(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－213003

(43)公開日　平成11年(1999)８月６日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ	17/30	Ｇ０６Ｆ	15/40	３７０Ｆ
	17/00	Ｃ０７Ｋ	14/195	
	19/00	Ｇ０６Ｆ	15/20	Ｆ
// Ｃ０７Ｋ	14/195		15/42	Ｚ
Ｃ１２Ｎ	15/09	Ｃ１２Ｎ	15/00	Ａ

審査請求　未請求　請求項の数６　ＯＬ　（全 14 頁）

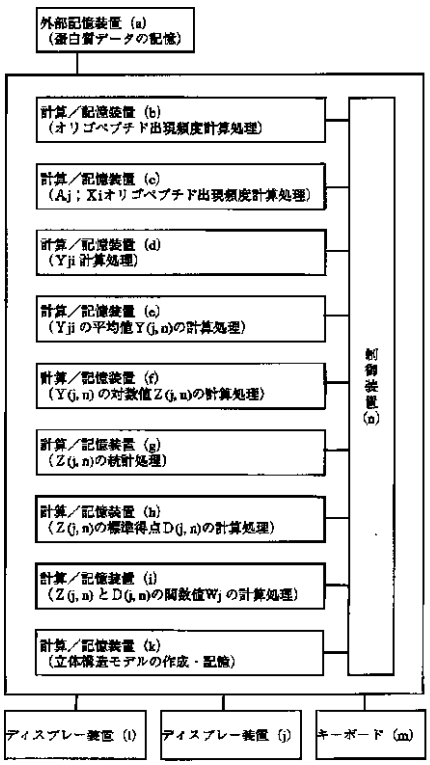
(21)出願番号　特願平10－18699

(22)出願日　平成10年(1998)１月30日

(71)出願人　396020800
科学技術振興事業団
埼玉県川口市本町４丁目１番８号
(71)出願人　597014682
土居　洋文
千葉県船橋市夏見５－29－４－515
(72)発明者　土居　洋文
千葉県船橋市夏見５－29－４－515
(72)発明者　平木　秀明
東京都文京区千駄木３丁目48－８－507
(72)発明者　金井　昭夫
茨城県つくば市松代３－17－13　ヴィラ松代301
(74)代理人　弁理士　西澤　利夫

(54)【発明の名称】　蛋白質機能部位の予測方法と予測装置

(57)【要約】
【課題】　ゲノム解析やｃＤＮＡ解析から得られた機能未知の蛋白質について、その機能部位を予測するための方法と装置を提供する。
【解決手段】　ゲノムデータまたはｃＤＮＡ解析データが既知である生物種の全蛋白質から、任意の蛋白質の機能部位を予測する方法であって、蛋白質のアミノ酸配列から、その蛋白質の機能部位を構成しているオリゴペプチドの出現頻度を算出し、この出現頻度に寄与しているアミノ酸残基の値を機能代表値として算出し、この機能代表値を指標として蛋白質機能部位を予測する方法と、機能部位予測装置。



【特許請求の範囲】

【請求項１】　ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質から、その生物種ａの任意の蛋白質の機能部位を予測する方法であって、(1) 生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を求め、(2) 生物種ａの任意の蛋白質について、(2') アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦Ｍ、ただしＭは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さＭのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；

$a_{j1}a_{j2}.....a_{ji}\ ..\ a_{jn}$（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）

の出現頻度と、

Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；

$a_{j1}a_{j2}.....X_i.....a_{jn}$（Xiは任意のアミノ酸残基を示す）

の出現頻度とを生物種ａの全蛋白質中で求め、(3) ＡｊオリゴペプチドとＸｉオリゴペプチドの出現頻度の比Ｙｊｉを求め、(4) Ｙｊｉの平均値Ｙ(j,n) ；

$Y(j,n) = \Sigma Yji / n$（１≦ｉ≦ｎ）

を求め、(5) Ｙ(j,n) の関数値Ｚ(j,n) ；

$Z(j,n) = -\log(Y(j,n))$

を求め、(6) 以下、上記ステップ(2')から(5) を順次繰り返し、アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にあるアミノ酸残基Ａｊについて各々のＺ(j,n) 値を求め、(7) 生物種ａの全蛋白質について上記ステップ(2) から(6) を順次繰り返し、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Av（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；

$g = (Z(j,n), Aj) = \{Z(j,n) - Av(Aa)\} / Sd(Aa)$

（ただしＡｊ＝Ａａ）

を求め、(8) アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にある全アミノ酸残基Ａｊについてステップ(7) で得られた関数ｇの値Ｄ(j,n) ；

