(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-92786
(P2005-92786A)
(43)公開日　平成17年4月7日(2005.4.7)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 19/00	G06F 19/00	600	4B024
C12N 15/09	C12Q 1/68	Z	4B063
C12Q 1/68	G06N 3/00	560A	
G06N 3/00	C12N 15/00	A	

審査請求 有　請求項の数 7　OL　（全 30 頁）

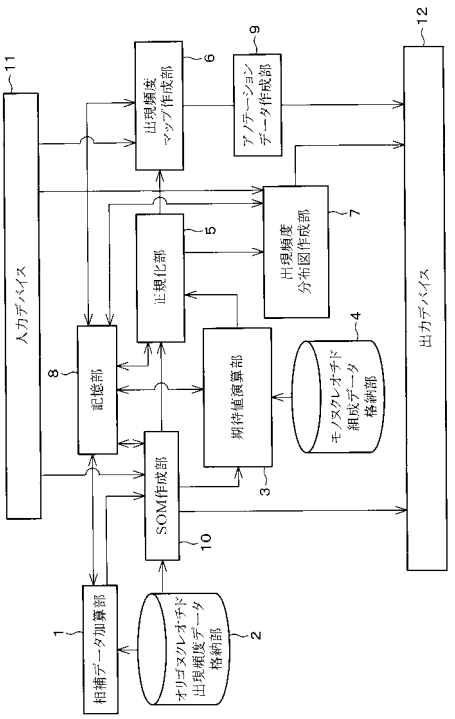
(21)出願番号　特願2003-328845 (P2003-328845)
(22)出願日　平成15年9月19日 (2003.9.19)

(71)出願人　504202472
大学共同利用機関法人情報・システム研究機構
東京都港区南麻布四丁目６番７号
(72)発明者　池村　淑道
静岡県三島市中１６９－１６
(72)発明者　阿部　貴志
静岡県三島市中田町１０－１１－１－Ｂ
(72)発明者　中川　智
東京都町田市中町３－９－９　協和アパートＥ－３
(72)発明者　上月　登喜男
千葉県松戸市岩瀬１２５－１－６０３

最終頁に続く

(54)【発明の名称】塩基配列の分類システムおよびオリゴヌクレオチド出現頻度の解析システム

(57)【要約】
【課題】　分類能力の目立った減少なしに短時間で自己組織化マップ（ＳＯＭ）を作成可能な塩基配列の分類システム、各生物学的分類における個々のオリゴヌクレオチドの出現頻度の把握や、ＤＮＡ塩基配列中のシグナル配列が多く存在する位置の予測を可能とするオリゴヌクレオチド出現頻度解析システムを提供する。
【解決手段】　オリゴヌクレオチド出現頻度解析システムは、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する相補データ加算部１と、各対ごとの出現頻度に基づいてＳＯＭを作成するＳＯＭ作成部と、オリゴヌクレオチドの出現頻度に関する情報を各格子点ごとに表した出現頻度マップを作成する出現頻度マップ作成部６と、ＤＮＡ配列上における個々のオリゴヌクレオチドの出現頻度の分布を示す出現頻度分布図を作成する出現頻度分布図作成部７とを備える。
【選択図】　図１



【特許請求の範囲】
【請求項１】
塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を複数の格子点が配置されたマップ上へ非線形に写像して上記塩基配列を各格子点に分類する自己組織化により、自己組織化マップを作成する塩基配列の分類システムであって、
相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部と、
各対ごとのオリゴヌクレオチドの出現頻度に基づいて上記自己組織化マップを作成する自己組織化マップ作成部とを備えることを特徴とする塩基配列の分類システム。
【請求項２】
塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を複数の格子点が配置されたマップ上へ非線形に写像して上記塩基配列を各格子点に分類する自己組織化により、自己組織化マップを作成する自己組織化マップ作成部と、
オリゴヌクレオチドの出現頻度に関する情報を各格子点ごとに表した出現頻度マップを個々のオリゴヌクレオチドについて作成する出現頻度マップ作成部とを備えることを特徴とするオリゴヌクレオチド出現頻度の解析システム。
【請求項３】
各格子点に分類された塩基配列中におけるモノヌクレオチド組成に基づいて、各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度の期待値を演算する期待値演算部と、
各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度を、上記期待値で除算することにより正規化する正規化部とをさらに備え、
上記出現頻度マップ作成部が、正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度マップを作成するようになっていることを特徴とする請求項２記載のオリゴヌクレオチド出現頻度の解析システム。
【請求項４】
相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、
上記自己組織化マップ作成部および出現頻度マップ作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップおよび出現頻度マップを作成するようになっていることを特徴とする請求項２または３に記載のオリゴヌクレオチド出現頻度の解析システム。
【請求項５】
同一のＤＮＡ配列から取り出した複数の断片塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を多次元空間から複数の格子点が配置されたマップ上へ自己組織化によって非線形に写像することにより、上記断片塩基配列が各格子点に分類された自己組織化マップを作成する自己組織化マップ作成部と、
各格子点に分類された断片塩基配列における個々のオリゴヌクレオチドの出現頻度に基づいて、ＤＮＡ配列上における個々のオリゴヌクレオチドの出現頻度の分布を示す出現頻度分布図を作成する出現頻度分布図作成部とを備えることを特徴とするオリゴヌクレオチド出現頻度の解析システム。
【請求項６】
各断片塩基配列中におけるモノヌクレオチド組成に基づいて、各断片塩基配列中におけるオリゴヌクレオチドの出現頻度の期待値を演算する期待値演算部と、
各断片塩基配列中におけるオリゴヌクレオチドの出現頻度を、上記期待値で除算することにより正規化する正規化部とをさらに備え、
上記出現頻度分布図作成部が、正規化されたオリゴヌクレオチドの出現頻度に基づいて

出現頻度分布図を作成するようになっていることを特徴とする請求項５記載のオリゴヌクレオチド出現頻度の解析システム。
【請求項７】
相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、
上記出現頻度分布図作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっていることを特徴とする請求項５または６に記載のオリゴヌクレオチド出現頻度の解析システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度に基づいて、塩基配列を生物学的分類に分類するための自己組織化マップを作成する塩基配列の分類システム、および、上記自己組織化マップを用いてオリゴヌクレオチドの出現頻度の偏り（種のような生物学的分類による偏りや、ＤＮＡ配列の位置による偏り）を解析するためのオリゴヌクレオチド出現頻度の解析システムに関するものである。
【背景技術】
【０００２】
因子対応分析や主成分分析（ＰＣＡ）のような多変量分析が、遺伝子配列の差異を調査するのに用いられ、成功を収めている。しかしながら、従来の多変量分析のクラスタリング能力は、多種多様なゲノムから得られた大量の配列データを集合的に分析する場合には、不十分である。
【０００３】
コホネンが開発した、競合ニューラルネットワークを利用した自己組織化マップ（Self Organizing Map；以下、「ＳＯＭ」と略記する）は、画像、音声や指紋等の認識や工業製品の生産プロセスの制御に利用されてきた（非特許文献１、非特許文献２）。ＳＯＭは、多次元データを結合重みベクトルの２次元配列上に非線形写像したものであり、高次元データ空間のトポロジーを効果的に保存する。ＳＯＭは、高次元の複雑なデータを二次元平面上にクラスタリングおよび視覚化するための強力なツールである。
【０００４】
近年、様々の生物のゲノム情報の解明に伴い、膨大な量の生命情報が蓄積しつつあり、コンピュータを用いてこれら生命情報から生命の謎を解くことも医薬開発等の面から重要になり、ＳＯＭの応用が盛んになっている。本願発明者等は、ゲノム情報科学のために従来のＳＯＭ作成法を改良した改良型のＳＯＭ作成法を提案した（特許文献１、非特許文献３・４参照）。この改良は、学習プロセスおよび作成されるマップ（ＳＯＭ）がデータ入力の順序に依存しないよう、データ入力および学習を一括処理する一括学習ＳＯＭ作成法に基づいている。また、改良型ＳＯＭ作成法では、主成分分析（ＰＣＡ）を使用して初期結合重みベクトルを定義している。したがって、改良型ＳＯＭは、データ入力の順序だけでなく初期条件にも依存しない。
【０００５】
例えば、特許文献１の実施例１では、高次元の入力データとしての１６種類の微生物のコドン（トリヌクレオチド）使用頻度に基づいて、改良型のＳＯＭを用いて微生物の遺伝子を分類したＳＯＭを作成する方法が開示されている。
【特許文献１】国際公開第ＷＯ　０２／５０７６７　Ａ１号（２００２年６月２７日公開）
【非特許文献１】自己組織化マップの応用－多次元情報の２次元可視化」（徳高平蔵、岸田悟、藤村喜久郎著、海文堂出版株式会社、１９９９年７月２０日初版発行、ＩＳＢＮ４－３０３－７３２３０－３）
【非特許文献２】「自己組織化マップ（Self Organizing-Map）」（Ｔ，コホネン著、徳高平蔵、岸田悟、藤村喜久郎訳、シュプリンガー・フェアラーク東京株式会社、１９９６

年６月１５日発行、ＩＳＢＮ４－４３１－７０７００－ＸＣ３０５５）
【非特許文献３】Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y, Yamada, V., Nishi, T., Marl, H. and Ikemura, T. (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the E. coli 0157 genome. Gene 276, 89-99.
【非特許文献４】Abe, T., Kanaya, S., Kinouchi, M., Ichiba, V., Kozuki, T. and Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. Genome Res. 13, 693-702.
【発明の開示】
【発明が解決しようとする課題】
【０００６】
公開ＤＮＡデータベース内には、相補的なＤＮＡ配列の対のうちの一方だけの配列データが登録されている。特許文献１や非特許文献３・４ではこの配列データにおけるオリゴヌクレオチドの出現頻度を用いてＳＯＭを作成している。
【０００７】
しかしながら、ゲノム内におけるオリゴヌクレオチド出現頻度の全体的な特徴を考慮すれば、相補的なＤＮＡ配列の対のうちの一方の配列における、相補的なオリゴヌクレオチド間（例えばＡＡＡＣ対ＧＴＴＴ間）の出現頻度の違いは、分類には重要ではない。むしろ、二本鎖ＤＮＡ全体においては、相補的なオリゴヌクレオチドの出現頻度は同一になるはずであるので、上記の出現頻度の違いは、作成されたＳＯＭが二本鎖ＤＮＡ全体における特徴を正確に反映しない結果となる可能性もある。また、ＳＯＭの作成は、長い演算時間を必要とするので、少しでも演算時間を短縮することが望まれる。
【０００８】
本発明の第１の目的は、分類能力の目立った減少なしに短時間で自己組織化マップを作成可能な塩基配列の分類システムを提供することにある。
【０００９】
また、特許文献１や非特許文献３・４においてオリゴヌクレオチドの出現頻度を用いて作成されたＳＯＭでは、塩基配列の由来する微生物を複数の生物学的分類に分類することができる。しかしながら、このＳＯＭでは、各生物学的分類に属する生物由来の塩基配列中における個々のオリゴヌクレオチドの出現頻度を把握することができない。各生物学的分類に属する生物由来の塩基配列中における個々のオリゴヌクレオチドの出現頻度を把握することができれば、例えば生物学的分類を分ける重要な鍵となるオリゴヌクレオチドを見つけ出すことができ、有用である。
【００１０】
本発明の第２の目的は、各生物学的分類に属する生物由来の塩基配列中における個々のオリゴヌクレオチドの出現頻度を把握することを可能とするオリゴヌクレオチド出現頻度の解析システムを提供することにある。
【００１１】
また、ヒトのゲノムは、全長のドラフト配列が決定されているが、その中で機能の分かっていない領域がまだまだ大量に残されている。ヒトのＤＮＡのような非常に長いＤＮＡ塩基配列の中から、シグナル配列（転写因子を認識する等の機能を有する塩基配列）を見つけ出し、その機能を解析するのは至難の業である。そのため、ＤＮＡ塩基配列の中からシグナル配列が多く存在する位置を予測できれば、機能の解析に有用である。
【００１２】
本発明の第３の目的は、ＤＮＡ塩基配列の中からシグナル配列が多く存在する位置を予測することを可能とするオリゴヌクレオチド出現頻度の解析システムを提供することにある。
【課題を解決するための手段】
【００１３】
本発明の塩基配列の分類システムは、上記の課題を解決するために、塩基配列中におい

て複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を複数の格子点が配置されたマップ上へ非線形に写像して上記塩基配列を各格子点に分類する自己組織化により、自己組織化マップを作成する塩基配列の分類システムであって、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部と、各対ごとのオリゴヌクレオチドの出現頻度に基づいて上記自己組織化マップを作成する自己組織化マップ作成部とを備えることを特徴としている。

【００１４】

本発明のオリゴヌクレオチド出現頻度の解析システムは、上記の課題を解決するために、塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を複数の格子点が配置されたマップ上へ非線形に写像して上記塩基配列を各格子点に分類する自己組織化により、自己組織化マップを作成する自己組織化マップ作成部と、オリゴヌクレオチドの出現頻度に関する情報を各格子点ごとに表した出現頻度マップを個々のオリゴヌクレオチドについて作成する出現頻度マップ作成部とを備えることを特徴としている。

【００１５】

上記解析システムは、各格子点に分類された塩基配列中におけるモノヌクレオチド組成に基づいて、各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度の期待値を演算する期待値演算部と、各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度を、上記期待値で除算することにより正規化する正規化部とをさらに備え、上記出現頻度マップ作成部が、正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度マップを作成するようになっていることが好ましい。

【００１６】

上記解析システムは、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、上記自己組織化マップ作成部および出現頻度マップ作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップおよび出現頻度マップを作成するようになっていることが好ましい。

【００１７】

本発明のオリゴヌクレオチド出現頻度の解析システムは、上記の課題を解決するために、同一のＤＮＡ配列から取り出した複数の断片塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を多次元空間から複数の格子点が配置されたマップ上へ自己組織化によって非線形に写像することにより、上記断片塩基配列が各格子点に分類された自己組織化マップを作成する自己組織化マップ作成部と、各格子点に分類された断片塩基配列における個々のオリゴヌクレオチドの出現頻度に基づいて、ＤＮＡ配列上における個々のオリゴヌクレオチドの出現頻度の分布を示す出現頻度分布図を作成する出現頻度分布図作成部とを備えることを特徴としている。

