(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2006 年3 月16 日 (16.03.2006)

PCT

(10) 国際公開番号
WO 2006/027913 A1

(51) 国際特許分類:
G06F 19/00 (2006.01) *C12N 15/09* (2006.01)
G06N 3/00 (2006.01) *C12Q 1/68* (2006.01)

(21) 国際出願番号: PCT/JP2005/014130

(22) 国際出願日: 2005 年8 月2 日 (02.08.2005)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2004-233848 2004 年8 月10 日 (10.08.2004) JP

(71) 出願人 (米国を除く全ての指定国について): 学校法人 早稲田大学 (WASEDA UNIVERSITY) [JP/JP]; 〒1698050 東京都新宿区戸塚町一丁目１０４番地 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 松山 泰男 (MATSUYAMA, Yasuo) [JP/JP]; 〒1698555 東京都新宿区大久保３丁目４番１号 学校法人 早稲田大学 理工学部内 Tokyo (JP). 川村 亮 (KAWAMURA, Ryo) [JP/JP]; 〒1420042 東京都品川区豊町４丁目１２番１号 マインハイム２０１ Tokyo (JP). 下田 啓太 (SHIMODA, Keita) [JP/JP]; 〒3330816 埼玉県川口市差間３丁目４番１５号 Saitama (JP).

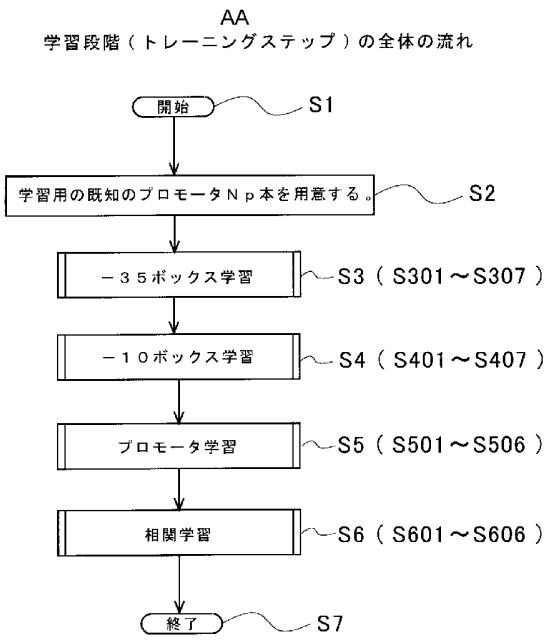
(74) 代理人: 中野 寛也 (NAKANO, Hironari); 〒1700013 東京都豊島区東池袋１丁目４５番４号 Ａ２ビル２階 エース特許事務所内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK,

[続葉有]

(54) **Title:** FEATURE PATTERN RECOGNIZING SYSTEM, METHOD, AND PROGRAM

(54) 発明の名称: 特徴パターン認識システムおよびその方法、並びにプログラム



AA FLOW OF WHOLE TRAINING STEPS
S1 START
S2 PREPARE KNOWN PROMOTER NP BOOK FOR TRAINING
S3 (S301 TO S307) -35 BOX TRAINING
S4 (S401 TO S407) -10 BOX TRAINING
S5 (S501 TO S506) PROMOTER TRAINING
S6 (S601 TO S606) CORRELATION TRAINING
S7 END

(57) **Abstract:** A feature pattern recognizing system, its method, and a program for recognizing a feature pattern present in a sequence with high accuracy. The discrete symbols (such as base symbols) constituting a sequence (such as a DNA sequence) to be judged are digitized by using the frequency determined for each sequence position and for each type of the discrete symbols to create examination data or an examination data matrix X_{test}. A matrix operation (such as $Y_{test}=W_{prom}X_{test}$) of multiplying a separation matrix (such as W_{prom}) determined by independent component analysis or main component analysis by the examination data or the examination data matrix X_{test} to determine separated data or a separated data matrix Y_{test}. By judging on which side of a threshold a feature judgment element (such as the element in the first line) of Y_{test} is present, it is judged by using the feature judgment element (elements in lines may be used) whether or not any feature pattern (such as a promoter) is present in the sequence.

(57) 要約: 配列中に存在する特徴パターンを高い精度で認識することができる特徴パターン認識システムおよびその方法、並びにプログラムを提供する。 判定対象の配列（ＤＮＡ配列等）を構成する各離散シンボル（塩基記号等）を、各配列位置毎でかつ離散シンボルの各種別毎に求められた出現頻度を用いて数値化して検査データまたは検査データ行列X_{test}を生成し、独立成分分析または主成分分析により得られた分離行列（W_{prom}等）と検査データまたは検査データ行列X_{test}とを乗じる行列演算（$Y_{test}=W_{prom}X_{test}$等）を行い、分離済データまたは分離済データ行列$Y_{test}$を求め、$Y_{test}$の特徴判定要素（第１行の要素等）が閾値に対し、いずれ

[続葉有]

WO 2006/027913 A1

DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書

２文字コード及び他の略語については、定期発行される各*PCT*ガゼットの巻頭に掲載されている「コードと略語のガイダンスノート」を参照。

の側の値であるかを判定する等、特徴判定要素（複数行の要素でもよい。）を用いて、配列中に特徴パターン（プロモータ等）が存在するか否かを判定する。

明 細 書

特徴パターン認識システムおよびその方法、並びにプログラム

技術分野

[0001] 本発明は、有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定する特徴パターン認識システムおよびその方法、並びにプログラムに係り、例えば、DNA配列中のプロモータ、アミノ酸配列中のモチーフ等を認識する場合に利用できる。

背景技術

[0002] 一般に、DNA配列の断片（以下、DNA配列断片と呼ぶ。）中には、プロモータ（promoter）と称される特徴的な塩基配列が遺伝子情報の周辺に存在している。このプロモータとは、DNA配列において特別なパターンを有している遺伝子制御部分であり、RNAポリメラーゼがこの部分を読み取ると、DNA配列の転写が始まる。従って、DNA配列断片にプロモータが含まれるか否かを判定することは非常に重要である。

[0003] 図41には、一例として、大腸菌塩基配列中でRNAポリメラーゼが大腸菌プロモータを探す様子が示されている。非常に長いDNA配列のうち、ある一部の塩基配列だけがRNA（リボ核酸：ribonucleic acid）に写し取られ、このRNAがアミノ酸に翻訳されてタンパク質が合成される。この際、RNAポリメラーゼという酵素は、DNAに衝突すると、弱く結合してDNAに沿って滑りながら動き、プロモータを見つけると強く結合し、DNA配列の転写を始める。

[0004] DNA配列断片にプロモータが含まれているかどうかを調べることは、もちろん試験管中の生物実験やX線解析を用いることにより可能である。しかしながら、これらの方法は、時間がかかる、高いコストがかかる、さらにX線使用の場合には安全面での配慮が必要になる等の不都合がある。従って、これらの方法に代えて、{A、T、G、C}の4つの塩基記号からなるDNA配列断片がコンピュータ内の離散値データとして与

えられたときに、その中にプロモータが存在するか否かを演算処理で判定することは、大量のDNA配列断片を高速かつ安価に取り扱うために、非常に重要な事項となっている。

[0005] このように、{A、T、G、C}の塩基記号列として与えられたDNA配列断片におけるプロモータの存否をコンピュータにより判定することは、すべて良いことずくめと思えるが、実際には簡単なことではない。これは、ひとえにプロモータのパターンにかなりの多様性があり、例えば、プロモータを構成する一部分の塩基が異なっていたり、あるいはプロモータ全体の長さが異なっていることがしばしばあるからである。このため、従来より、{A、T、G、C}による離散的な記号配列を連続量としての配列に変えてそのパターンを調べるという方法が採られ、これまでにも幾つかの提案がなされている。

[0006] 例えば、ニューラルネットワーク法を用いた方法(非特許文献1参照)と、さらに期待値最大化法を組み合わせた方法(非特許文献2参照)があり、これらはバイオインフォマティクスに関する著名文献(非特許文献3参照)にも収録されている。

[0007] なお、複数の種類の化学物質(遺伝子および遺伝子産物を含む。)の量の変化を表す情報を精度よく分類する化学物質分類装置として、主成分分析(PCA:Principal Component Analysis)若しくは独立成分分析(ICA:Independent Component Analysis)を行う装置があるが(特許文献1参照)、本発明のようなプロモータ等の特徴パターンの認識処理を行う装置ではない。

[0008] また、参考文献として、本発明の第2実施形態の処理で用いるホモロジ・スコアに関する文献を挙げておく(非特許文献4参照)。

[0009] 特許文献1:特開2003－141102号公報(請求項1、要約)

非特許文献1:I. マハデバン(I. Mahadevan)、I. ゴーシュ(I. Ghosh)、「ニューラルネットワーク法を用いた大腸菌プロモータ構造解析(Analysis of E.coli promoter structures using neural networks)」、ニュークレイック・アシッド・リサーチ(Nucleic Acids Research)、1994年、vol. 22、p. 2158－2165

非特許文献2:Q. マー(Q. Ma)、T. L. ワン(T. L. Wang)、D. シャシャ(D. Shasha)、C. H. ウー(C. H. Wu)、「期待値最大化法およびニューラルネットワーク法を用いたDNA配列分類法:事例研究(DNA sequence classification via an expectation m

aximization algorithm and neural networks:A case study)」、IEEEトランザクションズ・オン・システムズ(IEEE Transactions on Systems)、マン・アンド・サイバネティクス(Man and Cybernetics)、パートC:アプリケーションズ・アンド・レビューズ(Part-C:Applications and Reviews)、2001年、vol. 31、p. 468－475

非特許文献3:D. W. マウント(D. W. Mount)、「バイオインフォマティクス:配列およびゲノム解析(Bioinformatics:Sequence and Genome Analysis)」、コールド・スプリング・ハーバー・ラボラトリ・プレス(Cold Spring Harbor Laboratory Press)、2001年(岡崎康司・坊農秀雅監訳、「バイオインフォマティクス」、メディカルサイエンスインターナショナル、2002年)

非特許文献4:マーチン・E・ムリガン(Martin E. Mulligan)、ダイアン・K・ハーレー(Diane K. Hawley)、ロバート・エントリケン(Robert Entriken)、ウィリアム・R・マックルーア(William R. McClure)、「RNAポリメラーゼの選択で予測される大腸菌プロモータ配列(Eschrichia coli promoter sequences predict in vitro RNA polymerase selectivity)」、ニュークレイック・アシッド・リサーチ(Nucleic Acids Research)、1984年、vol. 12、p. 789－800

発明の開示

発明が解決しようとする課題

[0010] ところで、前述したように、試験管中の生物実験やX線解析によるプロモータ解析では、時間的・経済的要因等から大量のデータを処理することは困難であるため、これらの方法に代えて、ニューラルネットワーク法や期待値最大化法を用いたコンピュータによるプロモータ解析が行われている。しかし、これらの従来のコンピュータによる解析方法では、必ずしも高い認識率を得ることができないため、より優れた性能を発揮することができるコンピュータによる解析方法が望まれている。

[0011] また、コンピュータによるプロモータ解析の困難性は、DNA配列中のプロモータのパターンの多様性に起因するが、このことは、DNA配列中のプロモータ解析の場合に限らず、有限個の離散シンボルにより構成された配列中から特徴パターンを抽出する際に、その特徴パターンに多様性があれば、同様に生じる問題である。従って、DNA配列中のプロモータに限らず、広く一般に、有限個の離散シンボルにより構成

された配列中に存在する特徴パターンを高い精度で認識することができれば便利である。

[0012] 本発明の目的は、配列中に存在する特徴パターンを高い精度で認識することができる特徴パターン認識システムおよびその方法、並びにプログラムを提供するところにある。

課題を解決するための手段

[0013] 本発明は、有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定する処理を行う特徴パターン認識システムであって、複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置および離散シンボルの各種別とを対応させて形成された出現頻度テーブルを記憶する出現頻度テーブル記憶手段と、複数種類の既知の特徴パターンおよびこれらの既知の特徴パターンとは異なる複数種類の非特徴パターンを、出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う行列として得られた分離行列を記憶する分離行列記憶手段と、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う検査データ生成処理手段と、分離行列記憶手段に記憶された分離行列と検査データ生成処理手段により生成された検査データまたは複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列を求める処理を行う分離処理手段と、この分離処理手段により求めた分離済デ

ータまたは分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う判定処理手段とを備えたことを特徴とするものである。

[0014] ここで、「特徴パターン」としては、例えば、DNA配列中のプロモータやエンハンサ、タンパク質を構成するアミノ酸配列中のモチーフ(motif)、あるいはDNA配列とアミノ酸配列との混在型配列中のパターン等が挙げられるが、これらの生物学的な配列中の特徴パターンの他、例えば、言語学的、音響学的、地質学的、天文学的、物理学的等の様々な配列中の各種の特徴パターンであってもよく、要するに、有限個の離散シンボルにより構成された配列中に存在する特徴パターンであればよい。そして、この特徴パターンが「類似する」という意味は、生物学的な配列の場合には、例えば、ホモロジャスな(相同の、同類の)範囲内という意味であり、逆に、類似性があると判定されて新たに発見された特徴パターンについては、ホモロジャスな(相同の、同類の)範囲に加えることができる。

[0015] また、「離散シンボル」は、文字(アルファベット文字、ギリシャ文字、漢字、数字、ひらがな文字、かたかな文字等)、図形、記号、またはこれらの結合であり、例えば、DNA配列を構成する「離散シンボル」は、A(アデニン)、T(チミン)、G(グアニン)、C(シトシン)の4つの文字からなる標準的な塩基記号(勿論、標準的でない記号を処理対象としてもよい。)であり、パターンの長さを揃えるアラインメント(レングスアジャストメント)時に挿入されるギャップを示すマイナス記号等も「離散シンボル」に含まれる。また、アミノ酸配列を構成する「離散シンボル」は、A(アラニン)、R(アルギニン)、N(アスパラギン)等の20種類の文字からなる標準的なアミノ酸の略記号(勿論、標準的でない記号を処理対象としてもよい。)およびギャップを示すマイナス記号等である。

[0016] さらに、「配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列」とは、分離処理手段および判定処理手段による各処理

を行うための検査データを生成するにあたり、「配列若しくはその一部の配列断片」を直接に処理してもよく、あるいは「配列若しくは配列断片から切り出された検査配列」を処理してもよい趣旨である。

[0017] そして、「基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素」とは、基底行列において例えば第1列(実際には、推定された基底行列において検証することができる。)に特徴パターンの性質を示す特徴要素が現れる場合(つまり、第1列が特徴要素の行列内位置となる場合)には、分離済データ(縦ベクトル)または複数の分離済データが結合された分離済データ行列において、行列演算上、基底行列における第1列に配置された要素の影響を強く受ける第1行の各要素(縦ベクトルの場合には、第1番目の要素)をいう(後述する図39、図40参照)。また、基底行列に例えば第1例、第2列、第5列…という具合に、複数列の特徴要素が現れる場合には、特徴判定要素は、これに対応して第1行、第2行、第5行…という具合に、複数行の要素となる(後述する図44、図49参照)。なお、本願明細書において、基底行列(実際には、演算処理で推定された基底行列となる。)について第1列に特徴パターンの性質を示す特徴要素が「現れる」という場合には、システム設計者が、特徴パターンの性質を最も良く示す特徴要素が第1列に現れるような学習機構を強制的に付加した結果、現れる場合も含まれる。他の列に現れる場合も同様であり、その列に強制的に現れるようにする場合が含まれる。従って、これに伴って、基底行列内の特徴要素に対応する分離済データ(縦ベクトル)または分離済データ行列内の特徴判定要素についても、システム設計者が、第1行目(縦ベクトルの場合には、第1番目)等の特定位置に強制的に設定する場合が含まれる。

[0018] また、「判定処理手段」による判定処理には、特徴パターンが含まれるか否かを判定する硬判定(hard limited decision)と、特徴パターンの存否の程度(確からしさ)を判定する軟判定(soft decision)とがあり、例えば、特徴判定要素が1行だけの要素である場合には、前者の硬判定は、特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかを判定することにより行われ、後者の軟判定は、特徴判定要素の値の大きさの程度を判定することにより行われる。また、特徴判定要素が複数行の要素である場合には、例えば、分離処理手段により求めた分離済データま

たは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、前者の硬判定は、算出された類似度を示す値が、予め定められた閾値を境界としていずれの側の値であるかを判定することにより行われ、後者の軟判定は、算出された類似度を示す値の大きさの程度を判定することにより行われる。後者の軟判定としては、例えば、「特徴パターンである可能性が極めて高い」、「特徴パターンである可能性が高い」、「特徴パターンである可能性がある」、「特徴パターンでない可能性が極めて高い」、「特徴パターンでない可能性が高い」、「特徴パターンでない可能性がある」、「どちらか不明である」等の判定結果を出力することができ、これらの判定の程度を何段階に設定するかは任意であり、また、「特徴パターンである可能性が～％である」等の出力でもよい。後述する相関判定の場合も同様である。

[0019] このような本発明の特徴パターン認識システムにおいては、判定対象となる配列を構成する各離散シンボルを、各配列位置毎で、かつ、離散シンボルの各種別毎に求められた出現頻度を用いて数値化するとともに、独立成分分析（ICA：Independent Component Analysis）または主成分分析（PCA：Principal Component Analysis）を行って得られた分離行列を用いて配列中に特徴パターンが存在するか否かまたは存否の程度を判定するので、従来のニューラルネットワーク法や期待値最大化法を用いたコンピュータによるプロモータ解析の場合に比べ、プロモータ等の特徴パターンの認識精度を向上させることが可能となる。また、試験管中の生物実験やX線解析によるプロモータ解析の場合と比べても、時間的・経済的に優位な特徴パターンの解析を実現することが可能となり、これらにより前記目的が達成される。

[0020] また、前述した特徴パターン認識システムにおいて、特徴パターンは、異なる領域に位置する複数の部分パターンを含んで構成される多重化された特徴パターンであり、特徴パターン内での各部分パターンの領域位置および特徴パターンの全体の長さには、特徴パターンの種類の相違により複数種類の領域位置および複数種類の長さが存在し、出現頻度テーブル記憶手段は、複数種類の既知の特徴パターンのパターン全体の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの

除去により一定長に揃えた後、パターン全体の長さを揃えた複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎にギャップを含む離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置およびギャップを含む離散シンボルの各種別とを対応させて形成されたパターン全体用出現頻度テーブルを記憶するパターン全体用出現頻度テーブル記憶手段であり、分離行列記憶手段は、パターン全体の長さを揃えた複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた複数種類の非特徴パターンから生成されたパターン全体用学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られたパターン全体用分離行列を記憶するパターン全体用分離行列記憶手段であり、各領域の部分パターン毎に形成された部分パターン用出現頻度テーブルを記憶する各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段と、各領域の部分パターン毎に得られた部分パターン用分離行列を記憶する各領域の部分パターン毎の部分パターン用分離行列記憶手段と、各領域の部分パターン毎に学習段階で部分パターン用分離行列とともに得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を記憶する部分パターン用分離済データ行列記憶手段とを備え、各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段には、既知の特徴パターンに含まれる各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンを用いてこれらの部分パターン内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と部分パターン内での各配列位置および離散シンボルの各種別とを対応させて形成された各領域の部分パターン毎の部分パターン用出現頻度テーブルが記憶され、各領域の部分パターン毎の部分パターン用分離行列記憶手段には、既知の特徴パターンに含まれる各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンおよびこれらの既知の部分パターンとは異なる複数種類の非部分パターンを、部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して部分パターン用学習データ行列を生成し、この部分パターン用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、部分パターンの特徴を示す特徴要素を含

む部分パターン用基底行列の逆変換を行う行列として得られた部分パターン用分離行列が記憶され、部分パターン用分離済データ行列記憶手段には、独立成分分析または主成分分析を行うことにより部分パターン用分離行列を得る際に部分パターン用分離行列と部分パターン用学習データ行列とを乗じた結果として得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素が記憶され、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、各領域の部分パターンのそれぞれについて、前記部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について部分パターン用出現頻度テーブル記憶手段に記憶された部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成する処理を行う部分パターン候補データ生成処理手段と、部分パターン用分離行列記憶手段に記憶された部分パターン用分離行列と部分パターン候補データ生成処理手段により生成された複数の部分パターン候補データの各々または複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成する処理を行う部分パターン候補データ用分離済データ生成処理手段と、この部分パターン候補データ用分離済データ生成処理手段により生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた部分パターン候補データ用分離済データに対応する部分パターン候補配列を、特徴パターンの候補となる特徴パターン候補配列に含まれる部分パターン相当配列として決定する処理を行う部分パターン相当配列決定処理手段と、この部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列に基づき特徴パターン候補配列を決定する処理を行う特徴パターン候補配列決定処理手段と、部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列

をそれぞれ一塊の状態に保ったままで、特徴パターン候補配列決定処理手段により決定された特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行うアラインメント処理手段と、このアラインメント処理手段により長さを揃えた特徴パターン候補配列を、パターン全体用出現頻度テーブル記憶手段に記憶されたパターン全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う特徴パターン候補配列数値化処理手段とを含んで構成されていることが望ましい。

[0021] ここで、「部分パターン用分離済データ行列の真データ対応部分の各列ベクトル」とは、部分パターン用分離済データ行列を構成する真データ対応部分と偽データ対応部分とのうちの真データ対応部分であり(後述する図21参照)、真データ対応部分は、部分パターン用分離行列とともに部分パターン用分離済データ行列を得る際(後述する図10、図15参照)に用いられる部分パターン用学習データ行列のうちの部分パターン(つまり、真データ)を配置した部分に対応する部分であり、偽データ対応部分は、部分パターン用学習データ行列のうちの非部分パターン(つまり、偽データ)を配置した部分に対応する部分である。

[0022] また、「ギャップを含む離散シンボルの各種別毎の出現頻度」を求めるという意味は、ギャップの挿入によりアラインメント処理を行った場合に、挿入したギャップも離散シンボルに含めて考えて出現頻度を求めるという意味であり(後述する図19参照)、部分パターン以外の領域の離散シンボルの除去によりアラインメント処理を行った場合には、ギャップは挿入しないので、この場合には、ギャップの出現頻度を求める必要はない。

[0023] このように多重化された特徴パターンに含まれる複数の部分パターンのそれぞれについて部分パターン用出現頻度テーブルを作成し、独立成分分析(ICA)または主成分分析(PCA)を行って得られた部分パターン用分離行列を用いた処理を行う構成とした場合には、配列中から各部分パターンを見つけることにより、それらの部分パターンを含む特徴パターンを見つけることが可能となるので、特徴パターン内での各部分パターンの領域位置および特徴パターンの全体の長さに多様性があるような多

重化された特徴パターンであっても認識することが可能となる。また、配列中から各部分パターンを見つける際には、独立成分分析(ICA)または主成分分析(PCA)を行って得られた各部分パターン用分離行列を用いるので、高い精度で各部分パターンの位置を特定することが可能となり、多重化された特徴パターンを、高い精度で認識することが可能となる。

[0024] そして、上記のように、多重化された特徴パターンを認識することができる構成とした場合において、パターン全体用分離行列記憶手段に記憶されたパターン全体用分離行列を得る際に用いられるパターン全体用学習データ行列を構成するために、パターン全体の長さを揃えた複数種類の非特徴パターンを生成する際には、これらの生成にランダムデータを用いることができるという観点からは、部分パターン用分離行列および部分パターン用分離済データ行列を使って部分パターン相当配列を見つける処理を行わなくてもよいが、検査段階の処理と同様な処理を経たデータを学習データとするという観点からは、次のように、部分パターン用分離行列および部分パターン用分離済データ行列を使って部分パターン相当配列を見つける処理を行うことが好ましい。すなわち、検査段階では、判定処理の対象配列が、特徴パターンなのか非特徴パターンなのか不明な状態であるため、全ての判定処理の対象配列について、部分パターン用分離行列および部分パターン用分離済データ行列を使った部分パターン相当配列の決定処理(後述する図36参照)を含む一様な処理が行われて検査データが生成されるので、これに合わせて、学習段階でも、次のように、部分パターン用分離行列および部分パターン用分離済データ行列を使った部分パターン相当配列の決定処理(後述する図20～図23参照)を行うことが好ましい。

[0025] より具体的には、上記のように、多重化された特徴パターンを認識することができる構成とした場合において、パターン全体用分離行列記憶手段に記憶されたパターン全体用分離行列を得る際に用いられるパターン全体用学習データ行列を構成するためのパターン全体の長さを揃えた複数種類の非特徴パターンは、これらの非特徴パターンを生成するために用意された非特徴パターン生成用配列の中で、各領域の部分パターンのそれぞれについて、部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列

について部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成し、部分パターン用分離行列と複数の部分パターン候補データの各々または複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成し、生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた部分パターン候補データ用分離済データに対応する部分パターン候補配列を、非特徴パターンに含まれる部分パターン相当配列として決定し、さらに決定された各領域の部分パターン相当配列に基づき非特徴パターンを決定した後、決定された各領域の部分パターン相当配列をそれぞれ一塊の状態に保ったままで、決定された非特徴パターンの長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行って生成されたものであることが望ましい。

[0026] さらに、上記の場合において、複数種類の非特徴パターンは、複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行って生成されたものであることが望ましい。

[0027] また、上記のように、多重化された特徴パターンを認識することができる構成とした場合において、検査データ生成処理手段の部分パターン相当配列決定処理手段は、部分パターン候補データ用分離済データ生成処理手段により生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、部分パターン候補データ用分離済データと部分パターン用分離

済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行う構成とされていることが望ましい。

[0028] ここで、上記の「類似度」としては、例えば、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和等が挙げられる。

[0029] このように類似度、例えば内積の和を用いて複数の部分パターン候補配列から部分パターン相当配列を決定する処理を行う構成とした場合には、非特徴パターン生成用配列の中で定められた複数の部分パターン候補配列(後述する図20、図22参照)から、あるいは検査配列の中で定められた複数の部分パターン候補配列(後述する図36参照)から、部分パターン相当配列を決定する処理を行うにあたり、より確からしい部分パターン候補配列を部分パターン相当配列として決定することが可能となるので、高い精度で部分パターンを見つけ、ひいては高い精度で特徴パターンを見つけることが可能となる。

[0030] さらに、以上に述べた特徴パターン認識システムにおいて、判定処理手段は、基底行列への特徴要素の現れ方に応じて、1つ(1行)の特徴判定要素だけで判定処理を行う構成としてもよく、あるいは複数(複数行)の特徴判定要素を用いて判定処理を行う構成としてもよく、具体的には、例えば、次のような構成を採用することができる。

[0031] すなわち、以上に述べた特徴パターン認識システムにおいて、基底行列に1列の特徴要素が現れる場合には、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とすることができる。

[0032] また、以上に述べた特徴パターン認識システムにおいて、基底行列に複数列の特

徴要素が現れ、分離処理手段により求めた分離済データまたは分離済データ行列を構成する要素のうちの複数行の要素が特徴判定要素として採用される場合には、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値を用いて、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とすることができる。

[0033] ここで、「複数行の特徴判定要素の値を用いて」とは、例えば、複数の特徴判定要素の値の平均値を算出したり、和を算出する等、分離処理手段により求めた分離済データまたは分離済データ行列のうちの特徴判定要素の値（つまり、検査段階の処理で得られた特徴判定要素の値）だけを用いて判定処理を行ってもよく、あるいは分離処理手段により求めた分離済データまたは分離済データ行列のうちの特徴判定要素の値と、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値とを用いて判定処理を行ってもよい趣旨である。後者のように分離済データ行列のうちの真データ対応部分の特徴判定要素の値を用いる場合には、より具体的には、次のような構成を採用することができる。

[0034] 例えば、前述した特徴パターン認識システムにおいて、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とすることができる。

[0035] また、上記のような類似度を示す値を算出する構成とする場合において、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルのうち、特定の列ベクトルに近いか否か、つまり特定の特徴パターンに近いか否かの判定を行ってしまうことを避けるという観点からは、次のような構成とすることが望ましい。すなわち、判定処理手段は、類似度を示す値として、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの総体との類似度を示す値を算出する処理を行う構成とすることが望ましい。

[0036] そして、「各列ベクトルの総体との類似度を示す値を算出する処理」の具体的な例としては、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの重心ベクトルとの内積の値またはこれと等価な値を算出する処理等が挙げられる。

[0037] ここで、「重心ベクトルとの内積の値」と「等価な値を算出する処理」とは、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの内積の値をそれぞれ算出し、これらの内積の値の平均値を算出する処理である。

[0038] なお、類似度を示す値の算出処理は、上記のような重心計算に限らず、例えば、より一般的に、マハラノビス距離を求めるような処理等でもよい。

[0039] そして、前述したように、多重化された特徴パターンを認識することができる構成とした場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の特徴パターン候補配列の

長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の特徴パターン候補配列のそれぞれをパターン全体用出現頻度テーブル記憶手段に記憶されたパターン全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行う構成とされ、分離処理手段は、パターン全体用分離行列と検査データ生成処理手段により生成された複数の検査データの各々またはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行う構成とされ、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる各特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応する特徴パターン候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされていることが望ましい。

[0040] また、前述したように、多重化された特徴パターンを認識することができる構成とした場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の特徴パターン候補配列のそれぞれをパターン全体用出現頻度テーブル記憶手段に記憶されたパターン

全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行う構成とされ、分離処理手段は、パターン全体用分離行列と検査データ生成処理手段により生成された複数の検査データの各々またはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行う構成とされ、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる複数列の特徴要素の行列内位置に応じて定まる複数行の特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階でパターン全体用分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは特徴パターン寄りの程度を示す値であると判定された類似度を示す値について、閾値との差の絶対値が最も大きいかまたは特徴パターン寄りの程度が最も高い類似度を示す値を求め、求めた類似度を示す値を与えた検査データに対応する特徴パターン候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成としてもよい。

[0041] このように検査配列の中で特徴パターン候補配列を1離散シンボルずつずらしながら決定し(後述する図36参照)、これらの複数の特徴パターン候補配列の中から、特徴パターンであると認識される配列を1つ定める処理(但し、1つも特徴パターンであると認識されなくてもよく、ここでは、2以上の特徴パターン候補配列が特徴パターンであると認識されてしまうことがないという趣旨である。)を行う構成とした場合には、検査配列を適切な長さに設定することを前提とし、より高い精度で特徴パターンの認識を行うことが可能となる。

[0042] また、前述したように、多重化された特徴パターンを認識することができる構成とした場合において、多重化された特徴パターンに含まれる各領域の部分パターンが、配列中の特定点を含む特定点近傍領域配列と相関している場合に、各領域の部分パターンと特定点近傍領域配列との相関を判定するために用いられる情報として予め学習して得られた相関学習結果を含む情報を記憶する相関学習結果記憶手段を備え、検査データ生成処理手段の部分パターン候補データ生成処理手段は、各領域の部分パターンのそれぞれについて、特定点との相対的な位置関係または特定点との相対的な位置関係で定められた他の領域の部分パターン候補配列との相対的な位置関係で、複数の部分パターン候補配列を定める処理を行う構成とされ、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、相関学習結果記憶手段に記憶された相関学習結果を含む情報を用いて、検査データ生成処理手段の部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列と特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定し、相関がないか、または低いと判定された場合には、それらの各領域の部分パターン相当配列および特定点近傍領域配列中の特定点に基づく特徴パターン候補配列の決定処理は行わない構成とされていることが望ましい。

[0043] ここで、「特定点」とは、配列中に存在する何らかの性質または意味を持つ箇所に配置された離散シンボルであり、例えば、DNA配列中の遺伝子情報の周辺に存在する転写開始点等である。

[0044] このように各領域の部分パターン相当配列と特定点近傍領域配列との相関判定処理を行う構成とした場合には、部分パターン相当配列と特定点近傍領域配列との間に相関がないか、または低いときには、それらの部分パターン相当配列および特定点近傍領域配列中の特定点に基づいて定まる配列は、特徴パターン候補配列から除外することが可能となる。このため、特徴パターンの認識精度をより向上させることが可能となる。また、相関がないか、または低いと判定された場合には、それらの部分パターン相当配列および特定点近傍領域配列中の特定点に基づいて定まる配列は、特徴パターン候補配列として採用されないので、その配列については検査データを生成する必要はなく、また、分離処理手段による分離処理および判定処理手段

による判定処理を行う必要もないため、この点で、処理時間の短縮を図ることが可能となる。

[0045] なお、「相関学習結果記憶手段」に記憶される「相関学習結果を含む情報」は、例えば、ニューラルネットワークによる事前学習で得られた情報でもよいが、相関判定の精度向上等の観点から、特に、以下のような独立成分分析(ICA)または主成分分析(PCA)による事前学習で得られた情報とすることが好ましい。

[0046] すなわち、上記のように、各領域の部分パターン相当配列と特定点近傍領域配列との相関判定処理を行う構成とした場合において、相関学習結果記憶手段は、既知の特徴パターンに含まれる複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特徴点近傍領域配列を結合して形成された複数種類の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と相関結合配列内での各配列位置および離散シンボルの各種別とを対応させて形成された相関判定用出現頻度テーブルを記憶する相関判定用出現頻度テーブル記憶手段と、各領域の部分パターンと特定点近傍領域配列との相関を判定する相関判定処理に用いられる相関判定用分離行列を記憶する相関判定用分離行列記憶手段とを含んで構成され、相関判定用分離行列記憶手段には、複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列を結合して生成された相関結合配列、並びに複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列とは異なる非特定点近傍領域配列を結合して形成された非相関結合配列を、相関判定用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して相関判定用学習データ行列を生成し、この相関判定用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、相関結合配列の特徴を示す特徴要素を含む相関判定用基底行列の逆変換を行う行列として得られた相関判定用分離行列が記憶され、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、決定された各領域の部分パターン相当配列と特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定する際には、複数の領域のうちの少なくとも一つの領域の部分パ

ターン相当配列および特定点近傍領域配列を結合して相関判定配列を生成し、この相関判定配列を相関判定用出現頻度テーブル記憶手段に記憶された相関判定用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して相関判定データを生成した後、相関判定用分離行列記憶手段に記憶された相関判定用分離行列と、相関判定データまたは複数の相関判定データを結合した相関判定データ行列とを乗じる行列演算を行うことにより、相関判定用分離済データまたは複数の相関判定用分離済データが結合された相関判定用分離済データ行列を求める相関判定用分離処理を行い、この相関判定用分離処理で求めた相関判定用分離済データまたは相関判定用分離済データ行列を構成する要素のうち、相関判定用基底行列に含まれる特徴要素の行列内位置に応じて定まる相関判定要素の値を用いて、相関があるか否か、または相関の程度を判定する処理を行う構成とされていることが望ましい。

[0047] ここで、「複数の領域のうちの少なくとも一つの領域の既知の部分パターン」とは、相関判定用出現頻度テーブルを作成する際、あるいは学習段階(トレーニングステップ)で相関判定用分離行列を得る際には、全ての領域の既知の部分パターンと既知の特徴点近傍領域配列とを結合して相関結合配列を生成してもよく、あるいは一部の領域の既知の部分パターンと既知の特徴点近傍領域配列とを結合して相関結合配列を生成してもよい趣旨である。同様に、「複数の領域のうちの少なくとも一つの領域の部分パターン相当配列」とは、検査段階(テストステップ)で、全ての領域の部分パターン相当配列と特徴点近傍領域配列とを結合して相関判定配列を生成してもよく、あるいは一部の領域の部分パターン相当配列と特徴点近傍領域配列とを結合して相関判定配列を生成してもよい趣旨である。但し、相関判定の精度向上の観点からは、全ての領域の既知の部分パターンや、全ての領域の部分パターン相当配列を用いて、相関結合配列や相関判定配列を生成することが好ましい。

[0048] また、「相関判定用学習データ行列」は、複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列を結合して生成された相関結合配列、並びに複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列とは異なる非特定点近傍領域配列を結合して

形成された非相関結合配列を、相関判定用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して生成されるものとされているが（後述する図30参照）、少なくとも相関結合配列および非相関結合配列を数値化したものが相関判定用学習データ行列に含まれていればよい趣旨であり、例えば、相関結合配列および非相関結合配列を数値化したものに加え、さらにランダム配列（既知の部分パターンを含まない配列）を数値化したものを結合して相関判定用学習データ行列としてもよい。

[0049] このように独立成分分析（ICA）または主成分分析（PCA）による事前学習で得られた相関判定用分離行列を用いた相関判定処理を行う構成とした場合には、相関判定の精度向上が図られる。

[0050] また、上記のように、独立成分分析（ICA）または主成分分析（PCA）による事前学習で得られた相関判定用分離行列を用いた相関判定処理を行う構成とした場合において、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、相関判定要素の値が予め定められた相関判定用閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、相関があるか否か、または相関の程度を判定する処理を行う構成とすることができる。なお、複数行の相関判定要素の値を用いて、相関判定処理を行う構成としてもよい。

[0051] そして、以上に述べた特徴パターン認識システムは、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、特徴パターンが、DNA配列中のプロモータである場合に、好適に用いることができる。

[0052] また、以上に述べた特徴パターン認識システムのうち、多重化された特徴パターンを認識することができる構成としたシステムは、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、特徴パターンが、DNA配列中のプロモータであり、部分パターンが、プロモータに含まれる－35ボックスおよび－10ボックスである場合に、好適に用いることができる。

[0053] さらに、以上に述べた特徴パターン認識システムのうち、各領域の部分パターン相

当配列と特定点近傍領域配列との相関判定処理を行う構成としたシステムは、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、特徴パターンが、DNA配列中のプロモータであり、部分パターンが、プロモータに含まれる－35ボックスおよび－10ボックスであり、特定点が、DNA配列の転写開始点であり、特徴点近傍領域配列が、転写開始点近傍領域配列である場合に、好適に用いることができる。

[0054] そして、以上に述べた特徴パターン認識システムは、配列が、タンパク質を構成するアミノ酸配列であり、離散シンボルが、アミノ酸配列を構成するアミノ酸を示す記号であり、特徴パターンが、アミノ酸配列中のモチーフである場合にも、好適に用いることができる。

[0055] また、特徴パターンに複数種類の長さが存在する場合において、多重化されていない特徴パターンを認識するか、あるいは多重化されている特徴パターンであるが、多重化されていない場合と同様な処理でその特徴パターンを認識するときには、以下のような構成のシステムとすることができる。

[0056] すなわち、前述した特徴パターン認識システムにおいて、特徴パターンの長さには、特徴パターンの種類の相違により複数種類の長さが存在し、出現頻度テーブル記憶手段には、複数種類の既知の特徴パターンの長さをギャップの挿入により一定長に揃えた後、長さを揃えた複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎にギャップを含む離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置およびギャップを含む離散シンボルの各種別とを対応させて形成された出現頻度テーブルが記憶され、分離行列記憶手段には、長さを揃えた複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた複数種類の非特徴パターンから生成された学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られた分離行列が記憶され、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、特徴パターンの候補となる候補配列を決定した後、決定された候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた候補配列を出現頻

度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う構成とすることができる。

[0057] ここで、「候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理」は、既存の標準的なツールであるブラスト(BLAST)を用い、クラスタルダブリュー(ClustalW)という既存のアルゴリズムで行ってもよく、あるいは本発明のために新たに作成したルールに従って行ってもよい。なお、既知の特徴パターンの長さをギャップの挿入により一定長に揃える場合も同様である。

[0058] このようにアラインメント処理を行う構成とした場合には、特徴パターン(多重化されていない特徴パターン、あるいは多重化されているが、多重化されていない場合と同様な処理で取り扱われる特徴パターン)に複数種類の長さが存在しても、それらの特徴パターンを高い精度で認識することが可能となる。

[0059] さらに、上記の特徴パターン認識システムにおいて、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の候補配列のそれぞれを出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行う構成とされ、分離処理手段は、分離行列と検査データ生成処理手段により生成された複数の検査データまたはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行う構成とされ、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定し、

このうち特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応する候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされていることが望ましい。

[0060] また、上記の特徴パターン認識システムにおいて、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の候補配列のそれぞれを出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行う構成とされ、分離処理手段は、分離行列と検査データ生成処理手段により生成された複数の検査データまたはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行う構成とされ、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる複数列の特徴要素の行列内位置に応じて定まる複数行の特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは特徴パターン寄りの程度を示す値であると判定された類似度を示す値について、

閾値との差の絶対値が最も大きいかまたは特徴パターン寄りの程度が最も高い類似度を示す値を求め、求めた類似度を示す値を与えた検査データに対応する候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成としてもよい。

[0061] このように検査配列の中で候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、これらの複数の候補配列の中から、特徴パターンであると認識される配列を1つ定める処理(但し、1つも特徴パターンであると認識されなくてもよく、ここでは、2以上の候補配列が特徴パターンであると認識されてしまうことがないという趣旨である。)を行う構成とした場合には、検査配列を適切な長さに設定することを前提とし、より高い精度で特徴パターンの認識を行うことが可能となる。

[0062] また、以上に述べた本発明の特徴パターン認識システムにより実現される方法として、以下のような本発明の特徴パターン認識方法が挙げられる。

[0063] すなわち、本発明は、有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定するための特徴パターン認識方法であって、複数種類の既知の特徴パターンを用意するとともに、これらの既知の特徴パターンとは異なる複数種類の非特徴パターンを生成または用意し、複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置および離散シンボルの各種別とを対応させて出現頻度テーブルを作成し、この出現頻度テーブルを出現頻度テーブル記憶手段に記憶させておくとともに、複数種類の既知の特徴パターンおよび複数種類の非特徴パターンを、出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う

行列として得られた分離行列を分離行列記憶手段に記憶させておき、検査データ生成処理手段が、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行い、分離処理手段が、分離行列記憶手段に記憶された分離行列と、検査データ生成処理手段により生成された検査データまたは複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列を求める処理を行い、判定処理手段が、分離処理手段により求めた分離済データまたは分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行うことを特徴とするものである。

[0064] このような本発明の特徴パターン認識方法においては、前述した本発明の特徴パターン認識システムで得られる作用・効果がそのまま得られ、これにより前記目的が達成される。

[0065] また、前述した特徴パターン認識方法において、特徴パターンは、異なる領域に位置する複数の部分パターンを含んで構成される多重化された特徴パターンであり、特徴パターン内での各部分パターンの領域位置および特徴パターンの全体の長さには、特徴パターンの種類の相違により複数種類の領域位置および複数種類の長さが存在し、出現頻度テーブル記憶手段は、複数種類の既知の特徴パターンのパターン全体の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えた後、パターン全体の長さを揃えた複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎にギャップを含む離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置およびギャップを含む離散シンボルの各種別とを対応させて形成されたパ

ターン全体用出現頻度テーブルを記憶するパターン全体用出現頻度テーブル記憶手段であり、分離行列記憶手段は、パターン全体の長さを揃えた複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた複数種類の非特徴パターンから生成されたパターン全体用学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られたパターン全体用分離行列を記憶するパターン全体用分離行列記憶手段であり、各領域の部分パターン毎に形成された部分パターン用出現頻度テーブルを各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段に記憶させておくとともに、各領域の部分パターン毎に得られた部分パターン用分離行列を各領域の部分パターン毎の部分パターン用分離行列記憶手段に記憶させ、さらに、各領域の部分パターン毎に学習段階で部分パターン用分離行列とともに得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を部分パターン用分離済データ行列記憶手段に記憶させておき、各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段には、既知の特徴パターンに含まれる各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンを用いてこれらの部分パターン内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と部分パターン内での各配列位置および離散シンボルの各種別とを対応させて形成された各領域の部分パターン毎の部分パターン用出現頻度テーブルを記憶させておき、各領域の部分パターン毎の部分パターン用分離行列記憶手段には、既知の特徴パターンに含まれる各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンおよびこれらの既知の部分パターンとは異なる複数種類の非部分パターンを、部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して部分パターン用学習データ行列を生成し、この部分パターン用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、部分パターンの特徴を示す特徴要素を含む部分パターン用基底行列の逆変換を行う行列として得られた部分パターン用分離行列を記憶させておき、部分パターン用分離済データ行列記憶手段には、独立成分分析または主成分分析を行うことにより部分パターン用分離行列を得る際に部分パターン用分離行列と部分パターン用学習データ行列

とを乗じた結果として得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を記憶させておき、検査データ生成処理手段により検査データを生成する際には、部分パターン候補データ生成処理手段が、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、各領域の部分パターンのそれぞれについて、部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について部分パターン用出現頻度テーブル記憶手段に記憶された部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成する処理を行い、部分パターン候補データ用分離済データ生成処理手段が、部分パターン用分離行列記憶手段に記憶された部分パターン用分離行列と部分パターン候補データ生成処理手段により生成された複数の部分パターン候補データの各々または複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成する処理を行い、部分パターン相当配列決定処理手段が、部分パターン候補データ用分離済データ生成処理手段により生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた部分パターン候補データ用分離済データに対応する部分パターン候補配列を、特徴パターンの候補となる特徴パターン候補配列に含まれる部分パターン相当配列として決定する処理を行い、特徴パターン候補配列決定処理手段が、部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列に基づき特徴パターン候補配列を決定する処理を行い、アラインメント処理手段が、部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列をそれぞれ一塊の状態に保ったままで、特徴パターン候補配列決定処理手段により決定された特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメン

ト処理を行い、特徴パターン候補配列数値化処理手段が、アラインメント処理手段により長さを揃えた特徴パターン候補配列を、パターン全体用出現頻度テーブル記憶手段に記憶されたパターン全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行うことが望ましい。

[0066] さらに、上記のように、多重化された特徴パターンを認識することができるようにした場合において、パターン全体用分離行列記憶手段に記憶させるパターン全体用分離行列を得る際に用いられるパターン全体用学習データ行列を構成するためのパターン全体の長さを揃えた複数種類の非特徴パターンを生成する際には、これらの非特徴パターンを生成するために用意された非特徴パターン生成用配列の中で、各領域の部分パターンのそれぞれについて、部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について部分パターン用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成し、部分パターン用分離行列と複数の部分パターン候補データの各々または複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成し、生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた部分パターン候補データ用分離済データに対応する部分パターン候補配列を、非特徴パターンに含まれる部分パターン相当配列として決定し、さらに決定された各領域の部分パターン相当配列に基づき非特徴パターンを決定した後、決定された各領域の部分パターン相当配列をそれぞれ一塊の状態に保ったままで、決定された非特徴パターンの長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行うことが望ましい。

[0067] さらに、上記の場合において、複数種類の非特徴パターンの生成時に、複数の部

分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際には、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求めることが望ましい。この際、「類似度」としては、例えば、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和等が挙げられる。

[0068] また、前述したように、多重化された特徴パターンを認識することができるようにした場合において、検査データ生成処理手段の部分パターン相当配列決定処理手段は、部分パターン候補データ用分離済データ生成処理手段により生成された複数の部分パターン候補データ用分離済データのうち、部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行うことが望ましい。この際、「類似度」としては、例えば、部分パターン候補データ用分離済データと部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和等が挙げられる。

[0069] そして、以上に述べた特徴パターン認識方法において、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行うようにすることができる。

[0070] また、以上に述べた特徴パターン認識方法において、基底行列に複数列の特徴要素が現れるときには、分離処理手段により求めた分離済データまたは分離済データ

行列を構成する要素のうちの複数行の要素を特徴判定要素として採用し、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値を用いて、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行うようにすることができる。

[0071] さらに、上記のように複数行の特徴判定要素の値を用いて判定処理を行う場合において、判定処理手段は、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行うようにすることができる。

[0072] また、上記のように類似度を示す値を算出する場合において、判定処理手段は、類似度を示す値として、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの総体との類似度を示す値を算出する処理を行うようにすることが望ましい。より具体的には、判定処理手段は、類似度を示す値として、例えば、分離処理手段により求めた分離済データまたは分離済データ行列のうちの複数行の特徴判定要素の値により構成される列ベクトルと、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの重心ベクトルとの内積の値またはこれと等価な

値を算出することができる。

[0073] そして、前述したように、多重化された特徴パターンを認識することができるようにした場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の特徴パターン候補配列のそれぞれをパターン全体用出現頻度テーブル記憶手段に記憶されたパターン全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行い、分離処理手段は、パターン全体用分離行列と検査データ生成処理手段により生成された複数の検査データの各々またはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行い、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる各特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応する特徴パターン候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行うことが望ましい。

[0074] また、前述したように、多重化された特徴パターンを認識することができるようにした場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはそ

の一部の配列断片から切り出された検査配列の中で、特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の特徴パターン候補配列のそれぞれをパターン全体用出現頻度テーブル記憶手段に記憶されたパターン全体用出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行い、分離処理手段は、パターン全体用分離行列と検査データ生成処理手段により生成された複数の検査データの各々またはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行い、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる複数列の特徴要素の行列内位置に応じて定まる複数行の特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階でパターン全体用分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは特徴パターン寄りの程度を示す値であると判定された類似度を示す値について、閾値との差の絶対値が最も大きいかまたは特徴パターン寄りの程度が最も高い類似度を示す値を求め、求めた類似度を示す値を与えた検査データに対応する特徴パターン候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行うようにしてもよい。

[0075] また、前述したように、多重化された特徴パターンを認識することができるようにした場合において、多重化された特徴パターンに含まれる各領域の部分パターンが、配

列中の特定点を含む特定点近傍領域配列と相関している場合に、各領域の部分パターンと特定点近傍領域配列との相関を判定するために用いられる情報として予め学習して得られた相関学習結果を含む情報を相関学習結果記憶手段に記憶させておき、検査データ生成処理手段の部分パターン候補データ生成処理手段は、各領域の部分パターンのそれぞれについて、特定点との相対的な位置関係または特定点との相対的な位置関係で定められた他の領域の部分パターン候補配列との相対的な位置関係で、複数の部分パターン候補配列を定める処理を行い、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、相関学習結果記憶手段に記憶された相関学習結果を含む情報を用いて、検査データ生成処理手段の部分パターン相当配列決定処理手段により決定された各領域の部分パターン相当配列と特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定し、相関がないか、または低いと判定された場合には、それらの各領域の部分パターン相当配列および特定点近傍領域配列中の特定点に基づく特徴パターン候補配列の決定処理は行わないことが望ましい。

[0076] さらに、上記のように、各領域の部分パターン相当配列と特定点近傍領域配列との間の相関判定処理を行う場合において、相関学習結果記憶手段に、既知の特徴パターンに含まれる複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特徴点近傍領域配列を結合して形成された複数種類の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と相関結合配列内での各配列位置および離散シンボルの各種別とを対応させて形成された相関判定用出現頻度テーブルを記憶する相関判定用出現頻度テーブル記憶手段と、各領域の部分パターンと特定点近傍領域配列との相関を判定する相関判定処理に用いられる相関判定用分離行列を記憶する相関判定用分離行列記憶手段とを設けておき、相関判定用分離行列記憶手段には、複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列を結合して生成された相関結合配列、並びに複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特定点近傍領域配列とは異なる非特定点近傍領域配列を結合して形成された非相関結合配列

を、相関判定用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して相関判定用学習データ行列を生成し、この相関判定用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、相関結合配列の特徴を示す特徴要素を含む相関判定用基底行列の逆変換を行う行列として得られた相関判定用分離行列を記憶させておき、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、決定された各領域の部分パターン相当配列と特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定する際には、複数の領域のうちの少なくとも一つの領域の部分パターン相当配列および特定点近傍領域配列を結合して相関判定配列を生成し、この相関判定配列を相関判定用出現頻度テーブル記憶手段に記憶された相関判定用出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化して相関判定データを生成した後、相関判定用分離行列記憶手段に記憶された相関判定用分離行列と、相関判定データまたは複数の相関判定データを結合した相関判定データ行列とを乗じる行列演算を行うことにより、相関判定用分離済データまたは複数の相関判定用分離済データが結合された相関判定用分離済データ行列を求める相関判定用分離処理を行い、この相関判定用分離処理で求めた相関判定用分離済データまたは相関判定用分離済データ行列を構成する要素のうち、相関判定用基底行列に含まれる特徴要素の行列内位置に応じて定まる相関判定要素の値を用いて、相関があるか否か、または相関の程度を判定する処理を行うことが望ましい。

