修正例及び変更例は、先行する詳細な記載を読み、理解すると他者が思いつく。本発明が、添付の請求項又はその同等物の範囲内に入る限り、全てのこのような修正例及び変更例を含むと解釈されるべきである。

【図１】



【図２】



【図３】

クエリ染色体	染色体に対する相互情報量推定値													
	1	2	3	5	8	9	10	14						
1	-0.16883	-0.17662	-0.19062	-0.19116	-0.17525	-0.17702	-0.18031	-0.18617						
1	-0.0237	-0.02721	-0.03196	-0.03766	-0.03254	-0.02697	-0.03561	-0.03784						
1	-0.20133	-0.21518	-0.22017	-0.2222	-0.21824	-0.21132	-0.21977	-0.22331						
2	-0.00613	0.085012	-0.00542	-0.00982	-0.00741	-0.00223	-0.00618	-0.00994						
3	-0.02269	-0.01402	0.041464	-0.01267	-0.01713	-0.00675	-0.0218	-0.01881						
5	-0.07684	-0.06272	-0.06162	-0.00854	-0.06846	-0.05452	-0.07161	-0.07257						
5	-0.0971	-0.0648	-0.07114	-0.06463	-0.07918	-0.05804	-0.08502	-0.08603						
8	-0.01266	-0.01229	-0.01528	-0.01913	0.05676	-0.0103	-0.01446	-0.01544						
8	-0.02475	-0.02566	-0.04455	-0.03514	-0.02306	-0.02315	-0.02324	-0.03107						
9	-0.02467	-0.01264	-0.01365	-0.01563	-0.01858	0.073098	-0.02074	-0.02068						
10	-0.04395	-0.02693	-0.03615	-0.03762	-0.03394	-0.02079	-0.00575	-0.04144						
10	-0.04919	-0.03164	-0.03858	-0.04385	-0.0395	-0.02606	-0.00923	-0.04534						
10	-0.04458	-0.02924	-0.0371	-0.04147	-0.03617	-0.02454	-0.0071	-0.0417						
14	-0.05247	-0.05541	-0.05562	-0.05696	-0.05506	-0.05147	-0.05438	0.04525						

フロントページの続き

(72)発明者　イグナテンコ　ターニャ
オランダ国　５６５６　アーエー　アインドーフェン　ハイ　テック　キャンパス　ビルディング　５

審査官　塩田　徳彦

(56)参考文献　米国特許出願公開第２００４／００６８３３２（ＵＳ，Ａ１）
Z. Dawy ea al, Mutual information based distance measures for classification and content recognition with applications to genetics , Communications, 2005. ICC 2005. 2005 IEEE International Conference on, IEEE, ２００５年　5月１６日, pages 820-824, DOI: 10.1109/ICC.2005.1494466,
Kertesz-Farkas A, The Application of Data Compression-Based Distances to Biological Sequences, Springer, Boston, MA, ２００９年, pp 83-100, DOI: https://doi.org/10.1007/978-0-387-84816-7_4, Print ISBN: 978-0-387-84815-0, Online ISBN: 978-0-387-84816-7

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ　１９／１０　－　１９／２８
Ｃ１２Ｎ　１５／００
Ｃ１２Ｑ　１／６８