【００１８】

上記解析システムは、各断片塩基配列中におけるモノヌクレオチド組成に基づいて、各断片塩基配列中におけるオリゴヌクレオチドの出現頻度の期待値を演算する期待値演算部と、各断片塩基配列中におけるオリゴヌクレオチドの出現頻度を、上記期待値で除算することにより正規化する正規化部とをさらに備え、上記出現頻度分布図作成部が、正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっていることが好ましい。

【００１９】

上記解析システムは、各相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、上記出現頻度分布図作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっていることが好ましい。

【発明の効果】
【００２０】
本発明の分類システムによれば、各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップを作成することができる。
【００２１】
自己組織化マップは、多種類の塩基配列を生物学的分類に分類するための生物情報科学的ツールとして有用である。
【００２２】
また、自己組織化マップを利用すれば、成分が不明な細菌ＤＮＡの混合サンプル中にどのような種類の微生物由来のＤＮＡがどれだけの数存在するかを効率的に予測することが可能になる。したがって、自己組織化マップは、培養が困難な自然環境上の微生物の混合物等のような複数種の微生物を含む混合サンプルの成分分析に特に有用である。
【００２３】
また、自己組織化マップを利用すれば、生物学的分類に関するいかなる情報もない塩基配列の分類が可能になる。したがって、ＳＯＭは、そのゲノムの塩基配列の一部のみが分かっており、種が全く未知の生物（細菌等）が、どの系統群に属するかを特定するのに有用である。それゆえ、自己組織化マップは、新規で産業上有用な細菌等を探索するのに有用である。
【００２４】
また、自己組織化マップを利用すれば、ゲノム中から水平伝達を通じて他の種から導入されたと考えられるセグメントを見つけることも可能となる。
自己組織化マップによれば、詳しくは後段で述べるが、
本発明の分類システムによれば、加算部で、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより各対ごとのオリゴヌクレオチドの出現頻度を算出し、この各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップ作成部で自己組織化マップを作成するので、自己組織化マップ作成部における演算量を半減することができる。それゆえ、一般に長い演算時間を必要とする自己組織化マップ作成部における演算時間を約半分に短縮することができる。また、相補的な対をなす２つのオリゴヌクレオチドの間での出現頻度の違いは塩基配列の分類には重要ではないので、相補的な対をなすオリゴヌクレオチドを同一とみなして処理を行うことで、分類能力の目立った減少なしに自己組織化マップを作成することができる。
【００２５】
したがって、本発明によれば、分類能力の減少なしに短時間で自己組織化マップを作成可能な塩基配列の分類システムを提供できる。
【００２６】
自己組織化マップ作成部と出現頻度マップ作成部とを備える本発明のオリゴヌクレオチド出現頻度の解析システムでは、自己組織化マップを作成すると共に、自己組織化マップ上の格子点に対応してオリゴヌクレオチドの出現頻度に関する情報を表す出現頻度マップを作成することができる。自己組織化マップは、生物学的分類ごとの領域に分離されるので、自己組織化マップを参照しながら出現頻度マップを見れば、各生物学的分類に属する生物由来の塩基配列中における個々のオリゴヌクレオチドの出現頻度の特徴抽出ができる。例えば特定の生物学的分類に属する生物由来の塩基配列中において過剰に出現するオリゴヌクレオチドの種類や、特定の生物学的分類に属する生物由来の塩基配列中において過少に出現するオリゴヌクレオチドの種類等を把握することができる。それゆえ、例えば、生物学的分類を分ける鍵となる重要なオリゴヌクレオチドを見つけることができる。
【００２７】
上記解析システムは、期待値演算部と正規化部とをさらに備え、出現頻度マップ作成部が、正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度マップを作成するようになっている構成であれば、異なる格子点に分類された塩基配列間でのオリゴヌクレオチドの出現頻度の差を、各塩基配列におけるモノヌクレオチド組成の偏りから切り離して

、正確に検出することができる。したがって、異なる生物学的分類に属する生物間でのオリゴヌクレオチドの出現頻度の違いをより正確に反映した出現頻度マップを作成できる。

【００２８】

上記解析システムは、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、上記自己組織化マップ作成部および出現頻度マップ作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップおよび出現頻度マップを作成するようになっている構成であれば、さらに次の効果が得られる。すなわち、加算部で、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより各対ごとのオリゴヌクレオチドの出現頻度を算出し、この各対ごとのオリゴヌクレオチドの出現頻度に基づいて自己組織化マップ作成部で自己組織化マップを作成するので、自己組織化マップ作成部における演算量を半減することができる。それゆえ、一般に長い演算時間を必要とする自己組織化マップ作成部における演算時間を約半分に短縮することができる。また、相補的な対をなす２つのオリゴヌクレオチドの間での出現頻度の違いは塩基配列の分類には重要ではないので、相補的な対をなすオリゴヌクレオチドを同一とみなして処理を行うことで、分類能力の目立った減少なしに自己組織化マップを作成することができる。さらに、上記構成によれば、加算部で、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより各対ごとのオリゴヌクレオチドの出現頻度を算出し、この各対ごとのオリゴヌクレオチドの出現頻度に基づいて出現頻度マップ作成部で出現頻度マップを作成するので、出現頻度マップ作成部における演算量をも半減することができる。それゆえ、一般に長い演算時間を必要とする出現頻度マップ作成部における演算時間を約半分に短縮することができる。また、相補的な対をなす２つのオリゴヌクレオチドの間での出現頻度の違いは出現頻度マップには重要ではないので、相補的な対をなすオリゴヌクレオチドを同一とみなして処理を行うことで、正確性を低下させることなく出現頻度マップを作成することができる。

【００２９】

自己組織化マップ作成部と出現頻度分布図作成部とを備える本発明のオリゴヌクレオチド出現頻度の解析システムでは、ＤＮＡ配列上における個々のオリゴヌクレオチドの出現頻度の分布を示す出現頻度分布図を作成することができる。この出現頻度分布図により、ＤＮＡ配列上において、特定のオリゴヌクレオチドが過剰に出現する領域や、特定のオリゴヌクレオチドが過少に出現する領域を知ることができる。これらの領域の一部は、シグナル配列を多く含む領域や遺伝子リッチな領域に対応すると考えられる。それゆえ、上記出現頻度分布図により、シグナル配列を多く含む領域や遺伝子リッチな領域の位置を予測することが可能となる。

【００３０】

上記解析システムは、期待値演算部と正規化部とをさらに備え、出現頻度分布図作成部が、モノヌクレオチド出現頻度で正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっている構成であれば、異なる位置の断片塩基配列間でのオリゴヌクレオチドの出現頻度の差を、各塩基配列におけるモノヌクレオチド組成の偏りから切り離して、正確に検出することができる。したがって、異なる位置の断片塩基配列間でのオリゴヌクレオチドの出現頻度の違いをより正確に反映した出現頻度分布図を作成できる。

【００３１】

上記解析システムは、各相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより、各対ごとのオリゴヌクレオチドの出現頻度を算出する加算部をさらに備え、上記出現頻度分布図作成部が、各対ごとのオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっている構成であれば、さらに次の効果が得られる。すなわち、加算部で、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより各対ごとのオリゴヌクレオチドの出現頻度を算出し、この各対ごとのオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図作成部で出現頻度分布図を作成するので、出現頻度分布図作成部における演算量を半減することができる。それゆえ、一般に長い演算時間を必

要とする出現頻度分布図作成部における演算時間を約半分に短縮することができる。また、相補的な対をなす２つのオリゴヌクレオチドの間での出現頻度の違いは出現頻度分布図には重要ではないので、相補的な対をなすオリゴヌクレオチドを同一とみなして処理を行うことで、正確性を低下させることなく出現頻度分布図を作成することができる。

【発明を実施するための最良の形態】

【００３２】

本実施形態のオリゴヌクレオチド出現頻度解析システム（塩基配列の分類システム）は、図１に示すように、相補データ加算部（加算部）１と、オリゴヌクレオチド出現頻度データ格納部２と、期待値演算部３と、モノヌクレオチド組成データ格納部４と、正規化部５と、出現頻度マップ作成部６と、出現頻度分布図作成部７と、各種演算データを記憶するための記憶部８と、アノテーションデータ作成部９と、ＳＯＭ作成部１０と、データの入力や演算の実行指示のためのキーボードやマウスなどの入力デバイス１１と、表示や印刷等の出力を行うためのディスプレイやプリンタなどの出力デバイス１２とを備えている。

【００３３】

これらのうち、相補データ加算部１、期待値演算部３、正規化部５、出現頻度マップ作成部６、出現頻度分布図作成部７、アノテーションデータ作成部９、およびＳＯＭ作成部１０は、コンピュータと、相補データ加算部１、期待値演算部３、正規化部５、出現頻度マップ作成部６、出現頻度分布図作成部７、アノテーションデータ作成部９、およびＳＯＭ作成部１０としてコンピュータを機能させるためのコンピュータ・プログラムとによって実現される機能ブロックである。例えば、ＳＯＭ作成部１０は、コンピュータと、ＳＯＭ作成部１０としてコンピュータを機能させるための一括学習ＳＯＭプログラム、例えば株式会社ザナジェン製の“ＸａｎａＭｉｎｅ”とによって実現することができる。また、記憶部８はＲＡＭ等の書き換え可能なメモリによって実現でき、オリゴヌクレオチド出現頻度データ格納部２およびモノヌクレオチド組成データ格納部４は、ハードディスク等の大容量記憶装置によって実現できる。

【００３４】

オリゴヌクレオチド出現頻度データ格納部２は、公開されている塩基配列データベース等から供給された、複数の塩基配列中における１つずつのオリゴヌクレオチドの出現頻度のデータを格納している。上記複数の塩基配列は、異なる複数の生物種由来の塩基配列であってもよく、同一のＤＮＡ配列から取り出した複数の断片塩基配列であってもよい。断片塩基配列のデータを用いる場合、断片塩基配列の長さを一定長（例えば１０ｋｂや１００ｋｂ）に揃えることが好ましい。同一のＤＮＡ配列から取り出した複数の断片塩基配列を用いる場合、各断片塩基配列（領域）同士は、一部が重複していてもよく、全く重複していなくともよい。上記各断片塩基配列の長さは、１０ｋｂ以上が好ましく、１００ｋｂ以上がより好ましい。

【００３５】

オリゴヌクレオチド出現頻度データ格納部２に格納されたオリゴヌクレオチド出現頻度のデータは、例えば、次のような方法で用意すればよい。予め、既存のゲノム情報データベースや、各種の配列決定手法により新たに決定された遺伝子の全長配列あるいは断片塩基配列などから、複数のＤＮＡ全長塩基配列あるいは同一のＤＮＡ由来の複数の断片塩基配列のデータを入手する。同一のＤＮＡ由来の複数の断片塩基配列のデータを用いる場合、全長塩基配列から特定長の領域を複数切り出すことで得られた複数の断片塩基配列のデータを用いるとよい。次いで、このＤＮＡ塩基配列のデータに基づいて、ＤＮＡの全長塩基配列あるいは断片塩基配列におけるオリゴヌクレオチドの出現回数をカウントすることにより、オリゴヌクレオチドの出現頻度を求め、オリゴヌクレオチド出現頻度データ格納部２に記憶させる。分析対象の配列が既存のオリゴヌクレオチド出現頻度のデータベースに登録されている場合には、入力デバイス１１からの指示によって、そのデータをインターネット等を介して既存のデータベースからオリゴヌクレオチド出現頻度データ格納部２に取り込むだけでよい。また、オリゴヌクレオチド出現頻度データ格納部２へのオリゴヌ

クレオチド出現頻度のデータの格納は、入力デバイス１１を用いてオリゴヌクレオチド出現頻度の数値を直接的に入力する方法でも行うことができる。入力デバイス１１を用いて数値を入力する方法としては、入力デバイス１１としてのキーボード等を用いた手入力、音声入力、紙入力等を用いることができる。

【００３６】

上記オリゴヌクレオチド出現頻度は、分類能を向上させるために、３塩基以上のオリゴヌクレオチドの出現頻度であることが好ましい。一方、あまり塩基数の多いオリゴヌクレオチド出現頻度を用いると、演算されるデータ量が膨大となる。そのため、オリゴヌクレオチド出現頻度は、３～５塩基のオリゴヌクレオチド（トリヌクレオチド、テトラヌクレオチド、またはペンタヌクレオチド）の出現頻度であることが好ましい。また、オリゴヌクレオチド出現頻度のデータは、複数のオリゴヌクレオチドに関するオリゴヌクレオチド出現頻度の数値を含むものであればよいが、特定の塩基数のオリゴヌクレオチドの全種類（例えば６４種類のトリヌクレオチド）に関するオリゴヌクレオチド出現頻度の数値を含むものであることが好ましい。

【００３７】

また、上記オリゴヌクレオチド出現頻度には、パリンドロームに関係しない内部のｎ個の塩基Ｎ（ｎは１以上n_{max}以下の整数；n_{max}は自然数、例えば３）を持つパリンドロームのオリゴヌクレオチドの出現頻度を含めてもよい。すなわち、例えばオリゴヌクレオチド出現頻度としてのヘキサヌクレオチド出現頻度を分析する場合、パリンドロームに関係しない内部のｎ個の塩基Ｎ（ｎは１～３の整数）を持つパリンドロームのオリゴヌクレオチドの出現頻度、例えばＧＧＧＮＣＣＣ、ＧＧＧＮＮＣＣＣ、ＧＧＧＮＮＮＣＣＣの出現頻度を分析対象に含めてもよい。この場合、演算量は増えるが、近い種が異なる領域に分離されたＳＯＭを作成することが可能となる。

【００３８】

相補データ加算部１は、オリゴヌクレオチド出現頻度データ格納部２に格納された１つずつのオリゴヌクレオチドの出現頻度のデータに基づき、相補的な対をなすオリゴヌクレオチドの出現頻度を加算することにより各対ごとのオリゴヌクレオチドの出現頻度を算出し、算出結果のデータを出力する。