[0077] そして、上記のように、独立成分分析(ICA)または主成分分析(PCA)による事前学習で得られた相関判定用分離行列を用いた相関判定処理を行う場合において、検査データ生成処理手段の特徴パターン候補配列決定処理手段は、相関判定要素の値が予め定められた相関判定用閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、相関があるか否か、または相関の程度を判定する処理を行うようにすることができる。なお、複数行の相関判定要素の値を用いて、相関判定処理を行うようにしてもよい。

[0078] そして、以上に述べた特徴パターン認識方法は、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代

替記号であり、特徴パターンが、DNA配列中のプロモータである場合に、好適に用いることができる。

[0079] また、以上に述べた特徴パターン認識方法のうち、多重化された特徴パターンを認識することができるようにした方法は、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、特徴パターンが、DNA配列中のプロモータであり、部分パターンが、プロモータに含まれる－35ボックスおよび－10ボックスである場合に、好適に用いることができる。

[0080] さらに、以上に述べた特徴パターン認識方法のうち、各領域の部分パターン相当配列と特定点近傍領域配列との相関判定処理を行うようにした方法は、配列が、DNA配列であり、離散シンボルが、DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、特徴パターンが、DNA配列中のプロモータであり、部分パターンが、プロモータに含まれる－35ボックスおよび－10ボックスであり、特定点が、DNA配列の転写開始点であり、特徴点近傍領域配列が、転写開始点近傍領域配列である場合に、好適に用いることができる。

[0081] また、以上に述べた特徴パターン認識方法は、配列が、タンパク質を構成するアミノ酸配列であり、離散シンボルが、アミノ酸配列を構成するアミノ酸を示す記号であり、特徴パターンが、アミノ酸配列中のモチーフである場合にも、好適に用いることができる。

[0082] そして、特徴パターンに複数種類の長さが存在する場合において、多重化されていない特徴パターンを認識するか、あるいは多重化されている特徴パターンであるが、多重化されていない場合と同様な処理でその特徴パターンを認識するときには、以下のような方法をとることができる。

[0083] すなわち、前述した特徴パターン認識方法において、特徴パターンの長さには、特徴パターンの種類の相違により複数種類の長さが存在し、出現頻度テーブル記憶手段には、複数種類の既知の特徴パターンの長さをギャップの挿入により一定長に揃えた後、長さを揃えた複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎にギャップを含む離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置およびギャップを含む離散

シンボルの各種別とを対応させて形成された出現頻度テーブルを記憶させておくとともに、分離行列記憶手段には、長さを揃えた複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた複数種類の非特徴パターンから生成された学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られた分離行列を記憶させておき、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、特徴パターンの候補となる候補配列を決定した後、決定された候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた候補配列を出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行うことが望ましい。

[0084] また、上記のようにアラインメント処理を行う場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の候補配列のそれぞれを出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行い、分離処理手段は、分離行列と検査データ生成処理手段により生成された複数の検査データまたはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行い、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定し、このうち特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示

す値であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応する候補配列が、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行うことが望ましい。

[0085] さらに、上記のようにアラインメント処理を行う場合において、検査データ生成処理手段は、判定処理の対象となる配列若しくはその一部の配列断片から切り出された検査配列の中で、候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の候補配列のそれぞれを出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置およびギャップを含む離散シンボルの種別に従って数値化することにより、任意の一つの検査配列について複数の検査データを生成する処理を行い、分離処理手段は、分離行列と検査データ生成処理手段により生成された複数の検査データまたはこれらの複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を各検査配列毎に行い、判定処理手段は、分離処理手段により求めた任意の一つの検査配列についての複数の分離済データの各々または分離済データ行列を構成する要素のうち、基底行列に含まれる複数列の特徴要素の行列内位置に応じて定まる複数行の特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち特徴パターンであることを示す側の値であるかまたは特徴パターン寄りの程度を示す値であると判定された類似度を示す値について、閾値との差の絶対値が最も大きいかまたは特徴パターン寄りの程度が最も高い類似度を示す値を求め、求めた類似度を示す値を与えた検査データに対応する候補配列が、複

数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行うようにしてもよい。

[0086] また、本発明は、有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定する処理を行う特徴パターン認識システムとして、コンピュータを機能させるためのプログラムであって、複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と特徴パターン内での各配列位置および離散シンボルの各種別とを対応させて形成された出現頻度テーブルを記憶する出現頻度テーブル記憶手段と、複数種類の既知の特徴パターンおよびこれらの既知の特徴パターンとは異なる複数種類の非特徴パターンを、出現頻度テーブルを用いて配列位置および離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う行列として得られた分離行列を記憶する分離行列記憶手段と、判定処理の対象となる配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、出現頻度テーブル記憶手段に記憶された出現頻度テーブルを用いて配列位置および離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う検査データ生成処理手段と、分離行列記憶手段に記憶された分離行列と検査データ生成処理手段により生成された検査データまたは複数の検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列を求める処理を行う分離処理手段と、この分離処理手段により求めた分離済データまたは分離済データ行列を構成する要素のうち、基底行列に含まれる特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる配列若しくはその一部の配列

断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う判定処理手段とを備えたことを特徴とする特徴パターン認識システムとして、コンピュータを機能させるためのものである。

[0087] なお、上記のプログラムまたはその一部は、例えば、光磁気ディスク(MO)、コンパクトディスク(CD)を利用した読出し専用メモリ(CD－ROM)、CDレコーダブル(CD－R)、CDリライタブル(CD－RW)、デジタル・バーサタイル・ディスク(DVD)を利用した読出し専用メモリ(DVD－ROM)、DVDを利用したランダム・アクセス・メモリ(DVD－RAM)、フレキシブルディスク(FD)、磁気テープ、ハードディスク、読出し専用メモリ(ROM)、電気的消去および書換可能な読出し専用メモリ(EEPROM)、フラッシュ・メモリ、ランダム・アクセス・メモリ(RAM)等の記録媒体に記録して保存や流通等させることが可能であるとともに、例えば、ローカル・エリア・ネットワーク(LAN)、メトロポリタン・エリア・ネットワーク(MAN)、ワイド・エリア・ネットワーク(WAN)、インターネット、イントラネット、エクストラネット等の有線ネットワーク、あるいは無線通信ネットワーク、さらにはこれらの組合せ等の伝送媒体を用いて伝送することが可能であり、また、搬送波に載せて搬送することも可能である。さらに、上記のプログラムは、他のプログラムの一部分であってもよく、あるいは別個のプログラムと共に記録媒体に記録されていてもよい。

発明の効果

[0088] 以上に述べたように本発明によれば、判定対象となる配列を構成する各離散シンボルを、各配列位置毎で、かつ、離散シンボルの各種別毎に求められた出現頻度を用いて数値化するとともに、独立成分分析(ICA)または主成分分析(PCA)を行って得られた分離行列を用いて配列中に特徴パターンが存在するか否かまたは存否の程度を判定するので、従来のニューラルネットワーク法や期待値最大化法を用いたコンピュータによるプロモータ解析の場合に比べ、プロモータ等の特徴パターンの認識精度を向上させることができるうえ、試験管中の生物実験やX線解析によるプロモータ解析の場合と比べても、時間的・経済的に優位な特徴パターンの解析を実現する

ことができるという効果がある。

発明を実施するための最良の形態

[0089] 以下に本発明の各実施形態について図面を参照して説明する。

[0090] [第1実施形態]

図1には、本発明の特徴パターン認識システムである第1実施形態のプロモータ認識システム10の全体構成が示されている。図2には、プロモータ認識システム10の一部である学習処理手段20の詳細構成が示されている。また、図3には、学習段階(トレーニングステップ)の全体の流れがフローチャートで示され、図4には、－35ボックス学習の流れがフローチャートで示され、図5には、DNA配列から切り出された既知のプロモータを含む配列の例が示されている。図6～図10は、－35ボックス学習に伴う処理の説明図である。図11には、－10ボックス学習の流れがフローチャートで示されている。図12～図15は、－10ボックス学習に伴う処理の説明図である。図16～図18には、プロモータ学習の流れがフローチャートで示されている。図19～図26は、プロモータ学習に伴う処理の説明図である。図27には、相関学習の流れがフローチャートで示されている。図28～図31は、相関学習に伴う処理の説明図である。さらに、図32には、検査段階(テストステップ)の全体の流れがフローチャートで示され、図33～図35には、検査段階における詳細処理の流れがフローチャートで示されている。図36～図38は、検査段階における処理の説明図である。

[0091] 図1において、プロモータ認識システム10は、学習段階(トレーニングステップ)の各種処理を行う学習処理手段20と、検査段階(テストステップ)の各種処理を行う検査処理手段30とを備えている。このプロモータ認識システム10は、図5に示すようなA、T、G、Cの4つの塩基記号からなるDNA配列(例えば、大腸菌塩基配列)の中に、特徴パターンであるプロモータが存在するか否かを判定するシステムである。このプロモータ(例えば、大腸菌プロモータ)は、遺伝子情報の周辺に存在し、DNAからRNAへの塩基配列情報の転写が開始される特定点である転写開始点(＋1)およびその1つ上流の塩基(－1)からなる＋1領域と、転写開始点の約35塩基上流にある第1の部分パターンである6塩基からなる－35ボックス(－35box)と、転写開始点の約10塩基上流にある第2の部分パターンである6塩基からなる－10ボックス(－10b

ox)とを含んでいる。また、－35ボックスと－10ボックスとの間には、15～21塩基の比較的一定の距離の配列であるスペーサ35(Spacer35)があり、－10ボックスと転写開始点との間には、3～11塩基の配列であるスペーサ10(Spacer10)がある。なお、本願明細書では、スペーサ10は、－10ボックスと＋1領域(－1および＋1)との間の配列ではなく、－10ボックスと転写開始点(＋1)との間の配列として定義する。また、転写開始点は、前後の塩基とともに、特徴的なパターンを形成することから、本願明細書では、転写開始点(＋1)およびその前後の塩基(－1および＋2)の3塩基からなる配列を、特定点近傍領域配列である転写開始点近傍領域配列と定義する。但し、転写開始点近傍領域配列は、3塩基に限定されるものではなく、例えば、5塩基等としてもよく、要するに、転写開始点(＋1)を含んでいて特徴的なパターンが現れる領域であればよい。

[0092] また、プロモータ認識システム10は、第1の部分パターン用出現頻度テーブルである－35ボックス用出現頻度テーブルT_{-35}を記憶する第1の部分パターン用出現頻度テーブル記憶手段である－35ボックス用出現頻度テーブル記憶手段41と、第1の部分パターン用分離行列である－35ボックス用分離行列W_{-35}を記憶する第1の部分パターン用分離行列記憶手段である－35ボックス用分離行列記憶手段42と、第1の部分パターン用分離済データ行列である－35ボックス用分離済データ行列Y_{-35}を記憶する第1の部分パターン用分離済データ行列記憶手段である－35ボックス用分離済データ行列記憶手段43とを備えている。

[0093] さらに、プロモータ認識システム10は、第2の部分パターン用出現頻度テーブルである－10ボックス用出現頻度テーブルT_{-10}を記憶する第2の部分パターン用出現頻度テーブル記憶手段である－10ボックス用出現頻度テーブル記憶手段51と、第2の部分パターン用分離行列である－10ボックス用分離行列W_{-10}を記憶する第2の部分パターン用分離行列記憶手段である－10ボックス用分離行列記憶手段52と、第2の部分パターン用分離済データ行列である－10ボックス用分離済データ行列Y_{-10}を記憶する第2の部分パターン用分離済データ行列記憶手段である－10ボックス用分離済データ行列記憶手段53とを備えている。

[0094] そして、プロモータ認識システム10は、パターン全体用出現頻度テーブルであるプ

ロモータ用出現頻度テーブルT_{prom}を記憶するパターン全体用出現頻度テーブル記憶手段であるプロモータ用出現頻度テーブル記憶手段61と、パターン全体用分離行列であるプロモータ用分離行列W_{prom}を記憶するパターン全体用分離行列記憶手段であるプロモータ用分離行列記憶手段62と、パターン全体用分離済データ行列であるプロモータ用分離済データ行列Y_{prom}を記憶するパターン全体用分離済データ行列記憶手段であるプロモータ用分離済データ行列記憶手段63(図26参照)とを備えている。

[0095] また、プロモータ認識システム10は、相関学習結果を含む情報を記憶する相関学習結果記憶手段70を備えている。この相関学習結果記憶手段70には、相関判定用出現頻度テーブルT_{BBS}を記憶する相関判定用出現頻度テーブル記憶手段71と、相関判定用分離行列W_{BBS}を記憶する相関判定用分離行列記憶手段72と、相関判定用分離済データ行列Y_{BBS}を記憶する相関判定用分離済データ行列記憶手段73(図31参照)とが設けられている。

[0096] 学習処理手段20は、第1の部分パターン学習処理手段である－35ボックス学習処理手段21と、第2の部分パターン学習処理手段である－10ボックス学習処理手段22と、特徴パターン学習処理手段であるプロモータ学習処理手段23と、相関学習処理手段24とを含んで構成されている。

[0097] 図2において、－35ボックス学習処理手段21は、第1の部分パターン用出現頻度テーブル作成処理手段である－35ボックス用出現頻度テーブル作成処理手段21Aと、第1の非部分パターン生成処理手段であるランダムボックス生成処理手段21Bと、第1の部分パターン用数値化処理手段である－35ボックス用数値化処理手段21Cと、第1の部分パターン用学習データ行列生成処理手段である－35ボックス用学習データ行列生成処理手段21Dと、第1の部分パターン用分析処理手段である－35ボックス用分析処理手段21Eとを含んで構成されている。

[0098] －35ボックス用出現頻度テーブル作成処理手段21Aは、複数種類の既知の－35ボックスの塩基情報を用いて、これらの－35ボックス内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、－35ボックス内での各配列位置および各記号とを対応させて－35ボックス用出現頻度テーブルT

$_{-35}$（図6参照）を作成する処理を行うものである（図4のステップS302参照）。また、－35ボックス用出現頻度テーブル作成処理手段21Aは、作成したT_{-35}を－35ボックス用出現頻度テーブル記憶手段41（図1参照）に記憶させる。

[0099] ランダムボックス生成処理手段21Bは、第1の部分パターン用学習データ行列である－35ボックス用学習データ行列X_{-35}（図8参照）を構成する学習データ（列ベクトル）のうちの偽データ（図8のC_{-35}^{T}を構成する列ベクトル）を用意するために、ランダムボックスを生成する処理を行うものである（図4のステップS304参照）。

[0100] －35ボックス用数値化処理手段21Cは、－35ボックス用出現頻度テーブル記憶手段41（図1参照）に記憶された－35ボックス用出現頻度テーブルT_{-35}（図6参照）を用いて、既知の－35ボックスおよびランダムボックス生成処理手段21Bにより生成されたランダムボックスを、ボックス内での配列位置およびA、T、G、Cの記号種別に従ってそれぞれ数値化し、B_{-35}（図7参照）およびC_{-35}を生成する処理を行うものである（図4のステップS305参照）。

[0101] －35ボックス用学習データ行列生成処理手段21Dは、－35ボックス用数値化処理手段21Cにより数値化して得られたB_{-35}（図7参照）およびC_{-35}をそれぞれ転置してB_{-35}^{T}およびC_{-35}^{T}とし、これらを結合して－35ボックス用学習データ行列X_{-35}（図8参照）を生成する処理を行うものである（図4のステップS306参照）。なお、結合と数値化の順序は、逆にしてもよい。

[0102] －35ボックス用分析処理手段21Eは、－35ボックス用学習データ行列生成処理手段21Dにより生成された－35ボックス用学習データ行列X_{-35}（図8参照）について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後（図9参照）、前処理を経た－35ボックス用学習データ行列X_{-35}を用いて、独立成分分析（ICA）を行うことにより、－35ボックス用分離行列W_{-35}および－35ボックス用分離済データ行列Y_{-35}を求める処理を行うものである（図10および図4のステップS307参照）。また、－35ボックス用分析処理手段21Eは、求めたW_{-35}を－35ボックス用分離行列記憶手段42（図1参照）に記憶させるとともに、求めたY_{-35}を－35ボックス用分離済データ行列記憶手段43（図1参照）に記憶させる。

[0103] －10ボックス学習処理手段22は、第2の部分パターン用出現頻度テーブル作成

処理手段である－10ボックス用出現頻度テーブル作成処理手段22Aと、第2の非部分パターン生成処理手段であるランダムボックス生成処理手段22Bと、第2の部分パターン用数値化処理手段である－10ボックス用数値化処理手段22Cと、第2の部分パターン用学習データ行列生成処理手段である－10ボックス用学習データ行列生成処理手段22Dと、第2の部分パターン用分析処理手段である－10ボックス用分析処理手段22Eとを含んで構成されている。

[0104] －10ボックス用出現頻度テーブル作成処理手段22Aは、複数種類の既知の－10ボックスの塩基情報を用いて、これらの－10ボックス内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、－10ボックス内での各配列位置および各記号とを対応させて－10ボックス用出現頻度テーブルT_{-10}(図12参照)を作成する処理を行うものである(図11のステップS402参照)。また、－10ボックス用出現頻度テーブル作成処理手段22Aは、作成したT_{-10}を－10ボックス用出現頻度テーブル記憶手段51(図1参照)に記憶させる。

[0105] ランダムボックス生成処理手段22Bは、第2の部分パターン用学習データ行列である－10ボックス用学習データ行列X_{-10}(図14参照)を構成する学習データ(列ベクトル)のうちの偽データ(図14のC_{-10}^{T}を構成する列ベクトル)を用意するために、ランダムボックスを生成する処理を行うものである(図11のステップS404参照)。

[0106] －10ボックス用数値化処理手段22Cは、－10ボックス用出現頻度テーブル記憶手段51(図1参照)に記憶された－10ボックス用出現頻度テーブルT_{-10}(図12参照)を用いて、既知の－10ボックスおよびランダムボックス生成処理手段22Bにより生成されたランダムボックスを、ボックス内での配列位置およびA、T、G、Cの記号種別に従ってそれぞれ数値化し、B_{-10}(図13参照)およびC_{-10}を生成する処理を行うものである(図11のステップS405参照)。

[0107] －10ボックス用学習データ行列生成処理手段22Dは、－10ボックス用数値化処理手段22Cにより数値化して得られたB_{-10}(図13参照)およびC_{-10}をそれぞれ転置してB_{-10}^{T}およびC_{-10}^{T}とし、これらを結合して－10ボックス用学習データ行列X_{-10}(図14参照)を生成する処理を行うものである(図11のステップS406参照)。なお、結合と数値化の順序は、逆にしてもよい。

[0108] －10ボックス用分析処理手段22Eは、－10ボックス用学習データ行列生成処理手段22Dにより生成された－10ボックス用学習データ行列X_{-10}（図14参照）について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後（図9と同様）、前処理を経た－10ボックス用学習データ行列X_{-10}を用いて、独立成分分析（ICA）を行うことにより、－10ボックス用分離行列W_{-10}および－10ボックス用分離済データ行列Y_{-10}を求める処理を行うものである（図15および図11のステップS407参照）。また、－10ボックス用分析処理手段22Eは、求めたW_{-10}を－10ボックス用分離行列記憶手段52（図1参照）に記憶させるとともに、求めたY_{-10}を－10ボックス用分離済データ行列記憶手段53（図1参照）に記憶させる。

[0109] プロモータ学習処理手段23は、特徴パターン用アラインメント処理手段であるプロモータ用アラインメント処理手段23Aと、パターン全体用出現頻度テーブル作成処理手段であるプロモータ用出現頻度テーブル作成処理手段23Bと、非特徴パターン生成処理手段である非プロモータ生成処理手段23Cと、非特徴パターン用アラインメント処理手段である非プロモータ用アラインメント処理手段23Dと、パターン全体用数値化処理手段であるプロモータ用数値化処理手段23Eと、パターン全体用学習データ行列生成処理手段であるプロモータ用学習データ行列生成処理手段23Fと、パターン全体用分析処理手段であるプロモータ用分析処理手段23Gとを含んで構成されている。

[0110] プロモータ用アラインメント処理手段23Aは、複数種類の既知のプロモータの長さをギャップ（例えば「－」記号等で示される。）の挿入により一定長に揃えるアラインメント処理を行うものである（図16のステップS501参照）。このアラインメント処理は、予め定められたルールに従って行われるが、検査段階（テストステップ）において検査配列の中で決定されたプロモータ候補配列についてのアラインメント処理と同様のルールに従って行われるため、処理の詳細は後述する（図38参照）。なお、既知のプロモータの長さが、図38に示したルールで既に一定長に揃えられている場合には、このプロモータ用アラインメント処理手段23Aによる処理を行う必要はない。また、既知のプロモータの長さが、既に一定長に揃えられている場合であっても、図38に示したルールとは異なるルールでアラインメントされている場合には、プロモータ用アライン

メント処理手段23Aにより、ギャップを取り除き、図38に示したルールでアラインメント処理をし直す。この際、後述する図17のステップS50302～図18のステップS50312の場合と同様に、－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35}を用いて－10ボックスおよび－35ボックスを探してからアラインメント処理を行ってもよいが、－10ボックスおよび－35ボックスが既知の場合には、それらの既知の－10ボックスおよび－35ボックスをそのまま採用してアラインメント処理を行ってもよい。

[0111] プロモータ用出現頻度テーブル作成処理手段23Bは、プロモータ用アラインメント処理手段23Aにより一定長に揃えられた複数種類の既知のプロモータを用いて、これらのパターン全体内での各配列位置毎に、A、T、G、C、ギャップ（例えば「－」）の各記号毎の出現頻度（frequency）を求め、これらの出現頻度と、パターン全体内での各配列位置および各記号とを対応させてプロモータ用出現頻度テーブルT_{prom}（図19参照）を作成する処理を行うものである（図16のステップS502参照）。また、プロモータ用出現頻度テーブル作成処理手段23Bは、作成したT_{prom}をプロモータ用出現頻度テーブル記憶手段61（図1参照）に記憶させる。

[0112] 非プロモータ生成処理手段23Cは、全体の長さを揃えた複数種類の非特徴パターンである非プロモータを生成する処理を行うものである（図16のステップS503参照）。この非プロモータ生成処理手段23Cは、非プロモータを生成するために用意された非特徴パターン生成用配列である非プロモータ生成用配列（図20参照）の中で、各領域の部分パターンのそれぞれについて、すなわち－10ボックスおよび－35ボックスのそれぞれについて、各ボックスと同じ長さ（ここでは、6塩基）を有する互いに1塩基ずつ位置のずれた複数の部分パターン候補配列（図20のj＝1～9の複数の－10ボックス候補配列、および図22のj＝1～7の複数の－35ボックス候補配列）を定め（図17のステップS50303およびステップS50307参照）、これらの複数の部分パターン候補配列について部分パターン用出現頻度テーブル（－10ボックス用出現頻度テーブルT_{-10}および－35ボックス用出現頻度テーブルT_{-35}）を用いて配列位置およびA、T、G、Cの記号種別に従って数値化して複数の部分パターン候補データ（複数の－10ボックス候補データおよび複数の－35ボックス候補データ）を生成する（図17のステップS50304およびステップS50308参照）。

[0113] そして、非プロモータ生成処理手段23Cは、部分パターン用分離行列(－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35})と、複数の部分パターン候補データを結合した部分パターン候補データ行列(図20の－10ボックス候補データ行列Z_{-10}、および図22の－35ボックス候補データ行列Z_{-35})とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列(図21の－10ボックス候補データ用分離済データ行列Y_{-10can}、および図23の－35ボックス候補データ用分離済データ行列Y_{-35can})を生成する(図17のステップS50305および図18のステップS50309参照)。続いて、非プロモータ生成処理手段23Cは、生成された複数の部分パターン候補データ用分離済データ(－10ボックス候補データ用分離済データ$Y_{-10can}(j)$、および－35ボックス候補データ用分離済データ$Y_{-35can}(j)$)のうち、部分パターン候補データ用分離済データ($Y_{-10can}(j)$および$Y_{-35can}(j)$)と部分パターン用分離済データ行列(図21の－10ボックス用分離済データ行列Y_{-10}、および図23の－35ボックス用分離済データ行列Y_{-35})の真データ対応部分の各列ベクトル($Y_{-10}(k)$および$Y_{-35}(k)$)との内積の和が最大になる部分パターン候補データ用分離済データを求め、求めた部分パターン候補データ用分離済データに対応する部分パターン候補配列(図20のj＝1～9の複数の－10ボックス候補配列のうちのいずれか1つ、および図22のj＝1～7の複数の－35ボックス候補配列のうちのいずれか1つ)を、非特徴パターンである非プロモータに含まれる部分パターン相当配列(－10ボックス相当配列および－35ボックス相当配列)として決定し(図17のステップS50306および図18のステップS50310参照)、さらに決定された各領域の部分パターン相当配列(－10ボックス相当配列および－35ボックス相当配列)に基づき非プロモータを決定する(図18のステップS50311参照)。なお、ここでは、内積の和が最大になる部分パターン候補データ用分離済データを求める処理を行うものとされているが、真データ対応部分の各列ベクトル($Y_{-10}(k)$および$Y_{-35}(k)$)に最も近い部分パターン候補データ用分離済データを求めることができればよく、例えば、真データ対応部分の各列ベクトル($Y_{-10}(k)$および$Y_{-35}(k)$)の総体との類似度が最大になる部分パターン候補データ用分離済データを求めるにあたり、「内積の和」以外の類似度を採用してもよい。

[0114] 非プロモータ用アラインメント処理手段23Dは、非プロモータ生成処理手段23Cにより決定された－10ボックス相当配列および－35ボックス相当配列をそれぞれ一塊の状態に保ったままで、非プロモータの長さをギャップの挿入により一定長（ここでは、配列長：65）に揃えるアラインメント処理を行うものである（図16のステップS503、図18のステップS50312、および図24参照）。

[0115] プロモータ用数値化処理手段23Eは、プロモータ用出現頻度テーブル記憶手段61（図1参照）に記憶されたプロモータ用出現頻度テーブルT_{prom}を用いて、プロモータ用アラインメント処理手段23Aによりパターン全体の長さを一定長（ここでは、配列長：65）に揃えた複数種類の既知の特徴パターンであるプロモータ、および非プロモータ用アラインメント処理手段23Dにより全体の長さを一定長（ここでは、配列長：65）に揃えた複数種類の非特徴パターンである非プロモータを、配列位置およびA、T、G、C、ギャップ（例えば「－」）の記号種別に従って、それぞれ数値化する処理を行うものである（図16のステップS504および図25参照）。

[0116] プロモータ用学習データ行列生成処理手段23Fは、プロモータ用数値化処理手段23Eにより数値化して得られたB_{prom}およびC_{prom}（図25参照）をそれぞれ転置してB_{prom}^{T}およびC_{prom}^{T}とし、これらを結合してプロモータ用学習データ行列X_{prom}（図25参照）を生成する処理を行うものである（図16のステップS505参照）。なお、結合と数値化の順序は、逆にしてもよい。

[0117] プロモータ用分析処理手段23Gは、プロモータ用学習データ行列生成処理手段23Fにより生成されたプロモータ用学習データ行列X_{prom}（図25参照）について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後（図9と同様）、前処理を経たプロモータ用学習データ行列X_{prom}を用いて、独立成分分析（ICA）を行うことにより、プロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める処理を行うものである（図26および図16のステップS506参照）。また、プロモータ用分析処理手段23Gは、求めたW_{prom}をプロモータ用分離行列記憶手段62（図1参照）に記憶させるとともに、求めたY_{prom}をプロモータ用分離済データ行列記憶手段63（図26参照）に記憶させる。なお、Y_{prom}は、後に検査段階の処理で使用されるデータではないが、検査段階の処理で得られる分離済データ行列Y_{test}（図38参照

)の特徴判定要素(第1行の各要素)の値が、正負のいずれの場合にプロモータであると判定されるのかを確認するために求める。

[0118] 相関学習処理手段24は、相関結合配列生成処理手段24Aと、相関判定用出現頻度テーブル作成処理手段24Bと、非相関結合配列生成処理手段24Cと、相関結合配列用数値化処理手段24Dと、相関判定用学習データ行列生成処理手段24Eと、相関判定用分析処理手段24Fとを含んで構成されている。

[0119] 相関結合配列生成処理手段24Aは、既知の部分パターンである－35ボックス(6塩基)および－10ボックス(6塩基)と、既知の特定点近傍領域配列である転写開始点近傍領域配列(一例として3塩基とする。)とを結合して例えば15塩基の相関結合配列(図28参照)を生成する処理を行うものである(図27のステップS601参照)。なお、本実施形態では、一例として、15塩基の相関結合配列としているが、例えば、今後の研究の進歩・発展に伴い、スペーサ10やスペーサ35の部分、あるいはそれ以外の部分に、相関が見い出せる特徴部分が発見された場合には、その部分を含めて相関結合配列を生成してもよい。この場合には、その発見部分も、本発明における部分パターンとして捉えることができる。また、これに伴って、後述する非相関結合配列(図30参照)、相関判定配列(図37参照)も15塩基に限定されず、他の部分を含めて生成されてもよい。

[0120] 相関判定用出現頻度テーブル作成処理手段24Bは、相関結合配列生成処理手段24Aにより生成された複数種類の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、相関結合配列内での各配列位置および各記号とを対応させて相関判定用出現頻度テーブルT_{BBS}(図29参照)を作成する処理を行うものである(図27のステップS601参照)。また、相関判定用出現頻度テーブル作成処理手段24Bは、作成したT_{BBS}を相関判定用出現頻度テーブル記憶手段71(図1参照)に記憶させる。

[0121] 非相関結合配列生成処理手段24Cは、既知の部分パターンである－35ボックス(6塩基)および－10ボックス(6塩基)と、既知の特定点近傍領域配列である転写開始点近傍領域配列とは異なる非特定点近傍領域配列であるランダム配列(一例として3

塩基とする。)とを結合して例えば15塩基の非相関結合配列を生成する処理を行うものである(図27のステップS603および図30参照)。

[0122] 相関結合配列用数値化処理手段24Dは、相関判定用出現頻度テーブル記憶手段71(図1参照)に記憶された相関判定用出現頻度テーブルT_{BBS}(図29参照)を用いて、相関結合配列および非相関結合配列(図30参照)を、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化する処理を行うものである(図27のステップS604および図30参照)。

[0123] 相関判定用学習データ行列生成処理手段24Eは、相関結合配列用数値化処理手段24Dにより数値化して得られたB_{BBS}およびC_{BBS}(図30参照)をそれぞれ転置してB_{BBS}^{T}およびC_{BBS}^{T}とし、これらを結合して相関判定用学習データ行列X_{BBS}(図30参照)を生成する処理を行うものである(図27のステップS605および図30参照)。なお、結合と数値化の順序は、逆にしてもよい。

[0124] 相関判定用分析処理手段24Fは、相関判定用学習データ行列生成処理手段24Eにより生成された相関判定用学習データ行列X_{BBS}(図30参照)について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後(図9と同様)、前処理を経た相関判定用学習データ行列X_{BBS}を用いて、独立成分分析(ICA)を行うことにより、相関判定用分離行列W_{BBS}および相関判定用分離済データ行列Y_{BBS}を求める処理を行うものである(図27のステップS606および図31参照)。また、相関判定用分析処理手段24Fは、求めたW_{BBS}を相関判定用分離行列記憶手段72(図1参照)に記憶させるとともに、求めたY_{BBS}を相関判定用分離済データ行列記憶手段73(図31参照)に記憶させる。なお、Y_{BBS}は、後に検査段階の処理で使用されるデータではないが、検査段階の処理で得られる相関判定用分離済データ行列$Y_{BBS,test}$(図37参照)の相関判定要素(第1行の各要素)の値が、正負のいずれの場合に相関があると判定されるのかを確認するために求める。

[0125] 図1において、検査処理手段30は、検査データ生成処理手段31と、分離処理手段32と、判定処理手段33とを含んで構成されている。

[0126] 検査データ生成処理手段31は、第1の部分パターン候補データ生成処理手段である－35ボックス候補データ生成処理手段31Aと、第2の部分パターン候補データ

生成処理手段である－10ボックス候補データ生成処理手段31Bと、第1の部分パターン候補データ用分離済データ生成処理手段である－35ボックス候補データ用分離済データ生成処理手段31Cと、第2の部分パターン候補データ用分離済データ生成処理手段である－10ボックス候補データ用分離済データ生成処理手段31Dと、第1の部分パターン相当配列決定処理手段である－35ボックス相当配列決定処理手段31Eと、第2の部分パターン相当配列決定処理手段である－10ボックス相当配列決定処理手段31Fと、特徴パターン候補配列決定処理手段であるプロモータ候補配列決定処理手段31Gと、アラインメント処理手段31Hと、特徴パターン候補配列数値化処理手段であるプロモータ候補配列数値化処理手段31Jとを含んで構成されている。

[0127] －35ボックス候補データ生成処理手段31Aは、判定処理の対象となるDNA配列若しくはDNA配列断片から切り出された検査配列の中で、－35ボックスと同じ長さ(6塩基)を有する互いに1塩基ずつ位置のずれた複数の－35ボックス候補配列を定め(図33のステップS1306参照。図22と同様)、これらの複数の－35ボックス候補配列について、－35ボックス用出現頻度テーブル記憶手段41(図1参照)に記憶された－35ボックス用出現頻度テーブルT_{-35}を用いて、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化して複数の－35ボックス候補データを生成する処理を行うものである(図33のステップS1307参照。図22と同様であり、図36の－10ボックス候補データの生成の場合とも同様)。この－35ボックス候補データ生成処理手段31Aは、－10ボックス相当配列決定処理手段31Fにより決定した－10ボックス相当配列の位置を基準として、－10ボックス相当配列と各－35ボックス候補配列との間の隙間(スペーサ35に相当する隙間)が15～21塩基になるように複数の－35ボックス候補配列を定める。従って、－10ボックス候補データ生成処理手段31Bが、後述するように転写開始点を1塩基ずつずらしながら複数設定する構成である場合には、それに伴って－10ボックス相当配列決定処理手段31Fにより－10ボックス相当配列が複数決定されるので、－35ボックス候補データ生成処理手段31Aは、それらの複数の－10ボックス相当配列のそれぞれを基準として、各－10ボックス相当配列の1つ1つにつき、複数の－35ボックス候補配列を定める。

[0128] －10ボックス候補データ生成処理手段31Bは、判定処理の対象となるDNA配列若しくはDNA配列断片から切り出された検査配列の中で、転写開始点を設定し（図33のステップS1301参照）、設定した転写開始点を基準として、－10ボックスと同じ長さ（6塩基）を有する互いに1塩基ずつ位置のずれた複数の－10ボックス候補配列を、転写開始点と各－10ボックス候補配列との間の隙間（スペーサ10に相当する隙間）が3～11塩基になるように定め（図33のステップS1302および図36参照）、これらの複数の－10ボックス候補配列について、－10ボックス用出現頻度テーブル記憶手段51（図1参照）に記憶された－10ボックス用出現頻度テーブルT_{-10}を用いて、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化して複数の－10ボックス候補データを生成する処理を行うものである（図33のステップS1303および図36参照）。この－10ボックス候補データ生成処理手段31Bは、1つの検査配列の中で、1つの転写開始点を設定する構成としてもよいが、認識精度の向上の観点から、図36に示すように、1つの検査配列の中で、転写開始点を1塩基ずつずらしながら複数設定し、それらの1つ1つの転写開始点について、複数の－10ボックス候補配列を定めることが好ましい（図34のステップS1311，S1312参照）。

[0129] －35ボックス候補データ用分離済データ生成処理手段31Cは、－35ボックス用分離行列記憶手段42（図1参照）に記憶された－35ボックス用分離行列W_{-35}と、－35ボックス候補データ生成処理手段31Aにより生成された複数の－35ボックス候補データを結合した－35ボックス候補データ行列Z_{-35}とを乗じる行列演算を行うことにより、複数の－35ボックス候補データ用分離済データが結合された－35ボックス候補データ用分離済データ行列Y_{-35can}を生成する処理を行うものである（図34のステップS1308参照。図23と同様）。

[0130] －10ボックス候補データ用分離済データ生成処理手段31Dは、－10ボックス用分離行列記憶手段52（図1参照）に記憶された－10ボックス用分離行列W_{-10}と、－10ボックス候補データ生成処理手段31Bにより生成された複数の－10ボックス候補データを結合した－10ボックス候補データ行列Z_{-10}とを乗じる行列演算を行うことにより、複数の－10ボックス候補データ用分離済データが結合された－10ボックス候補データ用分離済データ行列Y_{-10can}を生成する処理を行うものである（図33のステップS1

304参照。図21と同様)。

[0131] －35ボックス相当配列決定処理手段31Eは、－35ボックス候補データ用分離済データ生成処理手段31Cにより生成された複数の－35ボックス候補データ用分離済データ(すなわち、－35ボックス候補データ用分離済データ行列Y_{-35can}を構成する各列ベクトル$Y_{-35can}(j)$)のうち、－35ボックス候補データ用分離済データ$Y_{-35can}(j)$と、－35ボックス用分離済データ行列Y_{-35}の真データ対応部分の各列ベクトル$Y_{-35}(k)$との内積の和$q(j)$が最大になる－35ボックス候補データ用分離済データを求め(図23と同様)、求めた－35ボックス候補データ用分離済データに対応する－35ボックス候補配列(図22の場合と同様に、j＝1～7の複数の－35ボックス候補配列のうちのいずれか1つ)を－35ボックス相当配列として決定する処理を行うものである(図34のステップS1309参照)。なお、ここでは、内積の和$q(j)$が最大になる－35ボックス候補データ用分離済データを求める処理を行うものとされているが、－35ボックス用分離済データ行列Y_{-35}の真データ対応部分の各列ベクトル$Y_{-35}(k)$に最も近い－35ボックス候補データ用分離済データを求めることができればよく、例えば、－35ボックス用分離済データ行列Y_{-35}の真データ対応部分の各列ベクトル$Y_{-35}(k)$の総体との類似度が最大になる－35ボックス候補データ用分離済データを求めるにあたり、「内積の和」以外の類似度を採用してもよい。

[0132] －10ボックス相当配列決定処理手段31Fは、－10ボックス候補データ用分離済データ生成処理手段31Dにより生成された複数の－10ボックス候補データ用分離済データ(すなわち、－10ボックス候補データ用分離済データ行列Y_{-10can}を構成する各列ベクトル$Y_{-10can}(j)$)のうち、－10ボックス候補データ用分離済データ$Y_{-10can}(j)$と、－10ボックス用分離済データ行列Y_{-10}の真データ対応部分の各列ベクトル$Y_{-10}(k)$との内積の和$q(j)$が最大になる－10ボックス候補データ用分離済データを求め(図21と同様)、求めた－10ボックス候補データ用分離済データに対応する－10ボックス候補配列(図20の場合と同様に、j＝1～9の複数の－10ボックス候補配列のうちのいずれか1つ)を－10ボックス相当配列として決定する処理を行うものである(図33のステップS1305参照)。なお、ここでは、内積の和$q(j)$が最大になる－10ボックス候補データ用分離済データを求める処理を行うものとされているが、－10ボックス用分離

済データ行列Y_{-10}の真データ対応部分の各列ベクトル$Y_{-10}(k)$に最も近い－10ボックス候補データ用分離済データを求めることができればよく、例えば、－10ボックス用分離済データ行列Y_{-10}の真データ対応部分の各列ベクトル$Y_{-10}(k)$の総体との類似度が最大になる－10ボックス候補データ用分離済データを求めるにあたり、「内積の和」以外の類似度を採用してもよい。

[0133] プロモータ候補配列決定処理手段31Gは、－35ボックス相当配列決定処理手段31Eにより決定された－35ボックス相当配列および－10ボックス相当配列決定処理手段31Fにより決定された－35ボックス相当配列、並びにこれらに対応する転写開始点に基づき、プロモータ候補配列を決定する処理を行うものである（図34のステップS1310および図37参照）。

[0134] また、プロモータ候補配列決定処理手段31Gは、決定された－35ボックス相当配列および－10ボックス相当配列と、これらに対応する転写開始点を含む転写開始点近傍領域配列との間に相関があるか否かを判定し、相関がない場合には、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点に基づくプロモータ候補配列の決定処理は行わない。すなわち、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点で定まる配列は、プロモータ候補配列として採用されない。

[0135] より具体的には、プロモータ候補配列決定処理手段31Gは、決定された－35ボックス相当配列および－10ボックス相当配列、並びにこれらに対応する転写開始点を含む転写開始点近傍領域配列を結合して相関判定配列を生成し、この相関判定配列を相関判定用出現頻度テーブル記憶手段71（図1参照）に記憶された相関判定用出現頻度テーブルT_{BBS}を用いて、配列位置およびA、T、G、Cの記号種別に従って数値化して相関判定データを生成した後、相関判定用分離行列記憶手段72（図1参照）に記憶された相関判定用分離行列W_{BBS}と、相関判定データまたは複数の相関判定データを結合した相関判定データ行列$X_{BBS,test}$とを乗じる行列演算を行うことにより、相関判定用分離済データまたは複数の相関判定用分離済データが結合された相関判定用分離済データ行列$Y_{BBS,test}$を求める相関判定用分離処理を行い、この相

関判定用分離処理で求めた相関判定用分離済データまたは相関判定用分離済データ行列$Y_{BBS,test}$を構成する要素のうち、推定された相関判定用基底行列W_{BBS}^{-1}に含まれる特徴要素(例えば、第1列の各要素)の行列内位置に応じて定まる相関判定要素(例えば、第1行の各要素、ベクトルの場合には、第1番目の要素)の値が、予め定められた相関判定用閾値(例えば、ゼロ)を境界としていずれの側の値であるかを判定することにより、相関があるか否かを判定する処理を行う(図35のステップS13101～S13104および図37参照)。

[0136] アラインメント処理手段31Hは、プロモータ候補配列決定処理手段31Gにより決定されたプロモータ候補配列について、予め定められたルール(詳細は後述する。)に従って、ギャップ(例えば「－」)の挿入により、長さを一定長(ここでは、一例として配列長:65とする。)に揃えるアラインメント処理を行うものである(図32のステップS14および図38参照)。

[0137] プロモータ候補配列数値化処理手段31Jは、プロモータ用出現頻度テーブル記憶手段61(図1参照)に記憶されたプロモータ用出現頻度テーブルT_{prom}(図19参照)を用いて、アラインメント処理手段31Hによりアラインメントされたプロモータ候補配列(図38参照)を、配列位置およびA、T、G、C、ギャップ(例えば「－」)の記号種別に従って数値化し、検査データを生成する処理を行うものである(図32のステップS15および図38参照)。

[0138] 分離処理手段32は、検査データ生成処理手段31により生成された検査データ(縦ベクトル)または複数の検査データを結合した検査データ行列X_{test}について前処理を行った後、プロモータ用分離行列記憶手段62(図1参照)に記憶されたプロモータ用分離行列W_{prom}と、前処理済みの検査データ(縦ベクトル)または複数の検査データを結合した検査データ行列X_{test}とを乗じる行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列Y_{test}を求める分離処理を行うものである(図32のステップS16および図38参照)。

[0139] また、分離処理手段32は、1つの検査配列の中で、複数のプロモータ候補配列が決定された場合(図36参照)には、それらの各プロモータ候補配列から得られる各検査データのそれぞれについて上記の分離処理を行い、1つの検査配列について複

数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列Y_{test}を求める処理を行う。

[0140] 判定処理手段33は、分離処理手段32により求めた分離済データまたは分離済データ行列Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}に含まれる特徴要素（例えば、第1列の各要素）の行列内位置に応じて定まる特徴判定要素（例えば、第1行の各要素、ベクトルの場合には、第1番目の要素）の値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかを判定することにより、判定処理の対象となるDNA配列若しくはDNA配列断片から切り出された検査配列の中に、複数種類の既知のプロモータのうちのいずれかのプロモータまたは既知のプロモータに類似する新規なプロモータが含まれるか否かを判定する処理を行う（図32のステップS17参照）。

[0141] また、判定処理手段33は、1つの検査配列の中で、複数のプロモータ候補配列が決定された場合（図36参照）には、分離処理手段32により求めた1つの検査配列についての複数の分離済データの各々またはこれらの複数の分離済データが結合された分離済データ行列Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}に含まれる特徴要素（例えば、第1列の各要素）の行列内位置に応じて定まる各特徴判定要素（例えば、第1行の各要素、ベクトルの場合には、第1番目の要素）の値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかをそれぞれ判定し、このうちプロモータであることを示す側の値（例えば、正の値）であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きい特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応するプロモータ候補配列が、複数種類の既知のプロモータのうちのいずれかのプロモータまたは既知のプロモータに類似する新規なプロモータであると認識する処理を行う。

[0142] 以上において、学習処理手段20および検査処理手段30は、プロモータ認識システム10を構成するコンピュータ（パーソナル・コンピュータのみならず、その上位機種または下位機種のものも含む。）の内部に設けられた中央演算処理装置（CPU）、およびこのCPUの動作手順を規定する一つまたは複数のプログラムにより実現される。

[0143] また、学習処理手段20と検査処理手段30とは、物理的に同じコンピュータで実現

する必要はなく、別のコンピュータにより実現してもよい。

[0144] さらに、学習処理手段20に含まれる各処理手段21(21A～21E), 22(22A～22E), 23(23A～23G), 24(24A～24F)は、一台のコンピュータあるいは一つのCPUにより実現されるものに限定されず、複数のコンピュータ等で分散処理(各処理手段21(21A～21E), 22(22A～22E), 23(23A～23G), 24(24A～24F)毎の機能的な分散処理、および並列処理を行って処理速度を向上させるための分散処理を含む。)を行うことにより実現されるものであってもよい。検査処理手段30に含まれる各処理手段31(31A～31J), 32, 33についても同様である。

[0145] なお、学習段階で行う各種処理は、必ずしもオンライン処理が要求されないことから、学習処理手段20により行われる各種処理の少なくとも一部は、手計算で行ってもよい。但し、ユーザの手間の軽減、新規なプロモータが発見された場合の再学習の円滑化等の観点から、学習段階で行う各種処理は、本実施形態のように、学習処理手段20によりコンピュータ処理で行うことが好ましい。

[0146] また、各記憶手段41～43, 51～53, 61～63, 70(71～73)としては、例えば、ハードディスク、ROM、EEPROM、フラッシュ・メモリ、RAM、MO、CD－ROM、CD－R、CD－RW、DVD－ROM、DVD－RAM、FD、磁気テープ、あるいはこれらの組合せ等を採用することができる。

[0147] このような第2実施形態においては、以下のようにしてプロモータ認識システム10によりDNA配列中のプロモータの認識が行われる。

[0148] ＜学習段階(トレーニングステップ)＞

図3において、先ず、プロモータ認識システム10を構成するコンピュータの電源を投入し、学習段階の各種処理のためのプログラムを立ち上げる(ステップS1)。次に、学習用の既知のプロモータNp本を用意する。ここでは、一例として、非特許文献1に記載された図5に示すような解析済みの大腸菌プロモータNp本(例えば106本とする。)を用いることにする(ステップS2)。そして、－35ボックス学習の処理(ステップS3)に移行する。

[0149] (－35ボックス学習)

図4において、図5に示すような学習用の既知のプロモータNp本(106本)から、－

35ボックスに相当する6塩基(既知であるものとする。)をNp個(106個)切り出す(ステップS301)。

[0150] 次に、－35ボックス用出現頻度テーブル作成処理手段21Aにより、Np個(106個)の－35ボックスの塩基情報を用いて、これらの－35ボックス内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、－35ボックス内での各配列位置(1～6)および各記号(A、T、G、C)とを対応させて、図6に示すような－35ボックス用出現頻度テーブルT_{-35}を作成し(図4のステップS302)、作成したT_{-35}を－35ボックス用出現頻度テーブル記憶手段41(図1参照)に記憶させる。

[0151] 図6において、$f_{A,-35(1)}$、$f_{A,-35(2)}$、$f_{A,-35(3)}$、…は、－35ボックス内の第1番目、第2番目、第3番目、…の位置におけるAの出現頻度を意味し、$f_{T,-35(1)}$、$f_{T,-35(2)}$、$f_{T,-35(3)}$、…は、－35ボックス内の第1番目、第2番目、第3番目、…の位置におけるTの出現頻度を意味する。G、Cについても同様である。また、－35ボックス用出現頻度テーブルT_{-35}において縦方向に並んだ各数値、すなわち－35ボックス内の同じ位置におけるA、T、G、Cの各出現頻度を加算すると、1．0になるように正規化されている。例えば、$f_{A,-35(1)}+f_{T,-35(1)}+f_{G,-35(1)}+f_{C,-35(1)}=1.0$となる。他の出現頻度テーブル$T_{-35}$、$T_{-10}$、$T_{prom}$、$T_{BS}$の場合も同様に正規化されている。但し、これらの正規化にあたっては、合計値を1．0とする必要はなく、例えば、1000等としてもよい。

[0152] 続いて、Np個の－35ボックスから、重なっている－35ボックスを取り除き、Np個からN_{-35}個(例えば、72個)にする(図4のステップS303)。なお、異なる種類のプロモータであっても、－35ボックスが同じになることがあるので、重なりが生じる。

[0153] それから、ランダムボックス生成処理手段21Bにより、図8に示すような－35ボックス用学習データ行列X_{-35}を構成する学習データ(列ベクトル)のうちの偽データ(図8のC_{-35}^{T}を構成する列ベクトル)を用意するために、6塩基のランダムボックスをN_{-35}個(72個)生成する(図4のステップS304)。この際、ランダムボックス生成処理手段21Bは、ボックス内の各配列位置での一致が2塩基以下となるようなランダムボックスを生成する。また、ランダムボックスは、A、T、G、Cの各出現頻度に偏りを持たせることなく、つまり、A、T、G、Cの各出現頻度が0．25ずつになるように生成される。なお、本

実施形態では、－35ボックス用学習データ行列X_{-35}を形成するために用意する－35ボックスの個数と、ランダムボックスとの個数は、同数のN_{-35}個（72個）としたが、同数に限定されるものではなく、例えば、－35ボックスの特徴が強い場合には、ランダムボックスの個数を－35ボックスの個数よりも多くしてもよい。他の学習データ行列X_{-10}、X_{prom}、X_{BBS}の場合も同様である。

[0154] 次に、－35ボックス用数値化処理手段21Cにより、－35ボックス用出現頻度テーブル記憶手段41（図1参照）に記憶された－35ボックス用出現頻度テーブルT_{-35}（図6参照）を用いて、既知の－35ボックスおよびランダムボックス生成処理手段21Bにより生成されたランダムボックスを、ボックス内での配列位置およびA、T、G、Cの記号種別に従ってそれぞれ数値化し、B_{-35}（図7参照）およびC_{-35}を生成する（図4のステップS305）。

[0155] 続いて、－35ボックス用学習データ行列生成処理手段21Dにより、－35ボックス用数値化処理手段21Cにより数値化して得られたB_{-35}（図7参照）およびC_{-35}をそれぞれ転置してB_{-35}^{T}およびC_{-35}^{T}とし、これらを結合して図8に示すような－35ボックス用学習データ行列X_{-35}を生成する（図4のステップS306）。図8において、－35ボックスおよびランダムボックスをT_{-35}でそれぞれ数値化したデータは、－35ボックス用学習データ行列X_{-35}を構成する各列ベクトルとなっている。このうち、B_{-35}^{T}は、真データである－35ボックスの部分であり、C_{-35}^{T}は、偽データであるランダムボックスの部分である。

