

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3928050号

(P3928050)

(45) 発行日 平成19年6月13日(2007.6.13)

(24) 登録日 平成19年3月16日(2007.3.16)

(51) Int. Cl.	F I
G06F 19/00 (2006.01)	G06F 19/00 600
C12N 15/09 (2006.01)	C12N 15/00 A
C12Q 1/68 (2006.01)	C12Q 1/68 Z
G06N 3/00 (2006.01)	G06N 3/00 560A

請求項の数 4 (全 30 頁)

(21) 出願番号	特願2003-328845 (P2003-328845)	(73) 特許権者	504202472
(22) 出願日	平成15年9月19日(2003.9.19)		大学共同利用機関法人情報・システム研究機構
(65) 公開番号	特開2005-92786 (P2005-92786A)		東京都港区南麻布四丁目6番7号
(43) 公開日	平成17年4月7日(2005.4.7)	(74) 代理人	110000338
審査請求日	平成15年9月19日(2003.9.19)		特許業務法人原謙三国際特許事務所
		(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) 【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、

複数の塩基配列中において複数種類のオリゴヌクレオチドの1つずつがそれぞれ出現する出現頻度のデータを格納しているオリゴヌクレオチド出現頻度データ格納部と、

上記オリゴヌクレオチド出現頻度データ格納部に格納された1つずつのオリゴヌクレオチドの出現頻度のデータを取り出し、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、算出された各対ごとのオリゴヌクレオチドの出現頻度のデータを出力する加算部と、

上記加算部から出力された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、各対ごとのオリゴヌクレオチドの出現頻度を上記入力ベクトル群として上記自己組織化を行うことにより上記自己組織化マップを作成し、作成された自己組織化マップのデータを出力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】

複数の塩基配列中において複数種類のオリゴヌクレオチドの1つずつがそれぞれ出現する出現頻度のデータを格納しているオリゴヌクレオチド出現頻度データ格納部と、

10

20

上記オリゴヌクレオチド出現頻度データ格納部に格納された1つずつのオリゴヌクレオチドの出現頻度のデータを取り出し、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、算出された各対ごとのオリゴヌクレオチドの出現頻度のデータを出力する加算部と、

上記加算部から出力された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、各対ごとのオリゴヌクレオチドの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し、作成された自己組織化マップのデータを出力する自己組織化マップ作成部と、

上記加算部から出力された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、各対ごとの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し、作成された出現頻度マップのデータを出力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】

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モノヌクレオチド組成データ格納部と、

上記モノヌクレオチド組成データ格納部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取り出し、取り出した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し、演算された期待値を出力する期待値演算部と、

上記加算部から出力された各対ごとのオリゴヌクレオチドの出現頻度のデータと上記期待値演算部で演算された期待値とを受け取り、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し、正規化された出現頻度のデータを出力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、上記正規化部で正規化された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】

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のデータを格納しているオリゴヌクレオチド出現頻度データ格納部と、

上記オリゴヌクレオチド出現頻度データ格納部に格納されたオリゴヌクレオチドの出現頻度のデータを取り出し、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、算出された各対ごとのオリゴヌクレオチドの出現頻度のデータを出力する加算部と、

上記加算部から出力された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、各対ごとのオリゴヌクレオチドの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し、作成された自己組織化マップのデータを出力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し、作成された出現頻度分布図のデータを出力する出現頻度分布図作成部とを備え、

上記出現頻度分布図作成部が、上記加算部から出力された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。

10

20

30

40

50

【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、様々な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、ISBN 4-303-73230-3）

【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、ISBN 4-431-70700-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 O157 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.

10

20

30

40

50

【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配列の対のうち的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A A A C対G T T T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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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と
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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

【 0 0 3 3 】

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エン製の“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。

【 0 0 3 4 】

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 10 kb や 100 kb）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、10 kb 以上が好ましく、100 kb 以上がより好ましい。

【 0 0 3 5 】

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。

【 0 0 3 6 】

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

10

20

30

40

50

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δρο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δρο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δρο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δροームのオリゴヌクレオチドの出現頻度、例えばGGGNCCC、GGGNCCC、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。

10

【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。

20

【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す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。

30

40

【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

50

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。

10

20

【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。

30

【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。

40

【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

50

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た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。

10

【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たことが好ましい。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。

20

30

【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。

40

【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

50

(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づくように行われる。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。

10

【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(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。

20

【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。

30

【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。

40

【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

50

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 $\{x_1, x_2, \dots, x_k, \dots, x_K\}$ ($k=1, 2, \dots, K$)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】

$$x_k = \{x_{k1}, x_{k2}, \dots, x_{kM}\} \quad (1)$$

(ここで、 $x_{k1}, x_{k2}, \dots, x_{kM}$ 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