【００３９】

アノテーションデータ作成部９は、塩基配列データベース等を参照して得られた断片配列の生物学的意味のアノテーションを行う。ＳＯＭ作成部１０は、塩基配列中において複数種類のオリゴヌクレオチドがそれぞれ出現する出現頻度を入力ベクトル群として多次元空間上に配置し、これら入力ベクトル群を複数の格子点が配置されたマップ上へ非線形に写像して上記塩基配列を各格子点に分類する自己組織化により、自己組織化マップを作成する。ＳＯＭ作成部１０は、好ましくは、相補データ加算部１から供給される相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータに基づき、このデータを入力ベクトルのデータとして自己組織化を行うことにより、自己組織化マップを作成する。相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータを用いることにより、演算時間を短縮できると共に相補性の影響を除去できる。ＳＯＭ作成部１０は、オリゴヌクレオチド出現頻度データ格納部２に格納された１つずつのオリゴヌクレオチドの出現頻度のデータに基づき、このデータを入力ベクトルのデータとして自己組織化を行うことにより、自己組織化マップを作成するものであってもよい。ＳＯＭ作成部１０は、作成したＳＯＭのデータを出力デバイス１２を介して出力することにより、ＳＯＭを２次元や３次元の画像として出力（表示や印刷等）することができるようになっている。画像の出力形態としては、例えば、２次元画像上において、各格子点に分類された塩基配列が属する生物学的分類を色で表現し、各格子点に分類された塩基配列の数を濃度で表現する形態；３次元画像上において、各格子点に分類された塩基配列の数を棒の高さで表現する形態等が挙げられる。ＳＯＭ作成部１０は、作成したＳＯＭのデータに加えて、ＳＯＭ上の各格子点に分類された塩基配列の識別情報（ＩＤ）、各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度のデータ、アノテーション情報を出力しうるように構成されている。相補デー

タ加算部１から供給される相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータに基づいてＳＯＭが作成された場合には、ＳＯＭ作成部１０から出力される各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度のデータは、相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータ（相補データ加算部１から供給される相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータに対応）である。

【００４０】

作成されるＳＯＭは、比較ゲノム解析のための、新規で、強力で、かつ高感度のツールである。特に細菌ゲノムに関するＳＯＭは、有用である。すなわち、細菌ゲノムに関するＳＯＭは、培養が困難な自然環境上の微生物の混合物から得られたＤＮＡ配列を生物学的分類に分類するための生物情報科学的ツールとして特に有用である。ＳＯＭは、分類能が非常に高いので、莫大な量の細菌配列からの多種多様なゲノム情報を抽出するための効率的で強力なツールである。

【００４１】

環境中の微生物の大部分は、依然として培養不可能である。したがって、環境中の支配的な微生物の個体群についての我々の理解は制限されている。環境中での微生物の多様性を研究し、完全に新規で産業上有用な遺伝子を探索するために、極限環境のような環境中の微生物の混合物に由来したＤＮＡ断片を配列を決定することが多くのグループによって行われている。そのような研究では、ＤＮＡは、これら微生物の混合物から、培養や種のクローニングを行うことなく抽出される。また、ＤＮＡサンプルは、配列決定ベクターなどのようなベクターに断片化およびクローニングされ、その後にベクターの配列が決定される。したがって、ＤＮＡ混合物中にどのようなタイプのゲノムＤＮＡがどれだけの数存在し、それらの配列がどれくらい新規であるかを知ることは重要である。本発明のシステムで教師なしニューラル・ネットワーク・アルゴリズムを用いて作成されるＳＯＭは、これらのＤＮＡ混合物中のゲノムＤＮＡに関する情報を得るための強力な生物情報科学的ツールである。すなわち、予め既知の細菌ゲノムを用いてＳＯＭを作成し、ＳＯＭ上の領域を生物学的分類ごとに分離しておけば、成分が不明な細菌ＤＮＡの混合サンプルから得た断片の配列を用い、その断片の配列がＳＯＭ上におけるどの位置に存在するかに基づいて、上記混合サンプル中にどのような種類の微生物がどれだけの数存在するかを効率的に予測することが可能になる。したがって、ＳＯＭは、培養が困難な自然環境上の微生物の混合物等のような複数種の微生物を含む混合サンプルの成分分析に特に有用である。それゆえ、ＳＯＭは、新規で産業上有用な細菌等を探索するのに有用である。

【００４２】

また、ＳＯＭは、各ゲノムの署名的特徴である種特異的特徴（オリゴヌクレオチド出現頻度の鍵となる組み合わせ）を認識することができるので、予め既知のＤＮＡ配列を用いてＳＯＭを作成し、ＳＯＭ上の領域を種ごとに分離しておけば、ＳＯＭを用いて種に関するいかなる情報もないゲノム配列の種特異的分類が可能になる。したがって、ＳＯＭは、そのゲノムの塩基配列の一部のみが分かっており、種が全く未知の生物（細菌等）が、どの種に属するかを特定するのに有用である。それゆえ、ＳＯＭは、新規で産業上有用な細菌等を探索するのに有用である。

【００４３】

以上のように、多種多様な分類グループのゲノム配列が蓄積されれば、ＳＯＭは、細菌の配列の分類のための広範囲に適用可能で、かつ強力で、培養不可能な微生物の混合ＤＮＡサンプル（例えば海底堆積物のサンプル）から得られた塩基配列の分類に対して極めて有用なツールになると考えられる。

【００４４】

また、予め既知のＤＮＡ配列を用いてＳＯＭを作成してＳＯＭ上におけるある種の塩基配列がほぼ含まれる専有領域を特定しておけば、ある１つの種のＤＮＡ配列から切り出した断片塩基配列ごとのオリゴヌクレオチド出現頻度をＳＯＭ上にマップし、特定された専有領域から外れた位置に対応する断片塩基配列を探せば、水平伝達を通じて他の種から導入されたと考えられるセグメントを見つけることができる。

【００４５】
モノヌクレオチド組成データ格納部４は、分析対象の各塩基配列中におけるモノヌクレオチド組成のデータを格納している。期待値演算部３は、モノヌクレオチド組成データ格納部４に格納された分析対象の各塩基配列中におけるモノヌクレオチド組成のデータを参照してＳＯＭ上の各格子点に分類された塩基配列中におけるモノヌクレオチド組成の値を取得し、取得した各格子点のモノヌクレオチド組成（各格子点に分類された塩基配列中における各モノヌクレオチドの含有率）の値に基づいて、各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度の期待値を演算する。正規化部５は、ＳＯＭ上の各格子点に分類された塩基配列中におけるオリゴヌクレオチドの出現頻度を、期待値演算部３で演算された期待値で除算することにより正規化する。

【００４６】
出現頻度マップ作成部６は、オリゴヌクレオチドの出現頻度に関する情報を各格子点ごとに表した出現頻度マップを個々のオリゴヌクレオチドについて作成する。出現頻度マップ作成部６は、作成した出現頻度マップのデータを出力デバイス１２を介して出力することにより、２次元や３次元の画像として出力（表示や印刷等）することができるようになっている。画像の出力形態としては、例えば、２次元画像上において、各格子点に分類された塩基配列のオリゴヌクレオチド出現頻度を色および濃度で表現する形態；３次元画像上において、各格子点に分類された塩基配列の数を棒の高さで表現する形態等が挙げられる。出現頻度マップ作成部６は、正規化部５で正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度マップを作成するようになっている。出現頻度マップ作成部６は、オリゴヌクレオチド出現頻度データ格納部２に格納された１つずつのオリゴヌクレオチドの出現頻度のデータに基づいて出現頻度マップを作成するようになっていてもよいが、ＳＯＭ作成部１０から供給される相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータに基づいて出現頻度マップを作成するようになっていることがより好ましい。これにより、演算時間を短縮できると共に相補性の影響を除去できる。

【００４７】
出現頻度分布図作成部７は、分析対象の塩基配列群が同一のゲノム配列から取り出した複数の断片塩基配列である場合に、各格子点に分類された断片塩基配列における個々のオリゴヌクレオチドの出現頻度に基づいて、ゲノム配列上における個々のオリゴヌクレオチドの出現頻度の分布を示す出現頻度分布図を作成する。出現頻度分布図作成部７は、作成した出現頻度分布図のデータを出力デバイス１２を介して出力することにより、出現頻度分布図を１次元や２次元の画像として出力（表示や印刷等）することができるようになっている。画像の出力形態としては、例えば、ゲノム配列に対応する棒グラフ上において、各格子点に分類された断片塩基配列が属する生物学的分類を色および濃度で表現する形態；２次元画像上において、ゲノム配列上での位置をｘ軸座標、その位置に対応する断片塩基配列のオリゴヌクレオチド出現頻度をｙ軸座標で表現する形態等が挙げられる。出現頻度分布図作成部７は、正規化部５で正規化されたオリゴヌクレオチドの出現頻度に基づいて出現頻度分布図を作成するようになっている。出現頻度分布図作成部７は、ＳＯＭ作成部１０から供給される相補的な各対ごとのオリゴヌクレオチドの出現頻度のデータに基づいて出現頻度分布図を作成するようになっていることが好ましい。これにより、演算時間を短縮できると共に相補性の影響を除去できる。

【００４８】
次に、ＳＯＭ作成部１０およびそれによって実行されるマップ作成ステップについて詳細に説明する。

【００４９】
マップ作成ステップにおけるＳＯＭ作成法としては、コホネンの自己組織化法によるＳＯＭ作成法（以下、「コホネン法」と呼ぶ）、特許文献１に記載の自己組織化法による改良型ＳＯＭ作成法（以下、単に「改良型ＳＯＭ作成法」と呼ぶ）等を用いることができるが、改良型ＳＯＭ作成法を用いることが好ましい。

【００５０】

ここで、ＳＯＭを作成する自己組織化法について、基本原理を説明する。この自己組織化法は、ニューラルネットワークを用いて多次元の入力データを高次元空間から低次元空間へ非線形に写像（マッピング)することで、高次元空間内での入力データ同士の類似関係（入力データの特徴）を保ったまま低次元空間へ写像を行うことができるものである。この自己組織化法には、多次元の入力データがプロットされる高次元空間の入力層と、低次元空間に格子状に配置された複数の出力ニューロン（格子点）で構成された出力層との２層からなるニューラルネットワークを用いる。そして、入力値に対応する入力ベクトル、および、入力値に対応する点と出力ニューロンとの結合の重みを表す結合重みベクトル（ニューロンベクトル）とを用い、結合重みベクトルを初期値（初期結合重みベクトル）に設定した後、結合重みベクトルを修正することで学習を行う。

【００５１】

自己組織化法は、出力ニューロンの位置関係を考慮し学習を行うものである。ニューラルネットワークの出力層では、出力ニューロン間に相対的な位置関係（距離関係)が存在する。そして、入力データベクトルと最も距離が近い結合重みベクトルに対応する出力ニューロンおよびその近傍の出力ニューロンの結合重みベクトルに対して、結合重みベクトルの修正を行う（入力データベクトル近傍以外の結合重みベクトルに対しては修正を行わない）。これによって、ニューラルネットワークの学習が行われる。入力ベクトルと結合重みベクトルの距離の計算には、ユークリッド距離が使われる。結合重みベクトルの修正は、結合重みベクトルが入力ベクトルに近付くように行われる。例えば、結合重みベクトルを入力ベクトルに近付けるために、入力ベクトルと勝者ニューロンの結合重みベクトルの差を学習係数（学習係数は０～１）倍してから元の結合重みベクトルに加える。

【００５２】

このようにして、自己組織化法では、出力ニューロンの位置関係を考慮することによって、入力データ空間（入力層）における入力データ間の距離関係を保ったまま、入力データを高次元空間（出力層）にマッピングすることができる。

【００５３】

コホネン法は、次の３工程よりなる。工程１：各ニューロン上のベクトル（結合重みベクトル）を、乱数値を用いて初期化する。工程２：入力ベクトルに対して最も近い結合重みベクトルを持つニューロンを選択する。工程３：選択されたニューロン及びその近傍の結合重みベクトルを更新する。工程２と工程３は入力ベクトルの数だけ繰り返される。これを１回の学習として、決められた回数の学習を行う。学習後には、入力ベクトルは最も近い結合重みベクトルを持つニューロンに分類されることになる。コホネンのＳＯＭでは，高次元空間上の入力ベクトル群から低次元のマップ上に配置されたニューロン群に、特徴を保ちつつ非線形な写像を行える。

【００５４】

このコホネン法では、工程２および工程３で一つの入力に対する結合重みベクトルへの分類をもとに結合重みベクトルへの更新を行うため、後で入力されるベクトルほど精細に分離され、入力ベクトルの学習順により異なるＳＯＭが作成される。そのため、再現性のあるＳＯＭを得ることができない恐れがある。また、工程１の初期結合重みベクトル設定では乱数値をとっているために、乱数値の構造が学習後に得られる自己組織化マップに影響を及ぼすことにより、入力ベクトル以外の因子が自己組織化マップに反映される。そのため、入力ベクトルの構造が正確にＳＯＭに反映できない恐れがある。さらに、工程１で乱数値をとっているために、初期値が入力ベクトルの構造と大きく異なるときには、非常に長い学習時間を要する。また、工程２および３で一つの入力に対する結合重みベクトルへの分類をもとに結合重みベクトルへの更新を行うため、入力ベクトルの数に比例して学習時間が長くなる。

【００５５】

一方、改良型ＳＯＭ作成法は、非線形写像法によりコンピュータを用いて入力ベクトルデータを結合重みベクトルに分類する方法であって、（ａ）オリゴヌクレオチド出現頻度のデータを多次元の入力ベクトルのデータとして入力するステップと、（ｂ）初期結合重

みベクトルを設定するステップと、（ｃ）入力ベクトルを各結合重みベクトルへ分類するステップと、（ｄ）各結合重みベクトルに分類された入力ベクトルおよび該結合重みベクトルの近傍に分類された入力ベクトルと類似の構造となるように結合重みベクトルを更新するステップと、（ｅ）学習回数（繰り返し回数）が設定学習回数に達するまでステップ（ｃ）および（ｄ）を繰り返すステップと、（ｅ）入力ベクトルを結合重みベクトルへ分類し、結果をＳＯＭとして出力するステップとを含んでいる。