$D(j,n) = g(Z(j,n), Aj)$

を求め、(9) アミノ酸配列（長さＬ）のｊ番目のアミノ酸残基の機能代表値をＺ(j,n) 値とＤ(j,n) 値の関数値Ｗｊ；

$Wj = h(Z(j,1), Z(j,2), ..., Z(j, M), D(j,1), D(j,2), ..., D(j, M))$

とする、ことによって、蛋白質の機能に対する各アミノ酸残基の責任の程度をＷｊ値の大きさを指標として予測すること特徴とする蛋白質の機能部位予測方法。

【請求項２】　各アミノ酸残基のＷｊ値を２次元的な分布図として表示する請求項１の方法。

【請求項３】　各アミノ酸残基のＷｊ値を、蛋白質の立体構造モデル上に分布図として表示する請求項１の方法。

【請求項４】　請求項１記載の方法を自動的に行なう装置であって、少なくとも以下の (a)から(i) の装置、

(a) ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質のアミノ酸配列データ、および既存の蛋白質データベースを記憶する外部記憶装置、(b) この生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を計算するＣＰＵと、その計算結果を記憶する記憶装置とからなる計算／記憶装置、(c) この生物種ａの任意の蛋白質について、アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦Ｍ、ただしＭは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さＭのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；

$a_{j1}a_{j2}.....a_{ji}..a_{jn}$（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）

の出現頻度と、

Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；

$a_{j1}a_{j2}.....X_i.....a_{jn}$（Xiは任意のアミノ酸残基を示す）

の出現頻度とを生物種ａの全蛋白質中で求めるＣＰＵと、その計算結果を記憶する記憶装置とからなる計算／記憶装置、(d) ＡｊオリゴペプチドとＸｉオリゴペプチドの出現頻度の比Ｙｊｉを求めるＣＰＵと、Ｙｊｉを記憶する記憶装置とからなる計算／記憶装置、(e) Ｙｊｉの平均値Ｙ(j,n) ；

$Y(j,n) = \Sigma Yji / n$（１≦ｉ≦ｎ）

を求めるＣＰＵと、Ｙ(j,n) を記憶する記憶装置とからなる計算／記憶装置、(f) Ｙ(j,n) の関数値Ｚ(j,n) ；

$Z(j,n) = -\log(Y(j,n))$

を求めるＣＰＵと、Ｚ(j,n) を記憶する記憶装置とからなる計算／記憶装置、(g) 生物種ａの全蛋白質のアミン酸配列について、各アミノ酸残基のＺ(j,n) を求め、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Av（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；

$g = (Z(j,n), Aj) = \{Z(j,n) - Av(Aa)\} /$

Sd（Ａａ）
（ただしＡｊ＝Ａａ）
を求めるＣＰＵと、ｇを記憶する計算装置とからなる計算／記憶装置、(h) アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にある全アミノ酸残基Ａｊについて、装置(g) に記憶された関数ｇの値Ｄ(j,n)；
Ｄ(j,n) ＝ｇ（Ｚ(j,n)，Ａｊ）
を求めるＣＰＵと、Ｄ(j,n) 値を記憶する記憶装置とからなる計算／記憶装置、(i) アミノ酸配列について、各アミノ酸残基のＺ(j,n) 値ｇとＤ(j,n) 値の任意の関数値Ｗｊ；
Ｗｊ＝ｈ（Ｚ(j,1)，Ｚ(j,2)，...，Ｚ(j，Ｍ)，Ｄ(j,1)，Ｄ(j,2)，...，Ｄ(j，Ｍ)）
を求める計算装置と、Ｗｊ値を記憶する記憶装置とからなる計算／記憶装置を備えていることを特徴とする蛋白質の機能部位予測装置。
【請求項５】　アミノ酸配列について、各アミノ酸残基のＷｊ値を２次元的な分布図として表示するディスプレー装置を備えている請求項４の装置。
【請求項６】　既存の蛋白質立体構造データベースを記憶し、または公知の方法に従ってアミノ酸配列から立体構造モデルを作成し記憶する計算／記憶装置と、アミノ酸配列について、各アミノ酸残基のＷｊ値を上記計算／記憶装置に記憶されている立体構造データベースまたは立体構造モデル上に分布図として表示するディスプレー装置を備えた請求項４の装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、蛋白質の機能部位を予測する方法と、この機能予測を行なうための装置に関するものである。さらに詳しくは、この発明は、ゲノム解析やｃＤＮＡ解析により得られた機能未知の蛋白質の機能部位の予測や、機能が既知である蛋白質であってもその蛋白質のもつ新規の機能と機能部位の予測に関するものである。
【０００２】
【従来の技術とその課題】病原微生物を含む種々の生物のゲノム解析やｃＤＮＡ解析の進展にともない、機能未知の新規遺伝子やそれによってコードされる蛋白質の数が急速に増加している。たとえば、これまでにマイコプラズマ・ジェニタリウム[Mycoplasma genitalium] (Fraser et al., Science 270, 397-403, 1995)、ヘモフィラス・インフルエンザエ[Heamophilus influenzae] (Fleischmann et. al., Science 269,496-512, 1995)、メタノコッカス・ヤナシイ[Methonococcus jannashcii] (Bultet. al., Science 273, 1058-1073, 1996) などの微生物の全ゲノムの核酸配列が解析され、遺伝子から予測される新規の蛋白質が数多く発見されている。またヒトやマウスではｃＤＮＡ解析がゲノム解析と同時に進行しており新規の蛋白質が多く発見されている。
【０００３】このような状況において、機能未知の蛋白質の機能または機能部位を予測することが重要な課題となってきている。また、新規の蛋白質のみならず、機能が既知の蛋白質についても、新規の機能あるいは機能部位が発見されれば、その蛋白質の産業上あるいは医療上の利用価値が判断可能となる。また、このような機能予測は、機能をさらに向上させた改変型蛋白質の作成をも可能とする。
【０００４】従来より、ゲノム解析やｃＤＮＡ解析によって明らかにされた遺伝子がコードする蛋白質が新規であるか機能既知であるかは、Swiss-Prot等の蛋白質データベースを用いたホモロジー検索によって行なわれてきた。また機能部位を予測するには、同じ機能をもった種々の生物由来の蛋白質を蛋白質データベースから抽出してアライメントを行い、両者に共通して保存されている領域を機能部位として予測していた。
【０００５】しかしながら、ゲノム解析やｃＤＮＡ解析から得られた蛋白質が全く新規の蛋白質であった場合、アライメント法は使えないという問題があった。また蛋白質データベース中の既知の蛋白質とホモロジーがあったとしても、近縁生物種の蛋白質とのホモロジーであった場合、保存領域がその蛋白質のアミノ酸配列のほとんどを占め、機能部位の予測が行えないという問題があった。さらに、機能が既知あるいは未知に関わらず、蛋白質の改変に関しては、アライメントによって機能部位の予測が行えたとしても、保存領域を変異させると一般的に機能が低下することが予想され、保存領域外のアミノ酸を変異させることにより機能向上を計らなければならない。すなわち、改変したい蛋白質において新規の機能部位を見い出す必要があり、新規の機能部位の発見やどのアミノ酸を変異させればよいかは従来のアライメント法では予測できないという問題があった。
【０００６】この発明は、以上のとおりの事情に鑑みてなされたものであって、ゲノム解析やｃＤＮＡ解析から得られた機能未知の蛋白質について、その機能部位を予測するための新しい方法を提供することを目的としている。また、この発明は、この機能予測を行なうための装置を提供することを目的としている。
【０００７】
【改題を解決するための手段】この出願は、上記の課題を解決する発明として、ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質から、その生物種ａの任意の蛋白質の機能部位を予測する方法であって、(1) 生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を求め、(2) 生物種ａの任意の蛋白質について、(2') アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列