10

20

30

40

50

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。

10

【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 $\{x_1, x_2, \dots, x_K\}$ 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 $\{W^0_1, W^0_2, \dots, W^0_P\}$ 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。

20

【0069】

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

式(2)において、iは $i = 1, 2, \dots, P$ である。また、式(2)中の $F\{x_1, x_2, \dots, x_K\}$ は入力ベクトル $\{x_1, x_2, \dots, x_K\}$ 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

30

(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 $\{x_1, x_2, \dots, x_K\}$ 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 $Z_{1k} = b_1 \cdot x_k$ および $Z_{2k} = b_2 \cdot x_k$ を求める($k = 1, 2, \dots, K$)。 $\{Z_{11}, Z_{12}, \dots, Z_{1k}, \dots, Z_{1K}\}$ および $\{Z_{21}, Z_{22}, \dots, Z_{2k}, \dots, Z_{2K}\}$ の標準偏差をそれぞれ σ_1 および σ_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 i, j ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$)で表現し、2次元の格子点(i, j)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、 $I \times \sigma_2 / \sigma_1$ 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。

40

【0072】

W^0_{ij} は、式(3)により定義することができる。

【0073】

【数 1】

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

(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 $Z_{1k} = b_1 x_k$ 、 $Z_{2k} = b_2 x_k$ 、および $Z_{3k} = b_3 x_k$ を求める。 $\{Z_{11}, Z_{12}, \dots, Z_{1k}, \dots, Z_{1K}\}$ 、 $\{Z_{21}, Z_{22}, \dots, Z_{2k}, \dots, Z_{2K}\}$ 、および $\{Z_{31}, Z_{32}, \dots, Z_{3k}, \dots, Z_{3K}\}$ 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 ($i = 1, 2, \dots, I$; $j = 1, 2, \dots, J$; $l = 1, 2, \dots, L$)で表現し、3次元の格子点(ijl)上に結合重みベクトル W_{ijl}^0 を置く。I、J、Lの値は、3以上の整数であれば良い。Jは、 $I \times \sigma_2 / \sigma_1$ よりも小さい整数のなかで最大のもの、Lは、 $I \times \sigma_3 / \sigma_1$ 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_{ijl}^0 は、下式(4)により定義することができる

【0074】

【数 2】

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

〔ステップ(c)〕

すべての入力ベクトル $\{x_1, x_2, \dots, x_K\}$ 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 $\{x_1, x_2, \dots, x_K\}$ 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 $W_{i1}^t, W_{i2}^t, \dots, W_{iP}^t$ 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)、 $t = 0, 1, 2, \dots, T$ 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、 W_{ti} で表すことができる。ここで、 $i = 1, 2, \dots, P$ 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 $W_{i1}^t, W_{i2}^t, \dots, W_{iP}^t$ 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_{ti} 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_{ti} は W_{tij}^t と表すことができる。

【0076】

入力ベクトル $\{x_1, x_2, \dots, x_K\}$ は、各入力ベクトル x_k 毎に並列処理して W_{ti} 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_{ti} 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_{ti} を更新する。

10

20

30

40

50

【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 $W^{t_{i'}}$ ($i' = 1, 2, \dots, P$) を更新する。

【0079】

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

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 $W^{t_{ij}}$ 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。

10

【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 $W^{t_{ij}}$ 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 $W^{t+1_{ij}}$ 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】

20

【数3】

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

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】

(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 ($0 < (t) < 1$) 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。

30

【0083】

【数4】

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

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 $10 \sim 1000$ 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、 $i - (t)$ $i + (t)$ かつ $j - (t)$ $j + (t)$ の条件を満たす格子点 $i' j'$ に分類された入力ベクトル $x_{i' j'}$ の集合である。 (t) は、近傍を決定する数であり、例えば式 (8) により求める。

40

【0085】

$$(t) = \max \{ 0, 25 - t \} \quad (8)$$

結合重みベクトル $\{ W^{t_1}, W^{t_2}, \dots, W^{t_p} \}$ 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

50

ブ(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_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、 W_i 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 $\{x_1, x_2, \dots, x_k\}$ 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。

10

【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。

20

【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、 T_1, T_2, T_3, T_4, T_5 および T_6 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、 $T_1, T_2, T_3/C, T_4/C, T_5/C, T_6$ となり、全体では、

30

$$T_1 + T_2 + T_3 + T_4 + T_5 + T_6 - \{T_1 + T_2 + (T_3 + T_4 + T_5) / C + T_6\} = (1 - 1/C)(T_3 + T_4 + T_5)$$

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。

40

【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 ($= 0, 1, \dots, I - 1$) および j ($= 0, 1, \dots, J - 1$) 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、 $\binom{2}{1} \times 250$ より大きく、かつ最も近い整数として定

50

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; 10 kbまたは100 kb)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。