【００５６】

上記改良型ＳＯＭ作成法では、コホネン法における「一つの入力ベクトルを（初期）結合重みベクトルへ分類する」という逐次処理アルゴリズムを、「すべての入力ベクトルを結合重みベクトルに分類した後、個々の結合重みベクトルを更新する」という一括処理学習アルゴリズムに変更したことで、再現性のあるＳＯＭを得ることができると共に、入力ベクトルの数が多くなっても演算時間を短く抑えることができる。

【００５７】

上記ステップ（ａ）において、入力ベクトルデータが、Ｍ次元（Ｍは正の整数）からなるＫ個の入力ベクトルのデータ（Ｋは３以上の正の整数）であってもよい。

【００５８】

また上記ステップ（ｂ）においては、教師なし多変量解析法により得られる多次元から成る入力ベクトルの多次元空間の分布の特徴を、初期結合重みベクトルの配置および要素に反映させることにより、初期結合重みベクトルを設定することが好ましい。これにより、入力ベクトルの構造を正確にＳＯＭに反映できると共に、学習時間を短縮することができる。教師なし多変量解析法としては、主成分分析または多次元尺度構成法等を用いることができる。

【００５９】

上記ステップ（ｃ）において、入力ベクトルを各結合重みベクトルに分類する方法としては、距離、内積および方向余弦からなる尺度より選ばれる類似性の尺度に基づいた分類方法等を用いることができる。上述の距離としては、ユークリッド距離等を挙げることができる。

【００６０】

上記ステップ（ｄ）において、各結合重みベクトルに分類された入力ベクトルおよび該結合重みベクトルの近傍に分類された入力ベクトルと類似の構造となるように結合重みベクトルを更新する処理にも、一括処理学習アルゴリズムを用いることができる。

【００６１】

上記各ステップの処理、特に一括処理学習アルゴリズムを用いた処理は、並列コンピュータを用いて演算処理することが好ましい。これにより、演算時間を短縮することができる。

【００６２】

以下、上記ＳＯＭ作成部１０を用いた改良型ＳＯＭ作成法の各ステップについて詳述する。

【００６３】

〔ステップ（ａ）〕

まず、オリゴヌクレオチド出現頻度データ格納部２に格納された複数の塩基配列のオリゴヌクレオチド出現頻度のデータを多次元の入力ベクトルデータとしてＳＯＭ作成部１０へ入力する。

【００６４】

入力ベクトルデータは、通常、Ｋ個（Ｋは、塩基配列の数）の入力ベクトル｛x_1，x_2，...，x_k，...，x_K｝（ｋ＝１，２，．．．，Ｋ）から構成されている。各入力ベクトルx_kは、Ｍ次元（Ｍは、オリゴヌクレオチドの区分の数）のベクトルであり、下式（１）で表すことができる。

【００６５】

$$x_k = \{x_{k1}, x_{k2}, \ldots, x_{kM}\} \qquad (1)$$

（ここで、x_{k1}，x_{k2}，...，x_{kM}は、オリゴヌクレオチド出現頻度）

例えば、オリゴヌクレオチド出現頻度を６４種のトリヌクレオチドの出現頻度とし、Ｋ種類のＤＮＡの全長塩基配列のトリヌクレオチド使用頻度に基づいて複数の微生物を分類する場合には、これら微生物由来のＫ種類のＤＮＡの全長塩基配列のトリヌクレオチド使用頻度を６４次元に数値化し、数値化された６４次元のデータを入力ベクトルとして設定する。

【００６６】

塩基配列を複数の生物学的分類へ分類する場合、各生物学的分類の特徴をより正確に分析するために、各生物学的分類（例えば種）ごとに十分な数のＤＮＡ塩基配列に関するオリゴヌクレオチド出現頻度のデータを入力ベクトルのデータとして入力することが好ましい。解析対象とするＤＮＡ塩基配列（全長配列あるいは断片塩基配列）の数は、通常、各生物学的分類ごとに数百個～数万個程度用意すればよい。なお、入力ベクトルは、通常、非特許文献１、非特許文献２等に記載の常法に準じて設定できる。

【００６７】

〔ステップ（ｂ）〕

次に、コンピュータを用いて、ニューラルネットワークを構築し、初期結合重みベクトルを設定する。ニューラルネットワークにおける出力層上には、作成しようとするＳＯＭの次元Ｄ（Ｄは正の整数；Ｄ＜Ｍ）に応じて、出力ニューロン（格子点）をＤ次元の格子状に配置する。各出力ニューロン（格子点）の初期結合重みベクトルも、Ｄ次元の格子状に配置する。

【００６８】

初期結合重みベクトルは、非特許文献１、非特許文献２等に記載されたＳＯＭの作成法と同じく乱数値に基づいて設定することができる。入力ベクトルの構造を正確にＳＯＭに反映させたい、あるいは学習時間を短縮させたい場合には、乱数値に基づいて初期結合重みベクトルを設定するよりも、主成分分析や多次元尺度構成法等多変量解析法を用いて、上記工程（ａ）で設定したＭ次元からなるＫ個の入力ベクトル$\{x_1, x_2, ..., x_K\}$データに基づいて、初期結合重みベクトルを設定することが好ましい。このようにして設定された初期結合重みベクトルが、Ｄ次元の格子状に配置されているＰ個の結合重みベクトル$\{W^0_1, W^0_2, ..., W^0_P\}$の集合よりなる場合には、各結合重みベクトルは、下式（２）で表すことができる。

【００６９】

$$W^0_i = F\{x_1, x_2, ..., x_K\} \qquad (2)$$

式（２）において、ｉはｉ＝１，２，...，Ｐである。また、式（２）中の$F\{x_1, x_2, ..., x_K\}$は入力ベクトル$\{x_1, x_2, ..., x_K\}$から初期結合重みベクトルへの変換関数を表わす。具体例として、２次元（Ｄ＝２）および３次元（Ｄ＝３）の格子状に初期結合重みベクトルを設定する方法について説明する。該方法に準じて、Ｄ次元の格子状への初期結合重みベクトルの設定を行うことができる。

（１）２次元（Ｄ＝２）の格子状に初期結合重みベクトルを設定する方法

（２次元のＳＯＭを作成する場合）

Ｍ次元からなるＫ個の入力ベクトル$\{x_1, x_2, ..., x_K\}$に対して主成分分析を行い、第１主成分ベクトルおよび第２主成分ベクトルを求め、得られたこれら主成分ベクトルをb_1およびb_2とする。これら２つの主成分ベクトルをもとに、Ｋ個の入力ベクトルに対する主成分$Z_{1k} = b_1 x_k$および$Z_{2k} = b_2 x_k$を求める（ｋ＝１，２，...，Ｋ）。$\{Z_{11}, Z_{12}, ..., Z_{1k}, ..., Z_{1K}\}$および$\{Z_{21}, Z_{22}, ..., Z_{2k}, ..., Z_{2K}\}$の標準偏差をそれぞれ$\sigma_1$および$\sigma_2$とする。

【００７０】

入力ベクトルの平均値を求め、得られた該平均値をx_{ave}とする。

【００７１】

出力層上の２次元の格子点を、２次元平面（出力層）上の座標で、すなわちｉｊ（ｉ＝１，２，...Ｉ；ｊ＝１，２，...，Ｊ）で表現し、２次元の格子点（ｉｊ）上に結合重み

ベクトルW^0_{ij}を置く。ＩとＪの値は、３以上の整数であれば良い。Ｊは、Ｉ×σ_2／σ_1よりも小さい整数のなかで最大のものが好ましい。Ｉの値は入力ベクトルのデータ数に応じて適宜設定すればよい。Ｉの値は、通常５０～１０００であり、例えば１００である。

【００７２】

W^0_{ij}は、式（３）により定義することができる。

【００７３】

【数１】

$$W^0_{ij}=x_{ave}+5\sigma_1\left\{b_1\left[\frac{i-I/2}{I}\right]+b_2\left[\frac{j-J/2}{J}\right]\right\} \qquad (3)$$

（２）３次元（Ｄ＝３）の格子状に初期結合重みベクトルを設定する方法

（３次元のＳＯＭを作成する場合）

上述（１）の主成分分析において、第１主成分ベクトルと第２主成分ベクトルに加えて、第３主成分ベクトルを求め、得られた第１主成分ベクトル、第２主成分ベクトル、および第３主成分ベクトルをそれぞれ、b_1、b_2、およびb_3とする。これら３つの主成分ベクトルをもとに、主成分$Z_{1k}=b_1x_k$、$Z_{2k}=b_2x_k$、および$Z_{3k}=b_3x_k$を求める。｛Z_{11}，Z_{12}，...，Z_{1k}，...，Z_{1K}｝、｛Z_{21}，Z_{22}，...，Z_{2k}，...，Z_{2K}｝、および｛Z_{31}，Z_{32}，...，Z_{3k}，...，Z_{3K}｝の標準偏差をそれぞれσ_1、σ_2、およびσ_3とする。３次元の格子点をｉｊｌ（ｉ＝１，２，...，Ｉ；ｊ＝１，２，...，Ｊ；ｌ＝１，２，...，Ｌ）で表現し、３次元の格子点（ｉｊｌ）上に結合重みベクトルW^0_{ijl}を置く。Ｉ、Ｊ、Ｌの値は、３以上の整数であれば良い。Ｊは、Ｉ×σ_2／σ_1よりも小さい整数のなかで最大のもの、Ｌは、Ｉ×σ_3／σ_1よりも小さい整数のなかで最大のものが好ましい。Ｉの値は、入力ベクトルのデータ数に応じて適宜設定すればよい。Ｉの値は、通常５０～１０００であり、例えば１００である。W^0_{ijl}は、下式（４）により定義することができる

【００７４】

【数２】

$$W^0_{ijl}=x_{ave}+5\sigma_1\left\{b_1\left[\frac{i-I/2}{I}\right]+b_2\left[\frac{j-J/2}{J}\right]+b_3\left[\frac{l-L/2}{L}\right]\right\} \qquad (4)$$

〔ステップ（ｃ）〕

すべての入力ベクトル｛x_1，x_2，...，x_K｝を、各結合重みベクトルへ分類する。

【００７５】

具体的には、すべての入力ベクトル｛x_1，x_2，...，x_K｝を、類似性の尺度（距離、内積、方向余弦等）を利用してｔ回の学習（修正）を行った後の、Ｄ次元の格子状に配置されたＰ個の結合重みベクトルW^t_1，W^t_2，...，W^t_Pのいずれかに、コンピュータを用いて分類する。ここで、ｔは、学習の回数（エポック）、すなわちステップ（ｃ）の前に何回ステップ（ｄ）が実行されているかを表わす。合計Ｔ回学習を行う場合（Ｔは設定学習回数を表す）、ｔ＝０，１，２，．．．Ｔである。第ｔエポック（第ｔ回目の学習時、すなわち第ｔ回目のステップ（ｄ）実行後）におけるｉ番目の結合重みベクトルは、Ｗｔｉで表すことができる。ここで、ｉ＝１，２，．．．，Ｐである。ｔ＝０の時（ステップ（ｃ）を初めて実行する時）には、結合重みベクトルW^t_1，W^t_2，...，W^t_Pは、ステップ（ｂ）で設定した初期結合重みベクトルに相当する。各入力ベクトルの分類は、各結合重みベクトルW^t_iとのユークリッド距離を計算し、該入力ベクトルを最小のユークリッド距離を有する結合重みベクトルに割り当てることにより行うことができる。なお、２次元の格子点（ｉｊ）上に配置された結合重みベクトルの場合には、W^t_iはW^t_{ij}と表すことができる。

【００７６】

入力ベクトル｛x_1，x_2，...，x_K｝は、各入力ベクトルx_k毎に並列処理してW^t_iに分類することが可能である。

【００７７】

〔ステップ（ｄ）〕

各ニューロンベクトルW^t_iについて、該結合重みベクトルに分類された入力ベクトル（x_k）および該結合重みベクトルの近傍に分類された入力ベクトルと類似の構造となるように、結合重みベクトルW^t_iを更新する。

【００７８】

即ち、ある特定の結合重みベクトル$W^t_{i'}$が位置づけられている格子点に帰属する入力ベクトルの集合を$S_{i'}$とする。$S_{i'}$に属するＮ個のベクトル$x^t_1(S_{i'})$，$x^t_2(S_{i'})$，...，$x^t_N(S_{i'})$と、$W^t_{i'}$とから、$S_{i'}$に属する入力ベクトルの構造を反映させた新たな結合重みベクトル（$W^{t+1}_{i'}$）を、次式（５）の関数Ｇにより求めることにより、ニューロンベクトル$W^t_{i'}$（ｉ’＝１，２，...，Ｐ）を更新する。

【００７９】

$$W^{t+1}_{i'}=G(W^t_{i'}, x^t_1(S_{i'}), x^t_2(S_{i'}), \ldots, x^t_N(S_{i'})) \qquad (5)$$

具体例として、２次元の格子状に設定された結合重みベクトルW_{ij}の更新について説明する。他のＤ次元の格子状に設定された結合重みベクトルについても同様に行うことができる。

【００８０】

入力ベクトルx_kが２次元の格子状に配置された結合重みベクトルのW^t_{ij}に帰属し、W^t_{ij}が位置づけられている格子点の近傍の格子点に帰属する入力ベクトルの集合をS_{ij}としたとき、S_{ij}に属するN_{ij}個の入力ベクトル$x^t_1(S_{ij})$，$x^t_2(S_{ij})$，...，$x^t_{N_{ij}}(S_{ij})$およびW^t_{ij}から、S_{ij}に属する入力ベクトル構造を反映させる新たな結合重みベクトルx^{t+1}_{ij}を下式（６）により求めることにより、結合重みベクトルW^t_{ij}を更新することができる。

【００８１】

【数３】

$$W^{t+1}_{ij}=W^t_{ij}+\alpha(t)\left(\frac{\sum_{x_k\in S_{ij}} x_k}{N_{ij}} - W^t_{ij}\right) \qquad (6)$$