の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦M、ただしMは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さMのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；

$a_{j1}a_{j2}.....a_{ji}..a_{jn}$（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）の出現頻度と、Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；

$a_{j1}a_{j2}.....X_i.....a_{jn}$（Xiは任意のアミノ酸残基を示す）

の出現頻度とを生物種ａの全蛋白質中で求め、(3) Ａｊオリゴペプチドと Ｘｉオリゴペプチドの出現頻度の比Ｙｊｉを求め、(4) Ｙｊｉの平均値Ｙ(j,n) ；

Ｙ(j,n) ＝Σ Ｙｊｉ／ｎ（１≦ｉ≦ｎ）

を求め、(5) Ｙ(j,n) の関数値Ｚ(j,n) ；

Ｚ(j,n) ＝-log（Ｙ(j,n) ）

を求め、(6) 以下、上記ステップ(2')から(5) を順次繰り返し、アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にあるアミノ酸残基Ａｊについて各々のＺ(j,n) 値を求め、(7) 生物種ａの全蛋白質について上記ステップ(2) から(6) を順次繰り返し、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Av（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；

ｇ＝（Ｚ(j,n) ，Ａｊ）＝｛Ｚ(j,n) －Ad（Ａａ）｝／Sd（Ａａ）

（ただしＡｊ＝Ａａ）

を求め、(8) アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にある全アミノ酸残基Ａｊについてステップ(7) で得られた関数ｇの値Ｄ(j,n) ；

Ｄ(j,n) ＝ｇ（Ｚ(j,n) ，Ａｊ）

を求め、(9) アミノ酸配列（長さＬ）のｊ番目のアミノ酸残基の機能代表値をＺ(j,n) 値とＤ(j,n) 値の関数値Ｗｊ；

Ｗｊ＝ｈ（Ｚ(j,1) ，Ｚ(j,2) ，... ，Ｚ(j，Ｍ) ，Ｄ(j,1) ，Ｄ(j,2) ，... ，Ｄ(j，Ｍ) ）

とする、ことによって、蛋白質の機能に対する各アミノ酸残基の責任の程度をＷｊ値の大きさを指標として予測することを特徴とする蛋白質の機能部位予測方法を提供する。

【０００８】またこの発明は、上記の方法を自動的に行なう装置であって、少なくとも以下の (a)から (i)の装置、(a) ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質のアミノ酸配列データ、および既存の蛋白質データベースを記憶する外部記憶装置、(b) この生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を計算するＣＰＵと、その計算結果を記憶する記憶装置とからなる計算／記憶装置、(c) この生物種ａの任意の蛋白質について、アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦M、ただしMは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さMのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；

$a_{j1}a_{j2}.....a_{ji}..a_{jn}$（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）の出現頻度と、Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；