[0156] その後、－35ボックス用分析処理手段21Eにより、－35ボックス用学習データ行列生成処理手段21Dにより生成された－35ボックス用学習データ行列X_{-35}（図8参照）について、前処理（pre-processing）として、平均値のゼロ化処理（zero-mean）およびホワイトニング処理（whitening：白色化処理）を行う。図9に示すように、先ず、－35ボックス用学習データ行列X_{-35}の各行について、行内の各要素の数値の平均値をゼロにする平均値のゼロ化処理を行う。次に、ホワイトニング処理として、X_{-35}を構成する各列ベクトル$X_{-35}(j)$（j＝1～144）と、これを転置したベクトル$X_{-35}(j)^{T}$とを乗じた$X_{-35}(j)X_{-35}(j)^{T}$をj＝1～144について加算して6行×6列の共分散行列Rを作成する。そして、この共分散行列Rの固有値λ_1～λ_6を対角線に大きい順に並べて持つ対角行列

D(対角線の要素以外は0の行列)、および固有値λ_1～λ_6に対応する固有ベクトルを列に持つ行列Gを得る。それから、$V=D^{-1/2}G^T$を用い、X_{-35}を構成する各列ベクトル$X_{-35}(j)$(j=1～144)について、Vと$X_{-35}(j)$とを乗じる行列演算を行い、得られたV$X_{-35}(j)$を新たに$X_{-35}(j)$(j=1～144)とし、これにより、前処理済みの－35ボックス用学習データ行列X_{-35}を生成する。なお、以上は、線形変換を用いた共分散行列Rによる処理を行う場合の説明であるが、本発明では、X_{-35}の非線形変換を用いた共分散行列による処理、すなわち$X_{-35}(j)$を非線形変換してから、新たに$X_{-35}(j)$と置き直す処理を行ってもよい。

[0157] それから、図10に示すように、－35ボックス用分析処理手段21Eにより、前処理済みの－35ボックス用学習データ行列X_{-35}を用いて、独立成分分析(ICA)を行い、$Y_{-35}=W_{-35}X_{-35}$を満たし、かつY_{-35}の各成分が独立になるような－35ボックス用分離行列W_{-35}および－35ボックス用分離済データ行列Y_{-35}を求め、求めたW_{-35}およびY_{-35}を－35ボックス用分離行列記憶手段42および－35ボックス用分離済データ行列記憶手段43(図1参照)に記憶させる(図4のステップS307)。ここで、独立成分分析(ICA)は、Y_{-35}を構成する各列ベクトル$Y_{-35}(j)$が独立になるような－35ボックス用分離行列W_{-35}を求める処理である。また、後述するプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際の独立成分分析(ICA)の場合(図16のステップS506)と同様に、いずれの種類の独立成分分析を行ってもよい。さらに、後述するプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際と同様に、－35ボックス用分析処理として主成分分析(PCA)を行い、次元縮約を行ってもよい。そして、前述したX_{-35}について非線形変換したものを入力として線形問題の解を求める独立成分分析(ICA)は、非線形ICAと呼ぶことができるものであるが、本発明には、このような非線形の独立成分分析(ICA)を行う場合も含まれる。なお、線形変換または非線形変換を行って得られた学習データを入力として非線形問題の解を求める独立成分分析(ICA)も、勿論、非線形ICAと呼ぶことができるものであるが、この場合でも本発明と同様な目的を達することができ、これを近似して線形化する処理も本発明に含まれる。

[0158] (－10ボックス学習)

図11において、図5に示すような学習用の既知のプロモータNp本(106本)から、－10ボックスに相当する6塩基(既知であるものとする。)をNp個(106個)切り出す(ステップS401)。

[0159] 次に、－10ボックス用出現頻度テーブル作成処理手段22Aにより、Np個(106個)の－10ボックスの塩基情報を用いて、これらの－10ボックス内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、－10ボックス内での各配列位置(1～6)および各記号(A、T、G、C)とを対応させて、図12に示すような－10ボックス用出現頻度テーブルT_{-10}を作成し(図11のステップS402)、作成したT_{-10}を－10ボックス用出現頻度テーブル記憶手段51(図1参照)に記憶させる。

[0160] 続いて、Np個の－10ボックスから、重なっている－10ボックスを取り除き、Np個からN_{-10}個(例えば、58個)にする(図11のステップS403)。なお、異なる種類のプロモータであっても、－10ボックスが同じになることがあるので、重なりが生じる。

[0161] それから、ランダムボックス生成処理手段22Bにより、図14に示すような－10ボックス用学習データ行列X_{-10}を構成する学習データ(列ベクトル)のうちの偽データ(図14のC_{-10}^{T}を構成する列ベクトル)を用意するために、6塩基のランダムボックスをN_{-10}個(58個)生成する(図11のステップS404)。この際、ランダムボックス生成処理手段22Bは、ボックス内の各配列位置での一致が2塩基以下となるようなランダムボックスを生成する。また、ランダムボックスは、A、T、G、Cの各出現頻度に偏りを持たせることなく、つまり、A、T、G、Cの各出現頻度が0．25ずつになるように生成される。

[0162] 次に、－10ボックス用数値化処理手段22Cにより、－10ボックス用出現頻度テーブル記憶手段51(図1参照)に記憶された－10ボックス用出現頻度テーブルT_{-10}(図12参照)を用いて、既知の－10ボックスおよびランダムボックス生成処理手段22Bにより生成されたランダムボックスを、ボックス内での配列位置およびA、T、G、Cの記号種別に従ってそれぞれ数値化し、B_{-10}(図13参照)およびC_{-10}を生成する(図11のステップS405)。

[0163] 続いて、－10ボックス用学習データ行列生成処理手段22Dにより、－10ボックス用数値化処理手段22Cにより数値化して得られたB_{-10}(図13参照)およびC_{-10}をそれ

ぞれ転置してB_{-10}^{T}およびC_{-10}^{T}とし、これらを結合して図14に示すような－10ボックス用学習データ行列X_{-10}を生成する(図11のステップS406)。図14において、－10ボックスおよびランダムボックスをT_{-10}でそれぞれ数値化したデータは、－10ボックス用学習データ行列X_{-10}を構成する各列ベクトルとなっている。このうち、B_{-10}^{T}は、真データである－10ボックスの部分であり、C_{-10}^{T}は、偽データであるランダムボックスの部分である。

[0164] その後、－10ボックス用分析処理手段22Eにより、－10ボックス用学習データ行列生成処理手段22Dにより生成された－10ボックス用学習データ行列X_{-10}(図14参照)について、－35ボックスの場合(図9参照)と同様に、前処理として、平均値のゼロ化処理およびホワイトニング処理を行う。それから、$V=D^{-1/2}G^{T}$を用い、X_{-10}を構成する各列ベクトル$X_{-10}(j)$(j＝1～116)について、Vと$X_{-10}(j)$とを乗じる行列演算を行い、得られた$VX_{-10}(j)$を新たに$X_{-10}(j)$(j＝1～116)とし、これにより、前処理済みの－10ボックス用学習データ行列X_{-10}を生成する。なお、－35ボックスの場合と同様に、X_{-10}の非線形変換を用いた共分散行列による処理、すなわち$X_{-10}(j)$を非線形変換してから、新たに$X_{-10}(j)$と置き直す処理を行ってもよい。

[0165] それから、図15に示すように、－10ボックス用分析処理手段22Eにより、前処理済みの－10ボックス用学習データ行列X_{-10}を用いて、独立成分分析(ICA)を行い、$Y_{-10}=W_{-10}X_{-10}$を満たし、かつY_{-10}の各成分が独立になるような－10ボックス用分離行列W_{-10}および－10ボックス用分離済データ行列Y_{-10}を求め、求めたW_{-10}およびY_{-10}を－10ボックス用分離行列記憶手段52および－10ボックス用分離済データ行列記憶手段53(図1参照)に記憶させる(図11のステップS407)。ここで、独立成分分析(ICA)は、Y_{-10}を構成する各列ベクトル$Y_{-10}(j)$が独立になるような－10ボックス用分離行列W_{-10}を求める処理である。また、後述するプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際の独立成分分析(ICA)の場合(図16のステップS506)と同様に、いずれの種類の独立成分分析を行ってもよい。さらに、後述するプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際と同様に、－10ボックス用分析処理として主成分分析(PCA)を行い、次元縮約を行ってもよい。そして、前述したX_{-10}について非線形変換したものを入力として線

形問題の解を求める独立成分分析(ICA)は、非線形ICAと呼ぶことができるものであるが、本発明には、このような非線形の独立成分分析(ICA)を行う場合も含まれる。なお、線形変換または非線形変換を行って得られた学習データを入力として非線形問題の解を求める独立成分分析(ICA)も、勿論、非線形ICAと呼ぶことができるものであるが、この場合でも本発明と同様な目的を達することができ、これを近似して線形化する処理も本発明に含まれる。

[0166] (プロモータ学習)

図16において、先ず、プロモータ用アラインメント処理手段23Aにより、既知のプロモータNp本(106本)の長さをギャップ(例えば「－」)の挿入により一定長(ここでは、配列長:65)に揃えるアラインメント処理、すなわち図5の上部から下部に示した状態にする処理を行う(ステップS501)。このアラインメント処理は、予め定められたルールに従って行われるが、検査段階(テストステップ)において検査配列の中で決定されたプロモータ候補配列についてのアラインメント処理と同様のルールに従って行われるため、処理の詳細は後述する(図38参照)。なお、Np本(106本)のプロモータは、全て異なる種類であるため、—35ボックスや－10ボックスの場合のように同じものは存在しないことから、重なりを取り除く処理は行わない。

[0167] 続いて、プロモータ用出現頻度テーブル作成処理手段23Bにより、プロモータ用アラインメント処理手段23Aにより一定長に揃えられた既知のプロモータNp本(106本)を用いて、これらのパターン全体内での各配列位置毎に、A、T、G、C、ギャップ(「－」)の各記号毎の出現頻度(frequency)を求め、これらの出現頻度と、パターン全体内での各配列位置および各記号とを対応させてプロモータ用出現頻度テーブルT_{pro}(図19参照)を作成し、作成したT_{prom}を$_{m}$プロモータ用出現頻度テーブル記憶手段61(図1参照)に記憶させる(図16のステップS502)。

[0168] それから、非プロモータ生成処理手段23Cにより、全体の長さを一定長(ここでは、配列長:65)に揃えた複数種類の非プロモータを生成する(図16のステップS503)。すなわち、非プロモータ生成処理手段23Cは、図20に示すように、先ず、非プロモータを生成するための非プロモータ生成用配列(例えば、50塩基)を用意する(図17のステップS50301)。この際、例えば、AおよびTの各出現頻度をそれぞれ0．3とし

、GおよびCの各出現頻度をそれぞれ0．2とし、複数本の非プロモータ生成用配列をランダムに自動生成する。

[0169] 次に、非プロモータ生成処理手段23Cは、非プロモータ生成用配列の中で、転写開始点を設定する（図17のステップS50302）。この際、非プロモータ生成処理手段23Cは、例えば、非プロモータ生成用配列において、下流側（図20中の右側）から見た最初のAまたはGを転写開始点とみなす。なお、図20では、転写開始点は、小文字で示されている。

[0170] 続いて、非プロモータ生成処理手段23Cは、図20に示すように、設定した転写開始点を基準として、－10ボックスと同じ長さ（6塩基）を有する互いに1塩基ずつ位置のずれた複数の－10ボックス候補配列（j＝1～9）を、各－10ボックス候補配列と転写開始点との間の隙間（スペーサ10に相当する隙間）が3～11塩基になるように定める（図17のステップS50303）。

[0171] それから、非プロモータ生成処理手段23Cは、複数（9個）の－10ボックス候補配列を、－10ボックス用出現頻度テーブル記憶手段51（図1参照）に記憶された－10ボックス用出現頻度テーブルT_{-10}を用いて、配列位置およびA、T、G、Cの記号種別に従って数値化し、複数（9個）の－10ボックス候補データを生成する（図17のステップS50304）。なお、これらの複数（9個）の－10ボックス候補データについては、前処理（平均値のゼロ化処理やホワイトニング処理）は行わなくてもよい。

[0172] そして、非プロモータ生成処理手段23Cは、図21に示すように、－10ボックス用分離行列W_{-10}と、複数（9個）の－10ボックス候補データを結合した－10ボックス候補データ行列Z_{-10}とを乗じる行列演算、すなわち$Y_{-10can} = W_{-10} Z_{-10}$という行列演算を行い、複数（9個）の－10ボックス候補データ用分離済データが結合された－10ボックス候補データ用分離済データ行列Y_{-10can}を生成する（図17のステップS50305）。

[0173] 続いて、非プロモータ生成処理手段23Cは、生成された複数（9個）の－10ボックス候補データ用分離済データ$Y_{-10can}(j)$（j＝1～9）のうち、$Y_{-10can}(j)$と図21の－10ボックス用分離済データ行列Y_{-10}の真データ対応部分の各列ベクトル$Y_{-10}(k)$との内積の和q（i）が最大になる$Y_{-10can}(j)$を求め、求めた$Y_{-10can}(j)$に対応する－10ボックス候補配列（図20のj＝1～9の複数の－10ボックス候補配列のうちのいずれか1つ）を、非

プロモータに含まれる－10ボックス相当配列として決定する（図17のステップS5030 6）。なお、－10ボックス用分離済データ行列Y_{-10}の真データ対応部分とは、図14の－10ボックス用学習データ行列X_{-10}を構成する各列ベクトルのうちの真データである－10ボックスの部分B_{-10}^{T}（58列）に対応する部分である。

[0174] すなわち、$q(j)=\Sigma<Y_{-10can}(j), Y_{-10}(k)>$という演算を行う。ここで、j＝1～9であり、k＝1～58であり、Σは、k＝1～58の和を意味し、＜…，…＞は、内積を意味する。そして、$J_{-10}=\arg\{\max\ q(j)\}$（但し、j＝1～9）を求め、求めた$J_{-10}$（j＝1～9のうちのいずれか1つ）を、非プロモータに含まれる－10ボックス相当配列の開始点（一番下流側の点）とする。なお、$J_{-10}=1$（j＝1）は、転写開始点と－10ボックス相当配列との間の隙間（スペーサ10に相当する隙間）が3塩基になるような－10ボックス相当配列の開始点であり、$J_{-10}=9$（j＝9）は、転写開始点と－10ボックス相当配列との間の隙間が11塩基になるような－10ボックス相当配列の開始点である（図20参照）。

[0175] その後、非プロモータ生成処理手段23Cは、図22に示すように、決定した－10ボックス相当配列の位置を基準として、－35ボックスと同じ長さ（6塩基）を有する互いに1塩基ずつ位置のずれた複数の－35ボックス候補配列（j＝1～7）を、各－35ボックス候補配列と－10ボックス相当配列との間の隙間（スペーサ35に相当する隙間）が15～21塩基になるように定める（図17のステップS50307）。

[0176] それから、非プロモータ生成処理手段23Cは、複数（7個）の－35ボックス候補配列を、－35ボックス用出現頻度テーブル記憶手段41（図1参照）に記憶された－35ボックス用出現頻度テーブルT_{-35}を用いて、配列位置およびA、T、G、Cの記号種別に従って数値化し、複数（7個）の－35ボックス候補データを生成する（図17のステップS50308）。なお、これらの複数（7個）の－35ボックス候補データについては、前処理（平均値のゼロ化処理やホワイトニング処理）は行わなくてもよい。

[0177] そして、非プロモータ生成処理手段23Cは、図23に示すように、－35ボックス用分離行列W_{-35}と、複数（7個）の－35ボックス候補データを結合した－35ボックス候補データ行列Z_{-35}とを乗じる行列演算、すなわち$Y_{-35can}=W_{-35}Z_{-35}$という行列演算を行い、複数（7個）の－35ボックス候補データ用分離済データが結合された－35ボックス候補データ用分離済データ行列Y_{-35can}を生成する（図18のステップS50309）。

[0178] 続いて、非プロモータ生成処理手段23Cは、生成された複数(7個)の－35ボックス候補データ用分離済データ$Y_{-35can}(j)$ (j＝1～7)のうち、$Y_{-35can}(j)$と図23の－35ボックス用分離済データ行列Y_{-35}の真データ対応部分の各列ベクトル$Y_{-35}(k)$との内積の和q(i)が最大になる$Y_{-35can}(j)$を求め、求めた$Y_{-35can}(j)$に対応する－35ボックス候補配列(図22のj＝1～7の複数の－35ボックス候補配列のうちのいずれか1つ)を、非プロモータに含まれる－35ボックス相当配列として決定する(図18のステップS50310)。なお、－35ボックス用分離済データ行列Y_{-35}の真データ対応部分とは、図8の－35ボックス用学習データ行列X_{-35}を構成する各列ベクトルのうちの真データである－35ボックスの部分B_{-35}^{T}(72列)に対応する部分である。

[0179] すなわち、$q(j)＝\Sigma＜Y_{-35can}(j), Y_{-35}(k)＞$という演算を行う。ここで、j＝1～7であり、k＝1～72であり、Σは、k＝1～72の和を意味し、＜…, …＞は、内積を意味する。そして、$J_{-35}＝\arg\{\max\ q(j)\}$ (但し、j＝1～7)を求め、求めたJ_{-35} (j＝1～7のうちのいずれか1つ)を、非プロモータに含まれる－35ボックス相当配列の開始点(一番下流側の点)とする。なお、$J_{-35}＝1$ (j＝1)は、－10ボックス相当配列と－35ボックス相当配列との間の隙間(スペーサ35に相当する隙間)が15塩基になるような－35ボックス相当配列の開始点であり、$J_{-35}＝7$ (j＝7)は、－10ボックス相当配列と－35ボックス相当配列との間の隙間が21塩基になるような－35ボックス相当配列の開始点である(図22参照)。

[0180] その後、非プロモータ生成処理手段23Cは、決定した－10ボックス相当配列および－35ボックス相当配列、並びに設定した転写開始点に基づき、非プロモータを決定する(図18のステップS50311)。

[0181] それから、非プロモータ用アラインメント処理手段23Dにより、図24に示すように、非プロモータ生成処理手段23Cにより決定された－10ボックス相当配列および－35ボックス相当配列をそれぞれ一塊の状態に保ったままで、非プロモータの長さをギャップの挿入により一定長(ここでは、配列長:65)に揃えるアラインメント処理を行う(図18のステップS50312)。すなわち、－10ボックス相当配列および－35ボックス相当配列は、そのままの長さとし、スペーサ10、スペーサ35の配列長をそれぞれ11、21に揃える。このアラインメント処理は、検査段階(テストステップ)で行われるアライン

メント処理(図38参照)と同様であるため、詳細は後述する。

[0182] そして、アラインメントされた非プロモータが、アラインメントされた学習用の既知のプロモータに対し、パターン全体内での各配列位置(1～65の各位置)での一致が25塩基以下であるか否かを判定し、25塩基以下の非プロモータのみ採用する(図18のステップS50313)。

[0183] 続いて、採用された非プロモータがNp本(106本)用意されたか否かを判定し(図18のステップS50314)、Np本(106本)用意されていない場合には、図17のステップS50301の処理に戻る。

[0184] 一方、Np本(106本)用意された場合には、図25に示すように、プロモータ用数値化処理手段23Eにより、プロモータ用出現頻度テーブル記憶手段61(図1参照)に記憶されたプロモータ用出現頻度テーブルT_{prom}を用いて、プロモータ用アラインメント処理手段23Aによりパターン全体の長さを一定長(ここでは、配列長:65)に揃えた既知のプロモータNp本(106本)、および非プロモータ用アラインメント処理手段23Dにより全体の長さを一定長(ここでは、配列長:65)に揃えた非プロモータNp本(106本)を、配列位置およびA、T、G、C、ギャップ(例えば「－」)の記号種別に従って、それぞれ数値化する処理を行う(図16のステップS504)。

[0185] それから、プロモータ用学習データ行列生成処理手段23Fにより、図25に示すように、プロモータ用数値化処理手段23Eによりプロモータ Np本(106本)および非プロモータNp本(106本)を数値化して得られたB_{prom}およびC_{prom}をそれぞれ転置してB_{prom}^{T}およびC_{prom}^{T}とし、これらを結合してプロモータ用学習データ行列X_{prom}(図25参照)を生成する処理を行う(図16のステップS505)。

[0186] 続いて、プロモータ用分析処理手段23Gにより、プロモータ用学習データ行列生成処理手段23Fにより生成されたプロモータ用学習データ行列X_{prom}(図25参照)について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後(図9と同様)、図26に示すように、前処理を経たプロモータ用学習データ行列X_{prom}を用いて、独立成分分析(ICA)を行うことにより、プロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求め、求めたW_{prom}およびY_{prom}をプロモータ用分離行列記憶手段62(図1参照)およびプロモータ用分離済データ行列記憶手段63(図

26参照）に記憶させる（図16のステップS506）。なお、前述した－35ボックスや－10ボックスの場合と同様に、前処理で、X_{prom}の非線形変換を用いた共分散行列による処理、すなわち$X_{prom}(j)$（j＝1～212）を非線形変換してから、新たに$X_{prom}(j)$と置き直す処理を行ってもよい。そして、X_{prom}について非線形変換したものを入力として線形問題の解を求める独立成分分析（ICA）は、非線形ICAと呼ぶことができるものであるが、本発明には、このような非線形の独立成分分析（ICA）を行う場合も含まれる。なお、線形変換または非線形変換を行って得られた学習データを入力として非線形問題の解を求める独立成分分析（ICA）も、勿論、非線形ICAと呼ぶことができるものであるが、この場合でも本発明と同様な目的を達することができ、これを近似して線形化する処理も本発明に含まれる。

[0187] また、認識精度向上等の観点からは、上記のように独立成分分析（ICA）を行うことが好ましいが、処理時間の短縮等の観点からは、次のように主成分分析（PCA）を行ってもよい。プロモータ学習では、図9において、共分散行列Rは、65行×65列の正方行列となり、この共分散行列Rの固有値λ_1～λ_{65}を対角線に大きい順に並べて持つ65行×65列の対角行列D（対角線の要素以外は0の行列）、および固有値λ_1～λ_{65}に対応する固有ベクトルを列に持つ65行×65列の行列Gが得られる。ここで、65次元からn次元への次元縮約を行うとすると、65行×65列の対角行列Dは、共分散行列Rの固有値λ_1～λ_nを対角線に大きい順に並べて持つn行×n列の対角行列D_nとなり、65行×65列の行列Gは、固有値λ_1～λ_nに対応する固有ベクトルを列に持つ65行×n列の行列G_nとなる。従って、$D_n^{-1/2}$は、n行×n列の行列となり、G_n^Tは、n行×65列の行列となるので、$V_n = D_n^{-1/2} G_n^T$であるから、V_nは、n行×65列の行列となる。そこで、このn行×65列の行列V_nを、独立成分分析（ICA）におけるプロモータ用分離行列W_{prom}に相当するプロモータ用分離行列とみなせば、検査段階（テストステップ）では、n行×65列のプロモータ用分離行列V_nと、アラインメントされた65次元の検査データ（縦ベクトル）または複数の検査データを結合した65行×r（任意数）列の検査データ行列X_{test}とを乗じる行列演算、$Y_{n,test} = V_n X_{test}$を行うことにより、n次元の分離済データ（縦ベクトル）または複数の分離済データが結合されたn行×r（任意数）列の分離済データ行列$Y_{n,test}$が得られる。そして、n次元の分離済データ（縦ベクト

ル)の特徴判定要素(第1番目の要素)の値、またはn行×r(任意数)列の分離済データ行列$Y_{n,test}$の特徴判定要素(第1行の各要素)の値が、閾値のいずれの側の値であるかを判定することにより、プロモータであるか否かを判定することができる。

[0188] さらに、n次元に次元縮約された$V_n X_{prom}$(X_{prom}が65次元の縦ベクトルであれば、$V_n X_{prom}$は、n次元の縦ベクトルとなり、X_{prom}が65行×r(任意数)列の行列であれば、$V_n X_{prom}$は、n行×r(任意数)列の行列となる。)を、新たな学習データ(縦ベクトル)または学習データ行列$X_{n,prom}$とし、この$X_{n,prom}$を用いて独立成分分析(ICA)を行い、n行×n列のプロモータ用分離行列$W_{n,prom}$を求めてもよい。この場合には、検査段階(テストステップ)において、n行×n列のプロモータ用分離行列$W_{n,prom}$と、n次元の検査データ(縦ベクトル)またはn行×r(任意数)列の検査データ行列$X_{n,test}$とを乗じる行列演算、$Y_{n,test}=W_{n,prom}X_{n,test}$を行うことにより、n次元の分離済データ(縦ベクトル)またはn行×r(任意数)列の分離済データ行列$Y_{n,test}$が得られる。そして、n次元の分離済データ(縦ベクトル)の特徴判定要素(第1番目の要素)の値、またはn行×r(任意数)列の分離済データ行列$Y_{n,test}$の特徴判定要素(第1行の各要素)の値が、閾値のいずれの側の値であるかを判定することにより、プロモータであるか否かを判定することができる。なお、上記において、n行×n列のプロモータ用分離行列$W_{n,prom}$と、n行×65列の行列V_nとを乗じた$W_{n,prom}V_n$、すなわち$W_{n,prom}D_n^{-1/2}G_n^T$をプロモータ用分離行列とみなしてもよい。

[0189] そして、独立成分分析(ICA)としては、(1)平均相互情報量の最小化を、勾配法により行う方法、(2)不動点法により非ガウス性の増大を行う方法、(3)高次統計量の対角化を行う方法のいずれの方法を適用してもよい。

[0190] なお、独立成分分析(ICA)または主成分分析(PCA)を行った結果、プロモータ用分離済データ行列Y_{prom}(または$Y_{n,prom}$、以下、同様である。)の特徴判定要素(第1行の各要素)の値の正負が、反転して得られることがあり得るため、プロモータ用分離済データ行列Y_{prom}の特徴判定要素の値が、正負のいずれであるかを確認する処理を行う。この確認処理は、機械的に行ってもよく、人間が行ってもよい。そして、プロモータ用学習データ行列X_{prom}のうちの真データであるB_{prom}^Tの部分の列ベクトルに対応するY_{prom}の列ベクトルの特徴判定要素の値が、正の値(例えば、1近傍の値)であれ

ば、検査段階の処理で得られる分離済データ行列Y_{test}（図38参照）の特徴判定要素（第1行の各要素）の値が、正の値（例えば、1近傍の値）になった場合に、プロモータであると判定することができ、負の値（例えば、－1近傍の値）になった場合に、プロモータではないと判定することができる。一方、プロモータ用学習データ行列X_{prom}のうちの真データである$B_{prom}{}^{T}$の部分の列ベクトルに対応するY_{prom}の列ベクトルの特徴判定要素の値が、負の値（例えば、－1近傍の値）であれば、検査段階の処理で得られる分離済データ行列Y_{test}（図38参照）の特徴判定要素（第1行の各要素）の値が、負の値（例えば、－1近傍の値）になった場合に、プロモータであると判定することができ、正の値（例えば、1近傍の値）になった場合に、プロモータではないと判定することができる。従って、これらの判定に必要な情報を、判定処理手段33に設定しておく。この設定処理は、機械的に行ってもよく、人間が行ってもよい。

[0191] （相関学習）

図27において、先ず、相関結合配列生成処理手段24Aにより、図28に示すように、既知のNp個（106個）の－35ボックス（6塩基）および既知のNp個（106個）の－10ボックス（6塩基）と、既知の転写開始点近傍領域配列（本実施形態では、一例として3塩基とする。）とを結合し、例えば15塩基の相関結合配列をNp個（106個）生成する（図27のステップS601）。なお、転写開始点近傍領域配列の長さは、例えば4塩基や5塩基等であってもよい。

[0192] 次に、相関判定用出現頻度テーブル作成処理手段24Bにより、相関結合配列生成処理手段24Aにより生成されたNp個（106個）の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に、A、T、G、Cの各記号毎の出現頻度（frequency）を求め、これらの出現頻度と、相関結合配列内での各配列位置および各記号とを対応させて図29に示すような相関判定用出現頻度テーブルT_{BBS}を作成し、作成したT_{BBS}を相関判定用出現頻度テーブル記憶手段71（図1参照）に記憶させる（図27のステップS601）。

[0193] 続いて、図30に示すように、Np個（106個）の相関結合配列を、例えば10倍の10×Np個（1060個）にし、ランダムに配置する。この拡大処理は、同じデータを例えば10倍に拡大するものであり、統計学におけるブーストラッピングの処理である。

[0194] それから、非相関結合配列生成処理手段24Cにより、図30に示すように、既知の10×Np個(1060個)の－35ボックス(6塩基)および既知の10×Np個(1060個)の－10ボックス(6塩基)と、既知の転写開始点近傍領域配列とは異なる10×Np個(1060個)のランダム配列(一例として3塩基とする。)とを結合し、例えば15塩基の非相関結合配列を10×Np個(1060個)生成する(図27のステップS603)。この際、10×Np個(1060個)の3塩基のランダム配列は、A、T、G、Cの4つの記号の順列である4^3＝64パターンのうち、既知の－35ボックスおよび既知の－10ボックスに対する既知の転写開始点近傍領域配列(つまり、真のデータである1パターン)および記号的にあり得ないパターンを除いた3字組の集合から、ランダムに選び出して生成される。なお、図30に示すように、15塩基の非相関結合配列は、相関結合配列における既知の3塩基の転写開始点近傍領域配列の部分だけ、3塩基のランダム配列に入れ替えればよいので、残りの6塩基の－35ボックスおよび6塩基の－10ボックスの部分は、相関結合配列と同じでよい。

[0195] そして、相関結合配列用数値化処理手段24Dにより、図30に示すように、相関判定用出現頻度テーブル記憶手段71(図1参照)に記憶された相関判定用出現頻度テーブルT_{BBS}(図29参照)を用いて、10×Np個(1060個)の相関結合配列および10×Np個(1060個)の非相関結合配列を、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化し、行列B_{BBS}および行列C_{BBS}を得る(図27のステップS604)。

[0196] その後、相関判定用学習データ行列生成処理手段24Eにより、図30に示すように、相関結合配列用数値化処理手段24Dにより数値化して得られたB_{BBS}およびC_{BBS}をそれぞれ転置してB_{BBS}^{T}およびC_{BBS}^{T}とし、これらを結合して相関判定用学習データ行列X_{BBS}を生成する(図27のステップS605)。

[0197] なお、相関判定用学習データ行列X_{BBS}を生成する際に、B_{BBS}^{T}およびC_{BBS}^{T}に加え、図30中の点線で示すように、配列長が15の例えば10×Np個(1060個)のランダム配列(既知の－35ボックスや既知の－10ボックスとは異なる配列)により形成されるD_{BBS}^{T}を結合してもよい。

[0198] 続いて、相関判定用分析処理手段24Fにより、相関判定用学習データ行列生成処

理手段24Eにより生成された相関判定用学習データ行列X_{BBS}（図30参照）について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行い、得られた結果を新たにX_{BBS}とした後（図9と同様）、図31に示すように、前処理を経た相関判定用学習データ行列X_{BBS}を用いて、独立成分分析（ICA）を行うことにより、相関判定用分離行列W_{BBS}および相関判定用分離済データ行列Y_{BBS}を求め、求めたW_{BBS}およびY_{BS}を相関判定用分離行列記憶手段72（図1参照）および相関判定用分離済データ行列記憶手段73（図31参照）に記憶させる（図27のステップS606）。独立成分分析（ICA）は、Y_{BBS}を構成する各列ベクトル$Y_{BBS}(j)$が独立になるような相関判定用分離行列W_{BBS}を求める処理である。また、前述したプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際の独立成分分析（ICA）の場合（図16のステップS506）と同様に、いずれの種類の独立成分分析を行ってもよい。さらに、前述したプロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める際と同様に、相関判定用分析処理として主成分分析（PCA）を行い、次元縮約を行ってもよい。そして、非線形の独立成分分析（ICA）や主成分分析（PCA）を行ってもよい。

[0199] なお、独立成分分析（ICA）または主成分分析（PCA）を行った結果、相関判定用分離済データ行列Y_{BBS}（またはn次元に縮約された$Y_{n,BBS}$、以下、同様である。）の相関判定要素（第1行の各要素）の値の正負が、反転して得られることがあり得るため、相関判定用分離済データ行列Y_{BBS}の相関判定要素の値が、正負のいずれであるかを確認する処理を行う。この確認処理は、機械的に行ってもよく、人間が行ってもよい。そして、相関判定用学習データ行列X_{BBS}のうちの真データである$B_{BBS}{}^{T}$の部分の列ベクトルに対応するY_{BBS}の列ベクトルの相関判定要素の値が、正の値（例えば、1近傍の値）であれば、検査段階の処理で得られる相関判定用分離済データ行列$Y_{BBS,test}$（図37参照）の相関判定要素（第1行の各要素）の値が、正の値（例えば、1近傍の値）になった場合に、相関があると判定することができ、負の値（例えば、－1近傍の値）になった場合に、相関がないと判定することができる。一方、相関判定用学習データ行列X_{BBS}のうちの真データである$B_{BBS}{}^{T}$の部分の列ベクトルに対応するY_{BBS}の列ベクトルの相関判定要素の値が、負の値（例えば、1近傍の値）であれば、検査段階

の処理で得られる相関判定用分離済データ行列$Y_{BBS,test}$（図37参照）の相関判定要素（第1行の各要素）の値が、負の値（例えば、1近傍の値）になった場合に、相関があると判定することができ、正の値（例えば、－1近傍の値）になった場合に、相関がないと判定することができる。従って、これらの相関判定に必要な情報を、プロモータ候補配列決定処理手段31Gに設定しておく。この設定処理は、機械的に行ってもよく、人間が行ってもよい。

[0200] 以上で、学習段階（トレーニングステップ）の処理は終了する（図3のステップS7）。その後、検査段階（テストステップ）の処理へ移行する。

[0201] ＜検査段階（テストステップ）＞

図32において、先ず、プロモータ認識システム10を構成するコンピュータの電源を投入し、必要なプログラムを立ち上げてプロモータ認識処理を開始し（ステップS11）、判定対象となるDNA配列またはDNA配列断片から、検査配列を切り出す（ステップS12）。なお、判定対象となるDNA配列またはDNA配列断片から既に切り出されている検査配列が与えられた場合には、切出処理は行わなくてよい。

[0202] この際、切り出す検査配列の長さは、複数種類の類似する既知のプロモータのうちの最長のプロモータの長さ以上とし、図36に示すように、少なくとも最長のプロモータの長さの分だけ重なるようにして順次位置をずらしながら、DNA配列またはDNA配列断片から検査配列を切り出していく。なお、類似する新規なプロモータ（新たに発見される可能性のあるプロモータ）が、既知の最長のプロモータよりも長くなることがある得るため、切り出す検査配列の長さは、最長のプロモータよりも若干長いプロモータを想定し、その想定された新規な最長のプロモータの長さ以上とすることが好ましく、同様に、切り出しの際の重なりは、少なくとも、想定された新規な最長のプロモータの長さの分だけ確保することが好ましい。また、切り出された検査配列の中に、複数のプロモータが存在する状況になると、1つの検査配列の中からプロモータであると認識される最も確からしいものを1つ決定する処理を行ううえで不都合が生じるので、検査配列の長さは、既知の最短のプロモータの長さの2倍未満の長さ、または想定される新規な最短のプロモータの長さの2倍未満の長さとすることが好ましい。例えば、非特許文献1に記載された既知の大腸菌プロモータの場合には、既知のプロモータ

の長さが、31～45塩基であるから、検査配列の長さは、45以上で、かつ、62(31×2)未満、すなわち45～61とすることができ、また、相関判定を考慮し、転写開始点近傍領域配列の3塩基を全て含めて(つまり、転写開始点の1つ下流の＋2の塩基も含めて)処理上必要なプロモータの長さとするならば、その処理上必要なプロモータの長さは、32～46塩基となるので、検査配列の長さは、46以上で、かつ、64(32×2)未満、すなわち46～63とすることができ、さらに、想定される新規な最長のプロモータを考慮するならば、検査配列の長さの下限値は、大きくなり、一方、想定される新規な最短のプロモータを考慮するならば、検査配列の長さの上限値は、小さくなる。本実施形態では、一例として検査配列の長さを50塩基として説明を行うものとする。

[0203] 次に、検査配列(50塩基)の中で、プロモータ候補配列を決定する(ステップS13)。この際、転写開始点を1塩基ずつずらして設定し、1つの検査配列の中で、図36中の二点鎖線で示すように複数のプロモータ候補配列を決定してもよい。具体的には、以下のようになる。

[0204] 図33において、先ず、－10ボックス候補データ生成処理手段31Bにより、検査配列の中で、転写開始点を設定する(ステップS1301)。この際、転写開始点は、下流側(図36中の右側)から見て最初のAまたはGに設定する等により、1つの検査配列につき1つだけ設定してもよいが、認識精度向上の観点から、以下に述べるステップS1302～S1312の処理を繰り返すことにより、1つの検査配列につき複数設定することが好ましいので、この場合について説明を行う。

[0205] また、－10ボックス候補データ生成処理手段31Bは、ステップS1301および後述するステップS1312で転写開始点を設定する際に、転写開始点近傍領域配列が、あり得ない配列になっているか否かを判定し、あり得ない配列になっている場合には、その転写開始点は採用しない処理を行ってもよい。なお、この処理を行う際には、転写開始点近傍領域配列としてあり得る配列の全てを、図示されないメモリに記憶しておけばよい。

[0206] 次に、－10ボックス候補データ生成処理手段31Bにより、図36に示すように、設定した転写開始点を基準として、－10ボックスと同じ長さ(6塩基)を有する互いに1塩基ずつ位置のずれた複数(9個)の－10ボックス候補配列を、転写開始点と各－10

ボックス候補配列との間の隙間(スペーサ10に相当する隙間)が3～11塩基になるように定める(図33のステップS1302)。

[0207] 続いて、－10ボックス候補データ生成処理手段31Bにより、図36に示すように、複数(9個)の－10ボックス候補配列について、－10ボックス用出現頻度テーブル記憶手段51(図1参照)に記憶された－10ボックス用出現頻度テーブルT_{-10}を用いて、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化して複数(9個)の－10ボックス候補データを生成する(図33のステップS1303)。

[0208] それから、－10ボックス候補データ用分離済データ生成処理手段31Dにより、図21の場合と同様にして、－10ボックス用分離行列記憶手段52(図1参照)に記憶された－10ボックス用分離行列W_{-10}と、－10ボックス候補データ生成処理手段31Bにより生成された複数(9個)の－10ボックス候補データを結合した－10ボックス候補データ行列Z_{-10}とを乗じる行列演算、すなわち$Y_{-10can} = W_{-10} Z_{-10}$を行い、複数(9個)の－10ボックス候補データ用分離済データが結合された－10ボックス候補データ用分離済データ行列Y_{-10can}を生成する(図33のステップS1304)。

[0209] その後、－10ボックス相当配列決定処理手段31Fにより、図21の場合と同様にして、－10ボックス候補データ用分離済データ生成処理手段31Dにより生成された複数(9個)の－10ボックス候補データ用分離済データ(すなわち、－10ボックス候補データ用分離済データ行列Y_{-10can}を構成する各列ベクトル$Y_{-10can}(j)$、但しj＝1～9)のうち、－10ボックス候補データ用分離済データ$Y_{-10can}(j)$と、－10ボックス用分離済データ行列Y_{-10}の真データ対応部分の各列ベクトル$Y_{-10}(k)$との内積の和$q(j)$が最大になる－10ボックス候補データ用分離済データを求め、求めた－10ボックス候補データ用分離済データに対応する－10ボックス候補配列(図20の場合と同様に、j＝1～9の複数の－10ボックス候補配列のうちのいずれか1つ)を－10ボックス相当配列として決定する(図33のステップS1305)。

[0210] すなわち、$q(j) = \Sigma < Y_{-10can}(j), Y_{-10}(k) >$という演算を行う。ここで、j＝1～9であり、k＝1～58であり、Σは、k＝1～58の和を意味し、＜…, …＞は、内積を意味する。そして、$J_{-10} = \arg\{\max\ q(j)\}$(但し、j＝1～9)を求め、求めた$J_{-10}$(j＝1～9のうちのいずれか1つ)を、プロモータ候補配列に含まれる－10ボックス相当配列の開始

点（一番下流側の点）とする。なお、$J_{-10}=1$（j＝1）は、転写開始点と－10ボックス相当配列との間の隙間（スペーサ10に相当する隙間）が3塩基になるような－10ボックス相当配列の開始点であり、$J_{-10}=9$（j＝9）は、転写開始点と－10ボックス相当配列との間の隙間が11塩基になるような－10ボックス相当配列の開始点である（図20と同様）。

[0211] 続いて、－35ボックス候補データ生成処理手段31Aにより、図22の場合と同様にして、検査配列の中で、－10ボックス相当配列決定処理手段31Fにより決定した－10ボックス相当配列の位置を基準として、－35ボックスと同じ長さ（6塩基）を有する互いに1塩基ずつ位置のずれた複数（7個）の－35ボックス候補配列を、－10ボックス相当配列と各－35ボックス候補配列との間の隙間（スペーサ35に相当する隙間）が15～21塩基になるように定める（図33のステップS1306）。

[0212] それから、－35ボックス候補データ生成処理手段31Aにより、図22の場合と同様にして、複数（7個）の－35ボックス候補配列について、－35ボックス用出現頻度テーブル記憶手段41（図1参照）に記憶された－35ボックス用出現頻度テーブルT_{-35}を用いて、配列位置およびA、T、G、Cの記号種別に従って、それぞれ数値化して複数（7個）の－35ボックス候補データを生成する（図33のステップS1307）。

[0213] その後、－35ボックス候補データ用分離済データ生成処理手段31Cにより、図23の場合と同様にして、－35ボックス用分離行列記憶手段42（図1参照）に記憶された－35ボックス用分離行列W_{-35}と、－35ボックス候補データ生成処理手段31Aにより生成された複数（7個）の－35ボックス候補データを結合した－35ボックス候補データ行列Z_{-35}とを乗じる行列演算、すなわち$Y_{-35can}=W_{-35}Z_{-35}$という行列演算を行い、複数（7個）の－35ボックス候補データ用分離済データが結合された－35ボックス候補データ用分離済データ行列Y_{-35can}を生成する（図34のステップS1308）。

[0214] 続いて、－35ボックス相当配列決定処理手段31Eにより、図23の場合と同様にして、－35ボックス候補データ用分離済データ生成処理手段31Cにより生成された複数（7個）の－35ボックス候補データ用分離済データ（すなわち、－35ボックス候補データ用分離済データ行列Y_{-35can}を構成する各列ベクトル$Y_{-35can}(j)$、但し、j＝1～7）のうち、－35ボックス候補データ用分離済データ$Y_{-35can}(j)$と、－35ボックス用分離済

データ行列Y_{-35}の真データ対応部分の各列ベクトル$Y_{-35}(k)$との内積の和$q(j)$が最大になる－35ボックス候補データ用分離済データを求め、求めた－35ボックス候補データ用分離済データに対応する－35ボックス候補配列（図22の場合と同様に、j＝1～7の複数の－35ボックス候補配列のうちのいずれか1つ）を－35ボックス相当配列として決定する（図34のステップS1309）。

[0215] すなわち、$q(j)=\Sigma<Y_{-35can}(j), Y_{-35}(k)>$という演算を行う。ここで、j＝1～7であり、k＝1～72であり、Σは、k＝1～72の和を意味し、＜…，…＞は、内積を意味する。そして、$J_{-35}=\arg\{\max\ q(j)\}$（但し、j＝1～7）を求め、求めた$J_{-35}$（j＝1～7のうちのいずれか1つ）を、プロモータ候補配列に含まれる－35ボックス相当配列の開始点（一番下流側の点）とする。なお、$J_{-35}=1$（j＝1）は、－10ボックス相当配列と－35ボックス相当配列との間の隙間（スペーサ35に相当する隙間）が15塩基になるような－35ボックス相当配列の開始点であり、$J_{-35}=7$（j＝7）は、－10ボックス相当配列と－35ボックス相当配列との間の隙間が21塩基になるような－35ボックス相当配列の開始点である（図22と同様）。

[0216] その後、プロモータ候補配列決定処理手段31Gにより、図37に示すように、－35ボックス相当配列決定処理手段31Eにより決定された－35ボックス相当配列および－10ボックス相当配列決定処理手段31Fにより決定された－35ボックス相当配列、並びにこれらに対応する転写開始点に基づき、プロモータ候補配列を決定する（図34のステップS1310）。この際、プロモータ候補配列決定処理手段31Gにより、－35ボックス相当配列および－10ボックス相当配列と、これらに対応する転写開始点を含む転写開始点近傍領域配列との間に相関があるか否かを判定し、相関がない場合には、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点に基づくプロモータ候補配列の決定処理は行わない。すなわち、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点で定まる配列は、プロモータ候補配列として採用されない。より具体的には、以下のようになる。

[0217] 図35において、プロモータ候補配列決定処理手段31Gにより、図37に示すように、－35ボックス相当配列および－10ボックス相当配列、並びにこれらに対応する転

写開始点を含む転写開始点近傍領域配列を結合して相関判定配列を生成する(図35のステップS13101)。

[0218] 次に、プロモータ候補配列決定処理手段31Gにより、図37に示すように、相関判定用出現頻度テーブル記憶手段71(図1参照)に記憶された相関判定用出現頻度テーブルT_{BBS}を用いて、相関判定配列を、配列位置およびA、T、G、Cの記号種別に従って数値化して相関判定データを生成する(図35のステップS13102)。

[0219] 続いて、プロモータ候補配列決定処理手段31Gにより、図37に示すように、相関判定用分離行列記憶手段72(図1参照)に記憶された相関判定用分離行列W_{BBS}と、相関判定データまたは複数の相関判定データを結合した相関判定データ行列$X_{BBS,test}$とを乗じる行列演算、すなわち$Y_{BBS,test} = W_{BBS} X_{BBS,test}$という行列演算を行うことにより、相関判定用分離済データまたは複数の相関判定用分離済データが結合された相関判定用分離済データ行列$Y_{BBS,test}$を求める相関判定用分離処理を行う(図35のステップS13103)。

[0220] それから、プロモータ候補配列決定処理手段31Gにより、相関判定用分離処理で求めた相関判定用分離済データまたは相関判定用分離済データ行列$Y_{BBS,test}$を構成する要素のうち、推定された相関判定用基底行列W_{BBS}^{-1}に含まれる特徴要素(例えば、第1列の各要素)の行列内位置に応じて定まる相関判定要素(例えば、第1行の各要素、ベクトルの場合には、第1番目の要素)の値が、予め定められた相関判定用閾値(例えば、ゼロ)を境界としていずれの側の値であるかを判定することにより、相関があるか否かを判定する。そして、プロモータ候補配列決定処理手段31Gは、相関があると判定した場合にのみ、それらの相関のある－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点に基づき、プロモータ候補配列を決定する(図35のステップS13104)。

[0221] その後、処理中の検査配列の中で、転写開始点を1塩基ずらせるか否かを判断し(図34のステップS1311)、ずらせると判断した場合には、－10ボックス候補データ生成処理手段31Bにより、転写開始点を1塩基ずらして設定した後(図34のステップS1312)、図33のステップS1302の処理に戻り、以降、転写開始点がずらせなくなるまで、図33のステップS1302～図34のS1312の処理を繰り返す。一方、図34のス

テップS1311で、ずらせないと判断した場合には、1つの検査配列の中での複数のプロモータ候補配列の決定処理が全て終了する。

[0222] そして、図32において、アラインメント処理手段31Hにより、プロモータ候補配列決定処理手段31Gにより決定された複数のプロモータ候補配列のそれぞれについて、予め定められたルールに従って、ギャップ「－」の挿入により、長さを一定長（ここでは、配列長：65）に揃えるアラインメント処理を行う（図32のステップS14）。

[0223] このアラインメント処理は、図38に示すように、次のようなルールで行われる。アラインメントされたプロモータ候補配列（配列長：65）を格納するメモリを、上流側より、M（1）～M（65）とすると、先ず、上流側より、M（1）～M（14）に－35ボックス相当配列の上流側（図中の左側）にある14塩基を格納する。この際、もし14塩基ない場合には、足りない分のギャップ「－」をM（1）から順に格納した後に、残りの塩基を格納する。

[0224] 次に、M（15）～M（20）に－35ボックス相当配列である6塩基を格納し、さらに、M（21）～M（25）に－35ボックス相当配列の下流側に続いている5塩基を格納する。この5塩基は、スペーサ35の一部である。

[0225] 続いて、アラインメント後のスペーサ35の配列長を21に揃えたいので、M（26）～M（41）に、（21－アラインメント前のスペーサ35の塩基数）個のギャップ「－」を格納した後、前述した5塩基を除いた残りのスペーサ35を格納する。

[0226] それから、M（42）～M（47）に－10ボックス相当配列である6塩基を格納し、さらに、M（48）～M（50）に－10ボックス相当配列の下流側に続いている3塩基を格納する。この3塩基は、スペーサ10の一部である。

[0227] 続いて、アラインメント後のスペーサ10の配列長を11に揃えたいので、M（51）～M（58）に、（11－アラインメント前のスペーサ10の塩基数）個のギャップ「－」を格納した後、前述した3塩基を除いた残りのスペーサ10を格納する。

[0228] そして、M（59）に転写開始点を格納し、さらに、M（60）～M（65）に転写開始点の下流側に続いている6塩基を格納する。この際、もし6塩基ない場合には、足りない分のギャップ「－」を最後に格納する。

[0229] 以上のようなアラインメントのルールは、前述した非特許文献1に記載されたアライ

ンメントと同じルールであるが、本発明におけるアラインメントは、これに限定されるものではなく、要するに、一定のルールを作り、部分パターンである－35ボックスや－10ボックス等に相当する配列が、同一若しくは略同一の位置に来るようにすればよい。従って、アラインメント後の配列長も65に限定されるものではなく、例えば、70、75、80等でもよい。

[0230] その後、プロモータ候補配列数値化処理手段31Jにより、図38に示すように、プロモータ用出現頻度テーブル記憶手段61（図1参照）に記憶されたプロモータ用出現頻度テーブルT_{prom}（図19参照）を用いて、アラインメント処理手段31Hによりアラインメントされたプロモータ候補配列を、配列位置およびA、T、G、C、ギャップ「－」の記号種別に従って数値化し、検査データを生成する（図32のステップS15）。

[0231] 続いて、分離処理手段32により、前述したプロモータ学習（図16のステップS506）で算出されて図示されないメモリに記憶されている平均値（平均値のゼロ化処理のときに得られた平均値）および$V=D^{-1/2}G^{T}$（ホワイトニング処理のときに得られた行列V）を用いて前処理を行う。すなわち検査データ生成処理手段31により生成された検査データ（縦ベクトル）または複数の検査データを結合した検査データ行列X_{test}を構成する各列ベクトル$X_{test}(j)$について、Vと$X_{test}(j)$とを乗じる行列演算を行い、得られた$VX_{test}(j)$を新たに$X_{test}(j)$とし、これにより、前処理済みの検査データ（縦ベクトル）または検査データ行列X_{test}を生成する。そして、分離処理手段32により、図38に示すように、プロモータ用分離行列記憶手段62（図1参照）に記憶されたプロモータ用分離行列W_{prom}と、前処理済みの検査データ（縦ベクトル）または複数の検査データを結合した検査データ行列X_{test}とを乗じる行列演算、すなわち$Y_{test}=W_{prom}X_{test}$という行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列Y_{test}を求める分離処理を行う（図32のステップS16）。