ここで、N_{ij}は、S_{ij}に分類された入力ベクトルの総数である。

【００８２】

α（ｔ）は、設定学習回数をＴと設定したときの第ｔエポックに対する学習係数（０＜α（ｔ）＜１）であり、単調減少関数を用いる。より好ましくは、下式（７）により求めることができる。

【００８３】

【数４】

$$\alpha(t)=\max\left\{0.01, 0.6\left[1-\frac{t}{T}\right]\right\} \qquad (7)$$

設定学習回数Ｔは、入力ベクトルのデータ数に応じて適宜設定すればよい。設定学習回数Ｔは、通常１０～１０００であり、例えば１００である。

【００８４】

近傍集合S_{ij}は、より好ましくは、ｉ－β（ｔ）≦ｉ’≦ｉ＋β（ｔ）かつｊ－β（ｔ）≦ｊ’≦ｊ＋β（ｔ）の条件を満たす格子点ｉ’ｊ’に分類された入力ベクトルx_{ij}の集合である。β（ｔ）は、近傍を決定する数であり、例えば式（８）により求める。

【００８５】

β（ｔ）＝ｍａｘ｛０，２５－ｔ｝　　（８）

結合重みベクトル｛W^t_1，W^t_2，...，W^t_P｝は、各結合重みベクトルW^t_i毎に並列処理して更新することが可能である。

【００８６】

〔ステップ（ｅ）〕

学習回数（ステップ（ｄ）を繰り返した回数）ｔが設定学習回数Ｔに達したか否かを判定し、学習回数ｔが設定学習回数Ｔに達していなければ、ステップ（ｃ）に戻り、ステップ（ｃ）およびステップ（ｄ）を再度行う。すなわち、学習回数ｔが設定学習回数Ｔに達するまで、ステップ（ｃ）およびステップ（ｄ）を繰り返し、学習を行う。そして、学習回数ｔが設定学習回数Ｔに達すると、次のステップ（ｆ）に移る。

【００８７】

〔ステップ（ｆ）〕

学習終了後、ステップ（ｃ）の方法に準じて、入力ベクトルx_kを結合重みベクトルW^t_iへ、コンピュータにより分類し、結果を出力する。入力ベクトルの構造を反映した、ザで表される分類の基準に基づいて、入力ベクトルx_kは分類される。即ち、複数の入力ベクトルが同一の結合重みベクトルに分類された場合には、これら入力ベクトルのベクトル構造は非常に類似していることがわかる。入力ベクトル｛x_1，x_2，...，x_K｝は、各入力ベクトルx_k毎に並列処理して分類することが可能である。

【００８８】

上記ステップで出力された分類結果にしたがってＳＯＭを作成する。作成したＳＯＭは、出力デバイス１２から出力（表示、印刷等）することにより可視化可能となる。ＳＯＭの作成および表示等は、非特許文献１、非特許文献２等に記載の方法に準じて行うことができる。例えば、２次元の格子点に結合重みベクトルを設定して得られた入力ベクトルの分類結果は、２次元のＳＯＭとして表示することができる。具体的には、２次元の格子点を有する結合重みベクトルの各格子点に帰属された入力ベクトルの属性に基づいて、各格子点に適当なラベルを付与した後、このラベルを２次元の格子に画面表示または印刷等により、ＳＯＭとして表示することができる。各格子点に帰属された入力ベクトルの総数の値を２次元の格子に画面表示または印刷等により、ＳＯＭとして表示することも可能である。

【００８９】

上記各ステップで使用するコンピュータとしては、計算速度の速いものが好ましい。上記ステップ（ａ）～（ｆ）は、同一のコンピュータを用いて行う必要はない。即ち、上記のあるステップで得られた結果を別のコンピュータに出力し、該コンピュータで次ステップの処理を行ってもよい。また、並列処理可能なステップ（ステップ（ｃ）～（ｆ））の演算処理は、マルチＣＰＵを有するコンピュータあるいは、複数台のコンピュータを用いて並列処理することも可能である。従来型のＳＯＭ作成法では、逐次処理学習アルゴリズムを採用しているために、並列処理することができないが、改良型のＳＯＭ作成法では、一括処理学習アルゴリズムを採用したことにより並列処理が可能である。並列処理が可能となることにより、入力ベクトルを分類するための演算時間を大幅に短縮することが可能となる。即ち、上記６ステップを一つのプロセッサで処理する時間をそれぞれ、T1,T2,T3,T4,T5およびT6とし、Ｃ個のプロセッサで並列処理すると、理想的には、それぞれのステップで要する時間は、T1,T2,T3/C,T4/C,T5/C,T6となり、全体では、

T1+T2+T3+T4+T5+T6-{T1+T2+(T3+T4+T5)/C+T6}=(1-1/C)(T3+T4+T5)

時間だけ、演算時間を短縮できる。

【実施例】

【００９０】

以下に、本発明の実施例を示す。

【００９１】

〔実施例１〕

（ＳＯＭ作成方法）

実施例１～３では、前記実施形態のシステムを用い、特許文献１に記載の改良型ＳＯＭ作成法にしたがって以下の方法で２次元および３次元のＳＯＭを作成した。初期結合重みベクトルは、乱数値の代わりに主成分分析（ＰＣＡ）によって定義した。これは、主成分分析が、比較的少量の配列を分析する場合には、遺伝子配列を既知の生物学的分類に分類できることに基づいている。結合重みベクトル（W_{ij}）は、ｉ（＝０，１，．．，Ｉ－１）およびｊ（＝０，１，．．，Ｊ－１）で表される２次元格子内に配列した。Ｉは、２５０に設定した。Ｊは、（σ_2／σ_1）×２５０より大きく、かつ最も近い整数として定義した（ここで、σ_1およびσ_2はそれぞれ第１および第２の主成分の標準偏差である）。結合重みベクトルは、非特許文献４に記載の方法で設定および更新した。この実施例で使用した一括学習ＳＯＭプログラム“ＸａｎａＭｉｎｅ”は、株式会社ザナジェンから入手した。

【００９２】

分析対象の配列のデータは、“ＧｅｎＢａｎｋ”(http://www.ncbi.nlm.nih.gov/Genbank/)から入手した。全長配列から切り出した断片塩基配列中における未決定ヌクレオチド（Ｎ）の数が、断片塩基配列の全長（窓サイズ；１０ｋｂまたは１００ｋｂ）の１０％を超えている場合には、当該断片塩基配列を分析対象から除外した。断片塩基配列中における未決定ヌクレオチド（Ｎ）の数が、窓サイズの１０％以下である場合には、未決定ヌクレオチド（Ｎ）を除く長さに対してオリゴヌクレオチドの出現頻度を正規化し、分析対象に含めた。

【００９３】

（１３種の真核生物のゲノムに対するＳＯＭ）

真核生物の配列に対するＳＯＭのクラスタリング能力を調査するために、本願発明者等は、まず初めに、１３種の真核生物のゲノム配列（合計３Ｇｂ）から切り出した、互いに重複していない３００，０００個の１０ｋｂの断片塩基配列と、１０ｋｂずつずれた約３００，０００個の１００ｋｂの断片塩基配列とにおけるトリヌクレオチド、テトラヌクレオチド、およびペンタヌクレオチドの出現頻度を分析した。これらのゲノム配列は、ヒト(Homo sapiens)、フグ(Fugu rubripes)、ゼブラフィッシュ(Danio rerio)、コメ(Oryza sativa)、シロイヌナズナ(Arabidopsis thaliana)、タルウマゴヤシ(Medicago truncatula)、キイロショウジョウバエ(Drosophila melanogaster)、線虫(Caenorhabditis elegans)、キイロタマホコリカビ(Dictyostelium discoideum)、熱帯熱マラリア原虫(Plasmodium falciparum)、赤痢アメーバ(Entamoeba histolytica)、分裂酵母(Schizosaccharomyces pombe)、およびパン酵母(Saccharomyces cerevisiae)のゲノム配列を含む。ヒトについては、ほぼ完全な配列データが入手できる染色体２，６，７，１３，１４，２０，２１，２２，Ｘ，およびＹ由来の配列を分析した。

【００９４】

そして、これら断片塩基配列について、ゲノム情報科学に適合させた改良型ＳＯＭを、非特許文献４に記載の方法で作成した。最初に、３００，０００個の１０ｋｂ断片塩基配列におけるオリゴヌクレオチド出現頻度（トリヌクレオチド、テトラヌクレオチド、およびペンタヌクレオチドの出現頻度）を主成分分析により分析し、第１および第２の主成分を用いて２次元格子として配列された初期結合重みベクトルを設定した。設定学習回数は８０に設定した。８０回の学習サイクルの後、１０ｋｂ断片塩基配列のオリゴヌクレオチド出現頻度を２次元の格子状に配置された最終結合重みベクトルで表すことができ、その結果としてＳＯＭが作成された。得られたＳＯＭは、明らかな種特異分離を示した。上記配列は、主として種特異領域にクラスタリングされた。

【００９５】

作成されたＳＯＭをカラー出力（カラー印刷やカラー表示等）により、単一の種由来の配列を含む格子点を有彩色で示し、複数の種由来の配列を含む格子点を黒色で示した（図示しない）。第１および第２の主成分によって設定された初期ベクトルによる分類（主成分分析）を、１０ｋｂの断片塩基配列のトリヌクレオチド出現頻度に関するＳＯＭ（以下、「１０ｋｂＴｒｉ－ＳＯＭ」と略記する）内で達成された分類と比較することにより、

１０ｋｂＴｒｉ－ＳＯＭ内では単一の種由来の配列が遥かに密にクラスタリングされていることが明確に分かった。種クラスタリングは、１０ｋｂの断片塩基配列のテトラヌクレオチド出現頻度に関するＳＯＭ（以下、「１０ｋｂＴｅｔｒａ－ＳＯＭ」と略記する）およびペンタヌクレオチドＳＯＭ（以下、「１０ｋｂＰｅｎｔａ－ＳＯＭ」と略記する）内ではさらに強まった。例えば、１０ｋｂＴｒｉ－ＳＯＭ内、１０ｋｂＴｅｔｒａ－ＳＯＭ内、および１０ｋｂＰｅｎｔａ－ＳＯＭ内ではそれぞれ、ヒト配列の９４％、９７％、および９８％がヒト領域に分類された。

【００９６】

ＤＮＡデータベース内では、相補的な配列の対のうちの一方だけが登録されている。ゲノム内におけるオリゴヌクレオチド出現頻度の全体的な特徴を考慮すれば、相補的なオリゴヌクレオチド間（例えばＡＡＡＣ対ＧＴＴＴ間）における出現頻度の違いは、重要ではない。テトラヌクレオチド出現頻度やペンタヌクレオチド出現頻度に関するＳＯＭの作成が長い演算時間を必要とすることも特筆すべきことである。

【００９７】

そこで、上記演算時間を削減する試みとして、相補的なオリゴヌクレオチドの対の出現頻度を加算した縮退セットの出現頻度を用いて、１０ｋｂの断片塩基配列のテトラヌクレオチド出現頻度に関するＳＯＭ、および１００ｋｂの断片塩基配列のペンタヌクレオチド出現頻度に関するＳＯＭ（以下、「１００ｋｂＤｅｇｅＰｅｎｔａ－ＳＯＭ」と略記する）を作成した。これにより、クラスタリング能力の目立った減少なしに演算時間を約半分にすることができた。

【００９８】

１０ｋｂＴｒｉ－ＳＯＭ内の各格子点に対する結合重みベクトルから得られたＧＣ含量（Ｇ＋Ｃ％）は、１０ｋｂＴｒｉ－ＳＯＭの横軸に反映され、１０ｋｂＴｒｉ－ＳＯＭの左から右へ増加する。高ＧＣ含量の配列は、１０ｋｂＴｒｉ－ＳＯＭの右側に位置する。１０ｋｂＴｅｔｒａ－ＳＯＭおよび１０ｋｂＰｅｎｔａ－ＳＯＭについても、類似の結果が得られた。同一のＧＣ含量を持つ配列が、オリゴヌクレオチド出現頻度の複合的な組み合わせによって分離され、結果として種特異分離が起こった。１０ｋｂＳＯＭ内では、種内分離が明らかである。例えば、ヒトは、１０ｋｂＴｒｉ－ＳＯＭ内および１０ｋｂＴｅｔｒａ－ＳＯＭ内において、２つの主要な領域に分離された。しかしながら、１０ｋｂＰｅｎｔａ－ＳＯＭ内においては、ヒト配列が単一の連続した領域に分類された。このことは、ヒト１０ｋｂ配列間における幅広い変化にもかかわらず、ＳＯＭがヒト配列内におけるペンタヌクレオチド出現頻度の共通した特徴を認識していることを示している。

【００９９】

次に、約３００，０００個の１００ｋｂの断片塩基配列におけるトリヌクレオチド、テトラヌクレオチド、およびペンタヌクレオチドの出現頻度に関するＳＯＭ（それぞれ「１００ｋｂＴｒｉ－ＳＯＭ」、「１００ｋｂＴｅｔｒａ－ＳＯＭ」、「１００ｋｂＰｅｎｔａ－ＳＯＭ」と略記し、これら３つを「１００ｋｂＳＯＭ」と総称する）を作成した。

【０１００】

そして、作成されたＳＯＭをカラー出力（カラー印刷やカラー表示等）により、単一の種由来の配列を含む格子点を種ごとに異なる有彩色で示し、複数の種由来の配列を含む格子点を黒色で示した。このカラー出力結果を白黒画像に変換したものを図２に示す。図２（ａ）、図２（ｂ）、図２（ｃ）、および図２（ｄ）はそれぞれ、１００ｋｂＴｒｉ－ＳＯＭ、１００ｋｂＴｅｔｒａ－ＳＯＭ、１００ｋｂＰｅｎｔａ－ＳＯＭ、および１００ｋｂＤｅｇｅＰｅｎｔａ－ＳＯＭを示す。また、図２において、Ｃは線虫の領域、Ａはシロイヌナズナの領域、Ｒはコメの領域、Ｄはキイロショウジョウバエの領域、Ｆはフグの領域、Ｚはゼブラフィッシュの領域、Ｈはヒトの領域を示す。