$a_{j1}a_{j2}.....X_i.....a_{jn}$（Xiは任意のアミノ酸残基を示す）

の出現頻度とを生物種ａの全蛋白質中で求めるＣＰＵと、その計算結果を記憶する記憶装置とからなる計算／記憶装置、(d) Ａｊオリゴペプチドと Ｘｉオリゴペプチドの出現頻度の比Ｙｊｉを求めるＣＰＵと、Ｙｊｉを記憶する記憶装置とからなる計算／記憶装置、(e) Ｙｊｉの平均値Ｙ(j,n) ；

Ｙ(j,n) ＝Σ Ｙｊｉ／ｎ（１≦ｉ≦ｎ）

を求めるＣＰＵと、Ｙ(j,n) を記憶する記憶装置とからなる計算／記憶装置、(f) Ｙ(j,n) の関数値Ｚ(j,n) ；

Ｚ(j,n) ＝-log（Ｙ(j,n) ）

を求めるＣＰＵと、Ｚ(j,n) を記憶する記憶装置とからなる計算／記憶装置、(g) 生物種ａの全蛋白質のアミン酸配列について、各アミノ酸残基のＺ(j,n) を求め、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Ad（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；

ｇ＝（Ｚ(j,n) ，Ａｊ）＝｛Ｚ(j,n) －Ad（Ａａ）｝／Sd（Ａａ）

（ただしＡｊ＝Ａａ）

を求めるＣＰＵと、ｇを記憶する計算装置とからなる計算／記憶装置、(h) アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にある全アミノ酸残基Ａｊについて、装置(g) に記憶された関数ｇの値Ｄ(j,n) ；

Ｄ(j,n) ＝ｇ（Ｚ(j,n) ，Ａｊ）

を求めるＣＰＵと、Ｄ(j,n) 値を記憶する記憶装置とからなる計算／記憶装置、(i) アミノ酸配列について、各アミノ酸残基のＺ(j,n) 値とＤ(j,n) 値の任意の関数値Ｗｊ；

Ｗｊ＝ｈ（Ｚ(j,1) ，Ｚ(j,2) ，... ，Ｚ(j，Ｍ) ，Ｄ(j,1) ，Ｄ(j,2) ，... ，Ｄ(j，Ｍ) ）

を求める計算装置と、Ｗｊ値を記憶する記憶装置とからなる計算／記憶装置を備えていることを特徴とする蛋白質の機能部位予測装置を提供する。

【０００９】すなわち、この発明の蛋白質機能部位予測方法は、以下のとおりの考えに立脚してなされたものである。すなわち、蛋白質は２０種類のアミノ酸残基の配列によって構成されているが、その並びはランダムではない。従って、任意の生物種において、アミノ酸配列の部分配列である特定のオリゴペプチドがゲノムでコードされる全蛋白質中に出現する頻度は均一ではなく、種々の蛋白質に高頻度で出現するオリゴペプチドや、まれにしか出現しないオリゴペプチドが存在する。このうち種々の蛋白質に共通して高頻度に出現するオリゴペプチドは、個々の蛋白質の独自性、すなわち機能を決める能力がなく、一方、低頻度で出現するオリゴペプチドが個々の蛋白質の独自性や機能を決定していると考えることができる。

【００１０】つまり、蛋白質の機能部位はその部分を構成しているオリゴペプチドの出現頻度と対応していると考えられる。この発明の方法においては、ステップ(3)において示されているＡｊオリゴペプチドとＸｉオリゴペプチドの出現頻度の比Ｙｊｉによって、アミノ酸残基ＡｊがＡｊオリゴペプチドの出現頻度に寄与している程度が評価され、従って蛋白質の任意の位置のアミノ酸残基Ａｊについて算出された関数値Ｚ(j,n) 値が、その位置にあるアミノ酸残基Ａｊの出現指数（すなわち、その機能代表値となる）。

【００１１】また、このＺ(j,n) 値はアミノ酸残基Ａｊの種類によって異なっている。この発明の方法におけるステップ(7) において、ある生物種ａの全蛋白質におけるＺ(j,n) 値の分布を２０種類のアミノ酸毎に求め、これらの分布より求めたアミノ酸毎の平均値と標準偏差値に基づいてＺ(j,n) 値を標準化したＤ(j,n) 値が、アミノ酸残基の種類によるバイアスを補正した機能代表値となる。

【００１２】さらにまた、オリゴペプチドの長さが長くなるほど、まれに出現するオリゴペプチドが多くなる。従って、一般に長さｎによってもＺ(j,n) 値やＤ(j,n) 値は異なるため、様々な長さｎで求めたＺ(j,n) 値とＤ(j,n) 値の関数値Ｗｊ値が機能代表値となる。以下、この発明の方法および装置について、発明の実施の形態をさらに詳しく説明する。

【００１３】

【発明の実施の形態】この発明の蛋白質機能部位予測方法は、ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質から、その生物種ａの任意の蛋白質の機能部位を予測する方法であって、以下のステップ (1)から(9) を構成要件としている。