[0232] この際、分離処理手段32は、1つの検査配列の中で、複数のプロモータ候補配列が決定された場合（図36参照）には、それらの各プロモータ候補配列から得られる各検査データのそれぞれについて上記の分離処理を行い、1つの検査配列について、複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列Y_{test}を求める。

[0233] それから、判定処理手段33は、分離処理手段32により求めた分離済データまたは複数の分離済データが結合された分離済データ行列Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}に含まれる特徴要素（例えば、第1列の各要素）の行列内位置に応じて定まる特徴判定要素（例えば、図38に示すように、第1行の各要素、ベクトルの場合には、第1番目の要素）の値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかを判定することにより、検査配列の中に、複数種類の既知のプロモータのうちのいずれかのプロモータまたは既知のプロモータに類似する新規なプロモータが含まれるか否かを判定する（図32のステップS17）。例えば、分離済データ行列Y_{test}の第1行の各要素、つまり分離済データ行列Y_{test}を構成する各列ベクトル$Y_{test}(j)$の第1番目の要素の値が、正の値（実際には、例えば1近傍の値となる。）の場合には、その列ベクトル$Y_{test}(j)$に対応するプロモータ候補配列は、プロモータであると判定され、一方、負の値（実際には例えば－1近傍の値となる。）の場合には、プロモータでないと判定される。なお、前述したように、独立成分分析（ICA）または主成分分析（PCA）を行って得られる結果として、正負が逆転することがあり得るので、その場合には、逆の判定を行う。

[0234] また、判定処理手段33は、1つの検査配列の中で、複数のプロモータ候補配列が決定された場合（図36参照）には、分離処理手段32により求めた1つの検査配列についての複数の分離済データの各々またはこれらの複数の分離済データが結合された分離済データ行列Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}に含まれる特徴要素（例えば、第1列の各要素）の行列内位置に応じて定まる各特徴判定要素（例えば、第1行の各要素、ベクトルの場合には、第1番目の要素）の値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかをそれぞれ判定し、このうちプロモータであることを示す側の値（例えば、正の値）であると判定された特徴判定要素の値について、閾値との差の絶対値が最も大きい特徴判定要素の値を求め、求めた特徴判定要素の値を与えた検査データに対応するプロモータ候補配列が、複数種類の既知のプロモータのうちのいずれかのプロモータまたは既知のプロモータに類似する新規なプロモータであると認識する処理を行う。例えば、分離済データ行列Y_{test}を構成する各列ベクトル$Y_{test}(j)$の第1番目の要素の値が、－0．9

3、－1．07、1．03、0．96…となった場合には、プロモータであることを示す側の値（正の値）である1．03、0．96のうち、最も大きい1．03という値を与えた検査データに対応するプロモータ候補配列が、プロモータであると認識される。

[0235] 以上により、検査段階（テストステップ）における処理が終了する（図32のステップS18）。なお、検査段階における処理結果として新規なプロモータが発見された場合には、そのプロモータを学習データとし、学習処理手段20により再学習処理を行うことが好ましい。

[0236] このような第1実施形態によれば、次のような効果がある。すなわち、プロモータ認識システム10は、判定対象となるDNA配列を構成する各塩基記号A、T、G、C、およびアラインメントで挿入されたギャップ「－」を、各配列位置毎で、かつ、各記号種別毎に求められてプロモータ用出現頻度テーブルT_{prom}に格納されている出現頻度を用いて数値化するとともに、独立成分分析（ICA）または主成分分析（PCA）を行って得られたプロモータ用分離行列W_{prom}（または次元縮約された$W_{n,prom}$、以下、同様である。）を用いてDNA配列中にプロモータが存在するか否かを判定するので、従来のニューラルネットワーク法や期待値最大化法を用いたコンピュータによるプロモータ解析の場合に比べ、プロモータの認識精度を向上させることができる。また、試験管中の生物実験やX線解析によるプロモータ解析の場合と比べても、時間的・経済的に優位なプロモータ解析を実現することができる。

[0237] また、プロモータに含まれる部分パターンである－35ボックスや－10ボックスのそれぞれについて部分パターン用出現頻度テーブルであるT_{-35}やT_{-10}を作成し、これらを用いて配列の数値化を行うとともに、独立成分分析（ICA）または主成分分析（PCA）を行って得られた部分パターン用分離行列であるW_{-35}（または次元縮約された$W_{n,-35}$、以下、同様である。）やW_{-10}（または次元縮約された$W_{n,-10}$、以下、同様である。）を用いて－35ボックスや－10ボックスに相当する配列を見つけるので、プロモータ内での－35ボックスや－10ボックスの各領域位置に多様性があり、かつ、これに伴ってプロモータの全体の長さに多様性があるという状況下においても、プロモータを精度よく認識することができる。

[0238] さらに、－35ボックスや－10ボックスに相当する配列を見つける際には、図21およ

び図23に示すように、内積の和q(j)を用いるので、より確からしい－35ボックス相当配列や－10ボックス相当配列を見つけることができ、これにより、より高い精度でプロモータを認識することができる。

[0239] そして、検査段階(テストステップ)では、1つの検査配列の中で、プロモータ候補配列を1塩基ずつずらしながら複数本決定し(図36参照)、これらの複数のプロモータ候補配列の中から、プロモータであると認識される配列を1つ定める処理を行うので、検査配列を適切な長さに設定することを前提とし、より高い精度でプロモータの認識を行うことができる。

[0240] また、検査段階(テストステップ)では、プロモータ候補配列を決定するにあたり、－35ボックス相当配列および－10ボックス相当配列と、転写開始点近傍領域配列との相関判定処理を行うので、これらの間に相関がないときには、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点に基づいて定まる配列は、プロモータ候補配列から除外することができる。このため、プロモータの認識精度をより向上させることができる。また、相関がないと判定された場合には、それらの－35ボックス相当配列および－10ボックス相当配列、並びに転写開始点近傍領域配列中の転写開始点に基づいて定まる配列は、プロモータ候補配列として採用されないので、その配列については検査データを生成する必要はなく、また、分離処理手段32による分離処理および判定処理手段33による判定処理を行う必要もないため、この点で、処理時間の短縮を図ることができる。

[0241] さらに、相関判定処理は、独立成分分析(ICA)または主成分分析(PCA)による事前学習で得られた相関判定用分離行列W_{BBS}を用いて行われるので、相関判定の精度向上を図ることができる。

[0242] ＜効果確認実験＞

また、本発明の効果を確かめるために、以下のような実験を行った。実験方法や実験条件は、本第1実施形態に基づくものとした。

[0243] 先ず、学習段階(トレーニングステップ)では、本第1実施形態で説明したように、前述した非特許文献1に記載された既知の106本の大腸菌プロモータを使って学習を行った。図39に、学習結果として得られた65行×65列のプロモータ用分離行列W_{pr}

(図26参照)の逆行列である65行×65列のW_{prom}^{-1}の第1列の各要素の値を示す。横軸は、行番号に対応している(但し、1～65ではなく、0～64としている)。なお、縦軸は、各値の大小関係が明らかになればよいため、特に示していない。また、図40に、学習結果として得られた65行×212列のプロモータ用分離済データ行列Y_{prom}(図26参照)の第1行の各要素の値を示す。横軸は、列番号(但し、1～212ではなく、0～211としている。)に対応している。

[0244] 図39によれば、プロモータ用分離行列W_{prom}の逆行列W_{prom}^{-1}、すなわちプロモータ用基底行列A_{prom}を推定した行列の第1列には、独立成分分析(ICA)を用いて得られるプロモータに特有な基底が現れていることがわかる。このようなプロモータの特徴を最も良く示す基底は、W_{prom}^{-1}内の第1列に現れ、他の列には、これ程顕著に特徴を示す基底は現れなかった。従って、W_{prom}^{-1}の第1列の各要素が、本発明における特徴パターンの特徴を示す特徴要素に該当することが確かめられた。このことは、以下のように考えることができる。

[0245] 一般に、未知の独立な情報源からのデータ(縦ベクトル)を集合させた行列をSとし、これらが重ね合わされた状態で観測された混合データ(縦ベクトル)を集合させた行列をXとし、入力Sに対してXを出力する際のフィルタになる混合行列、すなわち基底ベクトルを集合させた基底行列をAとすると、この関係は、X＝ASと表現することができる。これに対し、観測された混合データ(縦ベクトル)を集合させた行列Xが与えられたとき、それらを混合前の元のデータに復元するための分離行列をWとし、分離して得られる分離済データ(縦ベクトル)を集合させた行列をYとすると、Y＝WXと表現することができる。ここで、独立成分分析(ICA)は、Yの各成分(Yを構成する各列ベクトル)が独立になるような分離行列Wを求めていく処理であり、分離行列Wは、基底行列Aの逆行列を推定した行列であり、分離行列Wの逆行列W^{-1}は、基底行列Aを推定した行列に相当する。従って、複数種類の既知のプロモータを、本来的にプロモータが持つ特徴にノイズが乗った配列であると考え、複数種類の既知のプロモータから生成されたプロモータ用学習データ行列を、観測された混合データ(縦ベクトル)としてX_{prom}とおき、独立成分分析(ICA)を行うことで、プロモータ用基底行列A_{prom}を推定した行列として得られるW_{prom}^{-1}には、プロモータに特有な基底が現れることになり

、それが図39に示すW_{prom}^{-1}の第1列である。そして、$Y_{prom}=W_{prom}X_{prom}$（図26参照）であるから、$W_{prom}^{-1}Y_{prom}=X_{prom}$となり、行列演算上、$W_{prom}^{-1}$の第1列の各要素と乗じられるのは、$Y_{prom}$の第1行の各要素であり、$Y_{prom}$の第1行の各要素の値が大きければ、$W_{prom}^{-1}$の第1列への寄与度が大きくなることになる。従って、Y_{prom}の第1行目に現れた各要素の値が、本発明における特徴判定要素の値であることがわかる。

[0246] 図40によれば、Y_{prom}の第1行の各要素は、プロモータ用学習データ行列X_{prom}のうちの真データの部分B_{prom}^{T}（図25参照）に対応する部分の値が、1の近傍となっており、X_{prom}のうちの偽データの部分C_{prom}^{T}（図25参照）に対応する部分の値が、－1の近傍となっているので、上記のことを裏付ける結果となっている。従って、検査段階（テストステップ）では、Y_{test}（図38参照）の第1行目に現れた各要素の値が、本発明における特徴判定要素の値であり、この値の正負を見れば、プロモータであるか否かを判定することができることが確認できた。なお、図40の縦軸の各値は、＋1近傍か－1近傍の値となっているが、これらの近傍の値に限定されるものではなく、要するに、Y_{prom}の第1行の各要素の値に、プロモータに対応する部分と非プロモータに対応する部分とで差異が現れた場合に、その差異を利用して閾値を定め、Y_{test}（図38参照）の第1行の各要素の値で、プロモータであるか否かを判定することができればよい。従って、閾値は、必ずしもゼロである必要はない。また、判定結果としてプロモータの存否の程度を出力する軟判定を行う場合には、縦軸の値を何区画かに区切り（例えば、1，0．6，0．2，－0．2，－0．6，－1で区切り）、例えば、1以上は、「プロモータである可能性が極めて高い」、0．6～1は、「プロモータである可能性が高い」、0．2～0．6は、「プロモータである可能性がある」、－0．2～0．2は、「どちらであるか不明」等のように、それぞれの区画で異なる判定結果を出力するようにすることができる。

[0247] 次に、検査段階（テストステップ）では、テストデータとして、前述した非特許文献1に記載された既知の126本の大腸菌プロモータを使った。これらの126本の大腸菌プロモータは、学習段階で使った既知の106本の大腸菌プロモータとは異なるものである。なお、これらの126本のプロモータについては、非特許文献1では、既にギャップが挿入されていたため、それらのギャップを取り除き、ギャップのない配列長40～50のプロモータに戻してテストデータ（検査配列）とした。

[0248] また、テストデータ(検査配列)として、配列長が50のランダム配列を1000本生成して用意した。この際、AとTの出願頻度をそれぞれ0．3ずつとし、GとCの出願頻度をそれぞれ0．2ずつとした。従って、アラインメントされていないテストデータ(検査配列)が、合計で1126本用意された。

[0249] 続いて、1126本のテストデータについて、W_{-10}およびW_{-35}を用いて－10ボックス相当配列および－35ボックス相当配列を決定した後、図38に示した方法でアラインメントを施し、配列長を65に統一した。この際、転写開始点は、下流側から見て最初のAまたはGとし、1つのテストデータ(検査配列)につき、転写開始点を1つだけ設定した(つまり、この効果確認実験では、図36に示すような転写開始点のスライド設定は行わなかった)。また、－10ボックス相当配列および－35ボックス相当配列と、転写開始点近傍領域配列との相関判定処理は行わなかった。

[0250] それから、アラインメントされた1126本のプロモータ候補配列について、プロモータ用出現頻度テーブルT_{prom}を用いて数値化し、1126本の検査データを生成した後、図38に示すように、プロモータ用分離行列W_{prom}と、1126本の検査データを結合した検査データ行列X_{test}とを乗じる行列演算を行い、65行×1126列のプロモータ用分離済データ行列Y_{test}を求めた。

[0251] そして、プロモータ用分離済データ行列Y_{test}の第1行の要素(特徴判定要素)の値について、閾値をゼロとし、正負を判定した(つまり、硬判定を行った)。この結果、1126本のテストデータのうち、既知の126本のプロモータに対応する部分の特徴判定要素の値は、ほぼ全てが正の値となり、1000本のランダムデータ(非プロモータ)に対応する部分の特徴判定要素の値は、殆どが負の値となり、93．7%という高い認識率が得られ、これにより本発明の効果が顕著に示された。なお、この93．7%という認識率は、転写開始点のスライド設定処理や相関判定処理を行うことにより、更に向上させることが可能である。

[0252] [第2実施形態]

図42は、本発明の第2実施形態のアラインメントの方法を示す説明図である。図43は、第2実施形態のプロモータ用出現頻度テーブルT_{prom}を示す図である。図44には、第2実施形態における基底行例での特徴要素の現れ方と、検査段階の判定処理

で用いられるプロモータ用分離済データ行列Y_{prom}のうちの真データ対応部分の特徴判定要素との関係が示されている。図45および図46には、第2実施形態の学習段階の処理で得られたプロモータ用分離行列W_{prom}の逆行列を求めて推定された基底行列W_{prom}^{-1}の第1列の各要素の値と、W_{prom}とともに得られたプロモータ用分離済データ行列Y_{prom}の第1行の各要素の値とが、それぞれ採用例として示されている。また、図47および図48には、W_{prom}^{-1}の第27列の各要素の値と、Y_{prom}の第27行の各要素の値とが、それぞれ不採用の例として示されている。さらに、図49は、第2実施形態の検査段階の処理の説明図である。

[0253] 本第2実施形態では、アラインメント処理を行う際に、ギャップを挿入するのではなく、スペーサ35およびスペーサ10を構成する配列の一部を取り除くという点と、判定処理を行う際に、1つ(1行)の特徴判定要素だけを用いるのではなく、複数(複数行)の特徴判定要素を用いるという点で、前記第1実施形態の場合と異なるが、その他の処理やシステム構成は、前記第1実施形態の場合と略同様であるため、同一部分についての詳しい説明は省略し、以下では異なる部分を中心に説明を行う。

[0254] <学習段階(トレーニングステップ)>

学習段階(トレーニングステップ)の全体の流れは、図3に示した前記第1実施形態の場合と同様であり、学習用の既知のプロモータNp本(前記第1実施形態の場合と同様に、例えば106本の大腸菌プロモータとする。)を用意し、−35ボックス学習、−10ボックス学習、プロモータ学習、相関学習を行う。

[0255] (−35ボックス学習および−10ボックス学習)

前記第1実施形態の場合と同様である。なお、前記第1実施形態では、図4のステップS304および図11のステップS404のランダムボックスの生成処理で、A、T、G、Cの各出現頻度に偏りを持たせることなく、各出現頻度が0.25ずつになるように生成するという説明が記載されていたが、これに限定されるものではなく、例えば、AおよびTの各出現頻度をそれぞれ0.3とし、GおよびCの各出現頻度をそれぞれ0.2としてもよい。

[0256] (プロモータ学習)

前記第1実施形態の場合とは異なり、ギャップの挿入によるアラインメント処理は行

わず、従って、プロモータ用出現頻度テーブルT_{prom}には、ギャップの出現頻度は含まれない。また、スペーサ35およびスペーサ10を構成する配列の一部を取り除くことによるアラインメント処理を行うので、プロモータ用分離行列W_{prom}やプロモータ用分離済データ行列Y_{prom}の次元数が、前記第1実施形態よりも小さい。さらに、前記第1実施形態では、検査段階において、学習段階で得られたプロモータ用分離済データ行列Y_{prom}を用いることなく、検査段階で得られたY_{test}の1つ(1行)の特徴判定要素のみを用いて判定処理を行っていたが、本第2実施形態では、検査段階で得られたY_{test}の複数(複数行)の特徴判定要素と、学習段階で得られたY_{prom}とを用いて判定処理を行う点が異なる。より詳細には、以下の通りである。

[0257] 前記第1実施形態のプロモータ用アラインメント処理手段23Aは、複数種類の既知のプロモータの長さをギャップの挿入により一定長(例えば、配列長:65)に揃えるアラインメント処理を行うものであったが、本第2実施形態のプロモータ用アラインメント処理手段は、図42に示すようなルールでスペーサ35およびスペーサ10を構成する配列の一部を取り除くことにより、複数種類の既知のプロモータの長さを一定長(例えば、配列長:31)に揃えるアラインメント処理を行う。

[0258] 図42において、既知のプロモータにおいては、−35ボックスおよび−10ボックスの位置は既知であるため、スペーサ35を構成する塩基配列のうち、−35ボックス寄りの6塩基(−35ボックスの下流側に繋がる6塩基)および−10ボックス寄りの9塩基(−10ボックスの上流側に繋がる9塩基)を採用し、その間の配列を取り除く。従って、106本の既知の大腸菌プロモータのスペーサ35の配列長は、15～21塩基であるから、スペーサ35のうち除去される配列長は、0～6塩基となる。すなわち、スペーサ35において、ギャップの挿入を不要とし、かつ、なるべく除去する分を少なくするために、最短のスペーサ35の配列長である15塩基で配列長を揃えるようにした。なお、上記では、−35ボックス寄りの6塩基および−10ボックス寄りの9塩基を採用したが、これに限定されず、例えば、−35ボックス寄りの7塩基および−10ボックス寄りの8塩基を採用してもよく、要するに、採用される合計の配列長が15塩基になればよい。また、ギャップの挿入を不要とすることだけを目的とするならば、14塩基以下の配列長で揃えるようにしてもよい。

[0259] また、スペーサ10を構成する塩基配列のうち、－10ボックス寄りの2塩基（－10ボックスの下流側に繋がる2塩基）および転写開始点寄りの1塩基（転写開始点の上流側に繋がる1塩基）を採用し、その間の配列を取り除く。従って、106本の既知の大腸菌プロモータのスペーサ10の配列長は、3～11塩基であるから、スペーサ10のうち除去される配列長は、0～8塩基となる。すなわち、スペーサ10において、ギャップの挿入を不要とし、かつ、なるべく除去する分を少なくするために、最短のスペーサ10の配列長である3塩基で配列長を揃えるようにした。なお、上記では、－10ボックス寄りの2塩基および転写開始点寄りの1塩基を採用したが、これに限定されず、例えば、－10ボックス寄りの1塩基および転写開始点寄りの2塩基を採用してもよく、要するに、採用される合計の配列長が3塩基になればよい。また、ギャップの挿入を不要とすることだけを目的とするならば、2塩基以下の配列長で揃えるようにしてもよい。

[0260] 前記第1実施形態のプロモータ用出現頻度テーブル作成処理手段23Bは、プロモータ用アラインメント処理手段23Aにより一定長（例えば、配列長：65）に揃えられた複数種類の既知のプロモータにギャップが含まれているため、A、T、G、Cに、ギャップを加え、各記号毎の出現頻度（frequency）を求めていたが、本第2実施形態のプロモータ用出現頻度テーブル作成処理手段は、一定長（例えば、配列長：31）に揃えられた複数種類の既知のプロモータにギャップは含まれていないので、ギャップ無しで、A、T、G、Cについて、各記号毎の出現頻度を求める。従って、前記第1実施形態では、プロモータ用出現頻度テーブル記憶手段61（図1参照）に記憶させるプロモータ用出現頻度テーブルT_{prom}は、ギャップの出現頻度を含んでいたが（図19参照）、本第2実施形態では、プロモータ用出現頻度テーブル記憶手段に記憶させるプロモータ用出現頻度テーブルT_{prom}は、図43に示すように、ギャップの出現頻度を含んでいないので、記憶させるデータ量が減少している。また、配列長も65から31に減少しているので、その点でも記憶させるデータ量が減少している。

[0261] 前記第1実施形態の非プロモータ生成処理手段23Cは、複数本の非プロモータ生成用配列をランダムに自動生成し、これらの非プロモータ生成用配列の中で、部分パターン用分離行列（－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35}）と、部分パターン用分離済データ行列（図21の－10ボックス用分離済データ

行列Y_{-10}、および図23の－35ボックス用分離済データ行列Y_{-35}）の真データ対応部分の各列ベクトルとを用いて、非特徴パターンである非プロモータに含まれる部分パターン相当配列（－10ボックス相当配列および－35ボックス相当配列）を決定し（図17のステップS50305，S50306および図18のステップS50309，S50310参照）、さらに決定された各領域の部分パターン相当配列（－10ボックス相当配列および－35ボックス相当配列）に基づき非プロモータを決定していたが（図18のステップS50311参照）、本第2実施形態の非プロモータ生成処理手段は、部分パターン用分離行列（－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35}）や部分パターン用分離済データ行列（－10ボックス用分離済データ行列Y_{-10}および－35ボックス用分離済データ行列Y_{-35}）を用いた部分パターン相当配列（－10ボックス相当配列および－35ボックス相当配列）の決定処理を行わずに、非プロモータを生成する。

[0262] 先ず、本第2実施形態では、AおよびTの出現頻度をそれぞれ0．3とし、GおよびCの出現頻度をそれぞれ0．2として65塩基長のランダム配列を生成した後、ホモロジー・スコア（homology score）を計算し、求めたホモロジー・スコアの値の大小に基づき、プロモータに近い配列（ホモロジー・スコアが、適宜、定めた閾値を超える配列）を排除することにより、例えば、既知のプロモータの本数（106本）と同数本のランダム配列（ホモロジー・スコアが閾値以下の配列）を用意する。なお、ホモロジー・スコアの詳細については、前述した非特許文献4に記載されている。

[0263] ここで、ホモロジー・スコアは、拡大－35ボックス（extended −35 region）および拡大－10ボックス（extended −10 region）を定義し、次のような手順1～3で求めることができる。拡大－35ボックスは、－35ボックスの上流側に繋がる9塩基と、－35ボックスの下流側に繋がる1塩基とを、－35ボックス（6塩基）に加えた合計16塩基長の配列である。また、拡大－10ボックスは、－10ボックスの上流側に繋がる5塩基と、－10ボックスの下流側に繋がる3塩基とを、－10ボックス（6塩基）に加えた合計14塩基長の配列である。

[0264] 手順1では、前記第1実施形態と同様にしてギャップの挿入により65塩基長に揃えた例えば106本の既知のプロモータについて、拡大－35ボックスおよび拡大－10ボックスの各配列位置毎にA，T，G，Cの各出現頻度を求め、それぞれの出現頻度を

標準偏差(例えば、$(101/4)^{1/2}$となる。)で割ったスコアを求める。この際、標準偏差のルート内の分子の「101」は、拡大－35ボックスおよび拡大－10ボックスの中にギャップのある5本のプロモータを除いたため、「106」が「101」になったものであり、ルート内の分母の「4」は、シンボルの種別が、A, T, G, Cの4種別であることを意味している。

[0265] 手順2では、拡大－35ボックスおよび拡大－10ボックスの中にギャップのある5本のプロモータを除いた例えば101本の既知のプロモータのスペーサ35の配列長について、15～21塩基の各配列長毎に出現頻度を求め(つまり、スペーサ35の配列長が15塩基となる頻度、16塩基となる頻度、…、21塩基となる頻度を求める。)、それぞれの出現頻度を標準偏差(例えば、$(101/7)^{1/2}$となる。)で割ったスコアを求める。この際、標準偏差のルート内の分母の「7」は、配列長の種別が、15～21塩基の7種別であることを意味している。

[0266] 手順3では、既知のプロモータの本数(106本)と同数本用意されたランダム配列のそれぞれについて、ホモロジー・スコア＝100×{(ベース・ペア・スコアの総和＋スペーシング・スコア－ベースライン・スコア)／(マキシマム・スコア－ベースライン・スコア)}という式を用いて、ホモロジー・スコアを求める。

[0267] ここで、ベース・ペア・スコアの総和(sum of base pair scores)とは、ランダム配列における拡大－35ボックスおよび拡大－10ボックスの各塩基を、手順1で求めたスコア(A, T, G, Cの各出現頻度を標準偏差で割ったスコア)により数値化し、これらの値の和をとったものである。

[0268] スペーシング・スコア(spacing score)とは、ランダム配列におけるスペーサ35の配列長を、手順2で求めたスコア(15～21塩基の各配列長の出現頻度を標準偏差で割ったスコア)により数値化したものである。

[0269] マキシマム・スコア(maximum score)とは、拡大－35ボックスの各配列位置のスコアのうち最もよいスコア(最大スコア)と、拡大－10ボックスの各配列位置のスコアのうち最もよいスコア(最大スコア)と、スペーサ35の配列長のスコアとの和をとったものである。

[0270] ベースライン・スコア(baseline score)とは、拡大－35ボックスの各配列位置のスコア

の平均値と、拡大－10ボックスの各配列位置のスコアの平均値と、スペーサ35の配列長のスコアとの和をとったものである。

[0271] そして、手順3では、65塩基長のランダム配列における－35ボックス、－10ボックス、および転写開始点に相当する部分の各位置は、下流側の端から6塩基置いて、7塩基目を転写開始点と仮定し、ここから上流側に向かって、11塩基のスペーサ10、6塩基の－10ボックス、21塩基のスペーサ35、6塩基の－35ボックスをとり、上流側の端から14塩基を残すことにより決定する。つまり、前記第1実施形態でギャップの挿入により一定長(配列長:65)に揃えられた非プロモータ(図24参照)における－35ボックス、－10ボックス、および転写開始点に相当する部分の各位置と同じ位置とする。但し、前記第1実施形態の場合と同様に、部分パターン用分離行列(－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35})や部分パターン用分離済データ行列(－10ボックス用分離済データ行列Y_{-10}および－35ボックス用分離済データ行列Y_{-35})を用いて位置を決定してもよい。なお、ホモロジー・スコアの計算は、65塩基長に限定されるものではなく、従って、用意するランダム配列の長さも65塩基長に限定されるものではない。また、ホモロジー・スコアの計算によりプロモータに近い配列を排除する処理は省略してもよい。

[0272] それから、以上のようにホモロジー・スコアの計算をしてプロモータに近い配列を排除して得られた例えば106本の65塩基長のランダム配列を、本第2実施形態の非プロモータ用アラインメント処理手段により、図42に示したルールで配列長を31塩基長に揃える。従って、本第2実施形態では、前記第1実施形態の非プロモータ用アラインメント処理手段23Dのようなギャップの挿入によるアラインメント処理は行わない。

[0273] 続いて、前記第1実施形態では、プロモータ用数値化処理手段23Eにより、プロモータ用出現頻度テーブルT_{prom}を用いて、一定長(配列長:65)に揃えられた例えば106本のプロモータおよび106本の非プロモータを、それぞれ数値化する処理を行ったが(図25参照)、本第2実施形態のプロモータ用数値化処理手段も同様な処理を行う。但し、使用するプロモータ用出現頻度テーブルT_{prom}は、図19に示された前記第1実施形態のものとは異なり、図43に示すようなギャップの出現頻度がなく、配列位置も1～31とされたものである。

[0274] その後、前記第1実施形態では、プロモータ用数値化処理手段23Eにより数値化して得られたB_{prom}およびC_{prom}（図25参照）を、プロモータ用学習データ行列生成処理手段23Fにより、それぞれ転置してB_{prom}^{T}およびC_{prom}^{T}とし、これらを結合してプロモータ用学習データ行列X_{prom}（図25参照）を生成する処理を行ったが、本第2実施形態のプロモータ用学習データ行列生成処理手段も同様な処理を行う。但し、前記第1実施形態では、図25に示すように、B_{prom}^{T}およびC_{prom}^{T}が、それぞれ65行×106列であり、X_{prom}が、65行×212列であったのに対し、本第2実施形態では、B_{prom}^{T}およびC_{prom}^{T}が、それぞれ31行×106列であり、X_{prom}が、31行×212列となる（図44参照）。

[0275] そして、前記第1実施形態では、プロモータ用分析処理手段23Gにより、プロモータ用学習データ行列X_{prom}（図25参照）について、前処理として、平均値のゼロ化処理およびホワイトニング処理を行った後（図9と同様）、前処理を経たプロモータ用学習データ行列X_{prom}を用いて、独立成分分析（ICA）を行うことにより、プロモータ用分離行列W_{prom}およびプロモータ用分離済データ行列Y_{prom}を求める処理を行ったが（図26参照）、本第2実施形態のプロモータ用分析処理手段も同様な処理を行う。但し、前記第1実施形態では、図26に示すように、W_{prom}が、65行×65列であり、Y_{prom}が、65行×212列であったが、本第2実施形態では、図44に示すように、W_{prom}が、31行×31列であり、Y_{prom}が、31行×212列となる。

[0276] なお、前記第1実施形態では、Y_{prom}は、後に検査段階の処理で使用されるデータではないものの、検査段階の処理で得られる分離済データ行列Y_{test}（図38参照）の特徴判定要素（第1行の各要素）の値が、正負のいずれの場合にプロモータであると判定されるのかを確認するために求められてプロモータ用分離済データ行列記憶手段63（図26参照）に記憶されていたが、本第2実施形態では、Y_{prom}は、後に検査段階の判定処理で使用されるので、プロモータ用分離済データ行列記憶手段へ記憶させておくことが必要となる。但し、検査段階の判定処理で使用されるのは、後述するように、Y_{prom}のうち、真データ対応部分（B_{prom}^{T}に対応する部分）の特徴判定要素であるから、記憶させておくのは、Y_{prom}のうち真データ対応部分の特徴判定要素の値だけでもよい。また、Y_{prom}のうち真データ対応部分の特徴判定要素の値により構成される各列ベクトルの重心ベクトルを用いて判定処理を行う場合には、予め重心ベクト

ルを求めておき、それを記憶しておくだけでもよい。

[0277] (基底行列における特徴要素の決定)

前記第1実施形態では、基底行列のうちの1列にプロモータの特徴を示す基底が現れ、この1列のみが特徴要素となる場合(例えば、図39に示すように、W_{prom}^{-1}の第1列が特徴要素となる場合)、すなわちプロモータの特徴を最も顕著に示す唯一の基底を選択し、それを特徴要素とする場合について説明が行われていたが、本第2実施形態では、図44に示すように、基底行列のうちの複数列にプロモータの特徴を示す基底が現れ、これらの複数列が特徴要素となる場合について説明を行う。このような特徴要素となる列(基底)を決定するには、プロモータ用分離行列W_{prom}の逆行列W_{prom}^{-1}を求めて基底行列を推定し、この推定された基底行列であるW_{prom}^{-1}の各列(各基底)の要素の値をグラフ化してみることにより、目視による判断で、プロモータの特徴を示す列(基底)を選択して決定することができる。また、特徴要素の決定作業は、このようなW_{prom}^{-1}の各列(各基底)のグラフ化による目視判断で行ってもよいが、人為的な判断を少なくするという観点から、プロモータ用分離済データ行列Y_{prom}の各行の要素の値を、次のような方法で解析することにより行ってもよい。

[0278] W_{prom}^{-1}のいずれかの列(基底)に、プロモータの特徴を示す列(基底)が現れるとすれば、これに対応するプロモータ用分離済データ行列Y_{prom}の行の要素の値は、プロモータを結合したB_{prom}^{T}に対応する部分(真データ対応部分)と、非プロモータを結合したC_{prom}^{T}に対応する部分(偽データ対応部分)とで、相違が生じるはずである(図40参照)。従って、このことを利用し、プロモータ用分離済データ行列Y_{prom}の各行の要素の値を、真データ対応部分と偽データ対応部分とに分けて整理し、解析を行う。

[0279] 上記の解析にあたっては、X_{prom}を構成するB_{prom}^{T}におけるプロモータの並び順、およびX_{prom}を構成するC_{prom}^{T}における非プロモータの並び順は、特に意味があるものではなく、従って、プロモータ用分離済データ行列Y_{prom}の真データ対応部分(B_{prom}^{T}に対応する部分)における各列の並び順、およびY_{prom}の偽データ対応部分(C_{prom}^{T}に対応する部分)における各列の並び順も、特に意味があるものではないので、Y_{prom}の各行について、それぞれの行を構成する各要素の値を大小順に並べ替えてグラフ化する。ここでは、Y_{prom}の真データ対応部分では、左側から右側へ向かって値の大きい順

に並べ、Y_{prom}の偽データ対応部分では、左側から右側へ向かって値の小さい順に並べるものとする。また、Y_{prom}の各行のグラフ化にあたっては、W_{prom}^{-1}の各列(各基底)のグラフのうち、ピーク値(正負を問わず、最も大きな値)が正の値となる列のグラフの場合には、対応するY_{prom}の行を構成する各要素の値をそのまま用いてグラフ化し、一方、ピーク値が負の値となる列のグラフの場合には、対応するY_{prom}の行を構成する各要素の値に－1を乗じた値をグラフ化する。

[0280] 例えば、図45に示すように、W_{prom}^{-1}の第1列のグラフでは、－10ボックスに反応していると思われる山形部分があるが、これを目視で確認し、－10ボックスに反応している部分があるので、第1列は特徴要素であると人為的に決定してもよい。一方、図46に示すように、W_{prom}^{-1}の第1列に対応するY_{prom}の第1行のグラフでは、真データ対応部分(106列)で正の値が多く、偽データ対応部分(106例)で負の値が多く、真データ対応部分と偽データ対応部分とで、はっきりとした符号の相違が出ているので、この状態を、閾値を設定して自動的に判定することにより、W_{prom}^{-1}の第1列が特徴要素である、すなわちY_{prom}の第1行が特徴判定要素であると決定することができる。

[0281] また、図47に示すように、W_{prom}^{-1}の第27列のグラフでは、－35ボックス、－10ボックス、転写開始点のいずれかに反応していると思われる山形部分を目視で確認することができないので、第27列は特徴要素ではないと人為的に決定してもよい。一方、図48に示すように、W_{prom}^{-1}の第27列に対応するY_{prom}の第27行のグラフでは、真データ対応部分(106列)と偽データ対応部分(106例)とで、はっきりとした符号の相違が出ていないので、この状態を、閾値を設定して自動的に判定することにより、W_{prom}^{-1}の第27列が特徴要素ではない、すなわちY_{prom}の第27行が特徴判定要素ではないと決定することができる。W_{prom}^{-1}の他の列やY_{prom}の他の行の場合も同様である。

[0282] W_{prom}^{-1}の各列が特徴要素であるか否か、すなわちY_{prom}の各行が特徴判定要素であるか否かを決定するための閾値は、Y_{prom}の各行のグラフにおける零交差(zero-crossing)を調べ、次のように設定することができる。

[0283] 真データ対応部分(Np＝106列)において、各要素の値が正の値から負の値に転じる位置をZ_{down}とし、偽データ対応部分(Np＝106例)において、各要素の値が負の値から正の値に転じる位置をZ_{up}とする。このとき、$Z_{down} > (1 + \theta_{down}) \times Np / 2$であり

、かつ、$Z_{up}>\{(1+\theta_{up})\times Np/2\}+Np$であれば、$Y_{prom}$の当該行は、特徴判定要素として採用され、$W_{prom}^{-1}$の対応する列(基底)は、特徴要素であると決定される。この場合は、真データ対応部分および偽データ対応部分の双方を用いて判定が行われたことになる。プロモータの特徴を示す基底として採用されるには、Z_{down}が真データ対応部分の中央位置(Np/2)よりも、なるべく大きい方がよく、Z_{up}が偽データ対応部分の中央位置(Np/2+Np)よりも、なるべく大きい方がよいので、中央位置からの乖離度合いを定めるθ_{down}やθ_{up}を大きい値に設定すれば、より厳しい採用条件となり、良好な基底のみが採用されることになる。例えば、$\theta_{down}=\theta_{up}=0.1$と設定すれば、真データ対応部分の中央位置(Np/2=53)に対し、$(1+\theta_{down})\times Np/2=58$なので、$Z_{down}>58$という採用条件となり、偽データ対応部分の中央位置(Np/2+Np=159)に対し、$\{(1+\theta_{up})\times Np/2\}+Np=164$なので、$Z_{up}>164$という採用条件となる。

[0284] また、上記のように真データ対応部分および偽データ対応部分で設定された条件の双方を満たすことができない場合であっても、いずれか一方の条件を満たす場合には、その基底を採用するようにしてもよい。この場合には、いずれか一方の条件で判定するため、上記のように双方の条件で判定する場合に比べ、採用条件を厳しくすればよい。例えば、上記のような、$Z_{down}>(1+\theta_{down})\times Np/2$であり、かつ、$Z_{up}>\{(1+\theta_{up})\times Np/2\}+Np$であるという条件を満たさなくても、$Z_{down}>(1+\theta_{down}^{*})\times Np/2$であるか、または、$Z_{up}>\{(1+\theta_{up}^{*})\times Np/2\}+Np$であれば、$Y_{prom}$の当該行を、特徴判定要素として採用し、$W_{prom}^{-1}$の対応する列(基底)を、特徴要素であると決定してもよい。ここで、θ_{down}^{*}やθ_{up}^{*}の値は、θ_{down}やθ_{up}の値よりも大きくし、例えば、$\theta_{down}^{*}=\theta_{up}^{*}=0.15$と設定すれば、真データ対応部分の中央位置(Np/2=53)に対し、$(1+\theta_{down}^{*})\times Np/2=61$なので、$Z_{down}>61$という厳しい採用条件となり、偽データ対応部分の中央位置(Np/2+Np=159)に対し、$\{(1+\theta_{up}^{*})\times Np/2\}+Np=167$なので、$Z_{up}>167$という厳しい採用条件となる。

[0285] このようにしてθ_{down}やθ_{up}の値、あるいはθ_{down}^{*}やθ_{up}^{*}の値を設定しておき、Y_{prom}の各行が特徴判定要素であるか否か、すなわちW_{prom}^{-1}の各列(各基底)が特徴要素であるか否を決定した後、図44に示すように、Y_{prom}のうち真データ対応部分の特徴判

定要素だけを抽出し、これをY_{prom}'として、プロモータ用分離済データ行列記憶手段に記憶しておく。例えば、Y_{prom}の31行のうち、合計23行が特徴判定要素であると決定された場合には、Y_{prom}'は、23行×Np列(106列)の行列となる。このY_{prom}'は、後に行われる検査段階の判定処理で使用される。なお、Y_{prom}'を構成する各列ベクトルY_{prom}'(j)(j=1～106)の重心ベクトルを求め、この重心ベクトルを記憶しておいてもよい。

[0286] (相関学習)

本第2実施形態では、前記第1実施形態の場合と同様な相関学習を行うことができる。

[0287] <検査段階(テストステップ)>

前記第1実施形態の検査段階の処理では、検査データ生成処理手段31の各処理手段31A～31Gにより、部分パターン用分離行列(－10ボックス用分離行列W_{-10}および－35ボックス用分離行列W_{-35})や部分パターン用分離済データ行列(－10ボックス用分離済データ行列Y_{-10}および－35ボックス用分離済データ行列Y_{-35})を用いて部分パターン相当配列(－10ボックス相当配列および－35ボックス相当配列)を決定し、プロモータ候補配列の決定処理を行っていたが、本第2実施形態の検査段階でも同様な処理を行う。

[0288] なお、前記第1実施形態では、プロモータ候補配列の決定処理において、－10ボックス相当配列および－35ボックス相当配列と、転写開始点近傍領域配列との間に相関があるか否かを判定し、相関がないと判定された場合には、プロモータ候補配列として採用しない処理を行い(図34のステップS1310参照)、この相関判定処理の際には、1行のみの相関判定要素を用いて相関判定を行っていたが(図35のステップS13104、図37参照)、本第2実施形態の相関判定処理も同様の処理としてもよく、あるいは複数行の相関判定要素を用いて相関判定を行ってもよい。複数行の相関判定要素を用いて相関判定を行う場合には、例えば、後述するプロモータであるか否かの判定処理の場合と同様に、複数行の相関判定要素の値により構成される列ベクトル(相関判定データが1つの場合には、1つの縦ベクトルである。)と、学習段階で相関判定用分離行列W_{BBS}(図31参照)とともに得られた相関判定用分離済データ行

列Y_{BBS}の真データ対応部分(図30のB_{BBS}^{T}に対応する部分)の相関判定要素の値により構成される各列ベクトルとの類似度を示す値(例えば、各列ベクトルの総体との類似度を示す値、より具体的には、各列ベクトルの重心ベクトルとの内積の値またはこれと等価な値)を算出し、算出された類似度を示す値が、予め定められた閾値を境界としていずれの側の値であるか、または値の大きさの程度を判定することにより、相関があるか否か、あるいは相関の程度を判定することができる。

[0289] その後、前記第1実施形態の検査段階の処理では、検査データ生成処理手段31のアラインメント処理手段31Hにより、検査配列の中で決定されたプロモータ候補配列を、ギャップを挿入して一定長(配列長:65)に揃えるアラインメント処理を行っていたのに対し、本第2実施形態の検査段階の処理では、図42に示したルールでスペーサ35の一部の配列およびスペーサ10の一部の配列を取り除くことにより一定長(配列長:31)に揃えるアラインメント処理を行う点が異なる。

[0290] そして、前記第1実施形態では、プロモータ候補配列数値化処理手段31Jにより、プロモータ用出現頻度テーブルT_{prom}を用いて、アラインメントされたプロモータ候補配列を数値化する処理を行ったが(図38参照)、本第2実施形態のプロモータ候補配列数値化処理手段も同様な処理を行う(図49参照)。但し、使用するプロモータ用出現頻度テーブルT_{prom}は、図19に示された前記第1実施形態のものとは異なり、図43に示すようなギャップの出現頻度がなく、配列位置も1～31とされたものである。

[0291] それから、前記第1実施形態では、分離処理手段32により、検査データ生成処理手段31により生成された検査データ(縦ベクトル)または複数の検査データを結合した検査データ行列X_{test}について前処理を行った後、プロモータ用分離行列W_{prom}と、前処理済みの検査データ(縦ベクトル)または複数の検査データを結合した検査データ行列X_{test}とを乗じる行列演算を行うことにより、分離済データまたは複数の分離済データが結合された分離済データ行列Y_{test}を求める分離処理を行ったが(図38参照)、本第2実施形態の分離処理手段も同様な処理を行う(図49参照)。但し、使用するプロモータ用分離行列W_{prom}は、前記第1実施形態の65行×65列のものではなく、本第2実施形態の31行×31列のものである。

[0292] その後、前記第1実施形態の判定処理手段33による判定処理では、分離処理手

段32により求めた分離済データまたは分離済データ行列Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}に含まれる1列の特徴要素（例えば、第1列の各要素）の行列内位置に応じて定まる1行の特徴判定要素（例えば、第1行の各要素、ベクトルの場合には、第1番目の要素）のみを用い、この特徴判定要素の値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかを判定することにより、検査配列中のプロモータの存否を判定する処理を行っていた（図32のステップS17、図38参照）。これに対し、本第2実施形態では、図49に示すように、Y_{test}を構成する要素のうち、推定された基底行列W_{prom}^{-1}（図44参照）に含まれる複数列の特徴要素の行列内位置に応じて定まる複数行の特徴判定要素を用いて、検査配列中のプロモータの存否を判定する処理を行う。なお、前記1実施形態の場合も、本第2実施形態の場合も、プロモータの存否を判定する硬判定ではなく、プロモータの存否の程度を判定する軟判定を行ってもよい。

[0293] 具体的には、本第2実施形態の判定処理手段は、図49に示すように、31行（ベクトルの場合には31要素）のY_{test}を構成する要素のうち、複数行の特徴判定要素だけを抽出して、例えば23行（ベクトルの場合には23要素）のY_{test}'を生成し、このY_{test}'の各列ベクトルY_{test}'(k)（k＝1，2，3，…。但し、検査データが1つの場合には、1つの縦ベクトルである。）のそれぞれについて、Y_{prom}のうち真データ対応部分の特徴判定要素だけを抽出して生成されたY_{prom}'（図44参照）の各列ベクトルY_{prom}'(j)（j＝1～106）との類似度を示す値を求める。ここで、類似度を示す値とは、例えば、Y_{test}'の任意の1つの列ベクトルY_{test}'(k)と、Y_{prom}'（図44参照）の各列ベクトルY_{prom}'(j)（j＝1～106）とのそれぞれの内積の値を算出し、これらのNp個（106個）の内積の値の平均値をとったものとすることができる。なお、Y_{prom}'（図44参照）の各列ベクトルY_{prom}'(j)（j＝1～106）の重心ベクトルを学習段階で予め求めておき、記憶させておけば、Y_{test}'の任意の1つの列ベクトルY_{test}'(k)と、この重心ベクトルとの内積の値を算出することにより、等価な値を算出することができる。また、類似度を示す値は、以上のようなY_{prom}'の各列ベクトルY_{prom}'(j)（j＝1～106）との内積の値の平均値、あるいはY_{prom}'の各列ベクトルY_{prom}'(j)（j＝1～106）の重心ベクトルとの内積の値に限定されるものではなく、これら以外の値であっても、Y_{prom}'の各列ベクトルY_{prom}'(j)（j＝1～106）の

総体との類似度を示す値を採用すれば、特定の既知のプロモータに近いか否かを判定するといった偏った判定を回避することができる。そして、このようにして求めた類似度を示す値が、予め定められた閾値（例えば、ゼロ）を境界としていずれの側の値であるかを判定することにより、その類似度を示す値の算出に用いられた$Y_{test}'(k)$の元になった$Y_{test}(k)$に対応する検査データ$X_{test}(k)$の数値化前の状態であるプロモータ候補配列がプロモータであるか否かがわかるので、検査配列の中にプロモータが存在するか否かを判定することができる。例えば、類似度を示す値が、正の値であれば、プロモータが存在すると判定し、負の値であれば、プロモータは存在しないと判定することができる。なお、軟判定の場合には、求めた類似度を示す値の大きさの程度を判定することにより、検査配列の中のプロモータの存否の程度を判定することができる。

[0294] このような第2実施形態によれば、前記第1実施形態と同様な効果を得ることができることに加え、1行の特徴判定要素だけではなく、複数行の特徴判定要素を用いて判定処理を行うようにしたので、判定制度を向上させることができるうえ、ギャップの挿入ではなく、スペーサ35の一部の配列およびスペーサ10の一部の配列を取り除くことによりアラインメント処理を行うようにしたので、プロモータ用分離行列W_{prom}等の各行列の次元数を小さくすることができ、演算量の低減、処理速度の向上を図ることができる。

[0295] ＜効果確認実験＞

また、本発明の効果を確かめるために、以下のような実験を行った。実験方法や実験条件は、本第2実施形態に基づくものとした。

[0296] 先ず、学習段階（トレーニングステップ）では、本第2実施形態で説明したように、前述した非特許文献1に記載された既知の106本の大腸菌プロモータを使って学習を行った。この学習結果として得られた31行×31列のプロモータ用分離行列W_{prom}（図44参照）の逆行列である31行×31列のW_{prom}^{-1}の第1列の各要素の値が、前述した図45に示され、第27列の各要素の値が、前述した図47に示されている。これらの横軸は、行番号に対応している。また、学習結果としてW_{prom}とともに得られた31行×212列のプロモータ用分離済データ行列Y_{prom}（図44参照）の第1行の各要素の値が、

前述した図46に示され、第27行の各要素の値が、前述した図48に示されている。これらの横軸は、列番号に対応している。

[0297] そして、図45のW_{prom}^{-1}の第1列および図46のY_{prom}の第1行は、特徴要素および特徴判定要素として採用されたものであり、一方、図47のW_{prom}^{-1}の第27列および図48のY_{prom}の第27行は、採用されなかったものであり、このようにして、W_{prom}^{-1}を構成する31列(31基底)のうち、合計23列(23基底)が特徴要素として採用され、従って、31行のY_{prom}のうち、合計23行が特徴判定要素として採用された。なお、採用条件を、より厳しくして23行未満の特徴判定要素を採用してもよい。

[0298] 次に、検査段階(テストステップ)では、テストデータとして、前述した非特許文献1に記載された既知の126本の大腸菌プロモータを使った。これらの126本の大腸菌プロモータは、学習段階で使った既知の106本の大腸菌プロモータとは異なるものである。なお、これらの126本のプロモータについては、非特許文献1では、既にギャップが挿入されていたため、それらのギャップを取り除き、ギャップのない配列長40～50のプロモータに戻してテストデータ(検査配列)とした。

[0299] また、テストデータ(検査配列)として、配列長が65のランダム配列を5000本生成して用意した。この際、AとTの出願頻度をそれぞれ0．3ずつとし、GとCの出願頻度をそれぞれ0．2ずつとした。従って、アラインメントされていないテストデータ(検査配列)が、合計で5126本用意された。但し、ランダム配列の配列長を65にしたのは、前述したホモロジ・スコアの計算をしてプロモータに近い配列を排除するためである。従って、ホモロジ・スコアの計算を行わない場合には、プロモータであるか否かが未知の配列について判定を行う実際の検査段階の処理で採用される検査配列の配列長と同じ配列長とし、例えば配列長が50のランダム配列を5000本生成して用意してもよい。

[0300] 続いて、5126本のテストデータについて、W_{-10}およびW_{-35}を用いて－10ボックス相当配列および－35ボックス相当配列を決定した後、図42に示した方法でアラインメントを施し、配列長を31に統一した。この際、転写開始点は、下流側から見て7塩基目と仮定し、1つのテストデータ(検査配列)につき、転写開始点を1つだけ設定した(つまり、この効果確認実験では、図36に示すような転写開始点のスライド設定は行わ

なかった)。また、－10ボックス相当配列および－35ボックス相当配列と、転写開始点近傍領域配列との相関判定処理は行わなかった。

[0301] それから、アラインメントされた5126本のプロモータ候補配列について、プロモータ用出現頻度テーブルT_{prom}を用いて数値化し、5126本の検査データを生成した後、図49に示すように、プロモータ用分離行列W_{prom}と、5126本の検査データを結合した検査データ行列X_{test}とを乗じる行列演算を行い、31行×5126列のプロモータ用分離済データ行列Y_{test}を求めた。

[0302] そして、プロモータ用分離済データ行列Y_{test}のうち、学習段階でW_{prom}^{-1}およびY_{prom}のグラフ化および解析を行って既に決定されている複数行(ここでは、合計23行)の特徴判定要素を抽出して、23行×5126列のY_{test}'を生成した。それから、Y_{test}'の任意の1つの列ベクトルY_{test}'(k)(k＝1～5126)と、Y_{prom}'(図44参照)の各列ベクトルY_{prom}'(j)(j＝1～106)とのそれぞれの内積の値を算出し、これらのNp個(106個)の内積の値の平均値を求めた。このような類似度を示す値の算出は、Y_{test}'を構成する5126列の全ての列ベクトルY_{test}'(k)(k＝1～5126)について行う。なお、Y_{prom}'(図44参照)の各列ベクトルY_{prom}'(j)(j＝1～106)の重心ベクトルを学習段階で予め求めておいた場合には、Y_{test}'の任意の1つの列ベクトルY_{test}'(k)(k＝1～5126)と、この重心ベクトルとの内積の値を算出することにより、等価な値を算出することができる。