【０１０１】

１００ｋｂＳＯＭ内においては、１０ｋｂの断片塩基配列に関するＳＯＭ内よりも（種内分離でなく）種間分離が顕著であった。１００ｋｂＴｅｔｒａ－ＳＯＭ内および１００ｋｂＰｅｎｔａ－ＳＯＭ内においては、全ての種が１つの主要な領域を有していた（図２

（ｂ）および図２（ｃ））。さらに、上記種領域は、ゲノム配列を含まない白い連続した格子で囲まれていた。種特異格子のベクトルは、たとえ領域の境界に近くとも領域間で異なり、白い連続した格子に基づき種境界を主として自動的に描くことができる。１００ｋｂＳＯＭを詳細に調べると、特定の特徴を持つ少数の配列からなる小さな領域がいくつか存在した。例えば、コメ領域（図２のＲで示す領域）とフグ領域（図２のＦで示す領域）との間に位置するシロイヌナズナの小さな領域（図２のＡで示す領域）は、主として、動原体性領域および亜動原体性領域由来の配列からなっている。種内分離の分析は、個々のゲノムの詳細な構造に関する深い情報を与えることができる。

【０１０２】

〔実施例２〕

ＳＯＭは、各生物種のゲノムの代表的な特徴であるオリゴヌクレオチド出現頻度の種特異的な組み合わせを認識し、特徴的な出現頻度パターンを特定することができた。１００ｋｂＳＯＭ内の各格子点における各オリゴヌクレオチドの出現頻度（観測値）を計算し、各格子点におけるモノヌクレオチド組成から期待される各オリゴヌクレオチドの出現頻度の期待値で正規化した。そうして正規化した各格子点のオリゴヌクレオチドの出現頻度（観測値／期待値の比）を、ＳＯＭと同様の２次元の格子状のマップに表したもの（出現頻度マップ）を各オリゴヌクレオチドごとに作成した。２次元マップ上の各格子点における正規化したオリゴヌクレオチドの出現頻度（観測値／期待値の比Ｒ）の情報は、例えばカラー出力する場合、２次元マップ上の格子点の色で表現できる。

【０１０３】

一例として、オリゴヌクレオチドの出現頻度が期待値に対して過剰である（オリゴヌクレオチドが過剰に出現する）格子点、すなわち観測値／期待値の比が１より十分に大きい格子点を赤で示す。また、オリゴヌクレオチドの出現頻度が期待値に対して過少である（オリゴヌクレオチドが過少に出現する）格子点、すなわち観測値／期待値の比が１より十分に小さい格子点を青で示す。また、オリゴヌクレオチドの出現頻度が期待値と同程度である（オリゴヌクレオチドが期待値レベルで出現する）格子点、すなわち観測値／期待値の比が１付近である格子点を白で示す。そして、格子点の色の濃度は、観測値／期待値の比が１から離れるほど濃くなるようにする。

【０１０４】

このようにしてカラー出力した出現頻度マップを白黒画像に変換したものの代表例を図３および図４に示す。図３（ｂ）はＣＡＧＴの出現頻度マップ、図３（ｃ）はＡＡＴＴの出現頻度マップである。また、対照として、図３（ａ）に１００ｋｂＴｅｔｒａ－ＳＯＭを示す。また、図３には、観測値／期待値の比Ｒの値と黒濃度との関係を示すスケールを併せて示している。図４（ｂ）はＡＣＡＧＧＧとＣＣＴＧＴの合計の出現頻度を示す出現頻度マップ、図４（ｃ）はＣＧＡＣＧとＣＧＴＣＧの合計の出現頻度を示す出現頻度マップ、図４（ｄ）はＣＧＡＡＡとＴＴＴＣＧの合計の出現頻度を示す出現頻度マップである。また、対照として、図４（ａ）に１００ｋｂＤｅｇｅＰｅｎｔａ－ＳＯＭを示す。また、図４には、観測値／期待値の比Ｒの値と黒濃度との関係を示すスケールを併せて示している。

【０１０５】

なお、出現頻度マップにおいて、各格子点における正規化したオリゴヌクレオチドの出現頻度（観測値／期待値の比）の情報は、他の様式、例えば２次元格子状マップを３次元化した３次元の棒グラフにおける高さ等で表現してもよい。

【０１０６】

上記のオリゴヌクレオチド出現頻度の正規化は、各格子点におけるオリゴヌクレオチド出現頻度をモノヌクレオチド組成の差から切り離して調べることを可能にした。例えば、塩基配列間での、ＣＧおよびＧＣを含むオリゴヌクレオチドの出現頻度の差を、塩基配列間でのＧＣ含量の差から切り離して鋭敏に検出することができる。種々のテトラヌクレオチドおよびペンタヌクレオチドに関する出現頻度マップにおいて、過剰出現領域と過少出現領域との境界は、ほとんど種の境界と正確に一致した。図３に示したものは、種分離に

関する特徴的な例である。ＡＡＴＴは、コメ、キイロショウジョウバエ、および線虫では過剰に出現し、フグおよびゼブラフィッシュでは過少に出現し、ヒトおよびシロイヌナズナでは適度に出現した。ＣＡＧＴは、３種の脊椎動物（ヒト、フグ、およびゼブラフィッシュ）全てにおいて過剰に出現したが、２種の植物（コメおよびシロイヌナズナ）の両方において過少に出現した。

【０１０７】

また、相補的なテトラヌクレオチドの対の出現頻度を加算した値に基づいて出現頻度マップを作成した。作成された出現頻度マップは、相補的な対の一方のテトラヌクレオチドの出現頻度に基づく出現頻度マップとほとんど同一であった。それゆえ、相補的な対の一方のテトラヌクレオチドに基づく出現頻度マップのみを図示している。

【０１０８】

ペンタヌクレオチドの出現頻度マップについては、ＤｅｇｅＰｅｎｔａ－ＳＯＭに関する出現頻度マップの例を図示している（図４）。３種の脊椎動物（ヒト、フグ、およびゼブラフィッシュ）全てにおいて、（ＡＣＡＧＧ＋ＣＣＴＧＴ）および（ＣＧＡＣＧ＋ＣＧＴＣＧ）はそれぞれ過剰出現および過少出現した。（ＣＧＡＡＡ＋ＴＴＴＣＧ）は、キイロショウジョウバエおよび線虫では過剰に出現し、分裂酵母では適度に出現した。ＳＯＭは、配列分離に関する多くのオリゴヌクレオチドの複合的な組み合わせを利用することで、種による分類を実現できる。

【０１０９】

〔実施例３〕

（ヒトゲノム配列における種内の差）

細菌ゲノムに関するＴｅｔｒａ－ＳＯＭ内の特徴的なテトラヌクレオチドの生物学的な意味を明らかにするために、本願発明者等は、４塩基制限酵素を産生する細菌内における制限酵素系を用いてパリンドローム（回文構造）のテトラヌクレオチドの出現頻度の相関を調べた。制限酵素認識部位（切断部位）のテトラヌクレオチドは、４塩基制限酵素を産生する細菌のゲノムにおいては特徴的に過少に出現する。ＳＯＭは、この細菌ゲノムの生物学的特性を正しく認識した。ＳＯＭは、オリゴヌクレオチド出現頻度以外のいかなる情報を用いることなくゲノム配列を既知の生物学的分類に（図２の場合には種に）分類することができた。ＳＯＭは、分類能力が非常に高いので、多種多様なゲノム情報を抽出する強力な情報科学ツールになるはずである。１つのゲノム内における種内の差に関するＳＯＭの分類能力を調べるために、本願発明者等は、ヒトゲノム由来の２．８Ｇｂ高品質のドラフト配列を分析した。本願発明者等は、２．８Ｇｂヒト配列から得た、互いに重複していない１０ｋｂの配列と、１０ｋｂずつずれた１００ｋｂの配列に関するＴｅｔｒａ－ＳＯＭおよびＰｅｎｔａ－ＳＯＭを、各格子点に分類された配列の数を棒の高さで表した３次元画像として出力した。

【０１１０】

本願発明者等は、１０ｋｂＳＯＭの各格子点における個々のオリゴヌクレオチド（テトラヌクレオチド等）の出現頻度に対し、各格子点のモノヌクレオチド組成による正規化を行い、正規化されたオリゴヌクレオチドの出現頻度を計算した。正規化されたオリゴヌクレオチドの出現頻度は、モノヌクレオチド組成から期待されるオリゴヌクレオチドの出現頻度の期待値に対する、ＳＯＭ上におけるオリゴヌクレオチドの出現頻度の値（観測値）の比である。正規化されたオリゴヌクレオチドの出現頻度に基づいて、前述したのと同様にして、出現頻度マップを作成した。各テトラヌクレオチドの出現頻度マップは、全域にわたって過少に出現する出現頻度マップ（ＡＡＣＧの出現頻度マップ）、および全域にわたって過剰に出現する出現頻度マップ（ＴＴＣＣの出現頻度マップ）を含んでいた。次に、本願発明者等は、１０ｋｂＳＯＭ内の制限された部位（出現頻度マップにおける広域の過少出現領域に囲まれた小さな過剰出現領域）で顕著に出現するが、１００ｋｂＳＯＭ内の全域にわたって過少に出現するテトラヌクレオチドに着眼した。これら例の１タイプは、１つのＣＧと２つのＣまたはＧとを含む複数のテトラヌクレオチドに対応していた。このタイプについては、類似した局所的な過剰出現パターンが観測された（タイプＡ）。こ

れらのテトラヌクレオチドは、よく特徴づけられた転写シグナルであるＧＣボックスの構成成分に対応していた。ＴＡＴＡボックスの構成成分であるＴＴＡＡ、ＡＴＡＡ、およびＡＴＴＡは、タイプＡのテトラヌクレオチドのパターンに類似したパターンを持っていたが、その配列は全く異なっていた（タイプＢ）。他の特徴的なパターンも観測され、そのいくつかは類似していた。

【０１１１】

テトラヌクレオチドの局所的な過剰出現パターンによる生物学的な意味を調べるために、上記格子点における正規化した各テトラヌクレオチドの出現頻度（この格子点に分類された１０ｋｂ配列における、正規化した各テトラヌクレオチドの出現頻度を表す）を染色体２１ｑ配列に沿ってプロットした。結果を図５に示す。なお、図５において、各テトラヌクレオチドの出現頻度は、図３（ｂ）（ｃ）等と同様の表示色で表示した。また、図５には、遺伝子の存在位置も併せて示した。

【０１１２】

タイプＡおよびタイプＢのテトラヌクレオチドおよび他のいくつかのテトラヌクレオチドの分布パターンを染色体２１ｑ配列に沿ってプロットした。遺伝子リッチな領域において、タイプＡおよびタイプＢが類似のパターンを持ち、顕著な出現（赤および白）が観測されたという観測結果は、これらヌクレオチドが、転写調節シグナルの典型例であるＴＡＴＡボックスおよびＧＣボックスの核となる配列であるという見解と整合している。上記染色体のさまざまな部分においては、ＧＡＴＣおよびＡＧＴＡの分布パターンも、タイプＡおよびタイプＢの分布パターンと類似していた。これらテトラヌクレオチドの全てが、染色体２０および２２においても、遺伝子リッチ領域において高いレベルで出現した（図示しない）。この結果は、これらテトラヌクレオチドが、遺伝子の発現調節または機能に関連するシグナル配列またはその構成成分である可能性が高いことを示唆している。１つのゲノム内におけるジヌクレオチド、トリヌクレオチド、およびテトラヌクレオチドの出現頻度が、高い相関関係を持つことが見出されている。この基本的なゲノムの特徴のために、本願発明者等は、特徴的なテトラヌクレオチドを容易に特定することができた。そのテトラヌクレオチドは、Ｔｅｔｒａ－ＳＯＭにおいては多くの領域で過少に出現したが、Ｔｒｉ－ＳＯＭにおいてはその構成トリヌクレオチドはむしろ過剰に出現した。上記テトラヌクレオチドの過少出現は、その構成トリヌクレオチドの過少出現に起因するものではなかったので、上記過少出現は、細菌のゲノム内の制限酵素認識部位の配列について観測されたように、テトラヌクレオチドの生物学的な意味を反映していると考えられる。ＡＴＴＧは、このタイプに属しており、分布は遺伝子プアな領域内に集中していた（図５）。

【０１１３】

ゲノム中の特徴的なオリゴヌクレオチド、過少出現するものだけでなく過剰出現するものを考慮すると、ＤＮＡの立体構造、コンテキスト依存の突然変異、およびＤＮＡの修飾を含むさまざまな因子が、原因になっていると考えられる。過剰出現する配列に関し、多量に存在するＤＮＡ結合蛋白質によって認識される配列の優先傾向を考慮しなければならない。ＳＯＭを用いた種間分離および種内分離は非常に明瞭であるので、ＳＯＭは、進化の過程で個体のゲノムの配列の特徴を決定してきた詳細な分子機構を理解するための基礎的なガイドラインを提供するはずである。

【０１１４】

（遺伝子のシグナル配列の特徴決定）

多種多様なオリゴヌクレオチド配列が、遺伝子のシグナル配列（例えば遺伝子発現の調節シグナル配列）として機能する。種々のテトラヌクレオチド（例えばタイプＡおよびタイプＢ）が転写シグナル配列に適合した特徴を持つという図２等における知見は、ＳＯＭが遺伝子のシグナル配列の特徴決定およびコンピュータ予測（in-silico予測）を行うための新規なツールとなる可能性を示している。転写シグナル配列などのような遺伝子のシグナル配列は、典型的にはテトラヌクレオチドより長い。それゆえ、この可能性を試すにはより長いオリゴヌクレオチドの分析が必要になる。そこで、テトラヌクレオチドの場合と同様にしてヒトゲノム中において過少出現するペンタヌクレオチドを調べた。その結果

、遺伝子リッチな領域と遺伝子プアな領域との間で分布パターンが異なる例が存在した。上記の局所的で特異的な出現パターンは、多くの場合、テトラヌクレオチドのパターンよりも明瞭であった。このことは、上記の多くのテトラヌクレオチドが、より長い配列長を持つシグナル配列（例えばＧＣボックス）の構成成分であることを示唆している。