ステップ(1):生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を求める。

【００１４】たとえば、図１は、メタノコッカス・ヤナシイ[Methanococcus jannaschii] (Bult et. al., Science 273, 1058-1073, 1996) のゲノムデータをもとに、この微生物のゲノムがコードする全蛋白質中での長さ３のオリゴペプチド、長さ４のオリゴペプチド、長さ５のオリゴペプチドの頻度を求め、それぞれの長さについてある回数出現するオリゴペプチドの頻度分布をとったものである。

【００１５】図２は、このステップ(1) を実施するためのフローチャートの例である。

ステップ(2):生物種ａの任意の蛋白質について、

ステップ(2')アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦Ｍ、ただしＭは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さＭのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；

$a_{j1}a_{j2}.....a_{ji}..a_{jn}$（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）の出現頻度と、Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；

$a_{j1}a_{j2}.....X_{i}.....a_{jn}$（Xiは任意のアミノ酸残基を示す）

の出現頻度とを生物種ａの全蛋白質中で求める。

【００１６】このようなＡｊオリゴペプチドとＸｉオリゴペプチドは、例えば図３のように例示することができる。この図３の上段｛1｝は、メタノコッカス・ヤナシイ[Methanococcus jannaschii] (Bult et. al., Science 273, 1058-1073, 1996) の、α型ＤＮＡ合成酵素をコードしていると考えられる遺伝子MJ0885によって予想されるアミノ酸配列について、Ｎ末（アミノ端末）から20番目のアミノ酸残基までの部分配列をシングルレター・コードで表記したもので、中段｛2｝は、５番目のアミノ酸残基Met(M)を含む長さ４のＡｊオリゴペプチドの例を示し、さらにその下｛3｝～｛6｝に５番目のアミノ酸残基Ｍを含むＸｉオリゴペプチドの例を示している。

ステップ(3):ＡｊオリゴペプチドとＸｉオリゴペプチドの出現頻度の比Ｙｊｉを求める。

【００１７】図４は、以上のステップ(2')～(3) を実施するためのフローチャートの例である。

ステップ(4):Ｙｊｉの平均値Ｙ(j,n) を以下のとおりに求める。

Ｙ(j,n) ＝ΣＹｊｉ／ｎ（１≦ｉ≦ｎ）

ステップ(5):Ｙ(j,n) の対数値Ｚ(j,n) を以下のとおりに求める。

【００１８】Ｚ(j,n) ＝-log（Ｙ(j,n)）

図５は、以上のステップ(4) ～(5) を実施するためのフローチャートの例である。

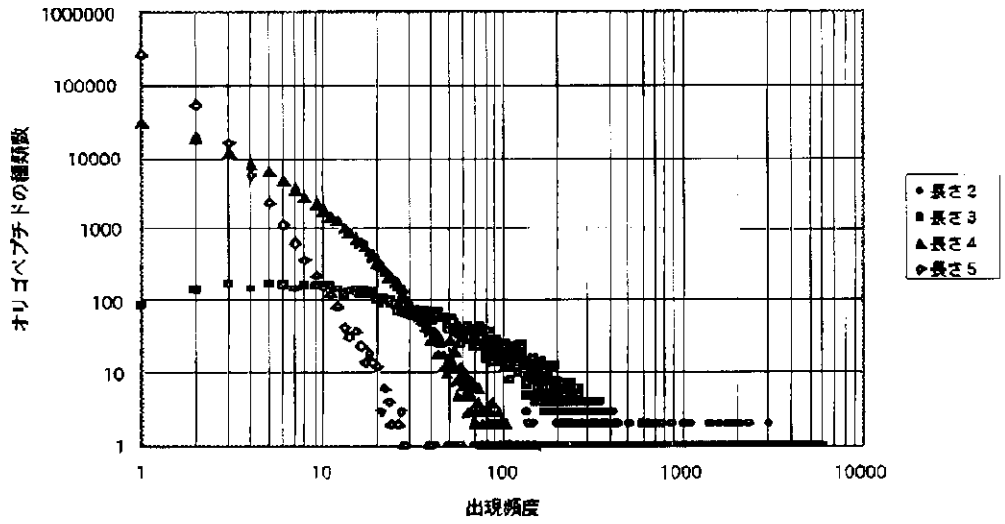
ステップ(6):以下、上記ステップ(2) から(5) を順次繰り返し、アミノ酸配列（長さＬ）のｎ≦ｊ≦Ｌ－ｎ＋１