10

【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。

20

【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。

30

【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。

40

50

【0096】

DNAデータベース内では、相補的な配列の対のうち的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A A A C対G T T T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、10 kb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100 kb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、「100 kb DegePenta-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。

10

【0098】

10 kb Tri-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、10 kb Tri-SOMの横軸に反映され、10 kb Tri-SOMの左から右へ増加する。高GC含量の配列は、10 kb Tri-SOMの右側に位置する。10 kb Tetra-SOMおよび10 kb Penta-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。10 kb SOM内では、種内分離が明らかである。例えば、ヒトは、10 kb Tri-SOM内および10 kb Tetra-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、10 kb Penta-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10 kb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。

20

【0099】

次に、約300,000個の100 kb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「100 kb Tri-SOM」、「100 kb Tetra-SOM」、「100 kb Penta-SOM」と略記し、これら3つを「100 kb SOM」と総称する）を作成した。

30

【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、100 kb Tri-SOM、100 kb Tetra-SOM、100 kb Penta-SOM、および100 kb DegePenta-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】

100 kb SOM内においては、10 kb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。100 kb Tetra-SOM内および100 kb Penta-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。100 kb SOM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

40

50

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。100 kb SOM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。

10

【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。

20

【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100 kb Tetra-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ACAGGとCCGT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100 kb Degen Pent-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。

30

【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。

40

【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

50

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。

10

【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δρο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。

20

【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。

30

40

【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

50

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100 kb SOM）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10 kb SOM）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。

10

【0116】

20

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と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CCAAAT、GATA-1因子結合部位GATAA、KLF結合部位CACCC、およびNF-1結合部位TGACAであった。

30

【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。CCAAAT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。CCAAAT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CCAAAT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）CCAAAT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TGACA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。

40

【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

50

10 kb Penta-SOMおよび100 kb Penta-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。CCAAAT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ため、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。

10

【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1 kb断片塩基配列(セグメント)および5 kb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)。5 kb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。

20

30

【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。言い換えれば、各ゲノム中の10 kb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。

40

【0121】

170,000個の互いに重複していない1 kb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、種特異的な分離が再び観察された。これは、種特異的な特徴(署名)が、1 kb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

50

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、プロテオバクテリア(Alphaproteobacteria)、プロテオバクテリア(Betaproteobacteria)、プロテオバクテリア(Gammaproteobacteria)、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が示された。

10

【0123】

20

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。

30

【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)を同一のもののみなし、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。

40

【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

50

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。

【 0 1 2 7 】

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 10 kb 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 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。

【 0 1 2 8 】

〔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。言い換えれば、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。

【 0 1 2 9 】

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、T t h 1 1 1 1 の認識配列は G A C N 3 G T C (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 G G G N C C C、G G G N N C C C、G G G N N N C C C 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。

【 0 1 3 0 】

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 (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

10

20

30

40

50

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CCGTG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】

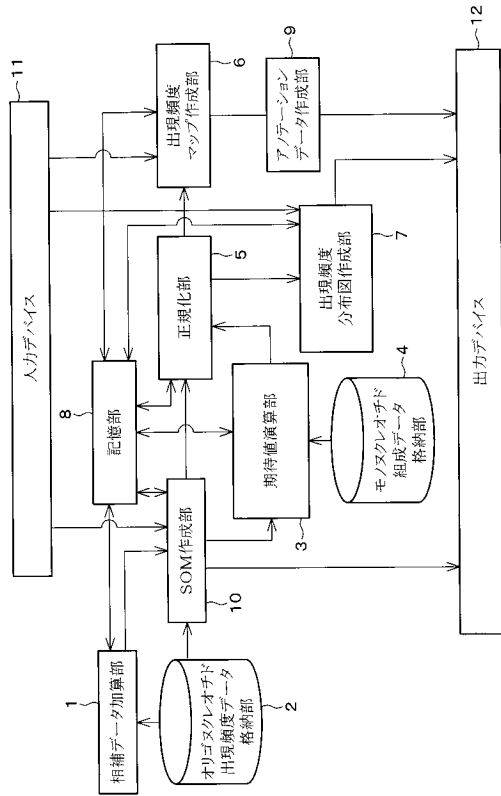
- 1 相補データ加算部(加算部)
- 2 オリゴヌクレオチド出現頻度データ格納部
- 3 期待値演算部
- 4 モノヌクレオチド組成データ格納部
- 5 正規化部
- 6 出現頻度マップ作成部
- 7 出現頻度分布図作成部
- 10 SOM作成部