【０１１５】

シグナル配列認識機構と、ゲノム全域でのそれぞれのオリゴヌクレオチド配列の出現レベルとは、互いに関連するものと考えられる。特定の標的蛋白質に結合する高い親和力などのような顕著な活性をオリゴヌクレオチド配列が持っている場合には、そのオリゴヌクレオチド配列の出現は、ランダムな出現から偏り、ゲノム全域で顕著に変化するであろう。例えば、転写因子に対して強い結合活性を持つオリゴヌクレオチド配列は、ゲノムの多くの領域にわたるランダムな出現と比較して過少に出現するが、遺伝子調節領域内ではより高頻度であろう。このようなシグナル配列は、広い窓を持つＳＯＭ（例えば１００ｋｂＳＯＭ）の全域にわたって過少に発現するが、より狭い窓を持つＳＯＭ（例えば１０ｋｂＳＯＭ）の制限された部位ではより高い頻度で出現する。既知の転写因子に対して結合活性を持つオリゴヌクレオチド配列がゲノム全域にわたってランダムな出現頻度で、あるいはそれより高い出現頻度で出現するような逆のケースでは、隣接する他のシグナル成分との組み合わせが、上記配列が調節シグナル配列として機能するための絶対的必須条件となるはずである。ゲノム全域にわたる因子結合活性を持つオリゴヌクレオチド配列の出現頻度は、転写調節のための組み合わせユニット内における各オリゴヌクレオチドの相互の役割を理解するための基礎的な情報（特異性決定における差異識別への貢献）を与えるだろう。因子結合活性を持つオリゴヌクレオチドのレベルに関するＳＯＭデータは、異なる分類のシグナル配列の異なる振る舞い（異なる窓のサイズを持つＳＯＭ上で可視化することができる）を分類することを可能にするであろう。よく研究された生物から収集された、分類されたシグナル配列の振る舞いを参照すれば、配列は決定されているが僅かな追加の実験データしかないゲノム内のシグナル配列の予測に有用なコンピュータによる方法（in-silico方法）を開発することができる。

【０１１６】

このアプローチの準備として、転写因子結合の可能性を持つことが知られているペンタヌクレオチドの特徴決定を次のようにして行った。まず、ＴＲＡＮＳＦＡＣデータベース(http://transfac.gbf.de/TRANSFAC)中でヒト転写因子に対する結合配列して知られているペンタヌクレオチドを検索した。上記データベース中では、因子結合配列として文献で報告されているペンタヌクレオチドは合計２２個あった。しかしながら、上記データベース中のＭＡＴＲＩＸテーブルを参照してこれらの配列を詳細にチェックしたところ、これらの多くはより長いシグナル配列の構成部分であり、転写因子結合のための主決定配列として選択できたものは、４つのペンタヌクレオチド、ＮＦ－Ｙ結合部位ＣＣＡＡＴ、ＧＡＴＡ－１因子結合部位ＧＡＴＡＡ、ＫＬＦ結合部位ＣＡＣＣＣ、およびＮＦ－１結合部位ＴＧＧＣＡであった。

【０１１７】

そして、ヒト２．８Ｇｂ配列の１０ｋｂＰｅｎｔａ－ＳＯＭおよび１００ｋｂＰｅｎｔａ－ＳＯＭのそれぞれにおける、これら４つのペンタヌクレオチドの出現頻度分布パターン（出現頻度マップ）を作成し、ＧＣボックスの核となる配列の分布パターンと比較した。ＧＡＴＡＡは、１０ｋｂＳＯＭおよび１００ｋｂＳＯＭの双方の多くの領域で過少に出現した。ＣＣＡＡＴは、１００ｋｂＳＯＭの多くの領域で過少に出現したが、１０ｋｂＳＯＭの制限された部位ではかなり出現した。ＣＣＡＡＴの分布の、ＧＣボックスの核となる成分の分布との詳細な比較により、１０ｋｂＳＯＭにおけるＣＣＡＡＴの多い領域がＧＣボックス配列の多い領域と明確に区別されることが示された。これは、（ＧＣボックス配列ではなく）ＣＣＡＡＴが、染色体２１ｑの遺伝子リッチな領域におけるよりも遺伝子プアな領域においてより優勢であったという知見（図５）と整合している。ＣＡＣＣＣおよびＴＧＧＣＡは、１０ｋｂＳＯＭおよび１００ｋｂＳＯＭの多くの領域にまたがって過剰に出現した。そのような頻繁に出現する配列は、標的遺伝子の正常な調節のために他の

追加の特定の配列を必要とする配列、または多量に存在するＤＮＡ結合因子に対する結合配列に対応しうる配列である。ゲノム全域にわたっての出現レベルに関する既知のシグナル配列の系統的な分類は、ゲノム配列決定以来可能になっており、また、正確なシグナル配列認識の基礎をなす分子機構を解析するための新規な方法を提供する。

【０１１８】

ＳＯＭは、１つのマップ上の多種多様なゲノムにおいてオリゴヌクレオチド出現頻度を可視化できる。１３種の真核生物に対する、上述した４種のペンタヌクレオチドの各々の１０ｋｂＰｅｎｔａ－ＳＯＭおよび１００ｋｂＰｅｎｔａ－ＳＯＭ（図２（ｃ））における出現頻度パターン（出現頻度マップ）を作成した。ＧＡＴＡＡは、熱帯熱マラリア原虫およびキイロタマホコリカビを除いて上記真核生物の全てにおいて過少出現した。ＣＣＡＡＴは、３種の脊椎動物（フグ、ゼブラフィッシュ、ヒト）全ての多くの領域において過少に出現したが、２種の植物（コメおよびシロイヌナズナ）および２種の無脊椎動物（線虫、キイロショウジョウバエ）においては過剰に出現した。これらの結果は、個々の種におけるシグナル配列認識の機構と、シグナル配列認識系を確立する進化の過程とを理解するための基礎的な情報を提供しうる。ＳＯＭは、２次元マップ上におけるランダムな分布から特徴的に偏ったオリゴヌクレオチドを明示する。さらに、特異的な特徴を持ったゲノム配列がマップ上で自己組織化されたので、そのような配列全てのゲノム位置を染色体に沿ってプロットすることができた（図５）。十分な実験データを持つ様々な種の既知のシグナル配列をＳＯＭを用いて特徴決定し、系統的に分類すれば、配列は決定されているがそれ以外には僅かな実験データしかないゲノムに対して最も有用なシグナル配列のコンピュータによる予測方法を開発することが可能になるであろう。そのようなゲノムの数は急速に増大しているので、そのようなコンピュータによる予測方法の開発の重要性も増大している。

【０１１９】

〔実施例４〕

以下の実施例では、全長配列が知られている８１種の細菌ゲノムから得た１７０００個強の互いに重複していない１ｋｂ断片塩基配列（セグメント）および５ｋｂ断片塩基配列について、前記実施形態のシステムを用い、トリヌクレオチド出現頻度およびテトラヌクレオチド出現頻度を用いて特許文献１に記載の改良型ＳＯＭ作成法にしたがって２次元および３次元のＳＯＭを構築した。初期結合重みベクトルを得るための第１のステップとして、非特許文献４に記載されているように、１７０００個強の互いに重複していないセグメントの出現頻度を主成分分析（ＰＣＡ）によって分析した。設定学習回数は１００に設定した。１００回の学習サイクルの後、断片塩基配列のオリゴヌクレオチド出現頻度は、ＳＯＭ中の結合重みベクトルに実質的に投影された。作成したＳＯＭは、単一種由来の断片塩基配列を含む格子点を有彩色で示し、複数の種の断片塩基配列を含む格子点を黒で示すカラー画像として出力した（図示しない）。５ｋｂ断片塩基配列におけるテトラヌクレオチド出現頻度を用いて作成したＳＯＭでは、ほとんどの種の断片塩基配列は、種特異的な重複しない複数の領域に分離された。

【０１２０】

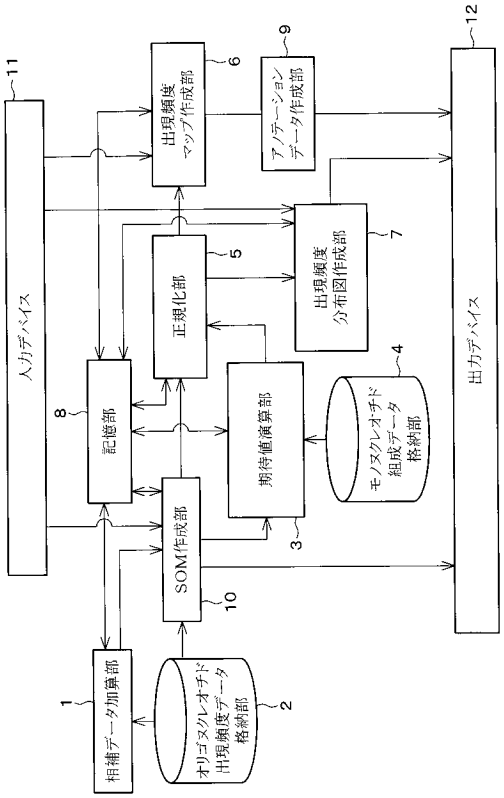
個々の格子点の結合重みベクトルの分析は、ジヌクレオチド出現頻度を用いて作成したＳＯＭ、トリヌクレオチド出現頻度を用いて作成したＳＯＭ、およびテトラヌクレオチド出現頻度を用いて作成したＳＯＭ（それぞれ「Ｄｉ－ＳＯＭ」、「Ｔｒｉ－ＳＯＭ」、および「Ｔｅｔｒａ－ＳＯＭ」と略記する）中における各結合重みベクトルに関するＧＣ含量（Ｇ＋Ｃ％）が主としてＳＯＭの横軸に投影され、ＳＯＭの左から右に行くにしたがって増加した。ＳＯＭにおいて、ＡＴリッチな細菌の配列は左側に、ＧＣリッチな細菌の配列は右側にそれぞれ分布した。重要なことには、同じＧＣ含量（Ｇ＋Ｃ％）を持つ配列は、オリゴヌクレオチド出現頻度の複合的な組み合わせによって分離され、その結果、種特異的な分離がなされた。言いかえれば、各ゲノム中の１０ｋｂ断片塩基配列の多くは、署名のようにそれぞれのゲノムを反映したオリゴヌクレオチドの組み合わせを持っている。ＳＯＭは、代表的な結合重みベクトルとして署名を明示することができる。

【０１２１】
１７０，０００個の互いに重複していない１ｋｂ断片塩基配列も同様にして分析し、ＳＯＭを作成した。この場合、分離の度合いは多少小さくなったが、種特異的分離が再び観察された。これは、種特異的特徴（署名）が、１ｋｂ断片塩基配列の主な構成グループ中でさえ検知できることを示している。
【０１２２】
個々の種への分類は、培養不可能な多様な微生物の分類の最初には重要でなく、系統学的な分類が重要になる。そこで、２次元および３次元のＳＯＭによる１２の主要な系統群への分類をテストしたところ、実際に、よい分類が検知できた。具体的には、オリゴヌクレオチド出現頻度に関するＳＯＭを約１００種の細菌の分類に適用したところ、５ｋｂの細菌配列の約９０％を、１２の系統群、すなわち、αプロテオバクテリア(Alphaproteobacteria)、βプロテオバクテリア(Betaproteobacteria)、γプロテオバクテリア(Gammaproteobacteria)、δプロテオバクテリア(Deltaproteobacteria)、アーケア（古細菌；Archaea)、クラミジア属(Chlamydia)、ファーミキューテス(Firmicutes)、アクチノバクテリア(Actinobacteria)、フゾバクテリウム属(Fusobacteria)、超好熱性グラム陰性嫌気性桿菌群(Thermotogae)等に粗分類することができた。上記で作成したＳＯＭを用いて、環境から得られた難培養性微生物の系統推定を行った。ＧｅｎＢａｎｋには、生物種が特定されていない環境微生物の塩基配列が登録されている。生物種が特定されていない塩基配列のうち、塩基長が１ｋｂ以上のリボソームＲＮＡ遺伝子（ｒＤＮＡ）配列以外の６６０件の非ｒＤＮＡ配列を用いて、系統群への分類を行った。ｒＤＮＡ配列と比べ、非ｒＤＮＡ配列では、従来の相同性解析による系統群への分類は難しい。非ｒＤＮＡ配列６６０件のうち３４３件は、反芻胃から採取された微生物由来の配列であった。これら３４３配列のＳＯＭ上での分類において、メタン生成菌と硫黄分解菌等の嫌気性微生物類の領域に大部分が分類されており、反芻胃という環境に生存する微生物として整合性の高い生物種であった。よって、環境中に複雑に混合している微生物群の系統推定ならびに多様性の推測が可能なことが示された。
【０１２３】
ＳＯＭは、多くの場合において、各々の種がほぼ等しい数の格子点からなる２つの主要な領域へ分離される結果となった。ＳＯＭによる同一種内での領域分離は、転写方向に関連していた。同一種内の分離は、ＳＯＭ中で使用される配列の長さに依存した。かなり長い範囲の近隣遺伝子が同一の転写方向を持っているゲノムについては、同一種内の分離が、１０ｋｂ配列内の出現頻度を分析したＳＯＭにおいてさえ顕著であった。しかしながら、転写方向が数ｋｂのような短い範囲のゲノムでさえ頻繁に変化するゲノムでは、同一種内の分離は１０ｋｂのＳＯＭにおいてはそれほど顕著ではなくなる。
【０１２４】
〔実施例５〕
１ｋｂのような短い断片のゲノム配列の場合には、転写方向は多くの場合知られておらず、転写方向の区別は、重要ではなく、生物種間の特定のための複雑さを引き起こすかもしれない。これを考慮に入れて、１対の相補的なテトラヌクレオチドの出現頻度を加算した。
【０１２５】
細菌８１種の１ｋｂ配列について、ジヌクレオチド（２連続塩基）、トリヌクレオチド（３連続塩基）、およびテトラヌクレオチド（４連続塩基）の各々の相補的な対をなすオリゴヌクレオチドの出現頻度を加算した出現頻度（縮退出現頻度）を用いて、実施例４と同様のＳＯＭ作成処理（頻度解析）を行った。すなわち、ゲノム配列の相補性の影響を除去するために、相補的な対をなす２つのオリゴヌクレオチド（例えば、ＡＡとＴＴ）を同一のものとみなし、これらの出現頻度を加算した出現頻度を用いてＳＯＭ作成処理を行った。例えば、６４個のトリヌクレオチドの出現頻度データの代わりに３２対の相補的なトリヌクレオチドの出現頻度データを使用する以外は、実施例４と同様にしてＳＯＭを作成した。その結果、実施例４と同様の結果が得られた。同一生物種内の分離も減少した。