の位置にある全アミノ酸残基について各々のＺ(j,n) 値を求める。
ステップ(7) ：生物種ａの全蛋白質について上記ステップ(2) から(6) を順次繰り返し、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Ad（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；
ｇ＝（Ｚ(j,n) ，Ａｊ）＝｛Ｚ(j,n) －Ad（Ａａ）｝／Sd（Ａａ）
（ただしＡｊ＝Ａａ）
を求める。
【００１９】例えば、図６は、メタノコッカス・ヤナシイ[Methanococcus jannaschii] (Bult et. al., Science 273, 1058-1073, 1996) のゲノムがコードする全蛋白質におけるＺ(j,n) 値の分布を３種類のアミノ酸、イソロイシン(Ile) 、アラニン(Ala) 、メチオニン(Met) について示している。この分布から例えば、アミノ酸イソロイシン(Ile) におけるＺ(j,n) 値の平均値Ad(Ile) = 3.16、標準偏差値Sd(Ile) = 0.17などが求められ、Ａｊ＝Ile の場合の関数ｇが以下のとおり求められる。
【００２０】
ｇ＝（Ｚ(j,n) ，Ａｊ）＝（Ｚ(j,n) －3.16）／0.17
図７は、このステップ(7) を実施するためのフローチャートの例である。
ステップ(8) ：アミノ酸配列（長さＬ）のｎ≦ｊ≦Ｌ－ｎ＋１の位置にある全アミノ酸残基Ａｊについてステップ(7) で得られた関数ｇの値；
Ｄ(j,n) ＝ｇ（Ｚ(j,n) ，Ａｊ）
を求める。
【００２１】図８は、ステップ(8) を実施するためのフローチャートの例である。
ステップ(9) ：Ｚ(j,n) 値とＤ(j,n) 値の関数値Ｗｊを以下のとおりに求める。
Ｗｊ＝Ｗｊ＝ｈ（Ｚ(j,1) ，Ｚ(j,2) ，... ，Ｚ(j, Ｍ) ，Ｄ(j,1) ，Ｄ(j,2) ，... ，Ｄ(j, Ｍ) ）
そしてこのＷｊの値を、アミノ酸配列（長さＬ）のｊ番目のアミノ酸残基の機能代表値とし、蛋白質の機能に対する各アミノ酸残基の責任の程度をＷｊ値の大きさを指標として予測する。
【００２２】図９は、ステップ(9) を実施するためのフローチャートの例である。なお、各アミノ酸残基のＷｊ値は、例えば、Ｘ軸にアミノ酸配列を、Ｙ軸にＷｊ値をプロットしたような分布図として表示することによって、一目で機能部位を確認することができ、この発明を実施する形態としては好ましい。また、機能部位予測対象の蛋白質の立体構造が既知である場合、または公知の方法（例えば、ホモロジーモデリング法：Peitsch, Proceedings of the fifthinternational cinference on intelligent systems for molecular biology 1997, 5, 234-236）等によって立体構造モデルが作成できる場合には、立体構造上で分布を表示することによって、新規の機能部位の候補となるアミノ酸残基の空間的な配置を確認することができ、この発明を実施する形態として好ましい。
【００２３】最後に、この発明の機能部位予測装置について説明する。すなわち、この発明の装置は、例えば図１０に構成例を示したように、少なくとも以下の (a)から(i) の装置を備えている。
外部記憶装置(a) ：ゲノムデータまたはｃＤＮＡ解析データが既知である生物種ａの予想される全蛋白質のアミノ酸配列データ、および既存の蛋白質データベースを記憶する外部記憶装置。
計算／記憶装置(b)
この生物種ａの全蛋白質のアミノ酸配列について、各アミノ酸残基の出現頻度および各アミノ酸残基を組み合わせて順に長さを長くした各オリゴペプチドの出現頻度を計算するＣＰＵと、その計算結果を記憶する記憶装置とからなる装置。
計算／記憶装置(c) ：この生物種ａの任意の蛋白質について、アミノ酸配列（長さＬ）のＮ末端からｊ番目アミノ酸残基をＡｊとし、この蛋白質のアミノ酸配列の部分配列でｊ番目のアミノ酸残基Ａｊ（ｎ≦ｊ≦Ｌ－ｎ＋１）を含む任意の長さｎ（１≦ｎ≦Ｍ、ただしＭは最初に以下の基準に合致するオリゴペプチドの長さＭ；長さＭのオリゴペプチドはすべて、出現頻度１である）のＡｊオリゴペプチド；
ａj1ａj2.....aji..ａjn（１≦ｉ≦ｎ＋１；Ａｊ＝ａｊｉでＡｊはこのオリゴペプチドのｉ番目の残基を示す）
の出現頻度と、Ａｊオリゴペプチドに対応する長さｎのＸｉオリゴペプチド；
ａj1ａj2.....Xi.....ajn（Xiは任意のアミノ酸残基を示す）
の出現頻度とを生物種ａの全蛋白質中で求めるＣＰＵと、その計算結果を記憶する記憶装置とからなる装置。
計算／記憶装置(d) ：ＡｊオリゴペプチドとＸｉオリゴペプチドの出現頻度の比Ｙｊｉを求めるＣＰＵと、Ｙｊｉを記憶する記憶装置とからなる装置。
計算／記憶装置(e) ：Ｙｊｉの平均値Ｙ(j,n) ；
Ｙ(j,n) ＝ΣＹｊｉ／ｎ（１≦ｉ≦ｎ）
を求めるＣＰＵと、Ｙ(j,n) を記憶する記憶装置とからなる装置。
計算／記憶装置(f) ：Ｙ(j,n) の関数値Ｚ(j,n) ；
Ｚ(j,n) ＝-log（Ｙ(j,n) ）
を求めるＣＰＵと、Ｚ(j,n) を記憶する記憶装置とからなる装置。
計算／記憶装置(g) ：生物種ａの全蛋白質のアミン酸配列について、各アミノ酸残基のＺ(j,n) を求め、アミノ酸残基の種類毎のＺ(j,n) 値の分布を求め、この分布に基づいて各アミノ酸Ａａに対するＺ(j,n) 値の平均値Ad