[0303] このようにして各列ベクトルY_{test}'(k)(k＝1～5126)毎に求めた類似度を示す値について、閾値をゼロとし、正負を判定した(つまり、硬判定を行った)。この結果、95.04%という高い認識率が得られ、これにより本発明の効果が顕著に示された。なお、この95.04%という認識率は、転写開始点のスライド設定処理や相関判定処理を行うことにより、更に向上させることが可能である。

[0304] [変形の形態]

なお、本発明は前記各実施形態に限定されるものではなく、本発明の目的を達成できる範囲内での変形等は本発明に含まれるものである。

[0305] すなわち、前記各実施形態では、特定点である転写開始点は、部分パターンである－35ボックスおよび－10ボックスよりも下流側に配置されていたが、特定点と各部

分パターンとの位置関係は、これに限定されるものではなく、本発明における特定点は、全ての部分パターンの下流側にあってもよく、全ての部分パターンの上流側にあってもよく、部分パターン同士の間に介在していてもよい。

[0306] また、前記各実施形態では、2つの－35ボックスおよび－10ボックスと、転写開始点近傍領域配列との相関を判定していたが、必ずしも全ての部分パターンと特定点近傍領域配列との相関を判定するのではなく、一部の部分パターンと特定点近傍領域配列との相関を判定してもよい。

[0307] さらに、前記各実施形態では、各部分パターンと特定点近傍領域配列との相関を判定していたが、多重化されていない特徴パターンの認識処理を行う場合には、その多重化されていない特徴パターンと、特定点近傍領域配列との相関を判定してもよい。

[0308] そして、前記各実施形態では、独立成分分析（ICA）または主成分分析（PCA）により相関判定を行っていたが、これに限定されるものではなく、例えば、ニューラルネットワーク法等により相関判定を行ってもよい。ニューラルネットワーク法の場合には、例えば、学習段階で、互いに相関のある既知の各部分パターンと既知の特定点近傍領域配列とを入力したときには、相関があることを示す数値（例えば1）が出力され、一方、相関のない配列（ランダム配列等）を入力したときには、相関がないことを示す数値（例えば0）が出力されるように係数を定めておき、検査段階で、特徴パターン候補配列に含まれる各部分パターン相当配列および特定点近傍領域配列を入力したときに、その出力結果が、相関があることを示す数値（例えば1）になるか、あるいは相関がないことを示す数値（例えば0）になるかを判定する方法等を採用することができる。

[0309] また、前記各実施形態では、転写開始点を設定（仮定）して－35ボックス相当配列および－10ボックス相当配列を決定した後に、これらの－35ボックス相当配列および－10ボックス相当配列と転写開始点近傍領域配列との間に相関があるか否かを判断するだけであったが、このような1回の往復処理だけではなく、例えば、転写開始点を設定（仮定）して－35ボックス相当配列および－10ボックス相当配列を決定した後に、これらの－35ボックス相当配列および－10ボックス相当配列と転写開始点

近傍領域配列との間の相関が高くなる転写開始点を見つけ、再度、その転写開始点に基づき－35ボックス相当配列および－10ボックス相当配列を決定するという具合に、複数回の反復処理を行ってもよい。

[0310] さらに、前記各実施形態では、図36の二点鎖線で示すように、1つの検査配列の中で、複数のプロモータ候補配列を決定する処理を行うが、それぞれのプロモータ候補配列を決定するにあたり、さらに複数(9個)の－10ボックス候補配列や複数(7個)の－35ボックス候補配列に関する演算処理を行うので、重複した演算処理を回避するために、一度算出した結果を適宜保存して再利用するか、または一度にまとめて実行できる演算については、まとめて行うようにしてもよい。例えば、ある点に設定された転写開始点についてのプロモータ候補配列を決定する際に行われるj＝5の－10ボックス候補配列に関する演算処理と、1つ上流側の点に設定された転写開始点についてのプロモータ候補配列を決定する際に行われるj＝4の－10ボックス候補配列に関する演算処理とは、重複する部分があるので、前者の演算処理の結果の一部(例えば、図21の$Y_{-10can}(j)$のうちの重複するデータ)を保存し、後者の演算処理の際に用いるようにしてもよい。また、図21の$Y_{-10can}(j)$のデータを、9個の－10ボックス候補配列に区切って求めるのではなく、一度にまとめて多数の－10ボックス候補配列に対応する$Y_{-10can}(j)$のデータを求めておき、そこから9個のデータを抜き出して(1個ずつずらしながら9個のデータを抜き出していく。)内積の和$q(j)$を求める処理に用いてもよい。

[0311] また、前記各実施形態のプロモータ認識システム10等は、多重化された特徴パターンであり、かつ、特徴パターンの全体の長さに多様性があるプロモータについての認識処理を行うシステムであったが、本発明の特徴パターン認識システムは、多重化されていない特徴パターンであり、かつ、特徴パターンの全体の長さが一定である場合の特徴パターンを認識するシステムにも適用することができ、また、特徴パターンの長さに多様性がある場合(複数種類の長さが存在する場合)において、多重化されていない特徴パターンを認識するシステム、あるいは多重化されている特徴パターンであるが、多重化されていない場合と同様な処理でその特徴パターンを認識するシステムにも適用することができる。

[0312] 例えば、特徴パターンの長さに多様性がある場合において、多重化されていない特徴パターンを認識するか、あるいは多重化されている特徴パターンであるが、多重化されていない場合と同様な処理でその特徴パターンを認識する処理を行う際には、クラスタルダブリュー(ClustalW)という既存のアルゴリズムを実行する既存の標準的なツールであるブラスト(BLAST)等を用いて候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行うことができる。また、この場合において、判定処理の対象となる配列若しくはその一部の配列断片から切り出された1つの検査配列の中で、候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の候補配列のそれぞれについて検査データを生成し、1つの検査配列から得られたそれらの複数の検査データについて、分離処理手段による分離処理および判定処理手段による判定処理を行うことにより、1つの検査配列の中で定められた複数の候補配列の中から、特徴パターンであると認識される配列を1つ定める(つまり、最も確からしい候補配列を特徴パターンであると認識する)か、あるいは1つも定めない(つまり、検査配列の中には特徴パターンは存在しないと判定する)処理を行うことができる。例えば、25～30の間で配列長が変動する特徴パターン(新規に発見される特徴パターンの想定長も考慮することが好ましい。)の場合であれば、検査配列の中で、ある点を基準(特徴パターンの右端)として、25～30の各長さの候補配列を決定し、次に、1つ上流側の点を基準(特徴パターンの右端)として、同様に25～30の各長さの候補配列を決定するという具合にして、1つの検査配列の中で、複数の候補配列を決定することができる。

産業上の利用可能性

[0313] 以上のように、本発明の特徴パターン認識システムおよびその方法、並びにプログラムは、例えば、DNA配列中のプロモータやエンハンサ、アミノ酸配列中のモチーフ、DNA配列とアミノ酸配列との混在型配列中のパターン等を認識する場合に用いるのに適している。

図面の簡単な説明

[0314] [図1]本発明の第1実施形態のプロモータ認識システムの全体構成図。

[図2]第1実施形態のプロモータ認識システムの一部である学習処理手段の詳細構成図。

[図3]第1実施形態の学習段階(トレーニングステップ)の全体の流れを示すフローチャートの図。

[図4]第1実施形態の－35ボックス学習の流れを示すフローチャートの図。

[図5]第1実施形態のDNA配列から切り出された既知のプロモータを含む配列の例示図。

[図6]第1実施形態の－35ボックス学習に伴う処理の第1の説明図。

[図7]第1実施形態の－35ボックス学習に伴う処理の第2の説明図。

[図8]第1実施形態の－35ボックス学習に伴う処理の第3の説明図。

[図9]第1実施形態の－35ボックス学習に伴う処理の第4の説明図。

[図10]第1実施形態の－35ボックス学習に伴う処理の第5の説明図。

[図11]第1実施形態の－10ボックス学習の流れを示すフローチャートの図。

[図12]第1実施形態の－10ボックス学習に伴う処理の第1の説明図。

[図13]第1実施形態の－10ボックス学習に伴う処理の第2の説明図。

[図14]第1実施形態の－10ボックス学習に伴う処理の第3の説明図。

[図15]第1実施形態の－10ボックス学習に伴う処理の第4の説明図。

[図16]第1実施形態のプロモータ学習の流れを示すフローチャートの図。

[図17]第1実施形態のプロモータ学習の流れの詳細(その1)を示すフローチャートの図。

[図18]第1実施形態のプロモータ学習の流れの詳細(その2)を示すフローチャートの図。

[図19]第1実施形態のプロモータ学習に伴う処理の第1の説明図。

[図20]第1実施形態のプロモータ学習に伴う処理の第2の説明図。

[図21]第1実施形態のプロモータ学習に伴う処理の第3の説明図。

[図22]第1実施形態のプロモータ学習に伴う処理の第4の説明図。

[図23]第1実施形態のプロモータ学習に伴う処理の第5の説明図。

[図24]第1実施形態のプロモータ学習に伴う処理の第6の説明図。

[図25]第1実施形態のプロモータ学習に伴う処理の第7の説明図。

[図26]第1実施形態のプロモータ学習に伴う処理の第8の説明図。

[図27]第1実施形態の相関学習の流れを示すフローチャートの図。

[図28]第1実施形態の相関学習に伴う処理の第1の説明図。

[図29]第1実施形態の相関学習に伴う処理の第2の説明図。

[図30]第1実施形態の相関学習に伴う処理の第3の説明図。

[図31]第1実施形態の相関学習に伴う処理の第4の説明図。

[図32]第1実施形態の検査段階(テストステップ)の全体の流れを示すフローチャートの図。

[図33]第1実施形態の検査段階における詳細処理(その1)の流れを示すフローチャートの図。

[図34]第1実施形態の検査段階における詳細処理(その2)の流れを示すフローチャートの図。

[図35]第1実施形態の検査段階における相関判定処理の流れを示すフローチャートの図。

[図36]第1実施形態の検査段階における処理の第1の説明図。

[図37]第1実施形態の検査段階における処理の第2の説明図。

[図38]第1実施形態の検査段階における処理の第3の説明図。

[図39]本発明の効果確認実験(第1実施形態に基づく実験)におけるプロモータ用分離行列W_{prom}の逆行列W_{prom}^{-1}の第1列の各要素の値を示すグラフの図。

[図40]本発明の効果確認実験(第1実施形態に基づく実験)におけるプロモータ用分離済データ行列Y_{prom}の第1行の各要素の値を示すグラフの図。

[図41]大腸菌塩基配列中でRNAポリメラーゼが大腸菌プロモータを探す様子を示す図。

[図42]本発明の第2実施形態のアラインメントの方法を示す説明図。

[図43]第2実施形態のプロモータ用出現頻度テーブルT_{prom}を示す図。

[図44]第2実施形態における基底行例での特徴要素の現れ方と、検査段階の判定処理で用いられるプロモータ用分離済データ行列Y_{prom}のうちの真データ対応部分の

特徴判定要素との関係を示す図。

[図45]第2実施形態の学習段階の処理で得られたプロモータ用分離行列W_{prom}の逆行列W_{prom}^{-1}の第1列の各要素の値を示すグラフの図。

[図46]第2実施形態の学習段階の処理で得られたプロモータ用分離済データ行列Y_{prom}の第1行の各要素の値を示すグラフの図。

[図47]第2実施形態の学習段階の処理で得られたプロモータ用分離行列W_{prom}の逆行列W_{prom}^{-1}の第27列の各要素の値を示すグラフの図。

[図48]第2実施形態の学習段階の処理で得られたプロモータ用分離済データ行列Y_{prom}の第27行の各要素の値を示すグラフの図。

[図49]第2実施形態の検査段階の処理の説明図。

符号の説明

[0315] 10 特徴パターン認識システムであるプロモータ認識システム

31 検査データ生成処理手段

31A 部分パターン候補データ生成処理手段である－35ボックス候補データ生成処理手段

31B 部分パターン候補データ生成処理手段である－10ボックス候補データ生成処理手段

31C 部分パターン候補データ用分離済データ生成処理手段である－35ボックス候補データ用分離済データ生成処理手段

31D 部分パターン候補データ用分離済データ生成処理手段である－10ボックス候補データ用分離済データ生成処理手段

31E 部分パターン相当配列決定処理手段である－35ボックス相当配列決定処理手段

31F 部分パターン相当配列決定処理手段である－10ボックス相当配列決定処理手段

31G 特徴パターン候補配列決定処理手段であるプロモータ候補配列決定処理手段

31H アラインメント処理手段

31J 特徴パターン候補配列数値化処理手段であるプロモータ候補配列数値化処理手段

32 分離処理手段

33 判定処理手段

41 部分パターン用出現頻度テーブル記憶手段である－35ボックス用出現頻度テーブル記憶手段

42 部分パターン用分離行列記憶手段である－35ボックス用分離行列記憶手段

43 部分パターン用分離済データ行列記憶手段である－35ボックス用分離済データ行列記憶手段

51 部分パターン用出現頻度テーブル記憶手段である－10ボックス用出現頻度テーブル記憶手段

52 部分パターン用分離行列記憶手段である－10ボックス用分離行列記憶手段

53 部分パターン用分離済データ行列記憶手段である－10ボックス用分離済データ行列記憶手段

61 パターン全体用出現頻度テーブル記憶手段であるプロモータ用出現頻度テーブル記憶手段

62 パターン全体用分離行列記憶手段であるプロモータ用分離行列記憶手段

70 相関学習結果記憶手段

71 相関判定用出現頻度テーブル記憶手段

72 相関判定用分離行列記憶手段

A, T, G, C 離散シンボル

T_{-35} 部分パターン用出現頻度テーブルである－35ボックス用出現頻度テーブル

T_{-10} 部分パターン用出現頻度テーブルである－10ボックス用出現頻度テーブル

T_{prom} パターン全体用出現頻度テーブルであるプロモータ用出現頻度テーブル

T_{BBS} 相関判定用出現頻度テーブル

W_{-35} 部分パターン用分離行列である－35ボックス用分離行列

W_{-10} 部分パターン用分離行列である－10ボックス用分離行列

W_{prom} パターン全体用分離行列であるプロモータ用分離行列

W_{BBS} 相関判定用分離行列

X_{-35} 部分パターン用学習データ行列である－35ボックス用学習データ行列

X_{-10} 部分パターン用学習データ行列である－10ボックス用学習データ行列

X_{prom} パターン全体用学習データ行列であるプロモータ用学習データ行列

X_{BBS} 相関判定用学習データ行列

$X_{BBS,test}$ 相関判定データ行列

X_{test} 検査データ行列

Y_{-35} 部分パターン用分離済データ行列である－35ボックス用分離済データ行列

Y_{-10} 部分パターン用分離済データ行列である－10ボックス用分離済データ行列

Y_{-35can} 部分パターン候補データ用分離済データ行列である－35ボックス候補データ用分離済データ行列

Y_{-10can} 部分パターン候補データ用分離済データ行列である－10ボックス候補データ用分離済データ行列

Y_{prom} パターン全体用分離済データ行列であるプロモータ用分離済データ行列

$Y_{BBS,test}$ 相関判定用分離済データ行列

Z_{-35} 部分パターン候補データ行列である－35ボックス候補データ行列

Z_{-10} 部分パターン候補データ行列である－10ボックス候補データ行列

請求の範囲

[1] 有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定する処理を行う特徴パターン認識システムであって、

前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記離散シンボルの各種別とを対応させて形成された出現頻度テーブルを記憶する出現頻度テーブル記憶手段と、

前記複数種類の既知の特徴パターンおよびこれらの既知の特徴パターンとは異なる複数種類の非特徴パターンを、前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う行列として得られた分離行列を記憶する分離行列記憶手段と、

判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う検査データ生成処理手段と、

前記分離行列記憶手段に記憶された前記分離行列と前記検査データ生成処理手段により生成された前記検査データまたは複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の前記分離済データが結合された分離済データ行列を求める処理を行う分離処理手段と、

この分離処理手段により求めた前記分離済データまたは前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる前記配列若しくはその

一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う判定処理手段と

を備えたことを特徴とする特徴パターン認識システム。

[2] 請求項1に記載の特徴パターン認識システムにおいて、

前記特徴パターンは、異なる領域に位置する複数の部分パターンを含んで構成される多重化された特徴パターンであり、前記特徴パターン内での前記各部分パターンの領域位置および前記特徴パターンの全体の長さには、前記特徴パターンの種類の相違により複数種類の領域位置および複数種類の長さが存在し、

前記出現頻度テーブル記憶手段は、前記複数種類の既知の特徴パターンのパターン全体の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えた後、パターン全体の長さを揃えた前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記ギャップを含む前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記ギャップを含む前記離散シンボルの各種別とを対応させて形成されたパターン全体用出現頻度テーブルを記憶するパターン全体用出現頻度テーブル記憶手段であり、

前記分離行列記憶手段は、パターン全体の長さを揃えた前記複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた前記複数種類の非特徴パターンから生成されたパターン全体用学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られたパターン全体用分離行列を記憶するパターン全体用分離行列記憶手段であり、

前記各領域の部分パターン毎に形成された部分パターン用出現頻度テーブルを記憶する前記各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段と、

前記各領域の部分パターン毎に得られた部分パターン用分離行列を記憶する前記各領域の部分パターン毎の部分パターン用分離行列記憶手段と、

前記各領域の部分パターン毎に学習段階で前記部分パターン用分離行列とともに得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を記憶する部分パターン用分離済データ行列記憶手段とを備え、

前記各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段には、

前記既知の特徴パターンに含まれる前記各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンを用いてこれらの部分パターン内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記部分パターン内での各配列位置および前記離散シンボルの各種別とを対応させて形成された前記各領域の部分パターン毎の部分パターン用出現頻度テーブルが記憶され、

前記各領域の部分パターン毎の部分パターン用分離行列記憶手段には、

前記既知の特徴パターンに含まれる前記各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンおよびこれらの既知の部分パターンとは異なる複数種類の非部分パターンを、前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して部分パターン用学習データ行列を生成し、この部分パターン用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記部分パターンの特徴を示す特徴要素を含む部分パターン用基底行列の逆変換を行う行列として得られた部分パターン用分離行列が記憶され、

前記部分パターン用分離済データ行列記憶手段には、

独立成分分析または主成分分析を行うことにより前記部分パターン用分離行列を得る際に前記部分パターン用分離行列と前記部分パターン用学習データ行列とを乗じた結果として得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素が記憶され、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、前記各領域の部分パターンの

それぞれについて、前記部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について前記部分パターン用出現頻度テーブル記憶手段に記憶された前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成する処理を行う部分パターン候補データ生成処理手段と、

前記部分パターン用分離行列記憶手段に記憶された前記部分パターン用分離行列と前記部分パターン候補データ生成処理手段により生成された前記複数の部分パターン候補データの各々または前記複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは前記複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成する処理を行う部分パターン候補データ用分離済データ生成処理手段と、

この部分パターン候補データ用分離済データ生成処理手段により生成された前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた前記部分パターン候補データ用分離済データに対応する前記部分パターン候補配列を、前記特徴パターンの候補となる特徴パターン候補配列に含まれる部分パターン相当配列として決定する処理を行う部分パターン相当配列決定処理手段と、

この部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列に基づき前記特徴パターン候補配列を決定する処理を行う特徴パターン候補配列決定処理手段と、

前記部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列をそれぞれ一塊の状態に保ったままで、前記特徴パターン候補配列決定処理手段により決定された前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行うアラインメント処理手段と、

このアラインメント処理手段により長さを揃えた前記特徴パターン候補配列を、前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、前記検査データを生成する処理を行う特徴パターン候補配列数値化処理手段と

を含んで構成されていることを特徴とする特徴パターン認識システム。

[3] 請求項2に記載の特徴パターン認識システムにおいて、

前記パターン全体用分離行列記憶手段に記憶された前記パターン全体用分離行列を得る際に用いられる前記パターン全体用学習データ行列を構成するためのパターン全体の長さを揃えた前記複数種類の非特徴パターンは、

これらの非特徴パターンを生成するために用意された非特徴パターン生成用配列の中で、前記各領域の部分パターンのそれぞれについて、前記部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成し、前記部分パターン用分離行列と前記複数の部分パターン候補データの各々または前記複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは前記複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成し、生成された前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた前記部分パターン候補データ用分離済データに対応する前記部分パターン候補配列を、前記非特徴パターンに含まれる部分パターン相当配列として決定し、さらに決定された各領域の前記部分パターン相当配列に基づき前記非特徴パターンを決定した後、決定された各領域の前記部分パターン相当配列をそれぞれ一塊の状態に保ったままで、決定された前記非特徴パターンの長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により

一定長に揃えるアラインメント処理を行って生成されたものである

ことを特徴とする特徴パターン認識システム。

[4] 請求項3に記載の特徴パターン認識システムにおいて、

前記複数種類の非特徴パターンは、

前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行って生成されたものである

ことを特徴とする特徴パターン認識システム。

[5] 請求項4に記載の特徴パターン認識システムにおいて、

前記類似度は、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和であることを特徴とする特徴パターン認識システム。

[6] 請求項2～5のいずれかに記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段の前記部分パターン相当配列決定処理手段は、

前記部分パターン候補データ用分離済データ生成処理手段により生成された前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[7] 請求項6に記載の特徴パターン認識システムにおいて、

前記類似度は、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和であることを

特徴とする特徴パターン認識システム。

[8] 請求項1～7のいずれかに記載の特徴パターン認識システムにおいて、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの前記特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[9] 請求項1～7のいずれかに記載の特徴パターン認識システムにおいて、

前記基底行列には、複数列の前記特徴要素が現れ、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列を構成する要素のうちの複数行の要素が前記特徴判定要素として採用され、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値を用いて、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[10] 請求項9に記載の特徴パターン認識システムにおいて、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴

判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[11] 請求項10に記載の特徴パターン認識システムにおいて、

前記判定処理手段は、

前記類似度を示す値として、前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの総体との類似度を示す値を算出する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[12] 請求項11に記載の特徴パターン認識システムにおいて、

前記判定処理手段は、

前記類似度を示す値として、前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの重心ベクトルとの内積の値またはこれと等価な値を算出する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[13] 請求項2～7のいずれかに記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記特徴パターン候補配列を1離散シンボルずつずらしながら決定

し、決定された複数の前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記特徴パターン候補配列のそれぞれを前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行う構成とされ、

前記分離処理手段は、

前記パターン全体用分離行列と前記検査データ生成処理手段により生成された複数の前記検査データの各々またはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行う構成とされ、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる各特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記特徴判定要素の値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記特徴判定要素の値を求め、求めた特徴判定要素の値を与えた前記検査データに対応する前記特徴パターン候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[14] 請求項2～7のいずれかに記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記特徴パターン候補配列のそれぞれを前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行う構成とされ、

前記分離処理手段は、

前記パターン全体用分離行列と前記検査データ生成処理手段により生成された複数の前記検査データの各々またはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行う構成とされ、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる複数列の前記特徴要素の行列内位置に応じて定まる複数行の前記特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で前記パターン全体用分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記類似度を示す値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記類似度を示す値を求め、求めた類似度を示す値を与えた前記検査データに対応する前記特徴パターン候

補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[15] 請求項2～7，13，14のいずれかに記載の特徴パターン認識システムにおいて、

多重化された前記特徴パターンに含まれる前記各領域の部分パターンが、前記配列中の特定点を含む特定点近傍領域配列と相関している場合に、

前記各領域の部分パターンと前記特定点近傍領域配列との相関を判定するために用いられる情報として予め学習して得られた相関学習結果を含む情報を記憶する相関学習結果記憶手段を備え、

前記検査データ生成処理手段の前記部分パターン候補データ生成処理手段は、

前記各領域の部分パターンのそれぞれについて、前記特定点との相対的な位置関係または前記特定点との相対的な位置関係で定められた他の領域の部分パターン候補配列との相対的な位置関係で、前記複数の部分パターン候補配列を定める処理を行う構成とされ、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

前記相関学習結果記憶手段に記憶された前記相関学習結果を含む情報を用いて、前記検査データ生成処理手段の前記部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列と前記特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定し、相関がないか、または低いと判定された場合には、それらの各領域の前記部分パターン相当配列および前記特定点近傍領域配列中の前記特定点に基づく前記特徴パターン候補配列の決定処理は行わない構成とされている

ことを特徴とする特徴パターン認識システム。

[16] 請求項15に記載の特徴パターン認識システムにおいて、

前記相関学習結果記憶手段は、

前記既知の特徴パターンに含まれる複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特徴点近傍領域配列を結合して形成された複数

種類の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記相関結合配列内での各配列位置および前記離散シンボルの各種別とを対応させて形成された相関判定用出現頻度テーブルを記憶する相関判定用出現頻度テーブル記憶手段と、

前記各領域の部分パターンと前記特定点近傍領域配列との相関を判定する相関判定処理に用いられる相関判定用分離行列を記憶する相関判定用分離行列記憶手段とを含んで構成され、

前記相関判定用分離行列記憶手段には、

複数の領域のうちの少なくとも一つの領域の前記既知の部分パターンおよび前記既知の特定点近傍領域配列を結合して生成された相関結合配列、並びに複数の領域のうちの少なくとも一つの領域の前記既知の部分パターンおよび前記既知の特定点近傍領域配列とは異なる非特定点近傍領域配列を結合して形成された非相関結合配列を、前記相関判定用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して相関判定用学習データ行列を生成し、この相関判定用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記相関結合配列の特徴を示す特徴要素を含む相関判定用基底行列の逆変換を行う行列として得られた相関判定用分離行列が記憶され、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

決定された各領域の前記部分パターン相当配列と前記特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定する際には、複数の領域のうちの少なくとも一つの領域の前記部分パターン相当配列および前記特定点近傍領域配列を結合して相関判定配列を生成し、この相関判定配列を前記相関判定用出現頻度テーブル記憶手段に記憶された前記相関判定用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して相関判定データを生成した後、前記相関判定用分離行列記憶手段に記憶された前記相関判定用分離行列と、前記相関判定データまたは複数の前記相関判定データを結合した相関判定

データ行列とを乗じる行列演算を行うことにより、相関判定用分離済データまたは複数の前記相関判定用分離済データが結合された相関判定用分離済データ行列を求める相関判定用分離処理を行い、この相関判定用分離処理で求めた前記相関判定用分離済データまたは前記相関判定用分離済データ行列を構成する要素のうち、前記相関判定用基底行列に含まれる前記特徴要素の行列内位置に応じて定まる相関判定要素の値を用いて、相関があるか否か、または相関の程度を判定する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[17] 請求項16に記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

前記相関判定要素の値が予め定められた相関判定用閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、相関があるか否か、または相関の程度を判定する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[18] 請求項1～17のいずれかに記載の特徴パターン認識システムにおいて、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは、前記DNA配列中のプロモータである

ことを特徴とする特徴パターン認識システム。

[19] 請求項2～7，13～17のいずれかに記載の特徴パターン認識システムにおいて、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは、前記DNA配列中のプロモータであり、前記部分パターンは、前記プロモータに含まれる－35ボックスおよび－10ボックスである

ことを特徴とする特徴パターン認識システム。

[20] 請求項15～17のいずれかに記載の特徴パターン認識システムにおいて、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは

、前記DNA配列中のプロモータであり、前記部分パターンは、前記プロモータに含まれる－35ボックスおよび－10ボックスであり、前記特定点は、前記DNA配列の転写開始点であり、前記特徴点近傍領域配列は、転写開始点近傍領域配列である

ことを特徴とする特徴パターン認識システム。

[21] 請求項1～17のいずれかに記載の特徴パターン認識システムにおいて、

前記配列は、タンパク質を構成するアミノ酸配列であり、前記離散シンボルは、前記アミノ酸配列を構成するアミノ酸を示す記号であり、前記特徴パターンは、前記アミノ酸配列中のモチーフである

ことを特徴とする特徴パターン認識システム。

[22] 請求項1に記載の特徴パターン認識システムにおいて、

前記特徴パターンの長さには、前記特徴パターンの種類の相違により複数種類の長さが存在し、

前記出現頻度テーブル記憶手段には、前記複数種類の既知の特徴パターンの長さをギャップの挿入により一定長に揃えた後、長さを揃えた前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記ギャップを含む前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記ギャップを含む前記離散シンボルの各種別とを対応させて形成された前記出現頻度テーブルが記憶され、

前記分離行列記憶手段には、長さを揃えた前記複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた前記複数種類の非特徴パターンから生成された前記学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られた前記分離行列が記憶され、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、前記特徴パターンの候補となる候補配列を決定した後、決定された前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた前記候補配列を前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置およ

び前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、前記検査データを生成する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[23] 請求項22に記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記候補配列のそれぞれを前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行う構成とされ、

前記分離処理手段は、

前記分離行列と前記検査データ生成処理手段により生成された複数の前記検査データまたはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行う構成とされ、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記特徴判定要素の値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記特徴判定要素の値を求め、求めた特徴判定要素の値を与えた前記検

査データに対応する前記候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[24] 請求項22に記載の特徴パターン認識システムにおいて、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記候補配列のそれぞれを前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行う構成とされ、

前記分離処理手段は、

前記分離行列と前記検査データ生成処理手段により生成された複数の前記検査データまたはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行う構成とされ、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる複数列の前記特徴要素の行列内位置に応じて定まる複数行の前記特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるか

または値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記類似度を示す値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記類似度を示す値を求め、求めた類似度を示す値を与えた前記検査データに対応する前記候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う構成とされている

ことを特徴とする特徴パターン認識システム。

[25] 有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定するための特徴パターン認識方法であって、

前記複数種類の既知の特徴パターンを用意するとともに、これらの既知の特徴パターンとは異なる複数種類の非特徴パターンを生成または用意し、

前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記離散シンボルの各種別とを対応させて出現頻度テーブルを作成し、この出現頻度テーブルを出現頻度テーブル記憶手段に記憶させておくとともに、

前記複数種類の既知の特徴パターンおよび前記複数種類の非特徴パターンを、前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う行列として得られた分離行列を分離行列記憶手段に記憶させておき、

検査データ生成処理手段が、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、前記

出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行い、

分離処理手段が、前記分離行列記憶手段に記憶された前記分離行列と、前記検査データ生成処理手段により生成された前記検査データまたは複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の前記分離済データが結合された分離済データ行列を求める処理を行い、

判定処理手段が、前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[26] 請求項25に記載の特徴パターン認識方法において、

前記特徴パターンは、異なる領域に位置する複数の部分パターンを含んで構成される多重化された特徴パターンであり、前記特徴パターン内での前記各部分パターンの領域位置および前記特徴パターンの全体の長さには、前記特徴パターンの種類の相違により複数種類の領域位置および複数種類の長さが存在し、

前記出現頻度テーブル記憶手段は、前記複数種類の既知の特徴パターンのパターン全体の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えた後、パターン全体の長さを揃えた前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記ギャップを含む前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記ギャップを含む前記離散シンボルの各種別とを対応させて形成されたパターン全体用出現頻度テーブルを記憶するパターン

全体用出現頻度テーブル記憶手段であり、

前記分離行列記憶手段は、パターン全体の長さを揃えた前記複数種類の既知の特徴パターンおよびこれらと同じ長さに揃えた前記複数種類の非特徴パターンから生成されたパターン全体用学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られたパターン全体用分離行列を記憶するパターン全体用分離行列記憶手段であり、

前記各領域の部分パターン毎に形成された部分パターン用出現頻度テーブルを前記各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段に記憶させておくとともに、

前記各領域の部分パターン毎に得られた部分パターン用分離行列を前記各領域の部分パターン毎の部分パターン用分離行列記憶手段に記憶させ、

さらに、前記各領域の部分パターン毎に学習段階で前記部分パターン用分離行列とともに得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を部分パターン用分離済データ行列記憶手段に記憶させておき、

前記各領域の部分パターン毎の部分パターン用出現頻度テーブル記憶手段には、

前記既知の特徴パターンに含まれる前記各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンを用いてこれらの部分パターン内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記部分パターン内での各配列位置および前記離散シンボルの各種別とを対応させて形成された前記各領域の部分パターン毎の部分パターン用出現頻度テーブルを記憶させておき、

前記各領域の部分パターン毎の部分パターン用分離行列記憶手段には、

前記既知の特徴パターンに含まれる前記各領域の部分パターンのそれぞれについて、複数種類の既知の部分パターンおよびこれらの既知の部分パターンとは異なる複数種類の非部分パターンを、前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して部分パターン用学習データ行列を生成し、この部分パターン用学習

データ行列を用いて独立成分分析または主成分分析を行うことにより、前記部分パターンの特徴を示す特徴要素を含む部分パターン用基底行列の逆変換を行う行列として得られた部分パターン用分離行列を記憶させておき、

前記部分パターン用分離済データ行列記憶手段には、

独立成分分析または主成分分析を行うことにより前記部分パターン用分離行列を得る際に前記部分パターン用分離行列と前記部分パターン用学習データ行列とを乗じた結果として得られた部分パターン用分離済データ行列のうちの少なくとも真データ対応部分の要素を記憶させておき、

前記検査データ生成処理手段により前記検査データを生成する際には、

部分パターン候補データ生成処理手段が、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、前記各領域の部分パターンのそれぞれについて、前記部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について前記部分パターン用出現頻度テーブル記憶手段に記憶された前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成する処理を行い、

部分パターン候補データ用分離済データ生成処理手段が、前記部分パターン用分離行列記憶手段に記憶された前記部分パターン用分離行列と前記部分パターン候補データ生成処理手段により生成された前記複数の部分パターン候補データの各々または前記複数の部分パターン候補データを結合した部分パターン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは前記複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成する処理を行い、

部分パターン相当配列決定処理手段が、前記部分パターン候補データ用分離済データ生成処理手段により生成された前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた前記

部分パターン候補データ用分離済データに対応する前記部分パターン候補配列を、前記特徴パターンの候補となる特徴パターン候補配列に含まれる部分パターン相当配列として決定する処理を行い、

特徴パターン候補配列決定処理手段が、前記部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列に基づき前記特徴パターン候補配列を決定する処理を行い、

アラインメント処理手段が、前記部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列をそれぞれ一塊の状態に保ったままで、前記特徴パターン候補配列決定処理手段により決定された前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、

特徴パターン候補配列数値化処理手段が、前記アラインメント処理手段により長さを揃えた前記特徴パターン候補配列を、前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、前記検査データを生成する処理を行う

ことを特徴とする特徴パターン認識方法。

[27] 請求項26に記載の特徴パターン認識方法において、

前記パターン全体用分離行列記憶手段に記憶させる前記パターン全体用分離行列を得る際に用いられる前記パターン全体用学習データ行列を構成するためのパターン全体の長さを揃えた前記複数種類の非特徴パターンを生成する際には、

これらの非特徴パターンを生成するために用意された非特徴パターン生成用配列の中で、前記各領域の部分パターンのそれぞれについて、前記部分パターンと同じ長さを有する互いに位置のずれた複数の部分パターン候補配列を定め、これらの複数の部分パターン候補配列について前記部分パターン用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して複数の部分パターン候補データを生成し、前記部分パターン用分離行列と前記複数の部分パターン候補データの各々または前記複数の部分パターン候補データを結合した部分パタ

ーン候補データ行列とを乗じる行列演算を行うことにより、複数の部分パターン候補データ用分離済データまたは前記複数の部分パターン候補データ用分離済データが結合された部分パターン候補データ用分離済データ行列を生成し、生成された前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求め、求めた前記部分パターン候補データ用分離済データに対応する前記部分パターン候補配列を、前記非特徴パターンに含まれる部分パターン相当配列として決定し、さらに決定された各領域の前記部分パターン相当配列に基づき前記非特徴パターンを決定した後、決定された各領域の前記部分パターン相当配列をそれぞれ一塊の状態に保ったままで、決定された前記非特徴パターンの長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行う

ことを特徴とする特徴パターン認識方法。

[28] 請求項27に記載の特徴パターン認識方法において、

前記複数種類の非特徴パターンの生成時に、

前記複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際には、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める

ことを特徴とする特徴パターン認識方法。

[29] 請求項28に記載の特徴パターン認識方法において、

前記類似度は、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和であることを特徴とする特徴パターン認識方法。

[30] 請求項26～29に記載の特徴パターン認識方法において、

前記検査データ生成処理手段の前記部分パターン相当配列決定処理手段は、

前記部分パターン候補データ用分離済データ生成処理手段により生成された前記

複数の部分パターン候補データ用分離済データのうち、前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルに最も近い部分パターン候補データ用分離済データを求める際に、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルの総体との類似度が最大になる部分パターン候補データ用分離済データを求める処理を行う

ことを特徴とする特徴パターン認識方法。

[31] 請求項30に記載の特徴パターン認識方法において、

前記類似度は、前記部分パターン候補データ用分離済データと前記部分パターン用分離済データ行列の真データ対応部分の各列ベクトルとの内積の和であることを特徴とする特徴パターン認識方法。

[32] 請求項25～31のいずれかに記載の特徴パターン認識方法において、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの前記特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[33] 請求項25～31のいずれかに記載の特徴パターン認識方法において、

前記基底行列に、複数列の前記特徴要素が現れるときには、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列を構成する要素のうちの複数行の要素を前記特徴判定要素として採用し、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値を用いて、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出され

た検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[34] 請求項33に記載の特徴パターン認識方法において、

前記判定処理手段は、

前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[35] 請求項34に記載の特徴パターン認識方法において、

前記判定処理手段は、

前記類似度を示す値として、前記分離処理手段により求めた前記分離済データまたは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの総体との類似度を示す値を算出する処理を行う

ことを特徴とする特徴パターン認識方法。

[36] 請求項35に記載の特徴パターン認識方法において、

前記判定処理手段は、

前記類似度を示す値として、前記分離処理手段により求めた前記分離済データま

たは前記分離済データ行列のうちの複数行の前記特徴判定要素の値により構成される列ベクトルと、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルの重心ベクトルとの内積の値またはこれと等価な値を算出する処理を行う

ことを特徴とする特徴パターン認識方法。

[37] 請求項26～31のいずれかに記載の特徴パターン認識方法において、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記特徴パターン候補配列のそれぞれを前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行い、

前記分離処理手段は、

前記パターン全体用分離行列と前記検査データ生成処理手段により生成された複数の前記検査データの各々またはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行い、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる各特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかま

たは前記特徴パターン寄りの程度を示す値であると判定された前記特徴判定要素の値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記特徴判定要素の値を求め、求めた特徴判定要素の値を与えた前記検査データに対応する前記特徴パターン候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う

ことを特徴とする特徴パターン認識方法。

[38] 請求項26～31のいずれかに記載の特徴パターン認識方法において、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記特徴パターン候補配列を1離散シンボルずつずらしながら決定し、決定された複数の前記特徴パターン候補配列の長さをギャップの挿入または部分パターン以外の領域の離散シンボルの除去により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記特徴パターン候補配列のそれぞれを前記パターン全体用出現頻度テーブル記憶手段に記憶された前記パターン全体用出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行い、

前記分離処理手段は、

前記パターン全体用分離行列と前記検査データ生成処理手段により生成された複数の前記検査データの各々またはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行い、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる複数列の前記特徴要素の行列内位置に応じて定まる複数行の前

記特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で前記パターン全体用分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記類似度を示す値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記類似度を示す値を求め、求めた類似度を示す値を与えた前記検査データに対応する前記特徴パターン候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う

ことを特徴とする特徴パターン認識方法。

[39] 請求項26～31, 37, 38のいずれかに記載の特徴パターン認識方法において、

多重化された前記特徴パターンに含まれる前記各領域の部分パターンが、前記配列中の特定点を含む特定点近傍領域配列と相関している場合に、

前記各領域の部分パターンと前記特定点近傍領域配列との相関を判定するために用いられる情報として予め学習して得られた相関学習結果を含む情報を相関学習結果記憶手段に記憶させておき、

前記検査データ生成処理手段の前記部分パターン候補データ生成処理手段は、

前記各領域の部分パターンのそれぞれについて、前記特定点との相対的な位置関係または前記特定点との相対的な位置関係で定められた他の領域の部分パターン候補配列との相対的な位置関係で、前記複数の部分パターン候補配列を定める処理を行い、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

前記相関学習結果記憶手段に記憶された前記相関学習結果を含む情報を用いて、前記検査データ生成処理手段の前記部分パターン相当配列決定処理手段により決定された各領域の前記部分パターン相当配列と前記特定点近傍領域配列との間

に相関があるか否か、または相関の程度を判定し、相関がないか、または低いと判定された場合には、それらの各領域の前記部分パターン相当配列および前記特定点近傍領域配列中の前記特定点に基づく前記特徴パターン候補配列の決定処理は行わない

ことを特徴とする特徴パターン認識方法。

[40] 請求項39に記載の特徴パターン認識方法において、

前記相関学習結果記憶手段に、

前記既知の特徴パターンに含まれる複数の領域のうちの少なくとも一つの領域の既知の部分パターンおよび既知の特徴点近傍領域配列を結合して形成された複数種類の相関結合配列を用いて、これらの相関結合配列内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記相関結合配列内での各配列位置および前記離散シンボルの各種別とを対応させて形成された相関判定用出現頻度テーブルを記憶する相関判定用出現頻度テーブル記憶手段と、

前記各領域の部分パターンと前記特定点近傍領域配列との相関を判定する相関判定処理に用いられる相関判定用分離行列を記憶する相関判定用分離行列記憶手段とを設けておき、

前記相関判定用分離行列記憶手段には、

複数の領域のうちの少なくとも一つの領域の前記既知の部分パターンおよび前記既知の特定点近傍領域配列を結合して生成された相関結合配列、並びに複数の領域のうちの少なくとも一つの領域の前記既知の部分パターンおよび前記既知の特定点近傍領域配列とは異なる非特定点近傍領域配列を結合して形成された非相関結合配列を、前記相関判定用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して相関判定用学習データ行列を生成し、この相関判定用学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記相関結合配列の特徴を示す特徴要素を含む相関判定用基底行列の逆変換を行う行列として得られた相関判定用分離行列を記憶させておき、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

決定された各領域の前記部分パターン相当配列と前記特定点近傍領域配列との間に相関があるか否か、または相関の程度を判定する際には、複数の領域のうちの少なくとも一つの領域の前記部分パターン相当配列および前記特定点近傍領域配列を結合して相関判定配列を生成し、この相関判定配列を前記相関判定用出現頻度テーブル記憶手段に記憶された前記相関判定用出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化して相関判定データを生成した後、前記相関判定用分離行列記憶手段に記憶された前記相関判定用分離行列と、前記相関判定データまたは複数の前記相関判定データを結合した相関判定データ行列とを乗じる行列演算を行うことにより、相関判定用分離済データまたは複数の前記相関判定用分離済データが結合された相関判定用分離済データ行列を求める相関判定用分離処理を行い、この相関判定用分離処理で求めた前記相関判定用分離済データまたは前記相関判定用分離済データ行列を構成する要素のうち、前記相関判定用基底行列に含まれる前記特徴要素の行列内位置に応じて定まる相関判定要素の値を用いて、相関があるか否か、または相関の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[41] 請求項40に記載の特徴パターン認識方法において、

前記検査データ生成処理手段の前記特徴パターン候補配列決定処理手段は、

前記相関判定要素の値が予め定められた相関判定用閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定することにより、相関があるか否か、または相関の程度を判定する処理を行う

ことを特徴とする特徴パターン認識方法。

[42] 請求項25～41のいずれかに記載の特徴パターン認識方法において、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは、前記DNA配列中のプロモータである

ことを特徴とする特徴パターン認識方法。

[43] 請求項26～31，37～41のいずれかに記載の特徴パターン認識方法において、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは、前記DNA配列中のプロモータであり、前記部分パターンは、前記プロモータに含まれる－35ボックスおよび－10ボックスである

ことを特徴とする特徴パターン認識方法。

[44] 請求項39～41に記載の特徴パターン認識方法において、

前記配列は、DNA配列であり、前記離散シンボルは、前記DNA配列を構成する塩基を示すA、T、G、Cの記号またはこれらの代替記号であり、前記特徴パターンは、前記DNA配列中のプロモータであり、前記部分パターンは、前記プロモータに含まれる－35ボックスおよび－10ボックスであり、前記特定点は、前記DNA配列の転写開始点であり、前記特徴点近傍領域配列は、転写開始点近傍領域配列である

ことを特徴とする特徴パターン認識方法。

[45] 請求項25～41のいずれかに記載の特徴パターン認識方法において、

前記配列は、タンパク質を構成するアミノ酸配列であり、前記離散シンボルは、前記アミノ酸配列を構成するアミノ酸を示す記号であり、前記特徴パターンは、前記アミノ酸配列中のモチーフである

ことを特徴とする特徴パターン認識方法。

[46] 請求項25に記載の特徴パターン認識方法において、

前記特徴パターンの長さには、前記特徴パターンの種類の相違により複数種類の長さが存在し、

前記出現頻度テーブル記憶手段には、前記複数種類の既知の特徴パターンの長さをギャップの挿入により一定長に揃えた後、長さを揃えた前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記ギャップを含む前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記ギャップを含む前記離散シンボルの各種別とを対応させて形成された前記出現頻度テーブルを記憶させておくとともに、

前記分離行列記憶手段には、長さを揃えた前記複数種類の既知の特徴パターン

およびこれらと同じ長さに揃えた前記複数種類の非特徴パターンから生成された前記学習データ行列を用いて独立成分分析または主成分分析を行うことにより得られた前記分離行列を記憶させておき、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中で、前記特徴パターンの候補となる候補配列を決定した後、決定された前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた前記候補配列を前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、前記検査データを生成する処理を行う

ことを特徴とする特徴パターン認識方法。

[47] 請求項46に記載の特徴パターン認識方法において、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記候補配列のそれぞれを前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行い、

前記分離処理手段は、

前記分離行列と前記検査データ生成処理手段により生成された複数の前記検査データまたはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理を前記各検査配列毎に行い、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる特徴判定要素の値が、予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度を判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記特徴判定要素の値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記特徴判定要素の値を求め、求めた特徴判定要素の値を与えた前記検査データに対応する前記候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う

ことを特徴とする特徴パターン認識方法。

[48] 請求項46に記載の特徴パターン認識方法において、

前記検査データ生成処理手段は、

判定処理の対象となる前記配列若しくはその一部の配列断片から切り出された検査配列の中で、前記候補配列を1離散シンボルずつずらし、かつ各箇所で長さを変えながら決定し、決定された複数の前記候補配列の長さをギャップの挿入により一定長に揃えるアラインメント処理を行い、長さを揃えた複数の前記候補配列のそれぞれを前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記ギャップを含む前記離散シンボルの種別に従って数値化することにより、任意の一つの前記検査配列について複数の前記検査データを生成する処理を行い、

前記分離処理手段は、

前記分離行列と前記検査データ生成処理手段により生成された複数の前記検査データまたはこれらの複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、任意の一つの前記検査配列について複数の分離済データまたはこれらの複数の分離済データが結合された分離済データ行列を求める処理

を前記各検査配列毎に行い、

前記判定処理手段は、

前記分離処理手段により求めた任意の一つの前記検査配列についての前記複数の分離済データの各々または前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる複数列の前記特徴要素の行列内位置に応じて定まる複数行の前記特徴判定要素の値により構成される各列ベクトルのそれぞれについて、学習段階で前記分離行列とともに得られた分離済データ行列のうちの真データ対応部分の特徴判定要素の値により構成される各列ベクトルとの類似度を示す値を算出し、算出された各類似度を示す値が予め定められた閾値を境界としていずれの側の値であるかまたは値の大きさの程度をそれぞれ判定し、このうち前記特徴パターンであることを示す側の値であるかまたは前記特徴パターン寄りの程度を示す値であると判定された前記類似度を示す値について、前記閾値との差の絶対値が最も大きいかまたは前記特徴パターン寄りの程度が最も高い前記類似度を示す値を求め、求めた類似度を示す値を与えた前記検査データに対応する前記候補配列が、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンであると認識する処理を行う

ことを特徴とする特徴パターン認識方法。

[49] 有限個の離散シンボルにより構成された配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、既に発見されている互いに類似する複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かを判定する処理を行う特徴パターン認識システムとして、コンピュータを機能させるためのプログラムであって、

前記複数種類の既知の特徴パターンを用いてこれらの特徴パターン内での各配列位置毎に前記離散シンボルの各種別毎の出現頻度を求め、これらの出現頻度と前記特徴パターン内での各配列位置および前記離散シンボルの各種別とを対応させて形成された出現頻度テーブルを記憶する出現頻度テーブル記憶手段と、

前記複数種類の既知の特徴パターンおよびこれらの既知の特徴パターンとは異な

る複数種類の非特徴パターンを、前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従ってそれぞれ数値化した後、これらを結合して学習データ行列を生成し、この学習データ行列を用いて独立成分分析または主成分分析を行うことにより、前記特徴パターンの特徴を示す特徴要素を含む基底行列の逆変換を行う行列として得られた分離行列を記憶する分離行列記憶手段と、

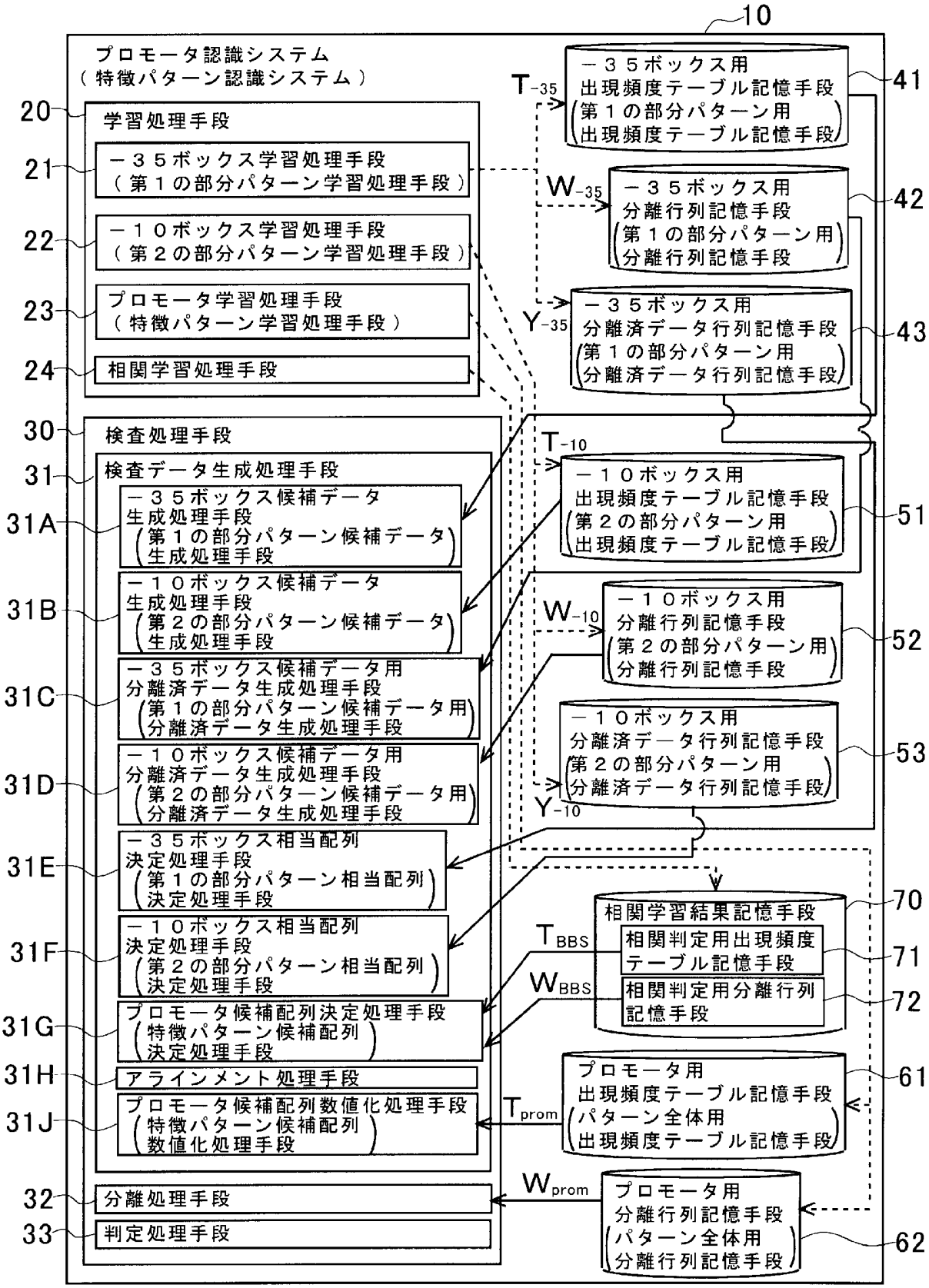
判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列を、前記出現頻度テーブル記憶手段に記憶された前記出現頻度テーブルを用いて前記配列位置および前記離散シンボルの種別に従って数値化することにより、検査データを生成する処理を行う検査データ生成処理手段と、