10

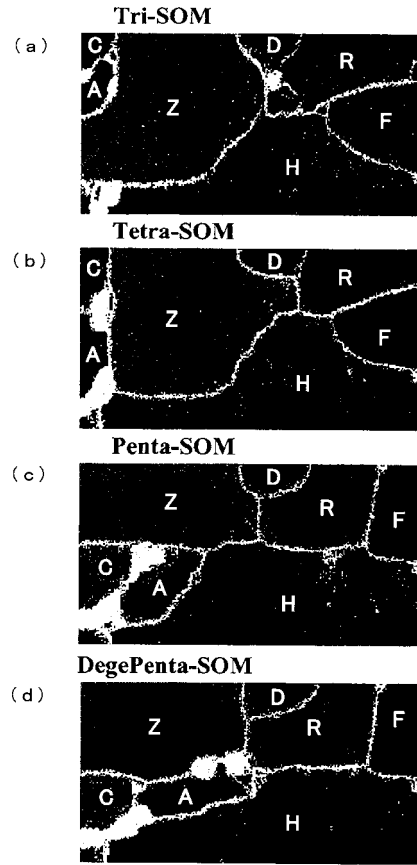
20

30

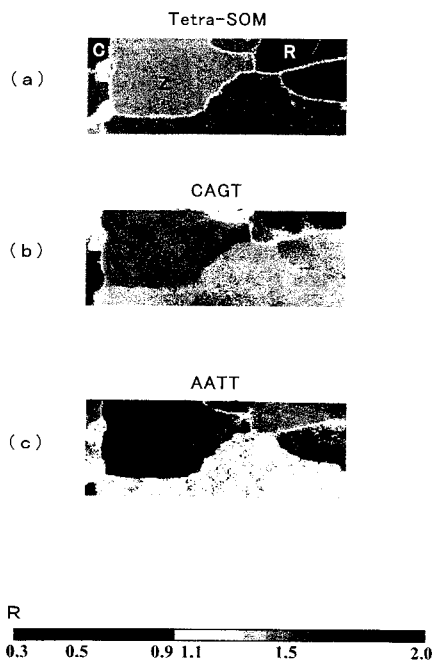
【 図 1 】



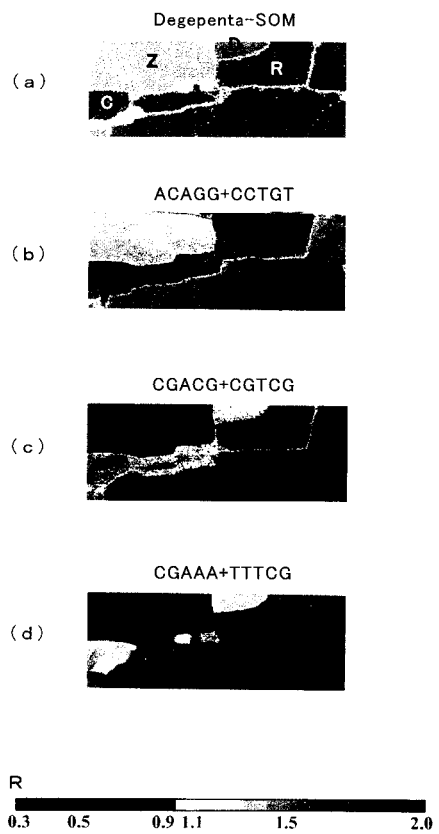
【 図 2 】



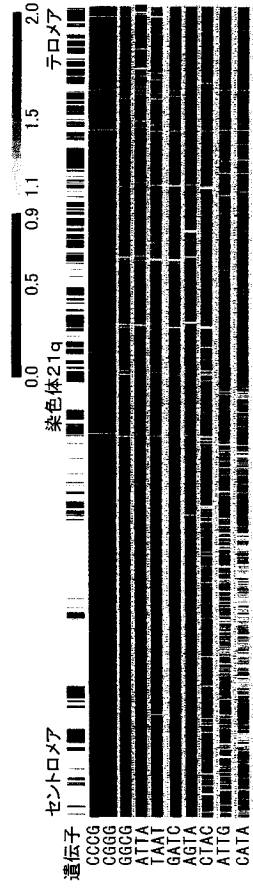
【 図 3 】



【 図 4 】



【 図 5 】



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

審査官 宮久保 博幸

- (56)参考文献 国際公開第02/050767(WO, A1)
特開平10-023893(JP, A)
特開平10-334104(JP, A)
Abe T, "A novel bioinformatic strategy for unveiling hidden genome signatures of eukaryotes: self-organizing map of oligonucleotide frequency, *Genome Inform.*, 2002年, Vol.13, p.12-20
Arnold J, Mono- through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis, *Nucleic Acids Res.*, 1998年, Vol.16, No.14, p.7145-58
Cuticchia AJ, The application of Markov chain analysis to oligonucleotide frequency prediction and physical mapping of *Drosophila melanogaster*, *Nucleic Acids Res.*, 1992年, Vol.20, No.14, p.3651-3657

- (58)調査した分野(Int.Cl., DB名)
G06F 19/00
G06N 3/00
PubMed
JSTPlus(JDream2)