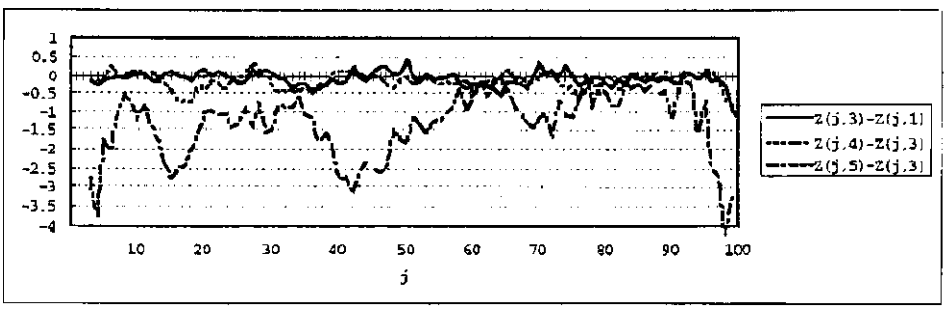
（Ａａ）と標準偏差値Sd（Ａａ）を求め、アミノ酸残基の種類による分布の違いを標準化する関数ｇ；
ｇ＝（Z(j,n)，Ａｊ）＝｛Z(j,n)－Ad（Ａａ）｝／Sd（Ａａ）
（ただしＡｊ＝Ａａ）
を求めるＣＰＵと、ｇを記憶する計算装置とからなる装置。
計算／記憶装置(h)：アミノ酸配列（長さＬ）のｊ番目（ｎ≦ｊ≦Ｌ－ｎ＋１）の位置にある全アミノ酸残基Ａｊについて、装置(g)に記憶された関数ｇの値Ｄ(j,n)；
Ｄ(j,n)＝ｇ（Z(j,n)，Ａｊ）
を求めるＣＰＵと、Ｄ(j,n)値を記憶する記憶装置とからなる装置。
計算／記憶装置(i)：アミノ酸配列について、各アミノ酸残基のZ(j,n)値とＤ(j,n)値の任意の関数値Ｗｊ；
Ｗｊ＝ｈ（Z(j,1)，Z(j,2)，...，Z(j，Ｍ)，Ｄ(j,1)，Ｄ(j,2)，...，Ｄ(j，Ｍ)）
を求める計算装置と、Ｗｊ値を記憶する記憶装置とからなる装置。
【００２４】さらに、この発明の機能部位予測装置においては、以下の(j)～(l)の装置を適宜に組み合わせて備えるようにすることもできる。
ディスプレー装置(j)：アミノ酸配列について、各アミノ酸残基のＷｊ値を分布図として表示する装置。
計算／記憶装置(k)：既存の蛋白質立体構造データベースを記憶し、または公知の方法に従ってアミノ酸配列から立体構造モデルを作成し記憶する装置。
ディスプレー装置(l)：アミノ酸配列について、各アミノ酸残基のＷｊ値を装置(k)に記憶されている立体構造データベースまたは立体構造モデル上に分布図として表示する装置。
【００２５】なお、これらの装置(a)～(l)以外にも、この発明の装置は、図１０にも例示したようなキーバード(m)および制御装置(n)等を備えるようにしてもよい。以下、実施例を示してこの発明のさらに詳細かつ具体的に説明するが、この発明は以下の例によって限定されるものではない。
【００２６】
【実施例】実施例１
メタノコッカス・ヤナシイ[Methanococcus jannaschii]（Bult et. al., Science 273, 1058-1073,(1996)）のゲノムデータをもとに、α型ＤＮＡ合成酵素をコードしていると考えられるこの微生物の遺伝子MJ0885によって予想されるＤＮＡ合成酵素のアミノ酸配列（Ｎ末からＣ末）の各アミノ酸残基について、この発明の方法でZ(j,1)＝-logＹ(j,1)、Z(j,3)＝-logＹ(j,3)、Z(j,4)＝-logＹ(j,4)、Z(j,5)＝-logＹ(j,5)を算出し、Ｗｊ＝Z(j,3)－Z(j,1)を算出した（ｈ＝Z(j,3)－Z(j,1)）。同様に、Ｗｊ＝Z(j,4)－Z(j,3)（ｈ＝Z(j,4)－Z(j,3)）、Ｗｊ＝Z(j,5)－Z(j,3)（ｈ＝Z(j,5)－Z(j,3)）を算出した。
【００２７】図１１は、Ｎ末から１００残基についてこれらの結果を分布図としてプロットしたものでである。ｈ＝Z(j,5)－Z(j,3)の場合、他の二つの場合と比べて大きく分布が異なる領域がＮ末から３５残基目から６０残基目にかけての領域等に存在している。この分布からＷｊ＝Z(j,5)－Z(j,3)が小さくなることによってアミノ酸配列が特徴づけられることが分かる。
【００２８】さらに、α型ＤＮＡ合成酵素の機能部位として知られているモチーフ部分のうち、エクソⅠ（exoⅠ）、エクソII(exoII)、モチーフＡ(motif A)，モチーフＢ(motif B)およびモチーフＣ(motif C)を含む領域を抜粋し、それぞれのアミノ酸残基のＷｊ値を図１２にプロットした。この図１２に示したように、Ｗｊ値が小さくなることによって特徴づけられる部分と機能部位が対応していることが分かる。
実施例２
図１３は、α型ＤＮＡ合成酵素の機能部位として知られているモチーフ部分のうち、エクソⅠ（exoⅠ）、エクソII(exoII)、モチーフＡ(motif A)，モチーフＢ(motif B)およびモチーフＣ(motif C)を含む領域を抜粋し、それぞれのアミノ酸残基のＷｊ＝Ｄ(j,3)値とＷｊ＝Ｄ(j,5)値をプロットしたものである（ｈ＝Ｄ(j,3)とｈ＝Ｄ(j,5)）。Ｗｊ＝Ｄ(j,n)が２以上または２以下になっているアミノ酸残基がモチーフ部分以外にも存在しており、これらのアミノ酸残基が新たな機能部位の候補となる。
実施例３
図１４は、メタノコッカス・ヤナシイ[Methanococcus jannaschii]のエノラーゼであると予想されるMJ0232のアミノ酸配列について、Ｗｊ＝Ｄ(j,3)値が２以上または２以下になっているアミノ酸残基の立体構造における位置を、出芽酵母菌のエノラーゼを基に公知の方法で作成した立体構造モデル上に濃色で表示したものである。アミノ酸配列上では離れた位置にある残基が立体構造では近くにあることが分かる。
【００２９】
【発明の効果】以上詳しく説明したとおり、この発明によって、ゲノム解析やｃＤＮＡ解析から得られた機能未知の蛋白質について、その機能部位を予測するが可能となる。また、機能既知の蛋白質についても、その新たな機能部位を予測することも可能となる。
【図面の簡単な説明】
【図１】長さ２のオリゴペプチド、長さ３のオリゴペプチド、長さ４のオリゴペプチド、長さ５のオリゴペプチドの各々の出現頻度を求め、それぞれの長さについてある回数出現するオリゴペプチドの頻度分布である。
【図２】この発明方法のステップ(1)を実施するためのフローチャートの例である。