前記分離行列記憶手段に記憶された前記分離行列と前記検査データ生成処理手段により生成された前記検査データまたは複数の前記検査データを結合した検査データ行列とを乗じる行列演算を行うことにより、分離済データまたは複数の前記分離済データが結合された分離済データ行列を求める処理を行う分離処理手段と、

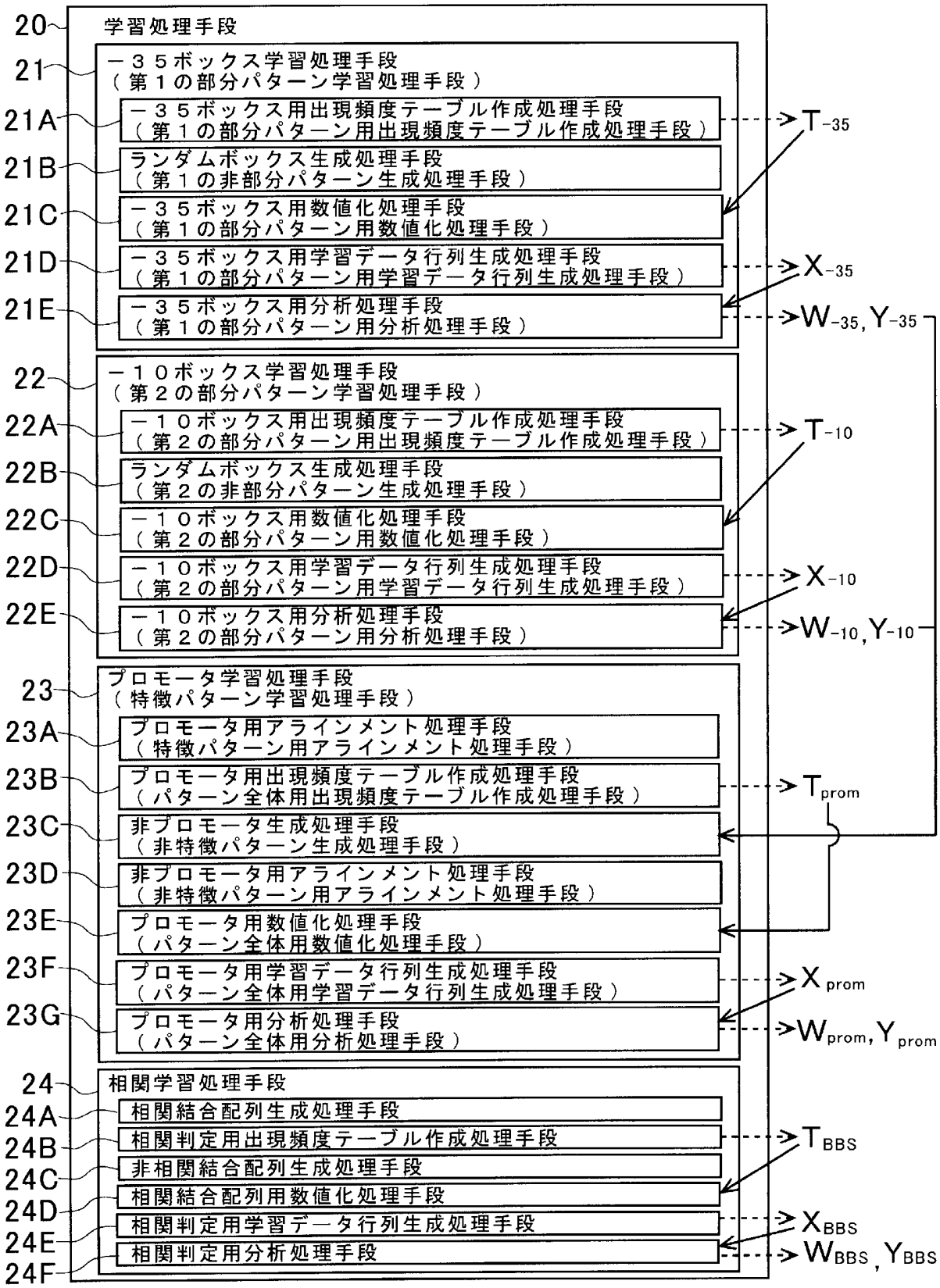
この分離処理手段により求めた前記分離済データまたは前記分離済データ行列を構成する要素のうち、前記基底行列に含まれる前記特徴要素の行列内位置に応じて定まる特徴判定要素の値を用いて、判定処理の対象となる前記配列若しくはその一部の配列断片またはこれらの配列若しくは配列断片から切り出された検査配列の中に、前記複数種類の既知の特徴パターンのうちのいずれかの特徴パターンまたは前記既知の特徴パターンに類似する新規な特徴パターンが含まれるか否かまたは存否の程度を判定する処理を行う判定処理手段と

を備えたことを特徴とする特徴パターン認識システムとして、コンピュータを機能させるためのプログラム。

[図1]

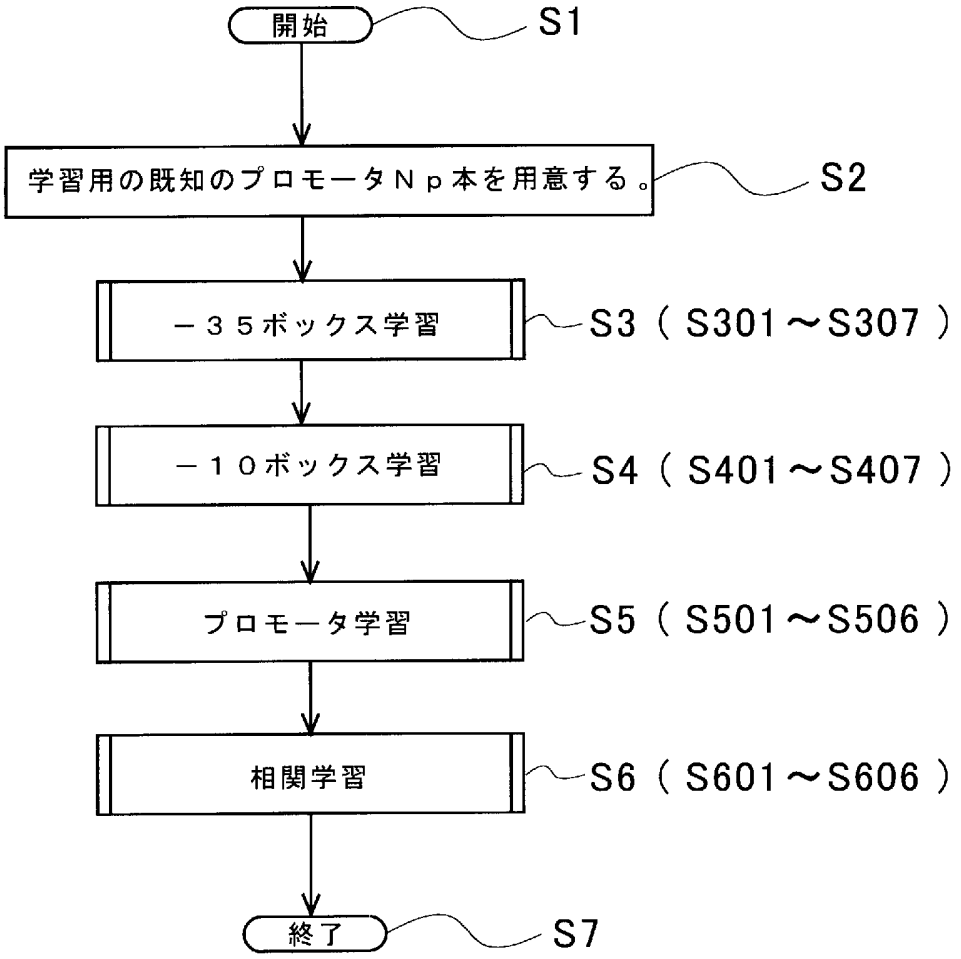

10
プロモータ認識システム
（特徴パターン認識システム）
20
学習処理手段
21
－３５ボックス学習処理手段
（第１の部分パターン学習処理手段）
22
－１０ボックス学習処理手段
（第２の部分パターン学習処理手段）
23
プロモータ学習処理手段
（特徴パターン学習処理手段）
24
相関学習処理手段
30
検査処理手段
31
検査データ生成処理手段
31A
－３５ボックス候補データ
生成処理手段
（第１の部分パターン候補データ
生成処理手段）
31B
－１０ボックス候補データ
生成処理手段
（第２の部分パターン候補データ
生成処理手段）
31C
－３５ボックス候補データ用
分離済データ生成処理手段
（第１の部分パターン候補データ用
分離済データ生成処理手段）
31D
－１０ボックス候補データ用
分離済データ生成処理手段
（第２の部分パターン候補データ用
分離済データ生成処理手段）
31E
－３５ボックス相当配列
決定処理手段
（第１の部分パターン相当配列
決定処理手段）
31F
－１０ボックス相当配列
決定処理手段
（第２の部分パターン相当配列
決定処理手段）
31G
プロモータ候補配列決定処理手段
（特徴パターン候補配列
決定処理手段）
31H
アラインメント処理手段
31J
プロモータ候補配列数値化処理手段
（特徴パターン候補配列
数値化処理手段）
32
分離処理手段
33
判定処理手段
T-35
－３５ボックス用
出現頻度テーブル記憶手段
（第１の部分パターン用
出現頻度テーブル記憶手段）
41
W-35
－３５ボックス用
分離行列記憶手段
（第１の部分パターン用
分離行列記憶手段）
42
Y-35
－３５ボックス用
分離済データ行列記憶手段
（第１の部分パターン用
分離済データ行列記憶手段）
43
T-10
－１０ボックス用
出現頻度テーブル記憶手段
（第２の部分パターン用
出現頻度テーブル記憶手段）
51
W-10
－１０ボックス用
分離行列記憶手段
（第２の部分パターン用
分離行列記憶手段）
52
－１０ボックス用
分離済データ行列記憶手段
（第２の部分パターン用
分離済データ行列記憶手段）
53
Y-10
相関学習結果記憶手段
70
TBBS
相関判定用出現頻度
テーブル記憶手段
71
WBBS
相関判定用分離行列
記憶手段
72
Tprom
プロモータ用
出現頻度テーブル記憶手段
（パターン全体用
出現頻度テーブル記憶手段）
61
Wprom
プロモータ用
分離行列記憶手段
（パターン全体用
分離行列記憶手段）
62

[図2]


20
学習処理手段
21
－３５ボックス学習処理手段
（第１の部分パターン学習処理手段）
21A
－３５ボックス用出現頻度テーブル作成処理手段
（第１の部分パターン用出現頻度テーブル作成処理手段）
21B
ランダムボックス生成処理手段
（第１の非部分パターン生成処理手段）
21C
－３５ボックス用数値化処理手段
（第１の部分パターン用数値化処理手段）
21D
－３５ボックス用学習データ行列生成処理手段
（第１の部分パターン用学習データ行列生成処理手段）
21E
－３５ボックス用分析処理手段
（第１の部分パターン用分析処理手段）
22
－１０ボックス学習処理手段
（第２の部分パターン学習処理手段）
22A
－１０ボックス用出現頻度テーブル作成処理手段
（第２の部分パターン用出現頻度テーブル作成処理手段）
22B
ランダムボックス生成処理手段
（第２の非部分パターン生成処理手段）
22C
－１０ボックス用数値化処理手段
（第２の部分パターン用数値化処理手段）
22D
－１０ボックス用学習データ行列生成処理手段
（第２の部分パターン用学習データ行列生成処理手段）
22E
－１０ボックス用分析処理手段
（第２の部分パターン用分析処理手段）
23
プロモータ学習処理手段
（特徴パターン学習処理手段）
23A
プロモータ用アラインメント処理手段
（特徴パターン用アラインメント処理手段）
23B
プロモータ用出現頻度テーブル作成処理手段
（パターン全体用出現頻度テーブル作成処理手段）
23C
非プロモータ生成処理手段
（非特徴パターン生成処理手段）
23D
非プロモータ用アラインメント処理手段
（非特徴パターン用アラインメント処理手段）
23E
プロモータ用数値化処理手段
（パターン全体用数値化処理手段）
23F
プロモータ用学習データ行列生成処理手段
（パターン全体用学習データ行列生成処理手段）
23G
プロモータ用分析処理手段
（パターン全体用分析処理手段）
24
相関学習処理手段
24A
相関結合配列生成処理手段
24B
相関判定用出現頻度テーブル作成処理手段
24C
非相関結合配列生成処理手段
24D
相関結合配列用数値化処理手段
24E
相関判定用学習データ行列生成処理手段
24F
相関判定用分析処理手段
T-35
X-35
W-35, Y-35
T-10
X-10
W-10, Y-10
Tprom
X prom
Wprom, Yprom
TBBS
XBBS
WBBS , YBBS

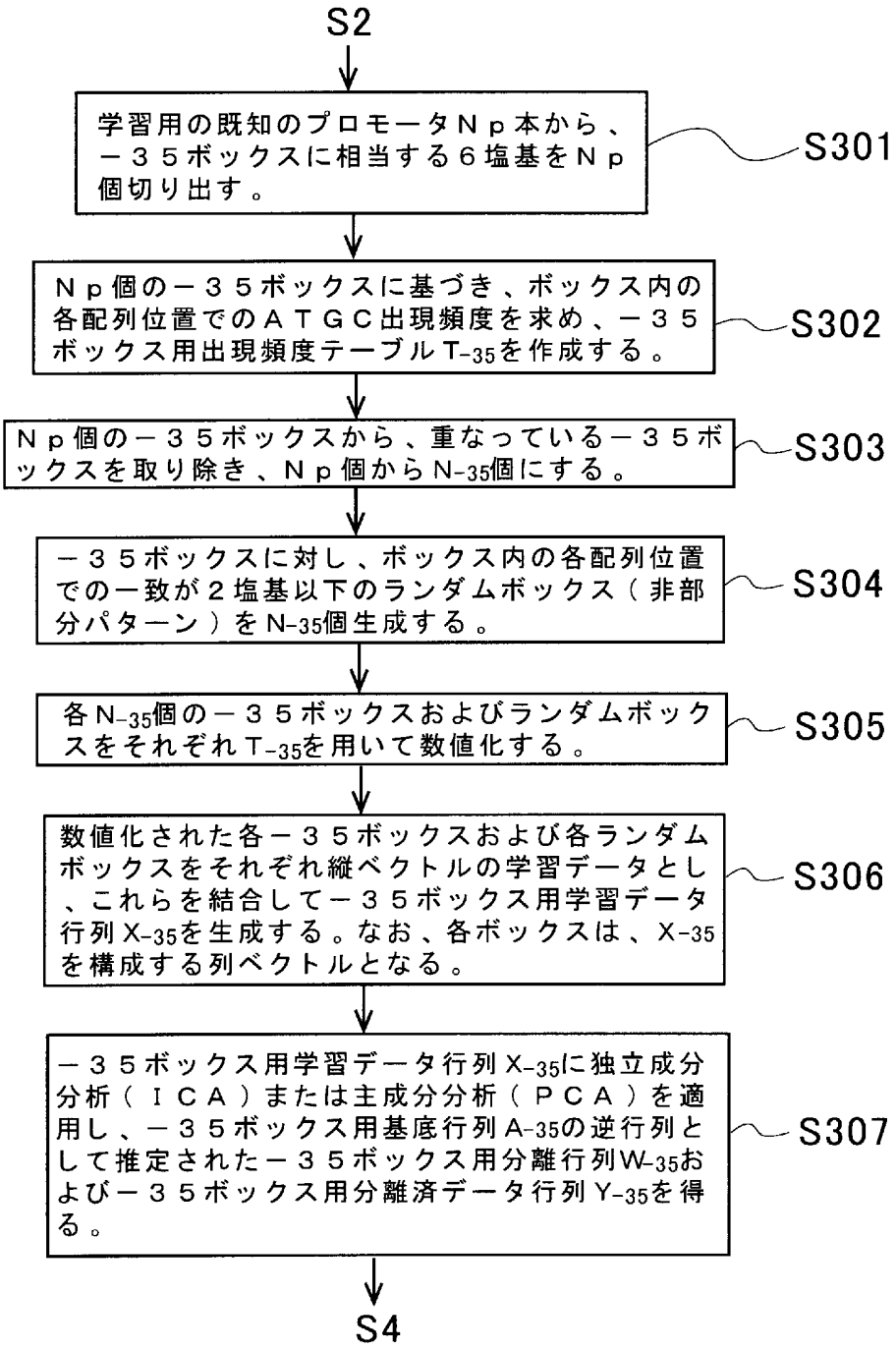
[図3]

学習段階（トレーニングステップ）の全体の流れ

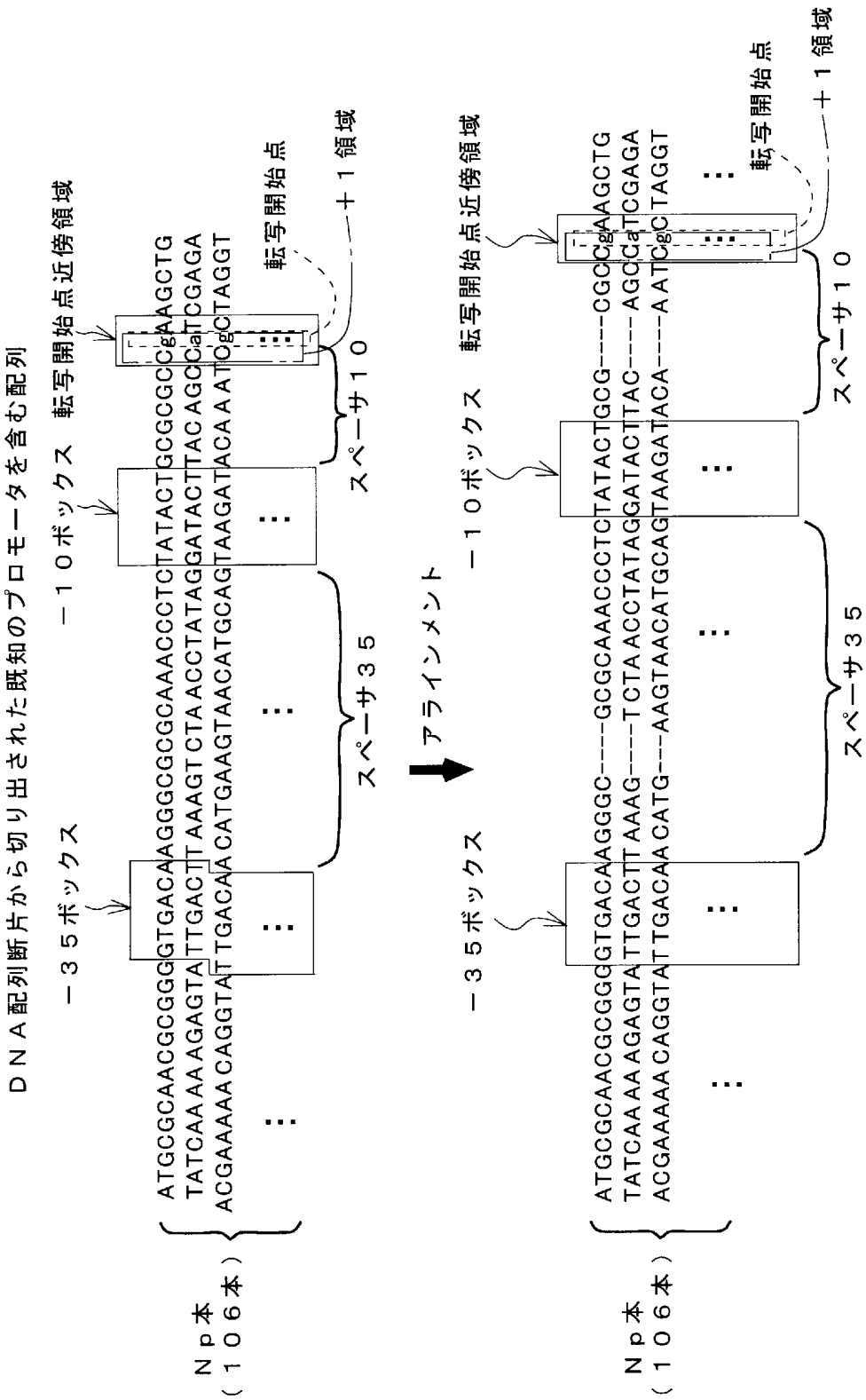


[図4]

－３５ボックス学習の流れ



[図5]

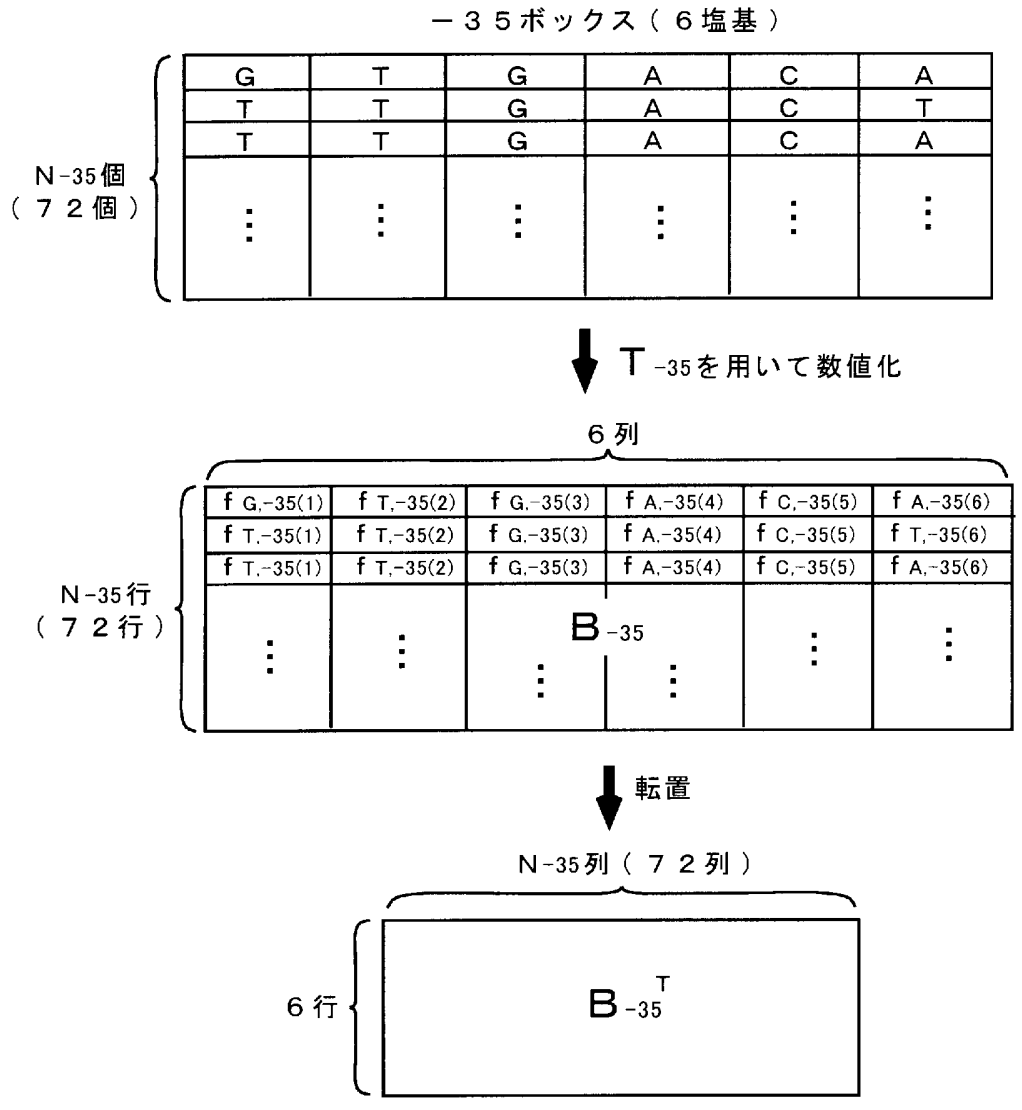

ＤＮＡ配列断片から切り出された既知のプロモータを含む配列
－３５ボックス
－１０ボックス
転写開始点近傍領域
転写開始点
＋１領域
スペーサ１０
スペーサ３５
ＮＰ本
（１０６本）
ATGCGCAACGCGGGGTGACAAGGGCGCGCAAACCCTCTATACTGCGCGCCGAAGCTG
TATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCATCGAGA
ACGAAAAACAGGTATTTGACAACATGAAGTAACATGCAGTAAGATACAAAATCGCTAGGT
アラインメント

[図6]

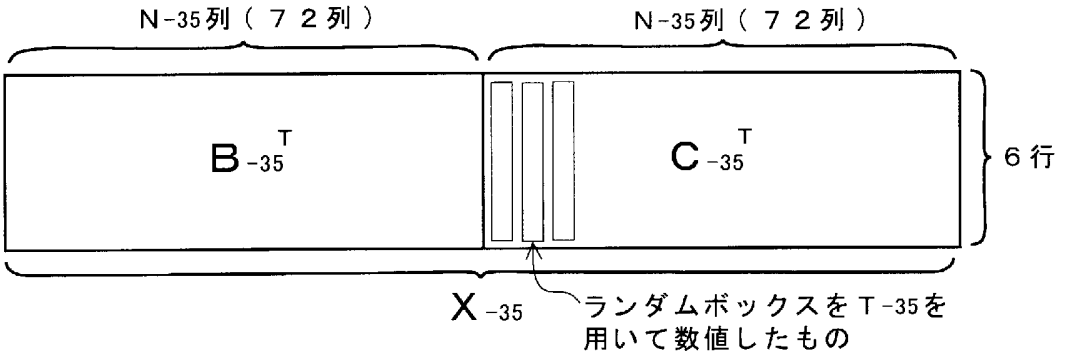
T_{-35} ：－３５ボックス用出現頻度テーブル

ボックス内の配列位置 / 塩基種別	1	2	3	4	5	6
A	$f_{A,-35(1)}$	$f_{A,-35(2)}$	$f_{A,-35(3)}$	$f_{A,-35(4)}$	$f_{A,-35(5)}$	$f_{A,-35(6)}$
T	$f_{T,-35(1)}$	$f_{T,-35(2)}$	$f_{T,-35(3)}$	$f_{T,-35(4)}$	$f_{T,-35(5)}$	$f_{T,-35(6)}$
G	$f_{G,-35(1)}$	$f_{G,-35(2)}$	$f_{G,-35(3)}$	$f_{G,-35(4)}$	$f_{G,-35(5)}$	$f_{G,-35(6)}$
C	$f_{C,-35(1)}$	$f_{C,-35(2)}$	$f_{C,-35(3)}$	$f_{C,-35(4)}$	$f_{C,-35(5)}$	$f_{C,-35(6)}$

[図7]

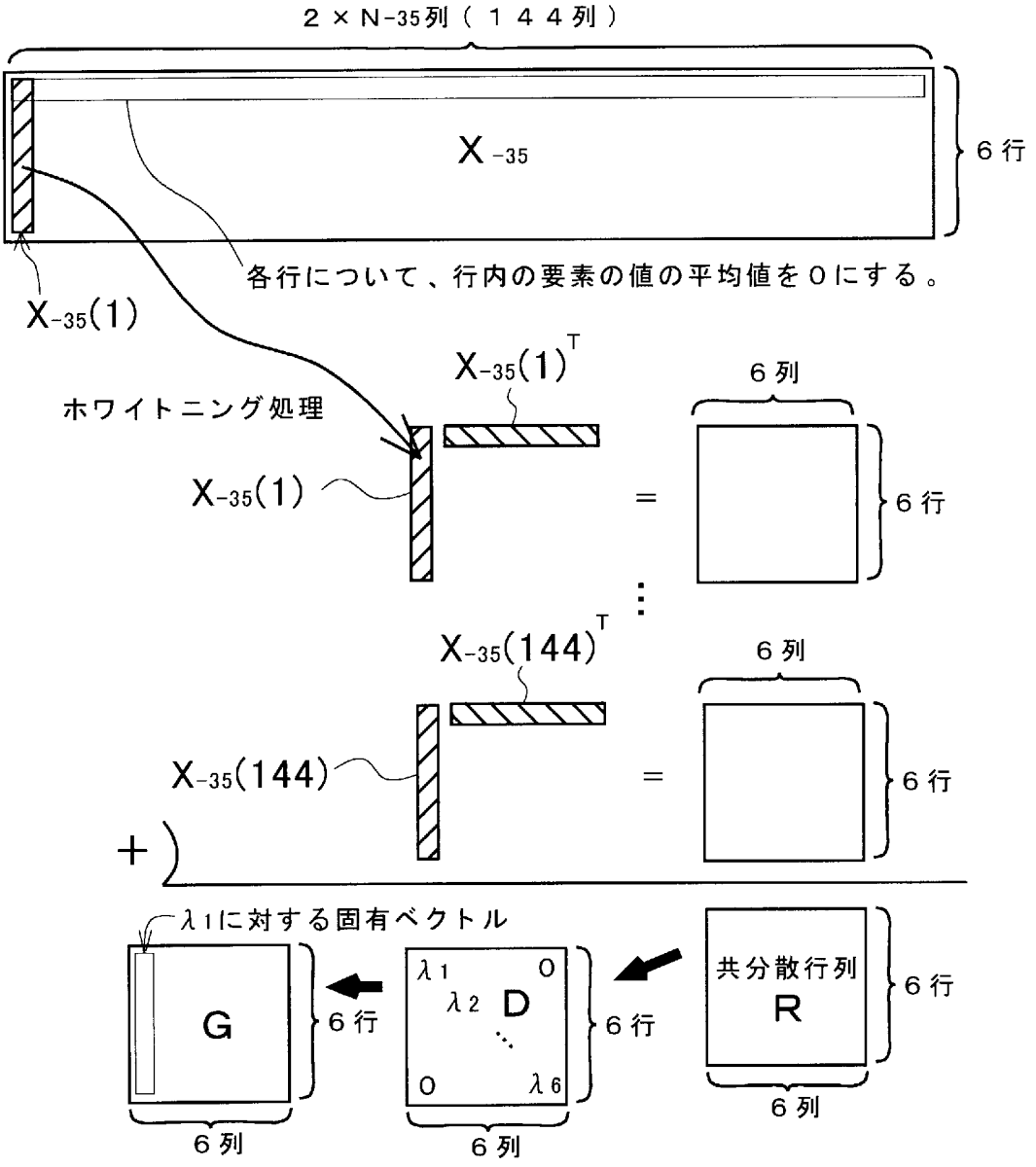


[図8]

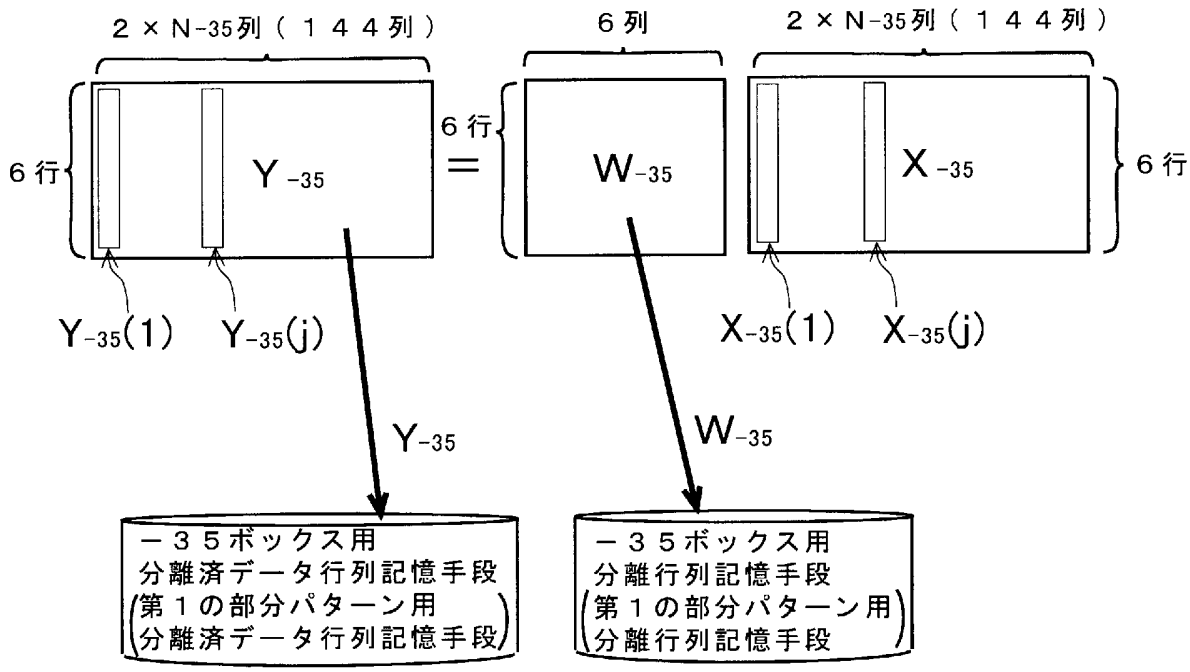

N-35列（７２列）
N-35列（７２列）
B -35 T
C -35 T
６行
X -35
ランダムボックスをT-35を用いて数値したもの

[図9]

前処理の流れ

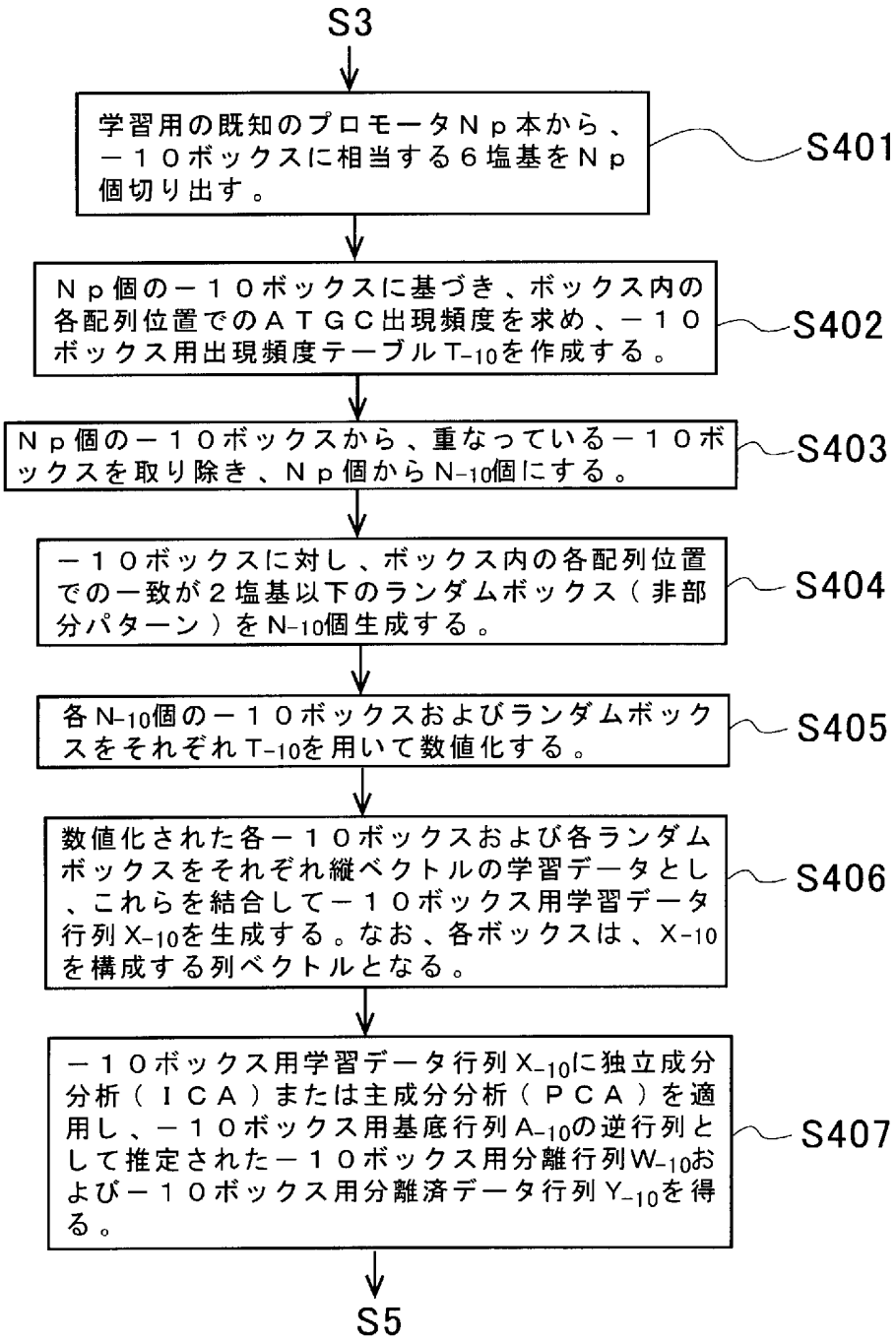

２×N-35列（１４４列）
X -35
６行
各行について、行内の要素の値の平均値を０にする。
X-35(1)
X-35(1) T
６列
ホワイトニング処理
X-35(1)
＝
６行
X-35(144) T
６列
X-35(144)
＝
６行
＋
λ1に対する固有ベクトル
G
６行
６列
λ1
λ2
D
λ6
０
０
６行
６列
共分散行列
R
６行
６列

[図10]


２×N-35列（１４４列）
６列
２×N-35列（１４４列）
６行
Y-35
６行
＝
W-35
X-35
６行
Y-35(1)
Y-35(j)
X-35(1)
X-35(j)
Y-35
W-35
－３５ボックス用
分離済データ行列記憶手段
（第１の部分パターン用
分離済データ行列記憶手段）
－３５ボックス用
分離行列記憶手段
（第１の部分パターン用
分離行列記憶手段）

[図11]

－１０ボックス学習の流れ

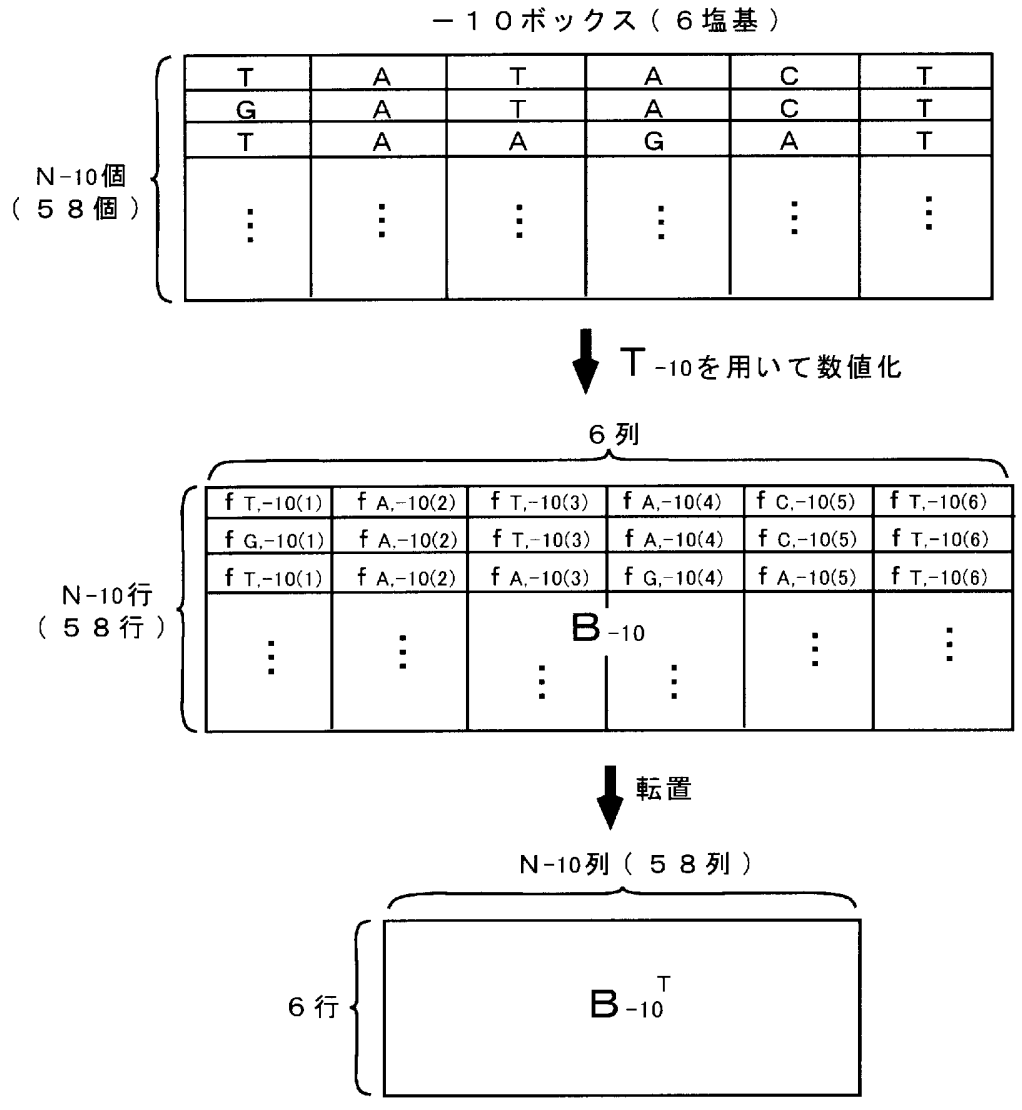


[図12]

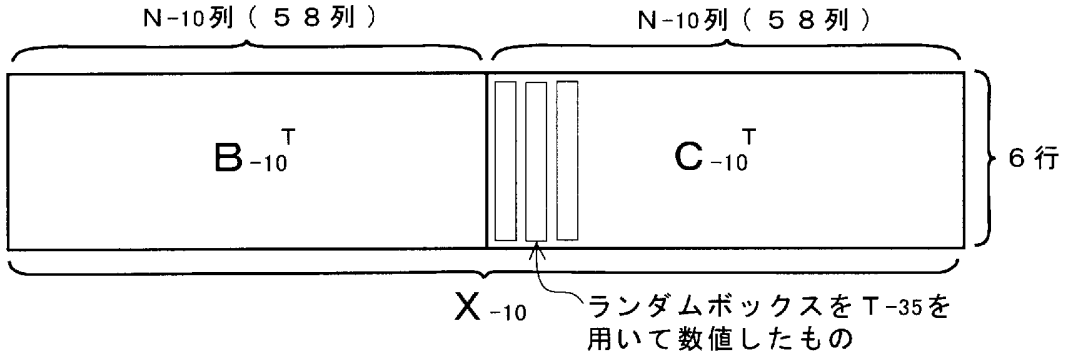
T_{-10} ：－１０ボックス用出現頻度テーブル

塩基種別＼ボックス内の配列位置	1	2	3	4	5	6
A	$f_{A,-10(1)}$	$f_{A,-10(2)}$	$f_{A,-10(3)}$	$f_{A,-10(4)}$	$f_{A,-10(5)}$	$f_{A,-10(6)}$
T	$f_{T,-10(1)}$	$f_{T,-10(2)}$	$f_{T,-10(3)}$	$f_{T,-10(4)}$	$f_{T,-10(5)}$	$f_{T,-10(6)}$
G	$f_{G,-10(1)}$	$f_{G,-10(2)}$	$f_{G,-10(3)}$	$f_{G,-10(4)}$	$f_{G,-10(5)}$	$f_{G,-10(6)}$
C	$f_{C,-10(1)}$	$f_{C,-10(2)}$	$f_{C,-10(3)}$	$f_{C,-10(4)}$	$f_{C,-10(5)}$	$f_{C,-10(6)}$

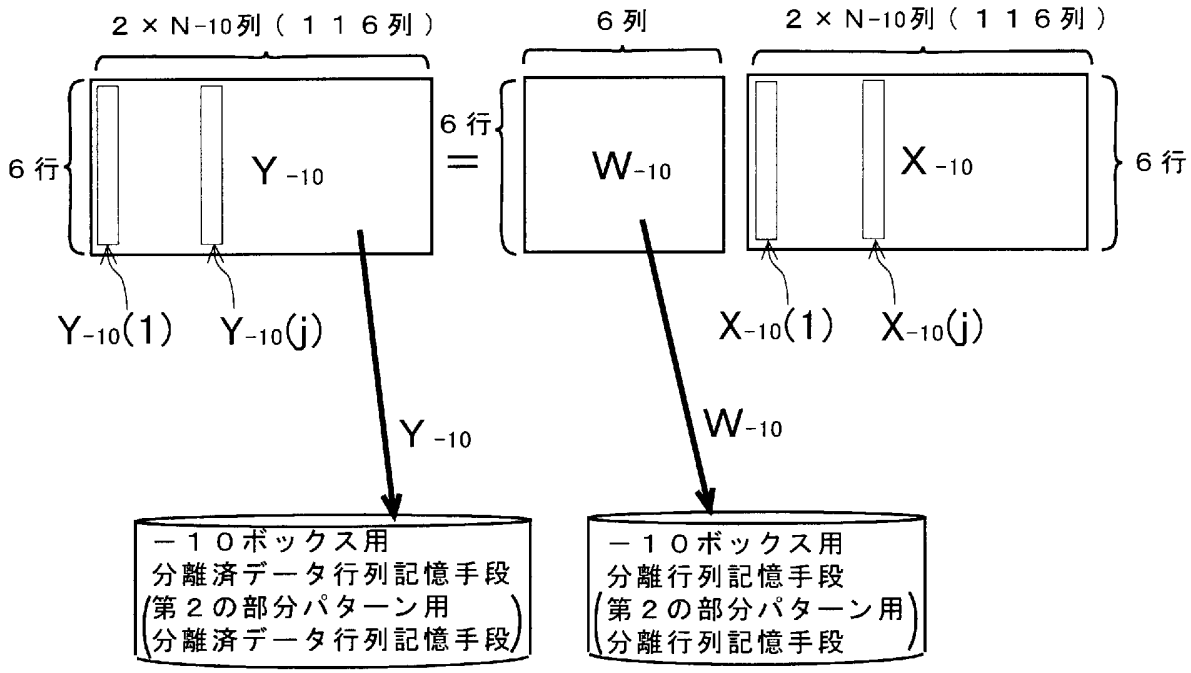
[図13]



[図14]

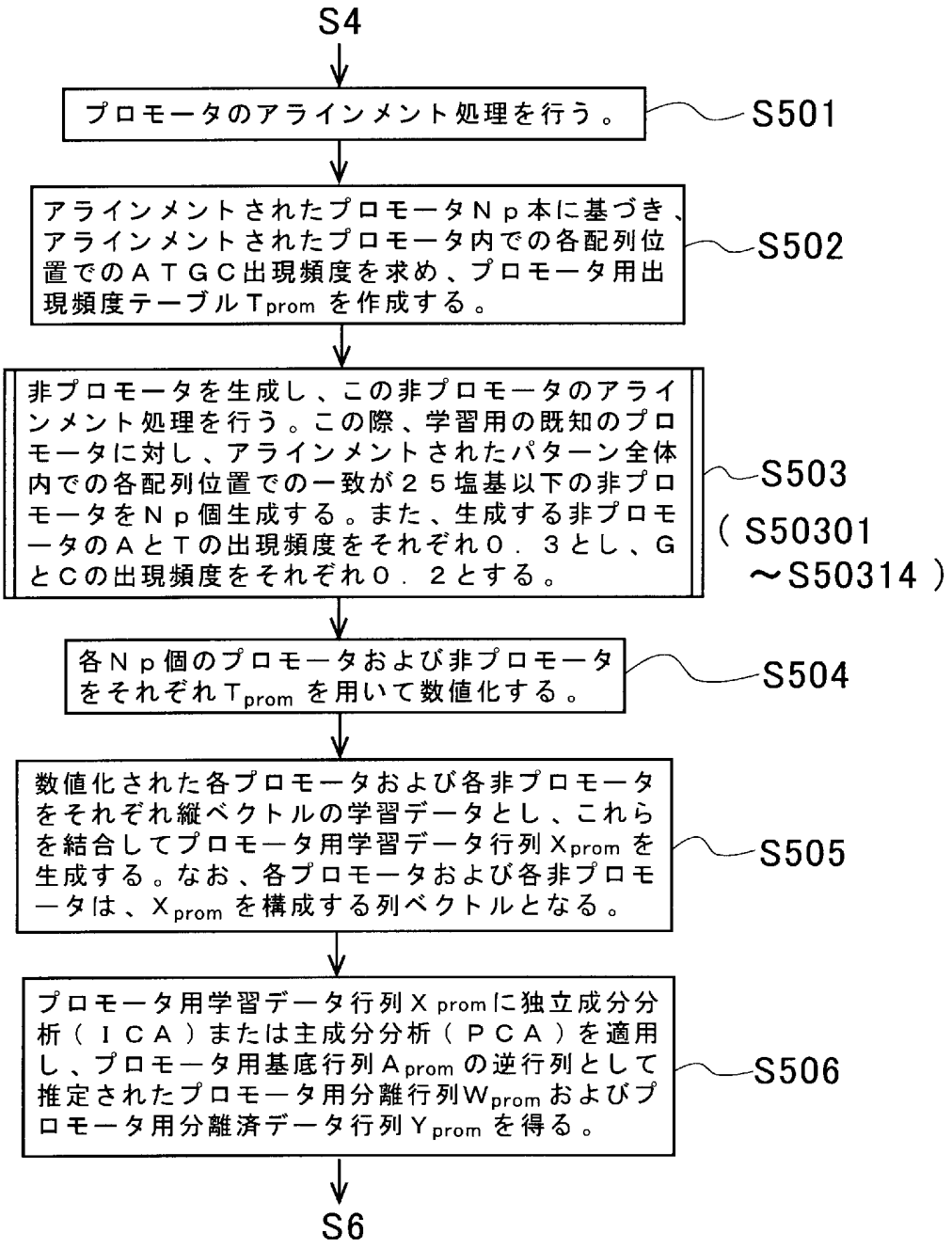

N-10列（５８列）
N-10列（５８列）
B -10 T
C -10 T
６行
X -10
ランダムボックスをT-35を
用いて数値したもの

[図15]


２×N-10列（１１６列）
６列
２×N-10列（１１６列）
６行
６行
Y -10
＝
W-10
X -10
６行
Y-10(1)
Y-10(j)
X-10(1)
X-10(j)
Y -10
W-10
－１０ボックス用
分離済データ行列記憶手段
（第２の部分パターン用
分離済データ行列記憶手段）
－１０ボックス用
分離行列記憶手段
（第２の部分パターン用
分離行列記憶手段）