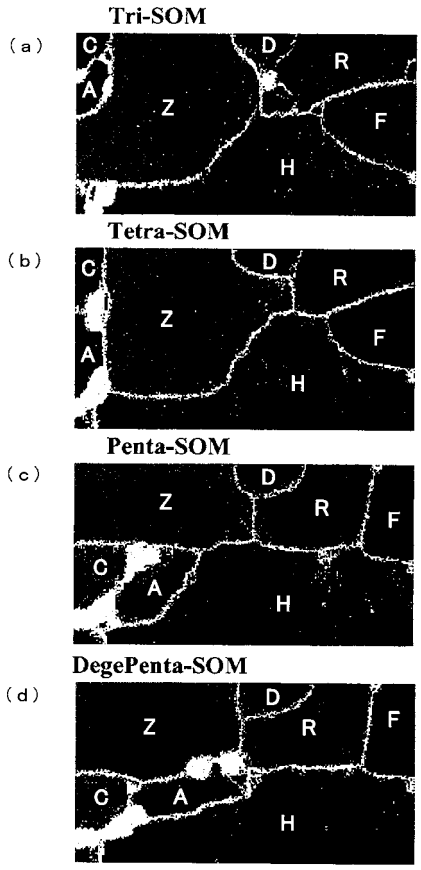
【０１２６】
〔実施例６〕
遠縁にあたる生物からの水平伝達を通じて導入されたゲノム・セグメントは、ドナーゲノムの配列の特徴を保持することが知られており、受容したゲノムのものと識別することができる。本願発明者等は、ＳＯＭが水平に伝達された遺伝子を識別するのに有用であり、重要なことには、伝達された遺伝子のドナーゲノムを予測するのに有用であることを以前に示した（非特許文献３）。ＳＯＭにおいて、個々の種の主要な領域から遠い所に位置する特有の格子点が存在する場合があった。主要な領域とは明白に異なるオリゴヌクレオチド出現頻度を持つ配列は、少なくとも一部は、他のゲノムから水平に伝達されたゲノム部分に対応するはずである。
【０１２７】
それらの種自体の領域とは異なるオリゴヌクレオチド組成を持つこれらの配列を視覚化するために、本願発明者等は、E. Coli（大腸菌）の領域およびそれと密接に関連する細菌S. typhimuriumの領域の両方の外に位置するE. Coli由来の１０ｋｂの配列を調べた。S. typhimurium領域の配列を除外すると、E. Coli配列の中で２番目に高い数の点がY pestis領域で見つかった。その後、ジヌクレオチド、トリヌクレオチド、およびテトラヌクレオチドで共通して見つかったY pestis領域内に存在する５つの配列に注目した。これらの配列内には、３７個の既知の遺伝子（それらのうちの２３個は、Y pestis遺伝子に対して顕著な相同性を持っていた）があった。例えば、アミノ酸レベルでは、これら遺伝子にコードされた２３個の蛋白質のうちの６個の蛋白質が６０％を越える同一性レベルを持っており、最も高い同一性レベルを持つものは８０％の同一性レベルを持っていた。この同一性は、垂直対であるE. Coli蛋白質およびY pestis蛋白質に関して計算された平均同一性の値４０％と比較して顕著に高かった。さらに、３つの遺伝子がファージにコードされた遺伝子と相同的であり、１つの遺伝子はトランスポゾン遺伝子と相同的であった。これらの発見は、これらの遺伝子が他の生物からE. Coliゲノムへ水平に伝達されたかもしれないという予測を支援する。
【０１２８】
〔実施例７〕
パリンドロームのテトラヌクレオチドの出現頻度の種特異的特徴が観察されることが見出された。例えば、それ自身のゲノムによってコードされた制限酵素の標的テトラヌクレオチドは、特に過少に出現した。さらに、パリンドロームのオリゴヌクレオチドは、転写調節蛋白質のような様々な蛋白質の標的部位であることが知られている。パリンドロームのオリゴヌクレオチドの出現頻度および分布は、これらのオリゴヌクレオチドがランダムに出現するものとして予測した結果と明白に異なる可能性がある。したがって、パリンドロームのオリゴヌクレオチドに注目することは興味深い。パリンドロームのヘキサヌクレオチドの場合にはオリゴヌクレオチドの種類が６４であり、パリンドロームのオクタヌクレオチドの場合には、オリゴヌクレオチドの種類が２５６であることに注意すべきである。言いかえれば、重要な生物学上の意味を持つ長い配列にさえ注目することができた。
【０１２９】
分析の結果、明らかに、種分離が明白になり、パリンドロームのオリゴヌクレオチドの出現頻度が、種特異的な特徴をより明白に表わした。制限酵素の認識配列の場合、およびさらにＤＮＡ結合蛋白質を備えた認識配列の場合、内部の塩基が認識に関係しない例がある。例えば、Ｔｔｈ１１１１の認識配列はＧＡＣＮ３ＧＴＣ（ここで「Ｎ」の位置ではどんな塩基も選ぶことができる）である。これを考慮に入れて、本願発明者等は、次のタイプの部分的にパリンドロームのオリゴヌクレオチドを分析対象に含めた。すなわち、パリンドロームに関係しない内部のｎ個の塩基（ｎは１～３の整数）を持つパリンドロームのオリゴヌクレオチド、例えば認識に使用されるパリンドロームのオリゴヌクレオチドとしてＧＧＧＮＣＣＣ、ＧＧＧＮＮＣＣＣ、ＧＧＧＮＮＮＣＣＣを分析対象に含めた。このパリンドロームのオリゴヌクレオチドの認識部位において、真のヘキサヌクレオチド（ｎ＝０）を分析対象に加えた場合、２５６（６４×４）個の変数を分析できた。

【０１３０】
この分析により、高い生物学的特異性を持っている可能性のあるオリゴヌクレオチドに注目できる。したがって、ＳＯＭは、同様の認識配列を持つ同じタイプの制限酵素あるいはＤＮＡ結合蛋白質を持つ種を、効率的に特定することが可能である。したがって、ＳＯＭは、系統学的な分類だけではなく近縁の種（γプロテオバクテリアに属する異なる種の細菌）を強力に分離できると考えられる。最初に通常のオリゴヌクレオチドＳＯＭを用いて塩基配列を系統学的なグループに分類し、それらをパリンドロームのオリゴヌクレオチドを使用して細分類すれば、多種多様な塩基配列を細分類することが可能になる。
【産業上の利用可能性】
【０１３１】
本発明の分類システムは、以上のように、塩基配列の生物学的分類への分類、複数種の生物を含む混合サンプルの成分分析、新規で産業上有用な細菌等の探索、ゲノム配列中における水平伝達を通じて他の種から導入されたセグメントの予測等に利用できる。
【０１３２】
また、本発明の解析システムは、以上のように、生物学的分類を分ける鍵となる重要なオリゴヌクレオチドの探索や、ゲノム配列中におけるシグナル配列を多く含む領域の探索等に利用できる。
【図面の簡単な説明】
【０１３３】
【図１】本発明の実施の一形態に係るオリゴヌクレオチド出現頻度解析システムの構成を示すブロック図である。
【図２】本発明の実施の一例において作成されたＳＯＭの例を示す図であり、（ａ）はトリヌクレオチド出現頻度を用いて作成されたＳＯＭ、（ｂ）はテトラヌクレオチド出現頻度を用いて作成されたＳＯＭ、（ｃ）はペンタヌクレオチド出現頻度を用いて作成されたＳＯＭ、（ｄ）相補的な対をなす２つのペンタヌクレオチドの出現頻度を加算したペンタヌクレオチド出現頻度を用いて作成されたＳＯＭを示す。
【図３】本発明の実施の一例において作成されたＳＯＭおよび出現頻度マップの例を示す図であり、（ａ）はテトラヌクレオチド出現頻度を用いて作成されたＳＯＭ、（ｂ）はＣＡＧＴの出現頻度マップ、（ｃ）はＡＡＴＴの出現頻度マップを示す。
【図４】本発明の実施の一例において作成されたＳＯＭおよび出現頻度マップの例を示す図であり、（ａ）は相補的な対をなす２つのペンタヌクレオチドの出現頻度を加算したペンタヌクレオチド出現頻度を用いて作成されたＳＯＭ、（ｂ）はＡＣＡＧＧおよびＣＣＴＧＴの合計の出現頻度マップ、（ｃ）はＣＧＡＣＧおよびＣＧＴＣＧの合計の出現頻度マップ、（ｄ）はＣＧＡＡＡおよびＴＴＴＣＧの合計の出現頻度マップを示す。
【図５】染色体２１ｑ上におけるいくつかのテトラヌクレオチドの正規化された出現頻度の分布を示すテトラヌクレオチド出現頻度分布図である。
【符号の説明】
【０１３４】
１　相補データ加算部（加算部）
２　オリゴヌクレオチド出現頻度データ格納部
３　期待値演算部
４　モノヌクレオチド組成データ格納部
５　正規化部
６　出現頻度マップ作成部
７　出現頻度分布図作成部
１０　ＳＯＭ作成部

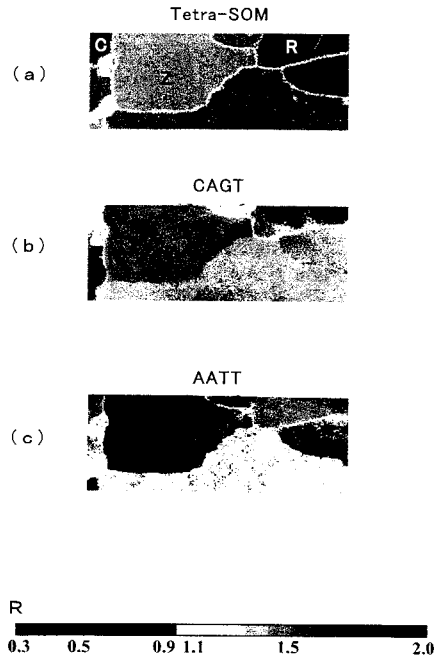
【図１】


入力デバイス
出現頻度マップ作成部
アノテーションデータ作成部
正規化部
出現頻度分布図作成部
記憶部
期待値演算部
モノヌクレオチド組成データ格納部
SOM作成部
相補データ加算部
オリゴヌクレオチド出現頻度データ格納部
出力デバイス
11
6
9
12
5
7
8
4
3
10
1
2

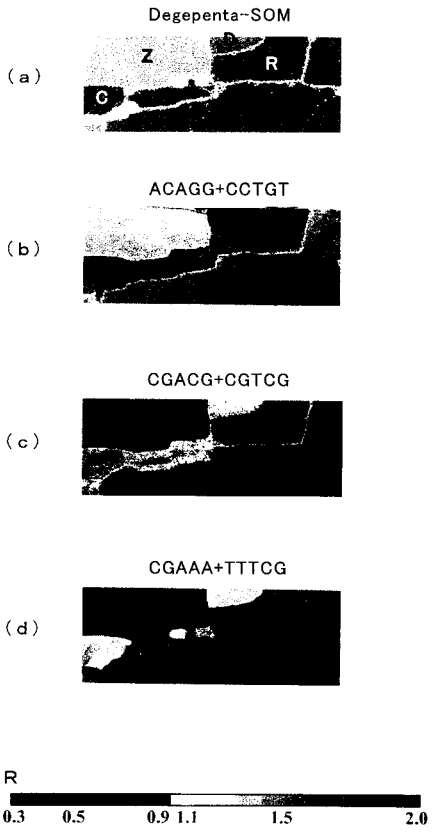
【図２】


Tri-SOM
(a)
Tetra-SOM
(b)
Penta-SOM
(c)
DegePenta-SOM
(d)
C
D
R
A
Z
F
H

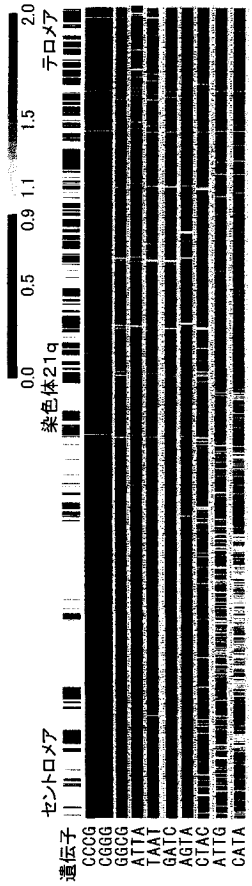
【図３】


Tetra-SOM
(a)
C
R
CAGT
(b)
AATT
(c)
R
0.3 0.5 0.9 1.1 1.5 2.0

【図４】


Degepenta-SOM
(a)
Z
D
R
C
ACAGG+CCTGT
(b)
CGACG+CGTCG
(c)
CGAAA+TTTCG
(d)
R
0.3 0.5 0.9 1.1 1.5 2.0

【図５】


0.0
0.5
0.9
1.1
1.5
2.0
染色体21q
テロメア
セントロメア
遺伝子
CCCG
CGGG
GGCG
ATTA
TAAT
GATC
AGTA
CTAC
ATTG
CATA

フロントページの続き

(72)発明者　金谷　重彦
奈良県生駒市高山町８９１６－５　大学宿舎Ｂ－３０６
(72)発明者　木ノ内　誠
山形県米沢市東１－１－２７　アズサハイツ２０１

Ｆターム(参考) 4B024 AA11 AA20 CA01 CA11 HA19
4B063 QA13 QA18 QQ42 QQ52 QS39