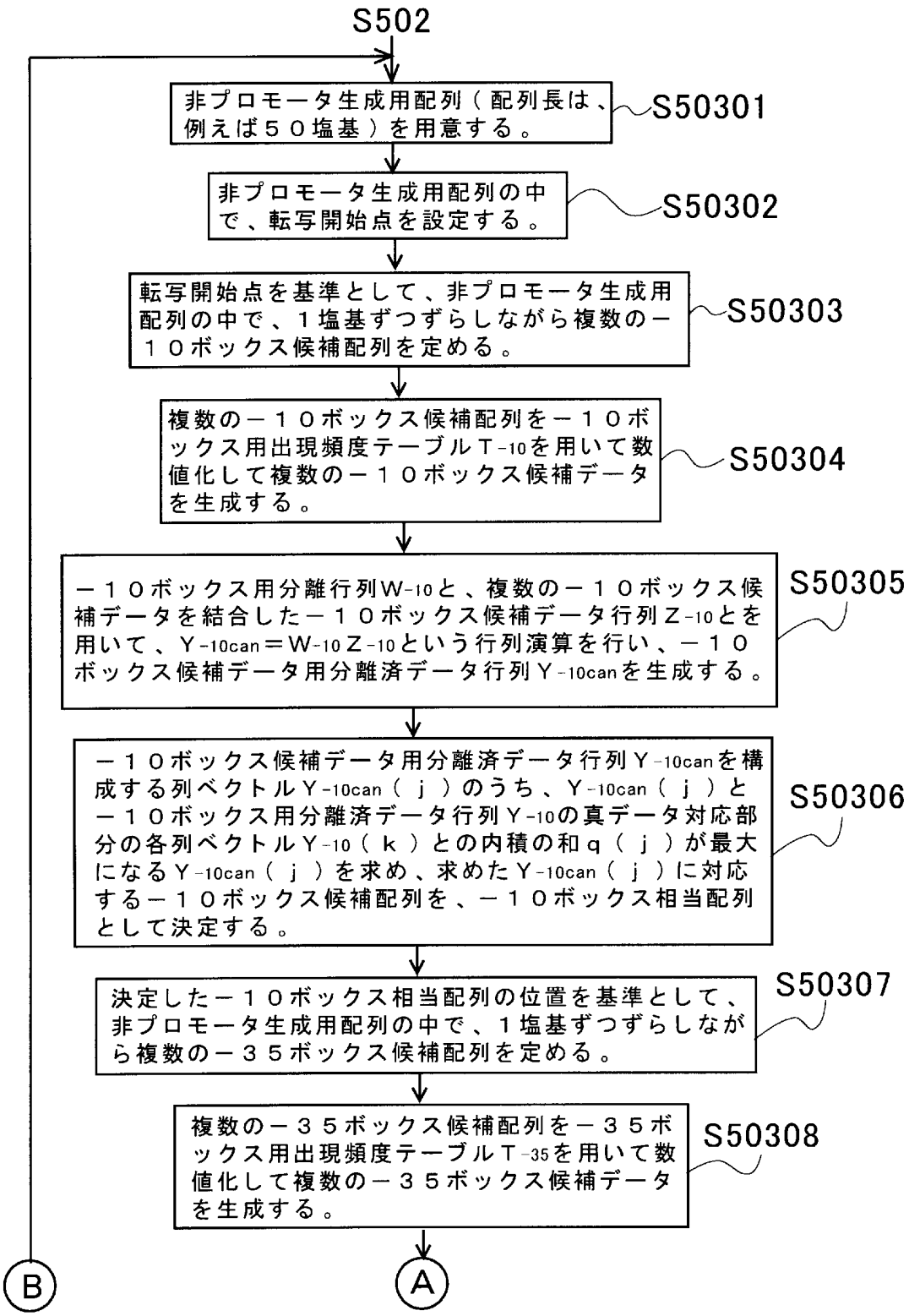
[図16]

プロモータ学習の流れ



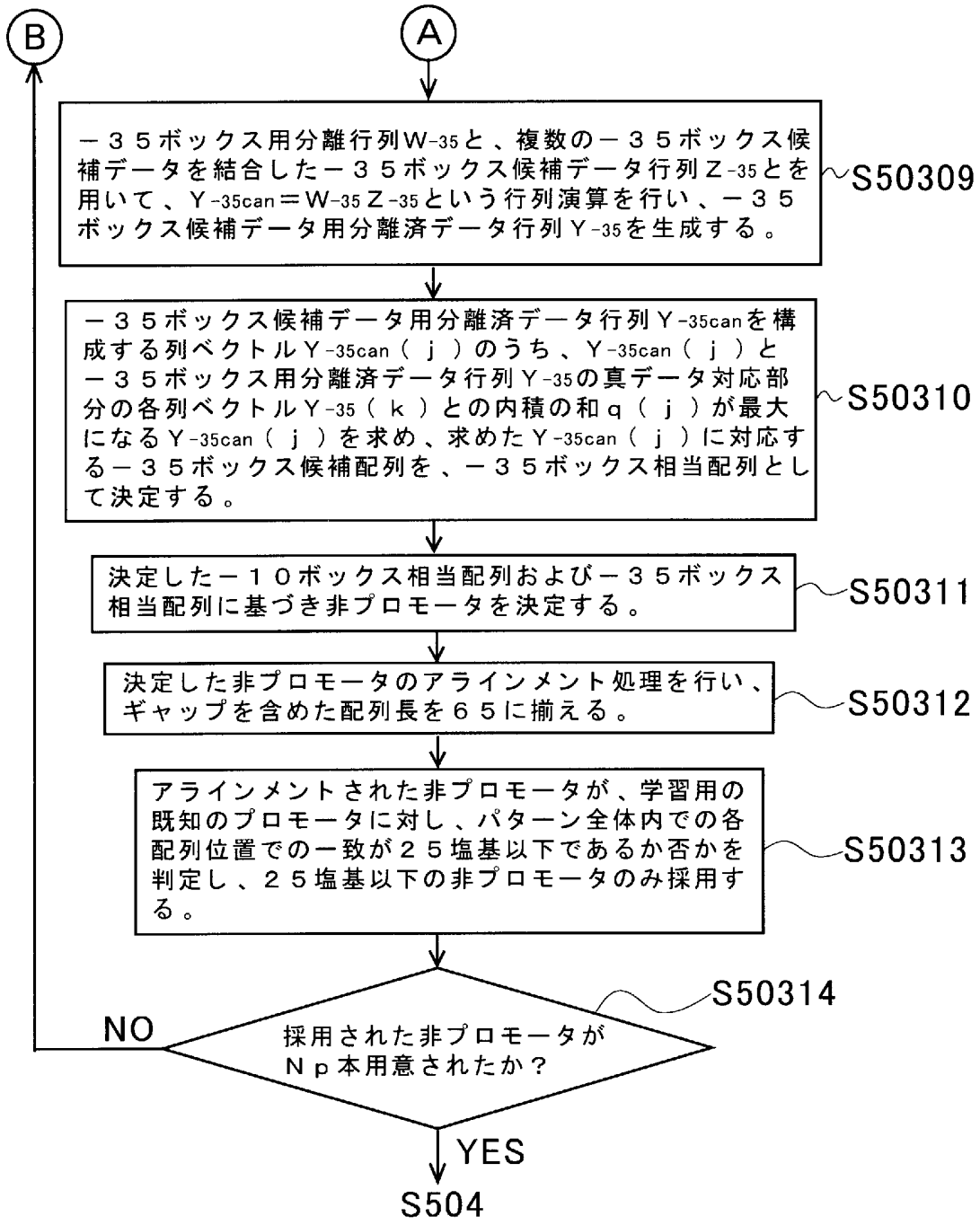
[図17]

プロモータ学習における非プロモータの生成および
生成した非プロモータのアラインメント処理の流れ（その１）



[図18]

プロモータ学習における非プロモータの生成および
生成した非プロモータのアラインメント処理の流れ（その２）

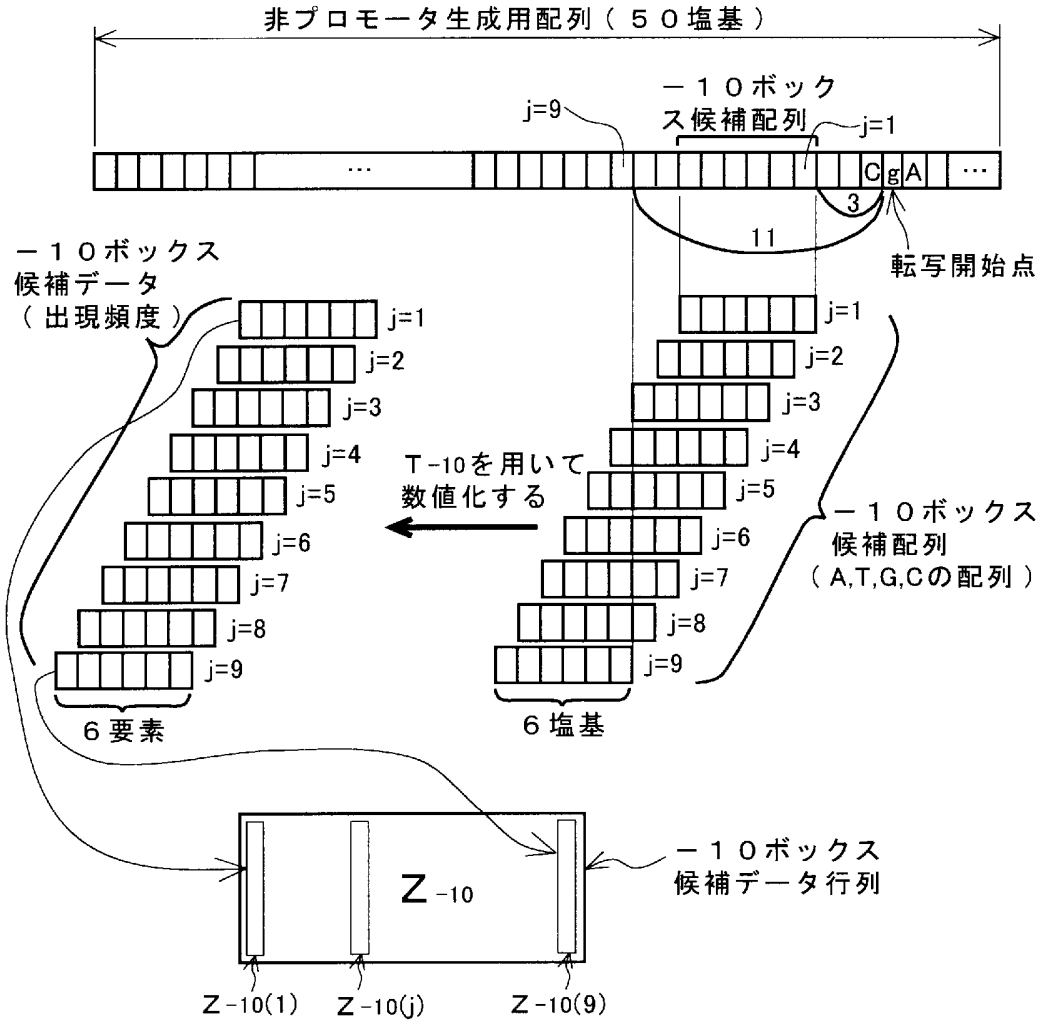


[図19]

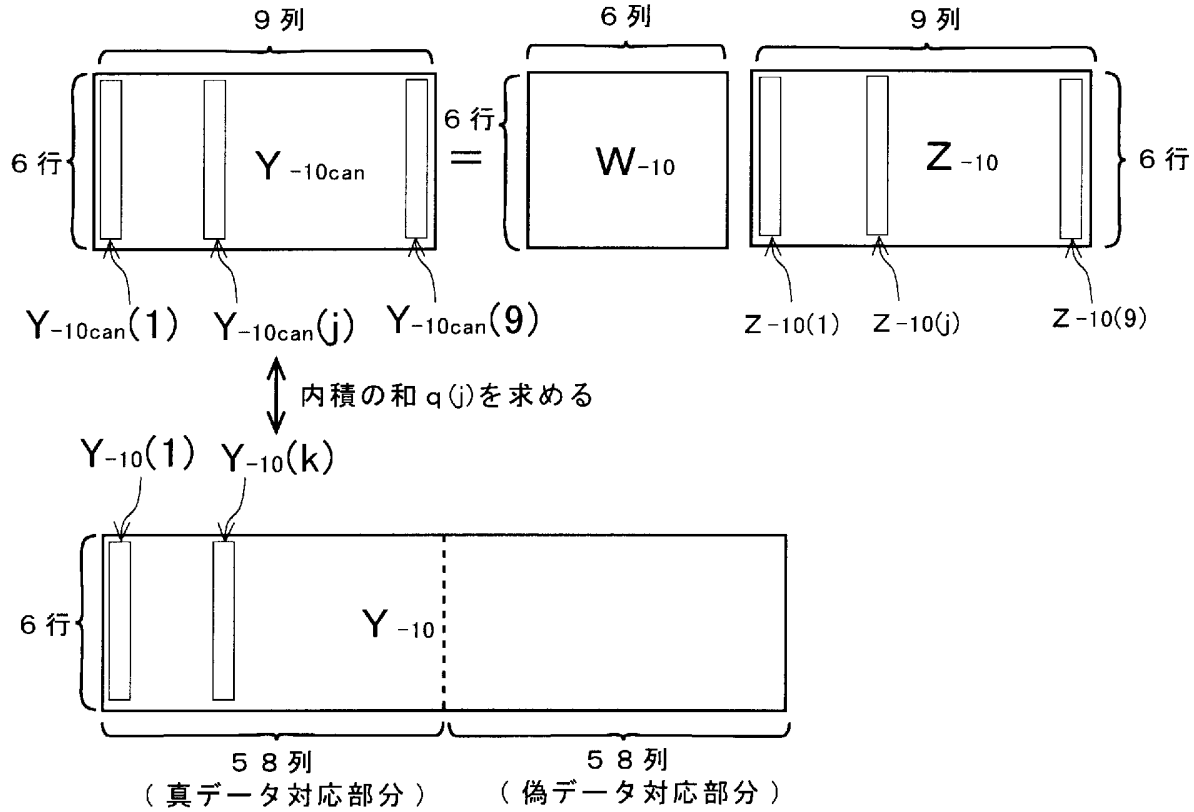
T_{prom}：プロモータ用出現頻度テーブル

パターン全体内の配列位置 / ギャップを含むシンボル種別	1	2	3	…	64	65
A	$f_{A,prom(1)}$	$f_{A,prom(2)}$	$f_{A,prom(3)}$	…	$f_{A,prom(64)}$	$f_{A,prom(65)}$
T	$f_{T,prom(1)}$	$f_{T,prom(2)}$	$f_{T,prom(3)}$	…	$f_{T,prom(64)}$	$f_{T,prom(65)}$
G	$f_{G,prom(1)}$	$f_{G,prom(2)}$	$f_{G,prom(3)}$	…	$f_{G,prom(64)}$	$f_{G,prom(65)}$
C	$f_{C,prom(1)}$	$f_{C,prom(2)}$	$f_{C,prom(3)}$	…	$f_{C,prom(64)}$	$f_{C,prom(65)}$
－	$f_{-,prom(1)}$	$f_{-,prom(2)}$	$f_{-,prom(3)}$	…	$f_{-,prom(64)}$	$f_{-,prom(65)}$

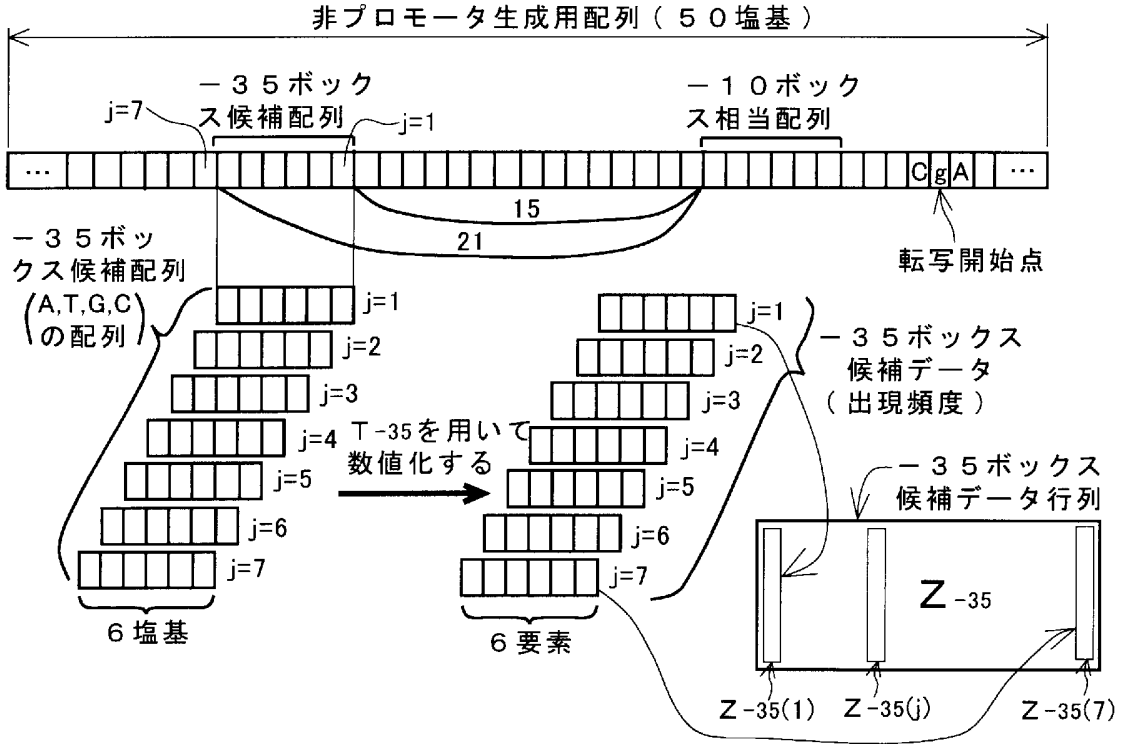
[図20]



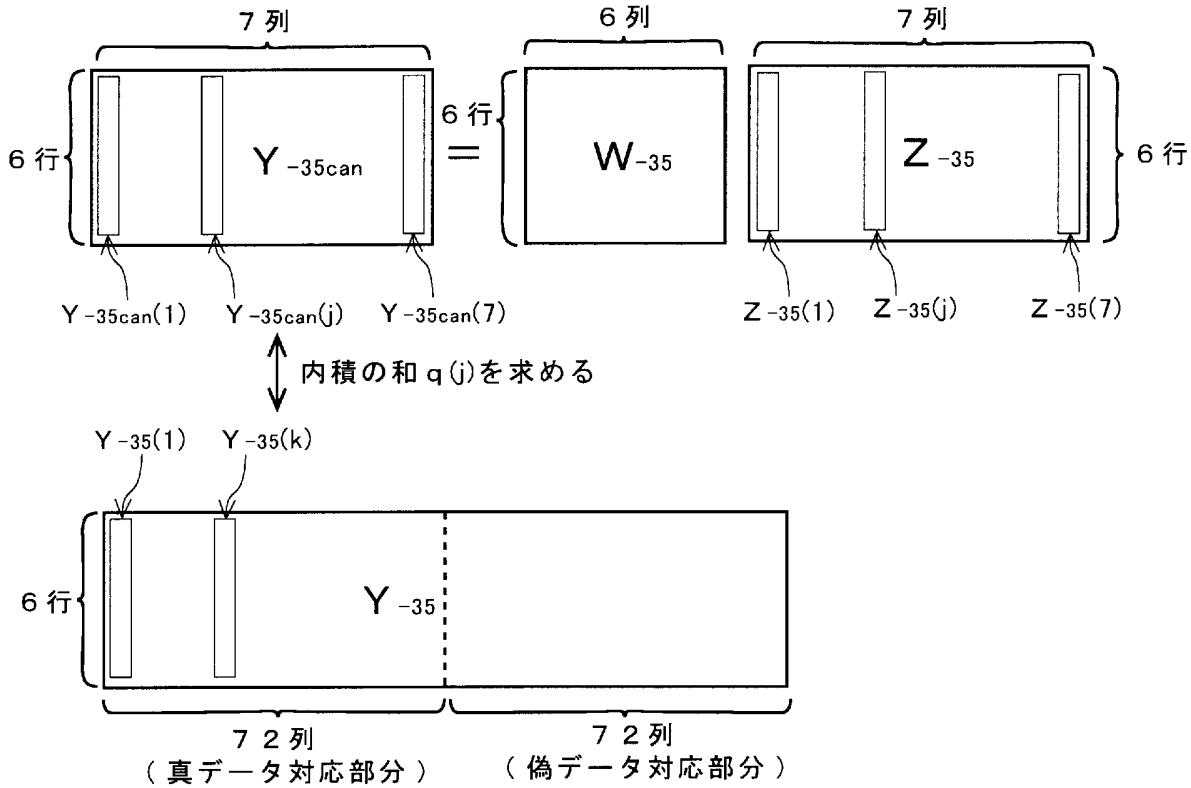
[図21]


９列
６列
９列
６行
Y -10can
６行
＝
W-10
Z -10
６行
Y-10can(1)
Y-10can(j)
Y-10can(9)
Z -10(1)
Z -10(j)
Z -10(9)
内積の和ｑ(j)を求める
Y-10(1)
Y-10(k)
６行
Y -10
５８列
（真データ対応部分）
５８列
（偽データ対応部分）

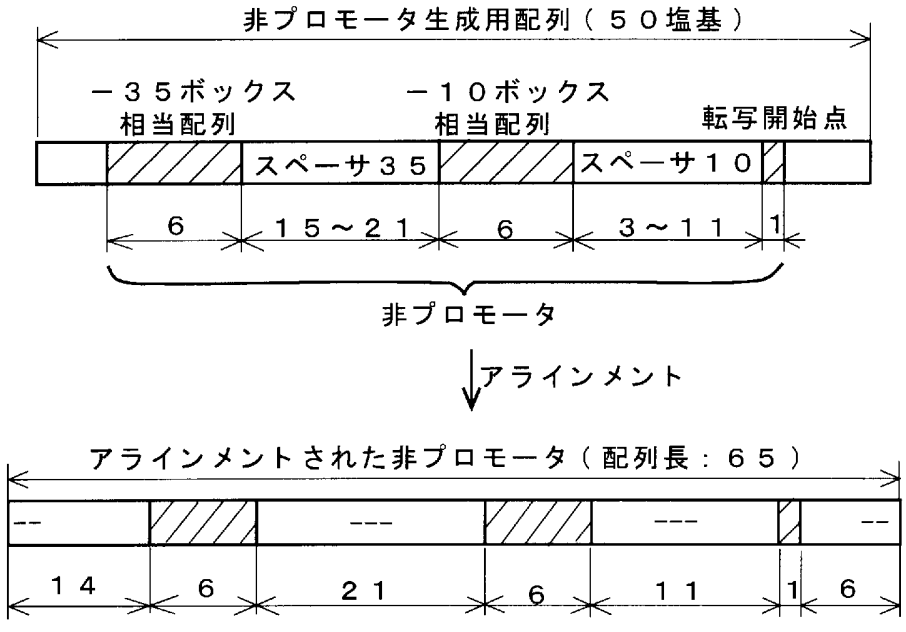
[図22]


非プロモータ生成用配列（５０塩基）
－３５ボックス候補配列
－１０ボックス相当配列
j=7
j=1
…
C
g
A
…
15
21
転写開始点
－３５ボックス候補配列
(A,T,G,C)の配列
j=1
j=2
j=3
j=4
j=5
j=6
j=7
６塩基
T -35を用いて数値化する
６要素
－３５ボックス候補データ（出現頻度）
－３５ボックス候補データ行列
Z -35
Z -35(1)
Z -35(j)
Z -35(7)

[図23]

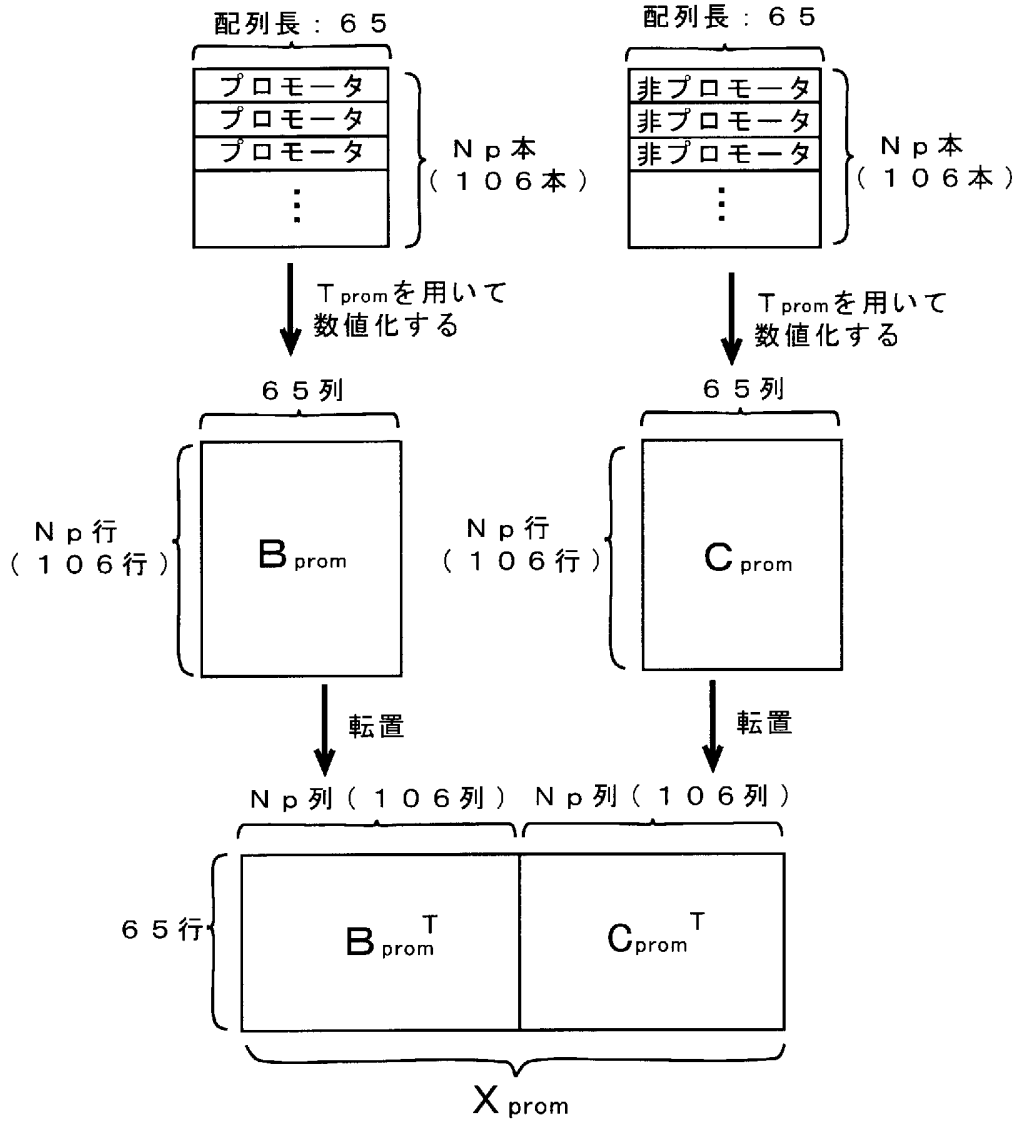

７列
６列
７列
６行
６行
６行
Y -35can
W-35
Z -35
Y -35can(1)
Y -35can(j)
Y -35can(7)
Z -35(1)
Z -35(j)
Z -35(7)
内積の和ｑ(j)を求める
Y -35(1)
Y -35(k)
６行
Y -35
７２列
（真データ対応部分）
７２列
（偽データ対応部分）

[図24]

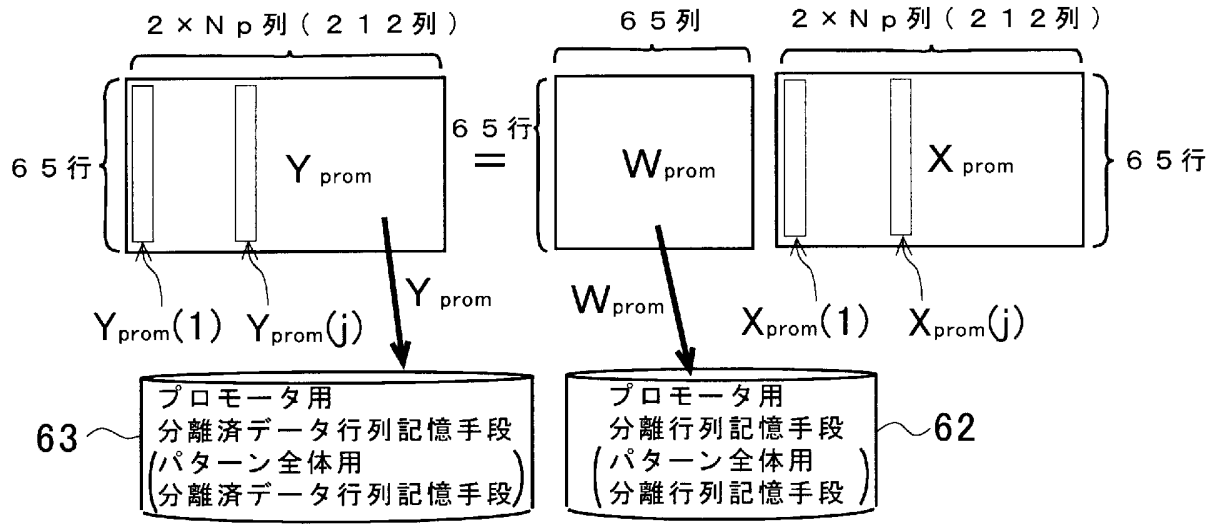

非プロモータ生成用配列（５０塩基）
－３５ボックス
相当配列
－１０ボックス
相当配列
転写開始点
スペーサ３５
スペーサ１０
6
１５～２１
6
３～１１
1
非プロモータ
アラインメント
アラインメントされた非プロモータ（配列長：６５）
--

--
１４
6
２１
6
１１
1
6

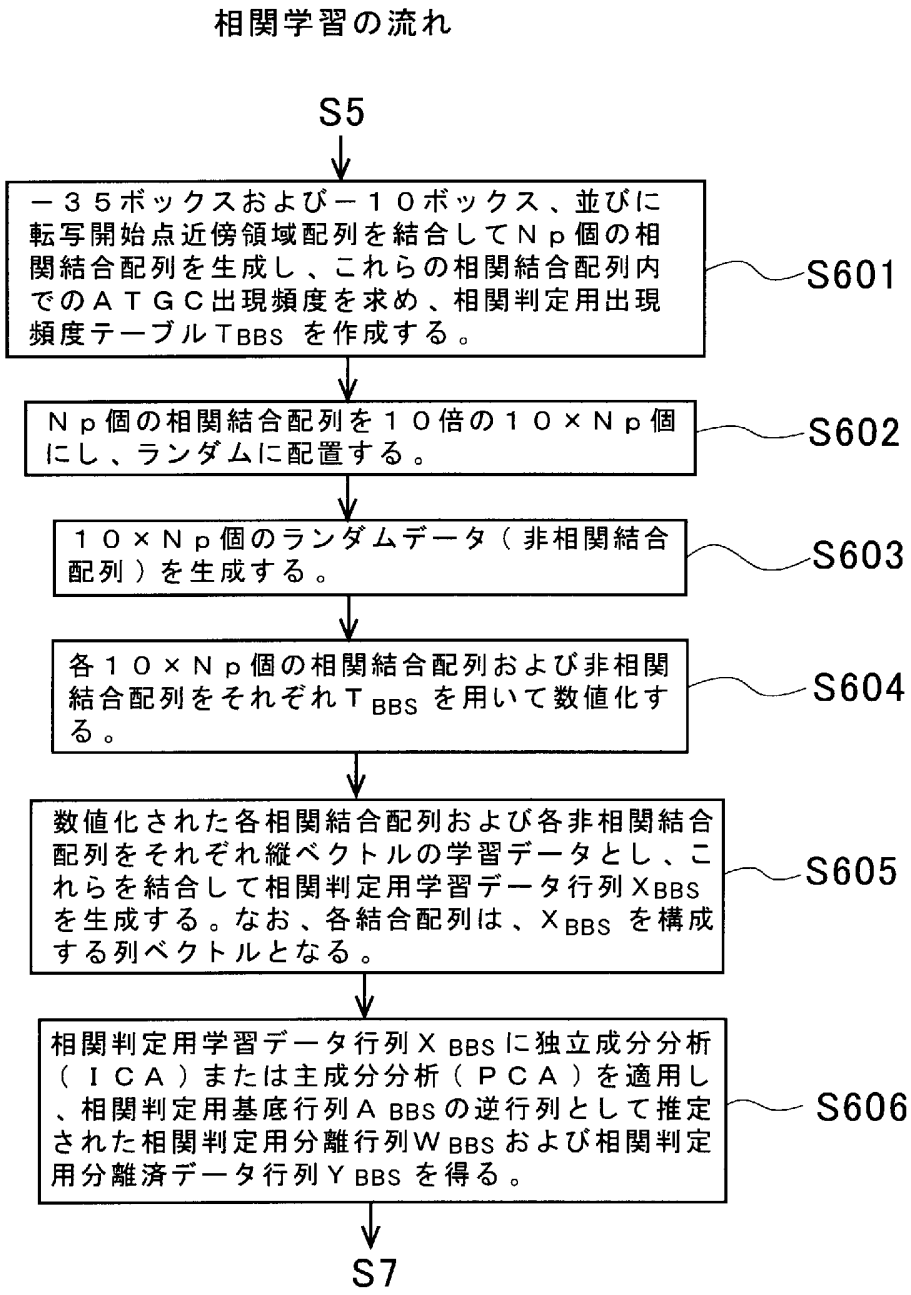
[図25]


配列長：６５
プロモータ
プロモータ
プロモータ
Ｎｐ本
（１０６本）
配列長：６５
非プロモータ
非プロモータ
非プロモータ
Ｎｐ本
（１０６本）
Ｔ prom を用いて
数値化する
Ｔ prom を用いて
数値化する
６５列
６５列
Ｎｐ行
（１０６行）
B prom
Ｎｐ行
（１０６行）
C prom
転置
転置
Ｎｐ列（１０６列）
Ｎｐ列（１０６列）
６５行
B prom T
C prom T
X prom

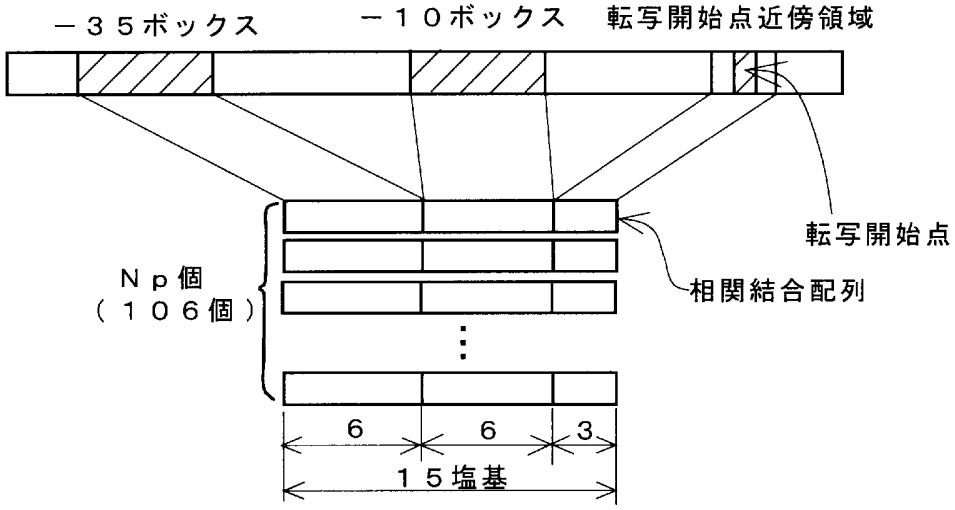
[図26]


２×Ｎｐ列（２１２列）
６５列
２×Ｎｐ列（２１２列）
６５行
Y prom
６５行
＝
W prom
X prom
６５行
Y prom(1)
Y prom(j)
Y prom
W prom
X prom(1)
X prom(j)
63
プロモータ用
分離済データ行列記憶手段
（パターン全体用
分離済データ行列記憶手段）
プロモータ用
分離行列記憶手段
（パターン全体用
分離行列記憶手段）
62

[図27]


相関学習の流れ
S5
－３５ボックスおよび－１０ボックス、並びに転写開始点近傍領域配列を結合してＮｐ個の相関結合配列を生成し、これらの相関結合配列内でのＡＴＧＣ出現頻度を求め、相関判定用出現頻度テーブルTBBS を作成する。
S601
Ｎｐ個の相関結合配列を１０倍の１０×Ｎｐ個にし、ランダムに配置する。
S602
１０×Ｎｐ個のランダムデータ（非相関結合配列）を生成する。
S603
各１０×Ｎｐ個の相関結合配列および非相関結合配列をそれぞれTBBS を用いて数値化する。
S604
数値化された各相関結合配列および各非相関結合配列をそれぞれ縦ベクトルの学習データとし、これらを結合して相関判定用学習データ行列XBBS を生成する。なお、各結合配列は、XBBS を構成する列ベクトルとなる。
S605
相関判定用学習データ行列XBBS に独立成分分析（ＩＣＡ）または主成分分析（ＰＣＡ）を適用し、相関判定用基底行列ABBS の逆行列として推定された相関判定用分離行列WBBS および相関判定用分離済データ行列YBBS を得る。
S606
S7

[図28]

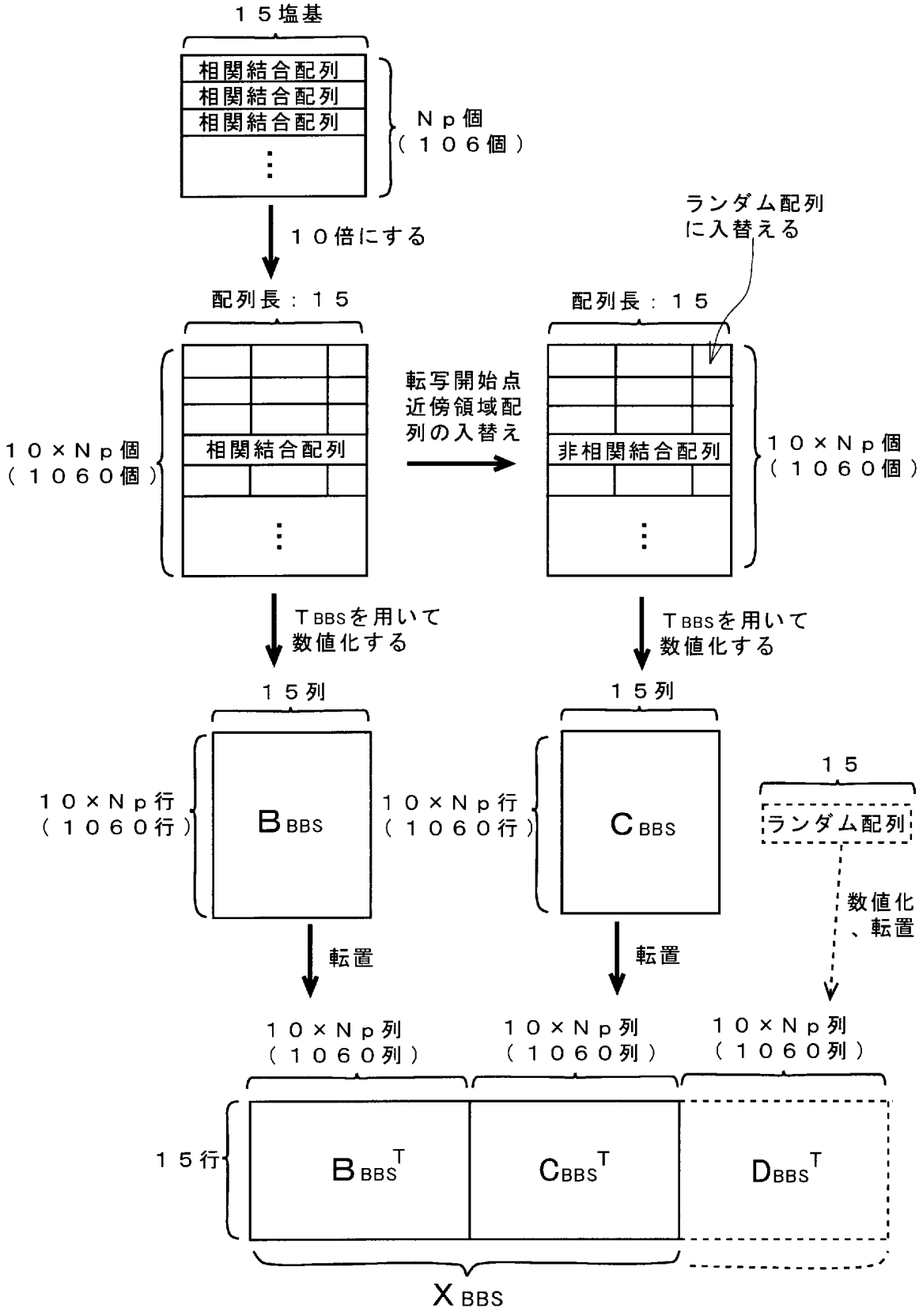


[図29]

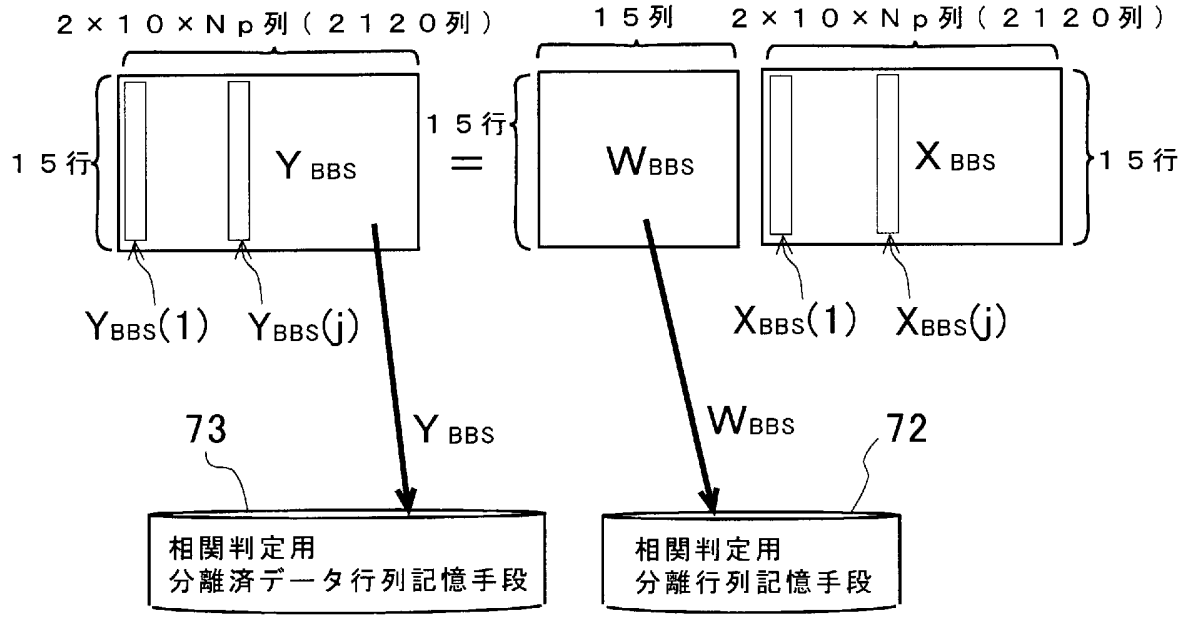
T_{BBS}：相関判定用出現頻度テーブル

塩基種別＼相関結合配列内の配列位置	1	2	3	…	１４	１５
A	$f_{A,BBS(1)}$	$f_{A,BBS(2)}$	$f_{A,BBS(3)}$	…	$f_{A,BBS(14)}$	$f_{A,BBS(15)}$
T	$f_{T,BBS(1)}$	$f_{T,BBS(2)}$	$f_{T,BBS(3)}$	…	$f_{T,BBS(14)}$	$f_{T,BBS(15)}$
G	$f_{G,BBS(1)}$	$f_{G,BBS(2)}$	$f_{G,BBS(3)}$	…	$f_{G,BBS(14)}$	$f_{G,BBS(15)}$
C	$f_{C,BBS(1)}$	$f_{C,BBS(2)}$	$f_{C,BBS(3)}$	…	$f_{C,BBS(14)}$	$f_{C,BBS(15)}$

[図30]

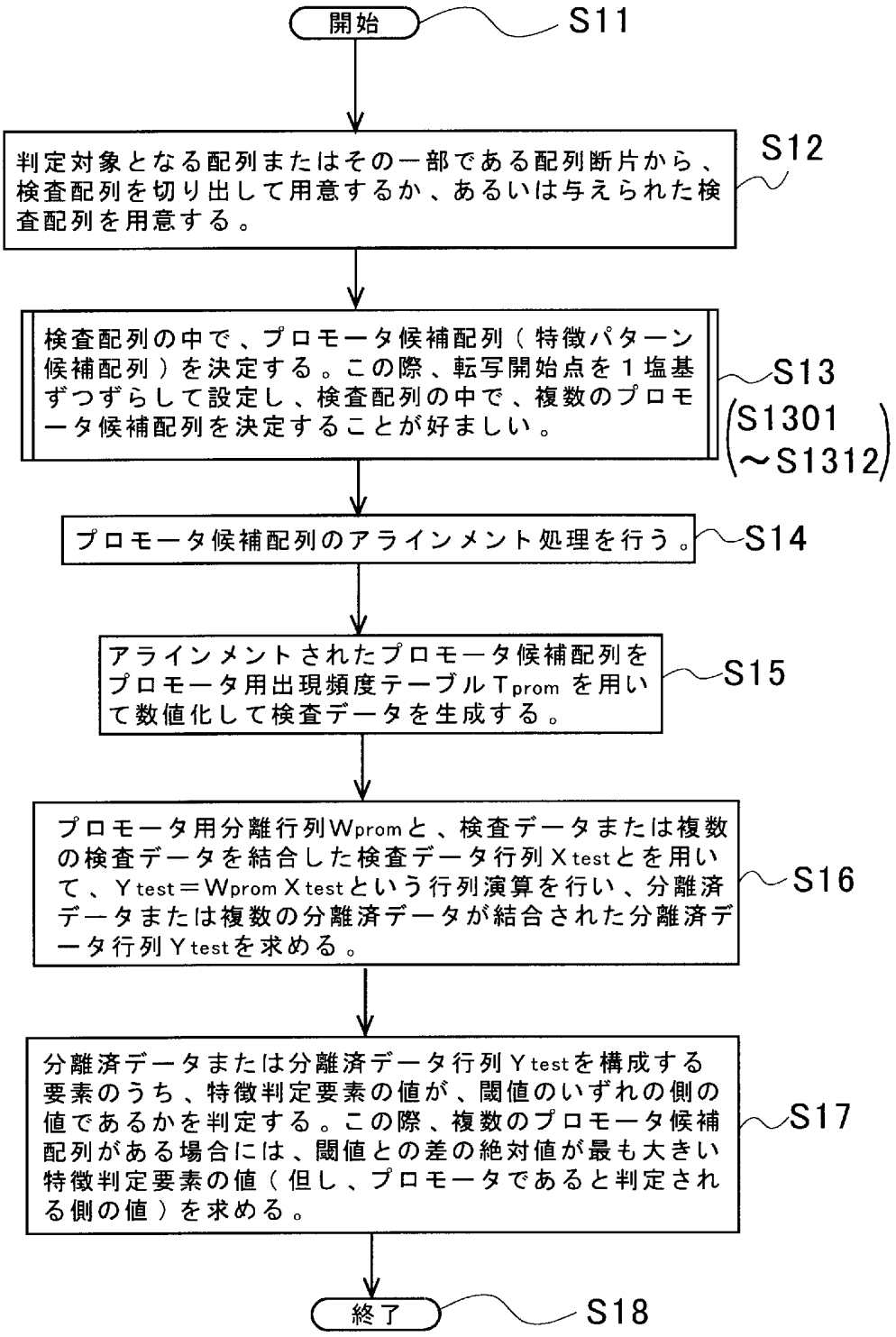

１５塩基
相関結合配列
相関結合配列
相関結合配列
Ｎｐ個
（１０６個）
１０倍にする
配列長：１５
１０×Ｎｐ個
（１０６０個）
相関結合配列
転写開始点
近傍領域配
列の入替え
ランダム配列
に入替える
配列長：１５
非相関結合配列
１０×Ｎｐ個
（１０６０個）
ＴBBSを用いて
数値化する
ＴBBSを用いて
数値化する
１５列
１０×Ｎｐ行
（１０６０行）
BBBS
１０×Ｎｐ行
（１０６０行）
１５列
CBBS
１５
ランダム配列
数値化
、転置
転置
転置
１０×Ｎｐ列
（１０６０列）
１０×Ｎｐ列
（１０６０列）
１０×Ｎｐ列
（１０６０列）
１５行
BBBS^T
CBBS^T
DBBS^T
XBBS

[図31]


２×１０×Ｎｐ列（２１２０列）
１５列
２×１０×Ｎｐ列（２１２０列）
１５行
Ｙ BBS
１５行
＝
WBBS
X BBS
１５行
YBBS(1)
YBBS(j)
XBBS(1)
XBBS(j)
73
Y BBS
WBBS
72
相関判定用
分離済データ行列記憶手段
相関判定用
分離行列記憶手段

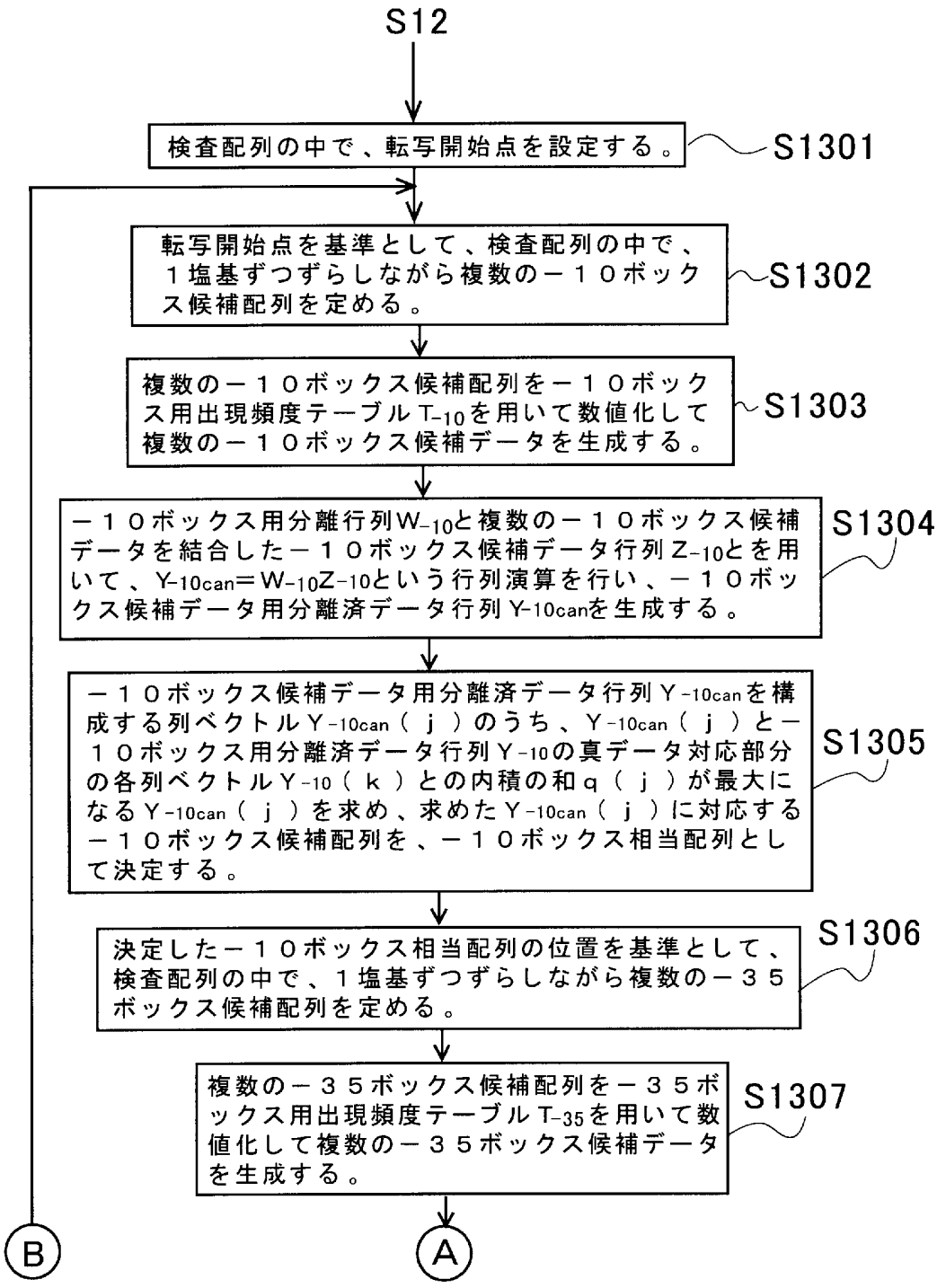
[図32]

検査段階（テストステップ）の全体の流れ



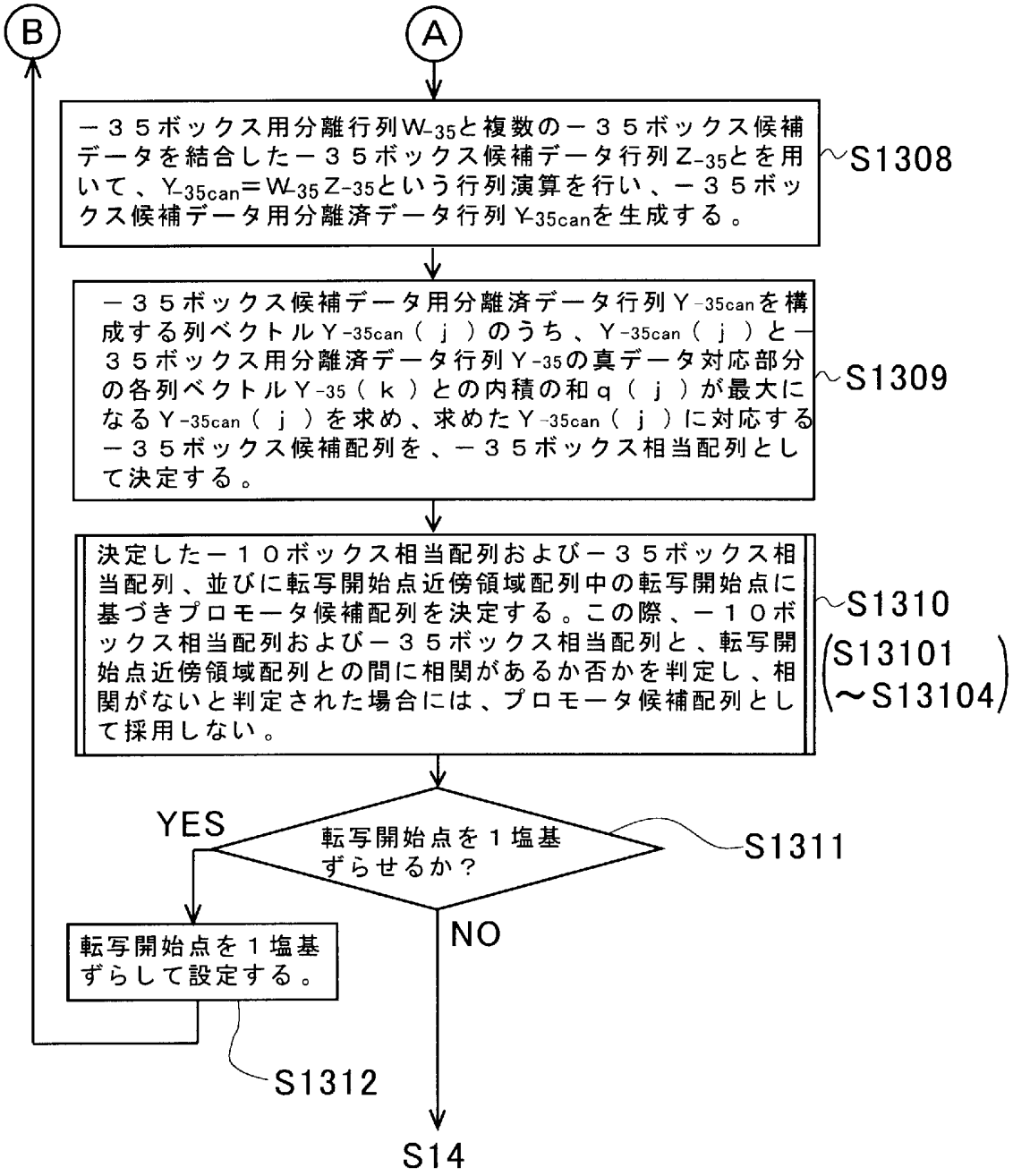
[図33]

検査段階（テストステップ）における
プロモータ候補配列の決定処理の流れ（その１）



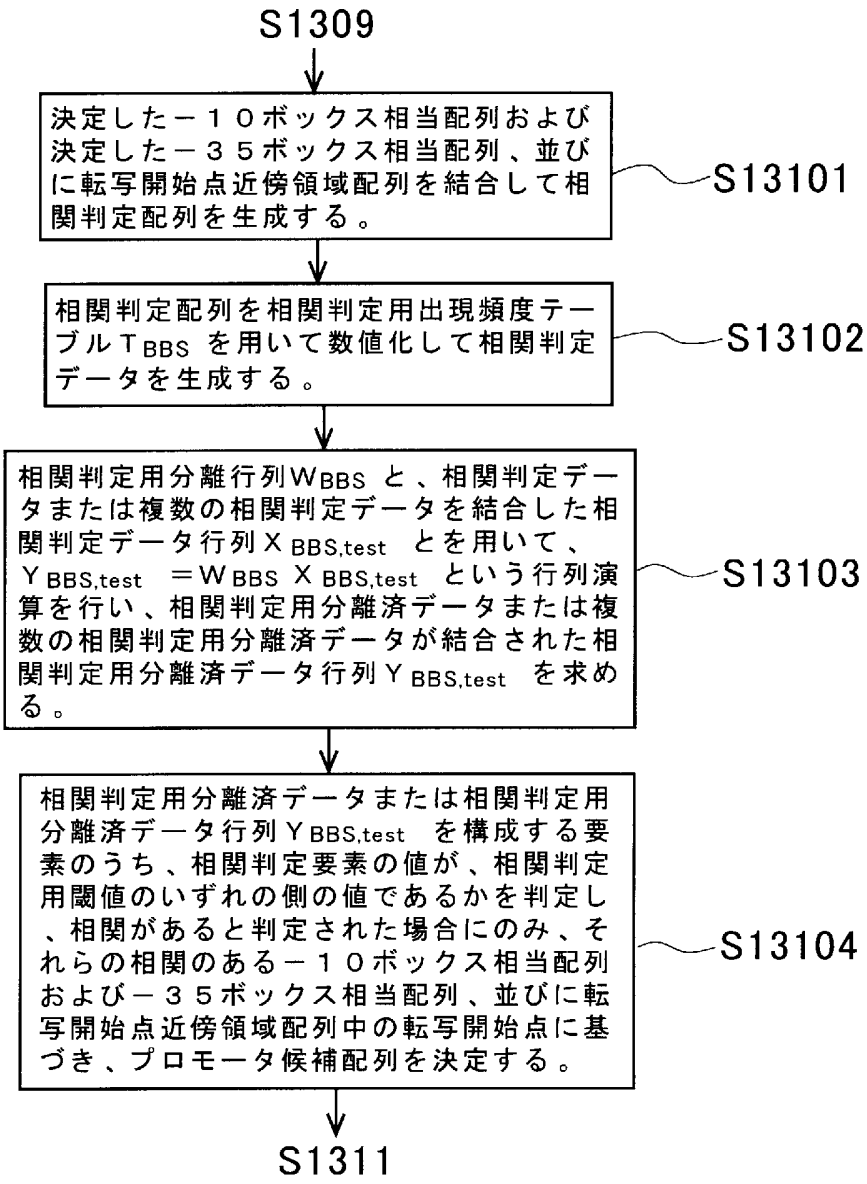
[図34]

検査段階（テストステップ）における
プロモータ候補配列の決定処理の流れ（その２）

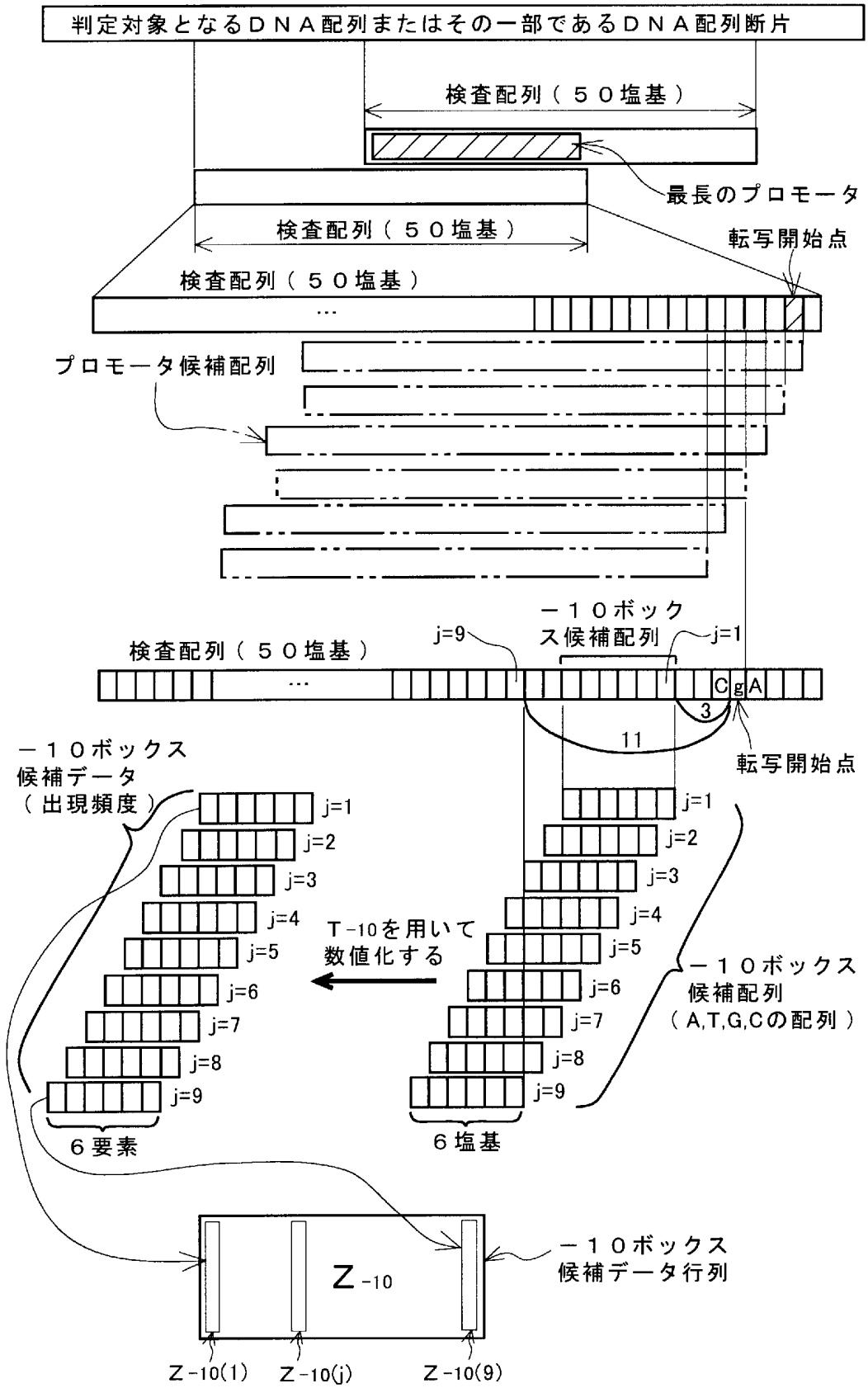


[図35]

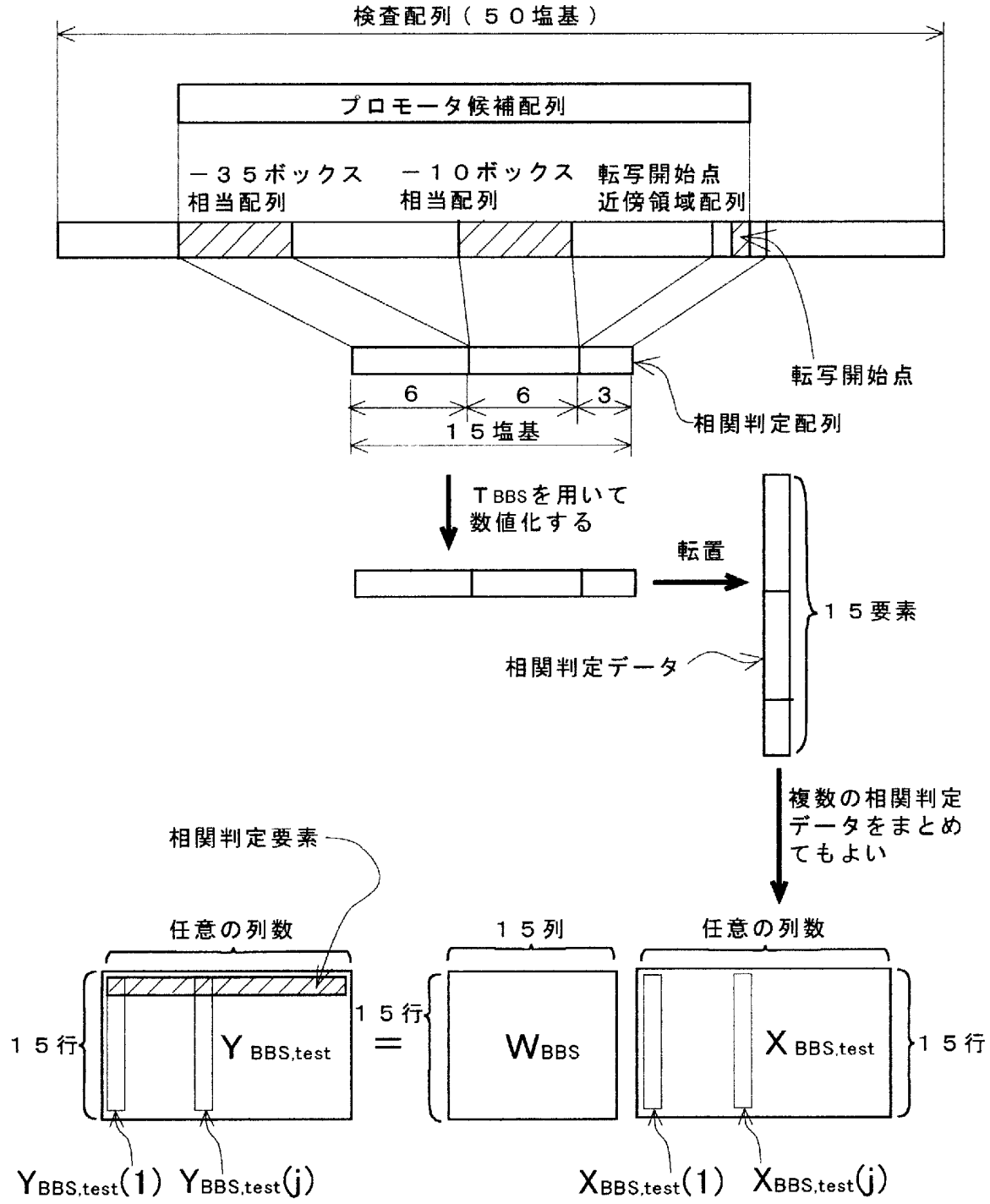
検査段階（テストステップ）におけるプロモータ候補配列の決定時の相関判定処理の流れ



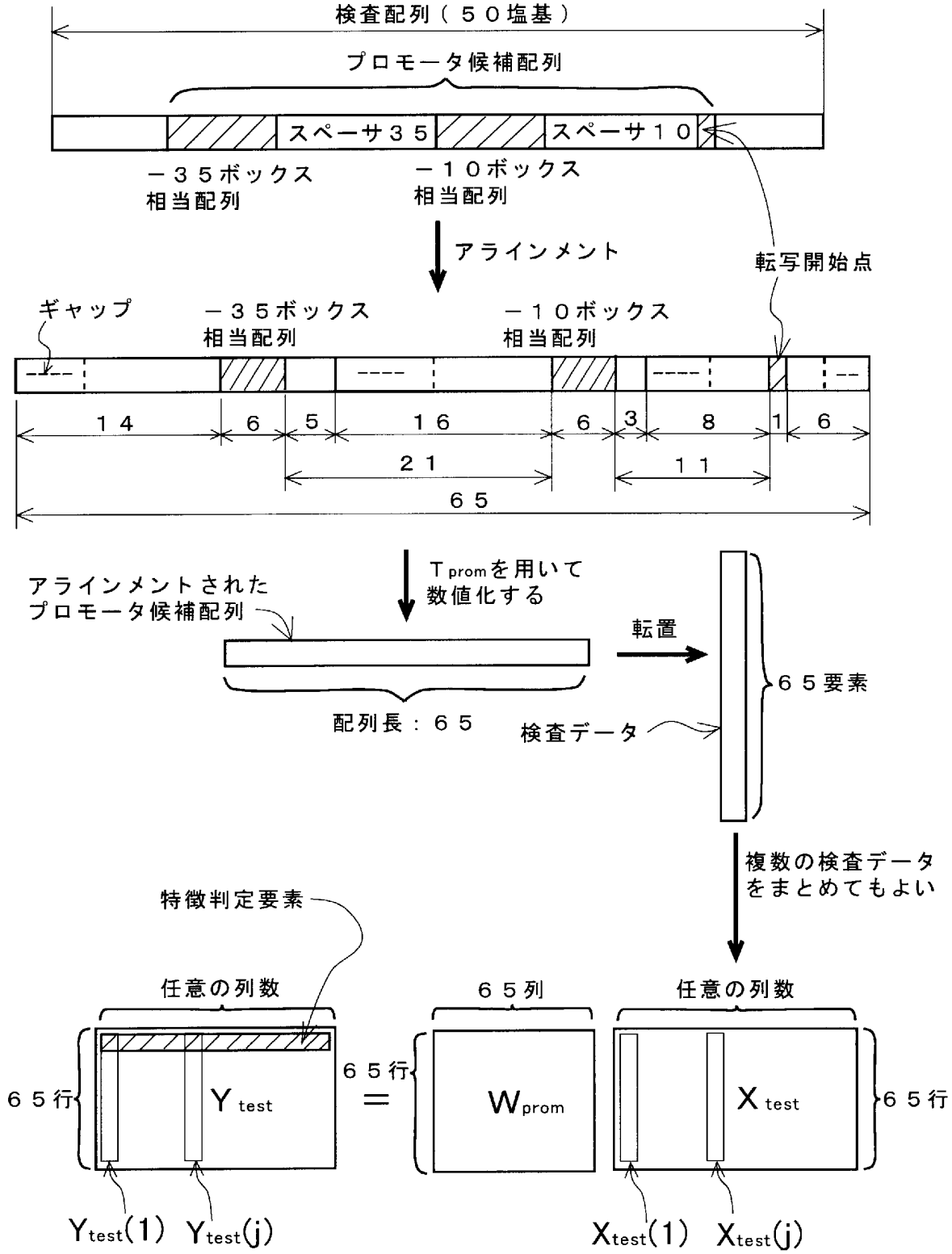
[図36]


判定対象となるＤＮＡ配列またはその一部であるＤＮＡ配列断片
検査配列（５０塩基）
最長のプロモータ
検査配列（５０塩基）
転写開始点
検査配列（５０塩基）
…
プロモータ候補配列
検査配列（５０塩基）
j=9
－１０ボックス候補配列
j=1
…
C
g
A
3
11
転写開始点
－１０ボックス候補データ（出現頻度）
j=1
j=2
j=3
j=4
j=5
j=6
j=7
j=8
j=9
６要素
T-10を用いて数値化する
j=1
j=2
j=3
j=4
j=5
j=6
j=7
j=8
j=9
６塩基
－１０ボックス候補配列（A,T,G,Cの配列）
Z-10
－１０ボックス候補データ行列
Z-10(1)
Z-10(j)
Z-10(9)

[図37]

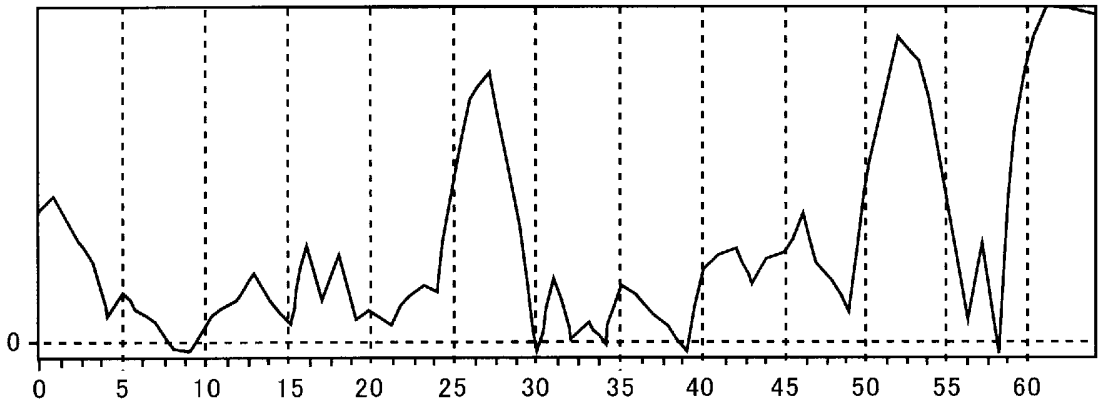

検査配列（５０塩基）
プロモータ候補配列
－３５ボックス相当配列
－１０ボックス相当配列
転写開始点近傍領域配列
転写開始点
6
6
3
１５塩基
相関判定配列
TBBSを用いて数値化する
転置
１５要素
相関判定データ
複数の相関判定データをまとめてもよい
相関判定要素
任意の列数
１５行
YBBS,test
１５行
＝
１５列
WBBS
任意の列数
XBBS,test
１５行
YBBS,test(1)
YBBS,test(j)
XBBS,test(1)
XBBS,test(j)

[図38]


検査配列（５０塩基）
プロモータ候補配列
スペーサ３５
スペーサ１０
－３５ボックス
相当配列
－１０ボックス
相当配列
アラインメント
転写開始点
ギャップ
－３５ボックス
相当配列
－１０ボックス
相当配列
１４
６
５
１６
６
３
８
１
６
２１
１１
６５
アラインメントされた
プロモータ候補配列
Ｔ prom を用いて
数値化する
転置
６５要素
配列長：６５
検査データ
複数の検査データ
をまとめてもよい
特徴判定要素
任意の列数
６５行
Y test
Ytest(1)
Ytest(j)
＝
６５列
任意の列数
W prom
X test
６５行
Xtest(1)
Xtest(j)

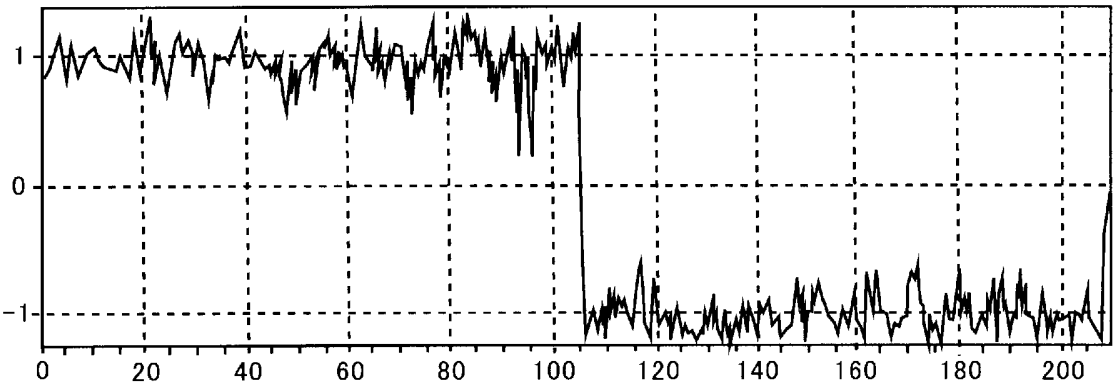
[図39]

W_{prom}^{-1} の第１列の各要素（特徴要素）の値

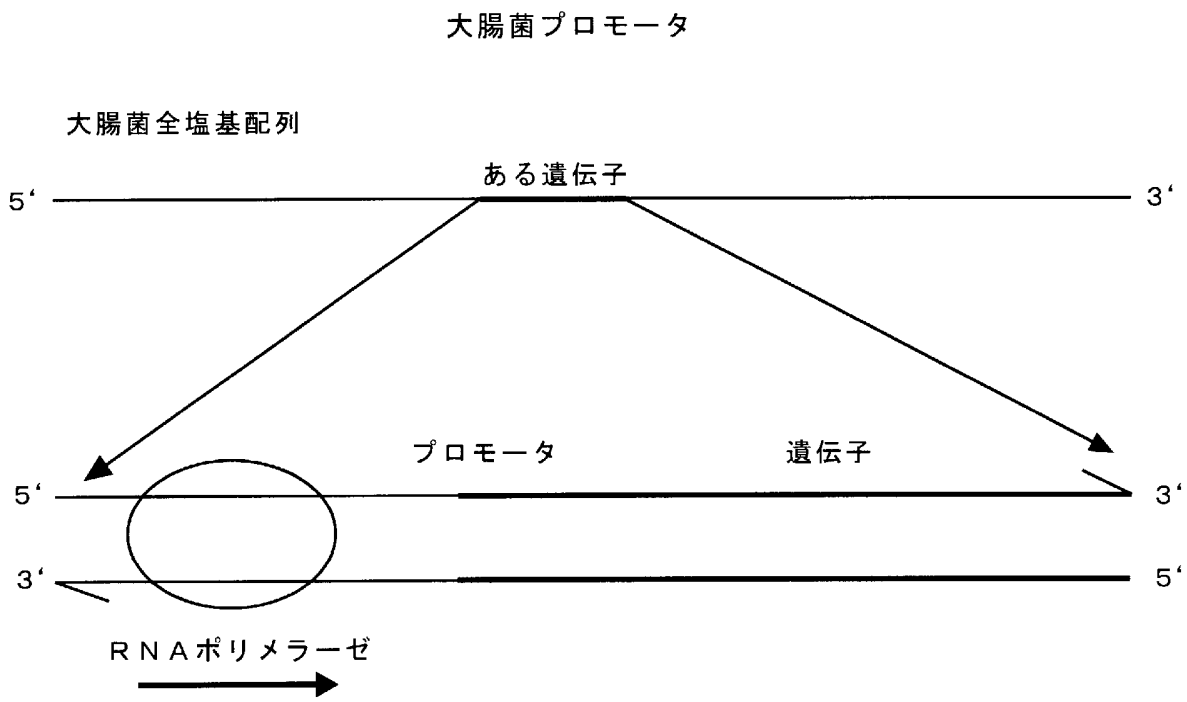


[図40]

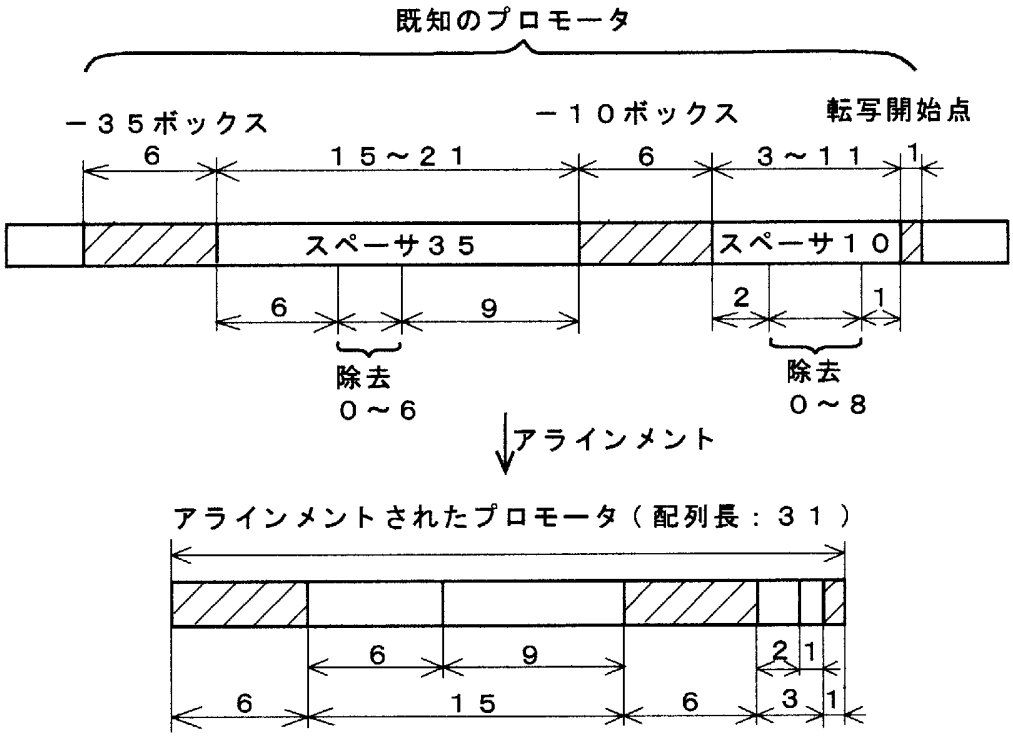
Y_{prom} の第１行の各要素（特徴判定要素）の値



[図41]


大腸菌プロモータ
大腸菌全塩基配列
ある遺伝子
5‘
3‘
プロモータ
遺伝子
5‘
3‘
3‘
5‘
ＲＮＡポリメラーゼ

[図42]

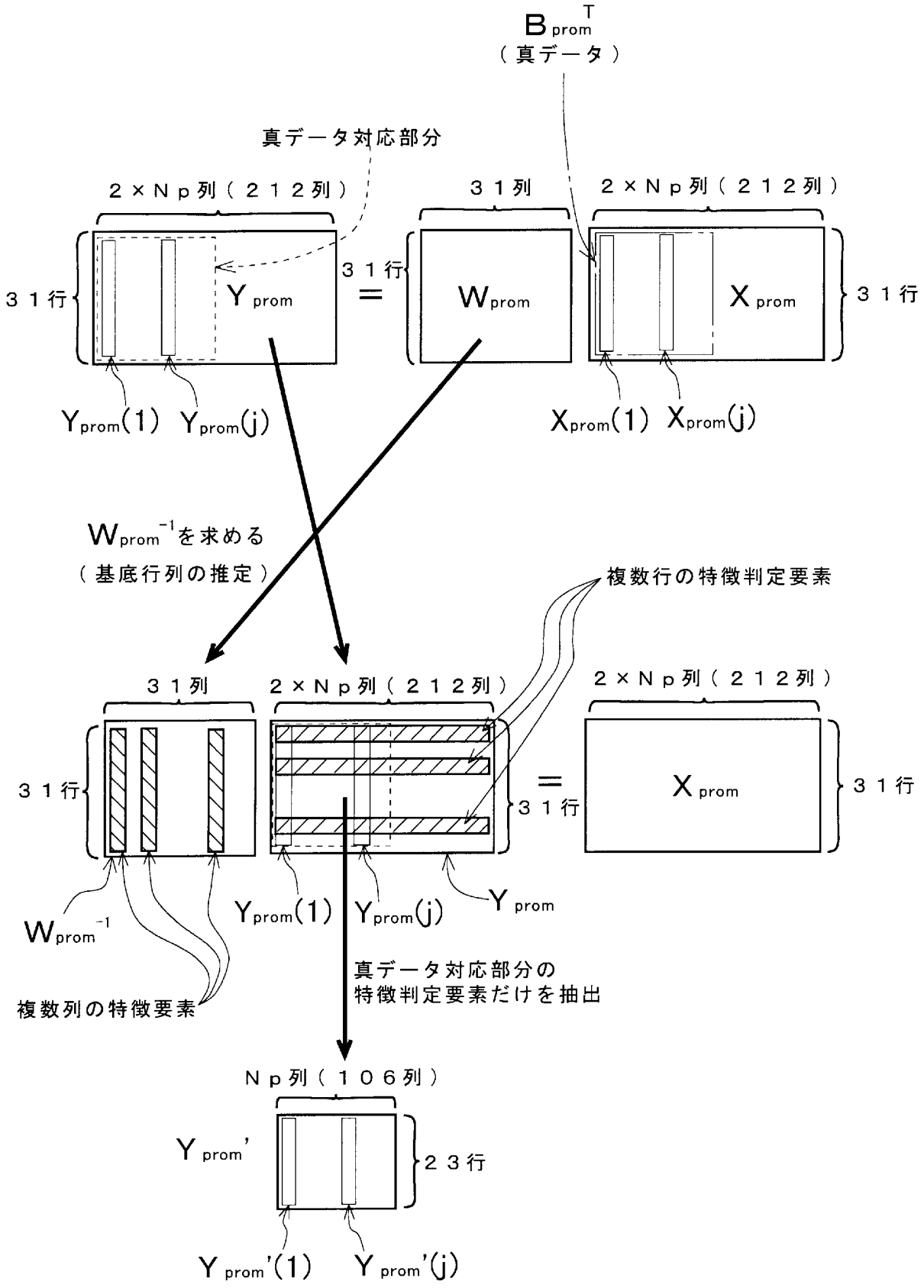

既知のプロモータ
－３５ボックス
－１０ボックス
転写開始点
6
１５～２１
6
３～１１
1
スペーサ３５
スペーサ１０
6
9
2
1
除去
０～６
除去
０～８
アラインメント
アラインメントされたプロモータ（配列長：３１）
6
9
2
1
6
１５
6
3
1

[図43]

T_{prom}：プロモータ用出現頻度テーブル

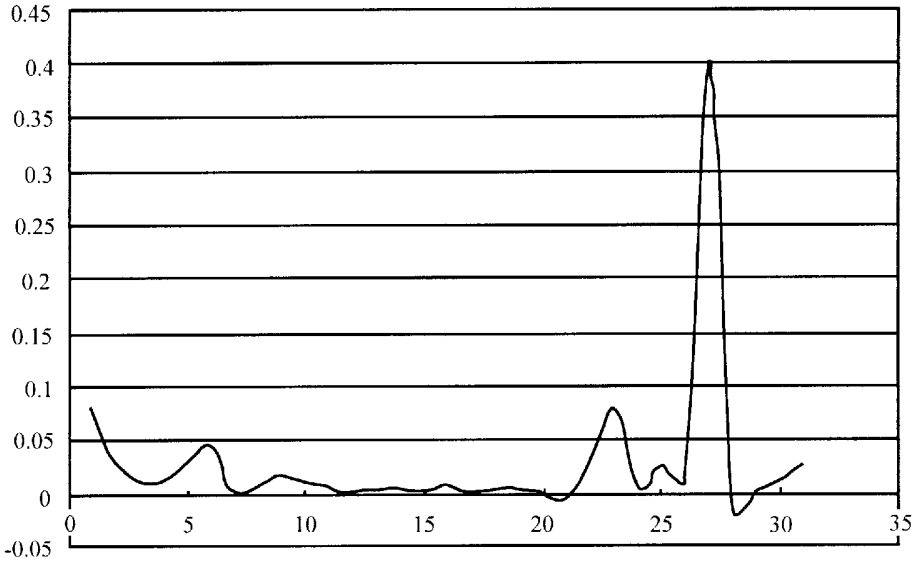
パターン全体内の配列位置 / シンボル種別	1	2	3	…	30	31
A	$f_{A,prom(1)}$	$f_{A,prom(2)}$	$f_{A,prom(3)}$	…	$f_{A,prom(30)}$	$f_{A,prom(31)}$
T	$f_{T,prom(1)}$	$f_{T,prom(2)}$	$f_{T,prom(3)}$	…	$f_{T,prom(30)}$	$f_{T,prom(31)}$
G	$f_{G,prom(1)}$	$f_{G,prom(2)}$	$f_{G,prom(3)}$	…	$f_{G,prom(30)}$	$f_{G,prom(31)}$
C	$f_{C,prom(1)}$	$f_{C,prom(2)}$	$f_{C,prom(3)}$	…	$f_{C,prom(30)}$	$f_{C,prom(31)}$

[図44]


B prom T
（真データ）
真データ対応部分
２×Ｎｐ列（２１２列）
３１列
２×Ｎｐ列（２１２列）
３１行
Y prom
＝
W prom
X prom
３１行
Y prom(1) Y prom(j)
X prom(1) X prom(j)
W prom -1を求める
（基底行列の推定）
複数行の特徴判定要素
３１列
２×Ｎｐ列（２１２列）
２×Ｎｐ列（２１２列）
３１行
＝
X prom
３１行
W prom -1
Y prom(1) Y prom(j)
Y prom
複数列の特徴要素
真データ対応部分の
特徴判定要素だけを抽出
Ｎｐ列（１０６列）
Y prom'
２３行
Y prom'(1) Y prom'(j)

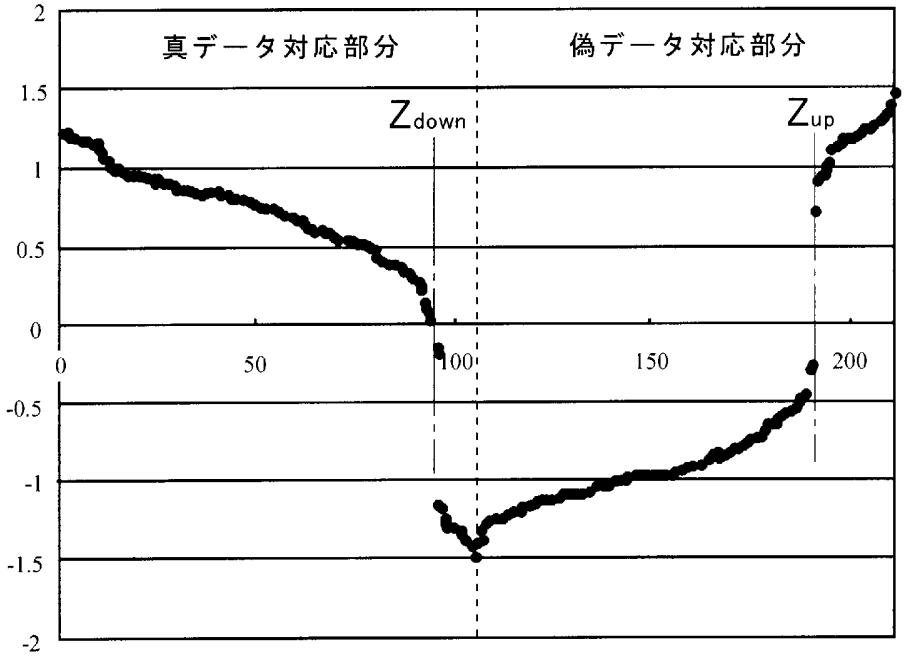
[図45]

W_{prom}^{-1} の第１列の各要素の値（特徴要素の採用例）



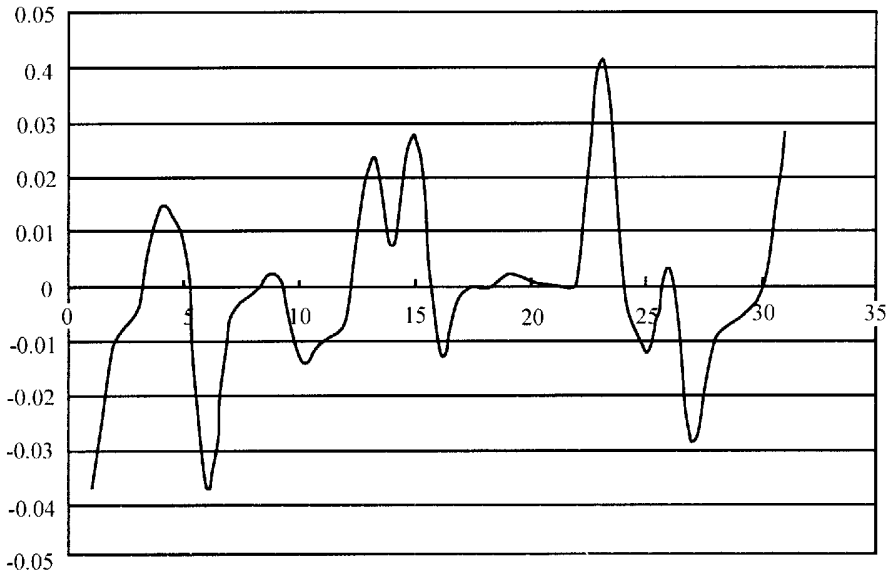
[図46]

Y_{prom}の第１行の各要素の値
（特徴判定要素の採用例）



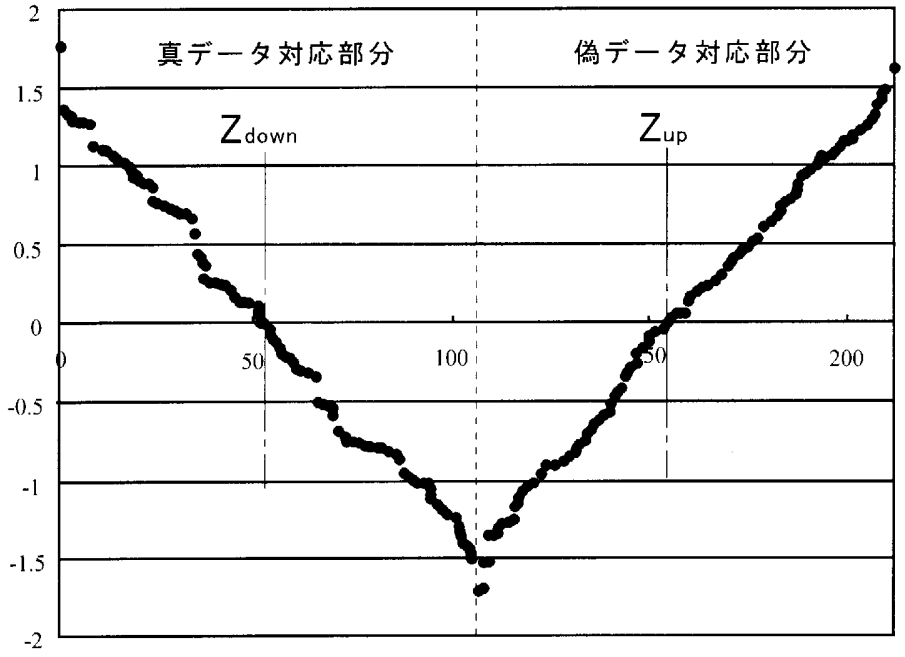
[図47]

W_{prom}^{-1} の第２７列の各要素の値（特徴要素の不採用例）

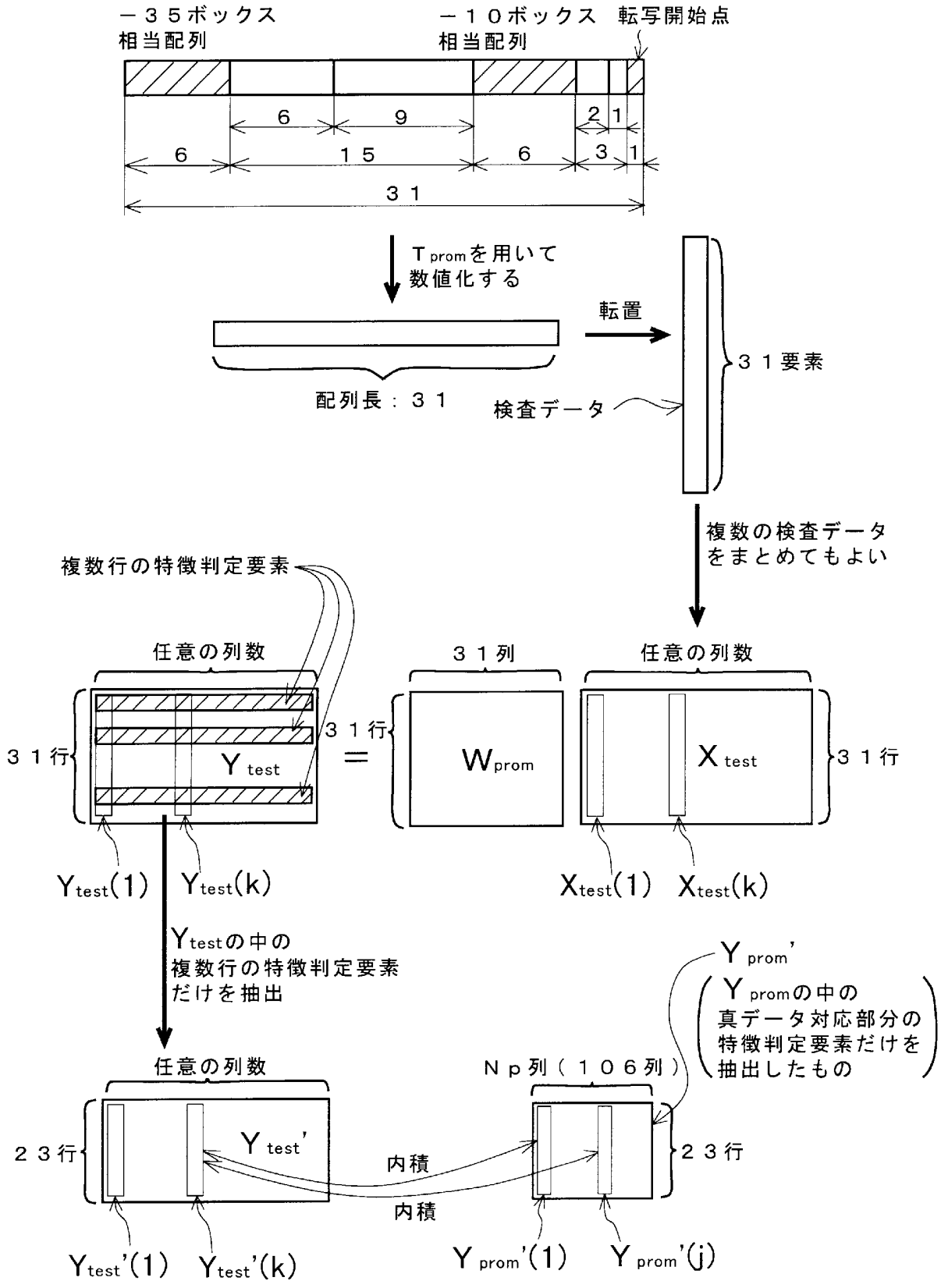


[図48]

Y_{prom}の第２７行の各要素の値
（特徴判定要素の不採用例）



[図49]


アラインメントされたプロモータ候補配列（配列長：３１）
－３５ボックス
相当配列
－１０ボックス
相当配列
転写開始点
6
9
2
1
6
１５
6
3
1
３１
Tpromを用いて
数値化する
転置
３１要素
配列長：３１
検査データ
複数の検査データ
をまとめてもよい
複数行の特徴判定要素
任意の列数
３１列
任意の列数
３１行
Ytest
３１行
＝
Wprom
Xtest
３１行
Ytest(1)
Ytest(k)
Xtest(1)
Xtest(k)
Ytestの中の
複数行の特徴判定要素
だけを抽出
Yprom'
Ypromの中の
真データ対応部分の
特徴判定要素だけを
抽出したもの
任意の列数
Ｎｐ列（１０６列）
２３行
Ytest'
内積
内積
２３行
Ytest'(1)
Ytest'(k)
Yprom'(1)
Yprom'(j)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/014130

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/00 (2006.01), ***G06N3/00*** (2006.01), ***C12N15/09*** (2006.01), ***C12Q1/68*** (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F19/00 (2006.01), ***G06N3/00*** (2006.01), ***C12N15/09*** (2006.01), ***C12Q1/68*** (2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2005
Kokai Jitsuyo Shinan Koho	1971-2005	Toroku Jitsuyo Shinan Koho	1994-2005

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
MEDLINE(STN), JSTPlus(JOIS)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JONSSON, J. Quantitative sequence-activity models (QSAM)-tools for sequence design, Nucleic Acids Res., Vol.21, No.3, 1993, pages 733 to 739	1-49
A	MA, Q. DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study, IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol.31, No.4, 2001, pages 468 to 475	1-49
A	SHAHMURADOV, I.A. PlantProm: a database of plant promoter sequences, Nucleic Acids Res., Vol.31, No.1, 2003, pages 114 to 117	1-49

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 October, 2005 (20.10.05)	01 November, 2005 (01.11.05)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2005/014130

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	MATSUYAMA, Y. Promoter Recognition for E. coli DNA Seqments by Independent Component Analysis, 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04), 16 August, 2004 (16.08.04), pages 686 to 691	1-49

国際調査報告

国際出願番号 PCT／JP2005／014130

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl.[7] *G06F19/00* (2006.01), *G06N3/00* (2006.01), *C12N15/09* (2006.01), *C12Q1/68* (2006.01)

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl.[7] *G06F19/00* (2006.01), *G06N3/00* (2006.01), *C12N15/09* (2006.01), *C12Q1/68* (2006.01)

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2005年
日本国実用新案登録公報	1996－2005年
日本国登録実用新案公報	1994－2005年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

MEDLINE(STN), JSTPlus(JOIS)

C. 関連すると認められる文献

引用文献のカテゴリー＊	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	JONSSON, J. Quantitative sequence-activity models (QSAM)-tools for sequence design, Nucleic Acids Res. , Vol.21, No. 3, 1993, p. 733-739	1-49
A	MA, Q. DNA Sequence Classification via an Expectation Maximization Algorithm and Neural Networks: A Case Study, IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol.31, No. 4, 2001, p. 468-475	1-49

☑ C欄の続きにも文献が列挙されている。　☐ パテントファミリーに関する別紙を参照。

＊ 引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日 20. 10. 2005	国際調査報告の発送日 01. 11. 2005
国際調査機関の名称及びあて先 日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	特許庁審査官（権限のある職員） 宮久保 博幸 電話番号 03－3581－1101 内線 3562

5L	3136

C（続き）．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
A	SHAHMURADOV, I. A. PlantProm: a database of plant promoter sequences, Nucleic Acids Res., Vol. 31, No. 1, 2003, p. 114-117	1-49
P, X	MATSUYAMA, Y. Promoter Recognition for E. coli DNA Segments by Independent Component Analysis, 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04), 2004.08.16, p. 686-691	1-49