【図３】長さ２０のアミノ酸配列、この配列の５番目のアミノ酸残基Metを含む長さ４のＡｊオリゴペプチド、およびＸｉオリゴペプチドの例である。
【図４】この発明方法のステップ(2')～(3) を実施するためのフローチャートの例である。
【図５】この発明方法のステップ(4) ～(5) を実施するためのフローチャートの例である。
【図６】アミノ酸の種類毎のＺ(j,3) 値の頻度分布である。実線はイソロイシン(Ile)、破線はアラニン(Ala) 、一点鎖線はメチオニン(Met) における分布を示している。
【図７】この発明方法のステップ(7) を実施するためのフローチャートの例である。
【図８】この発明方法のステップ(8) を実施するためのフローチャートの例である。
【図９】この発明方法のステップ(9) を実施するためのフローチャートの例である。
【図１０】この発明の装置を例示した構成図である。
【図１１】MJ0885でコードされるα型ＤＮＡ合成酵素をコードする全アミノ酸配列のＮ末から１００残基について、この発明の方法により算出したＷｊ＝Ｚ(j,3) －Ｚ(j,1) 値（実線）、Ｗｊ＝Ｚ(j,4) －Ｚ(j,3) 値（破線）、Ｗｊ＝Ｚ(j,5) －Ｚ(j,3) 値（一点鎖線）をプロットした分布図である。
【図１２】MJ0885でコードされるα型ＤＮＡ合成酵素のアミノ酸配列の部分配列（エクソＩ（exoI）、エクソII(exoII) 、モチーフＡ(motif A) , モチーフＢ(motif B)およびモチーフＣ(motif C) を含む領域）について、Ｗｊ＝Ｚ(j,5) －Ｚ(j,3)の値をプロットした分布図である。
【図１３】MJ0885でコードされるα型ＤＮＡ合成酵素のアミノ酸配列の部分配列（エクソＩ（exoI）、エクソII(exoII) 、モチーフＡ(motif A) , モチーフＢ(motif B)およびモチーフＣ(motif C) を含む領域）について、Ｗｊ＝Ｄ(j,3) 値（濃色）とＷｊ＝Ｄ(j,5) 値（淡色）をプロットした分布図である。
【図１４】MJ0232でコードされるエノラーゼのアミノ酸配列について、Ｗｊ＝Ｄ(j,3) 値が２以上または２以下になっているアミノ酸残基の立体構造における位置を、立体構造モデル上に濃色で示した分布図である。

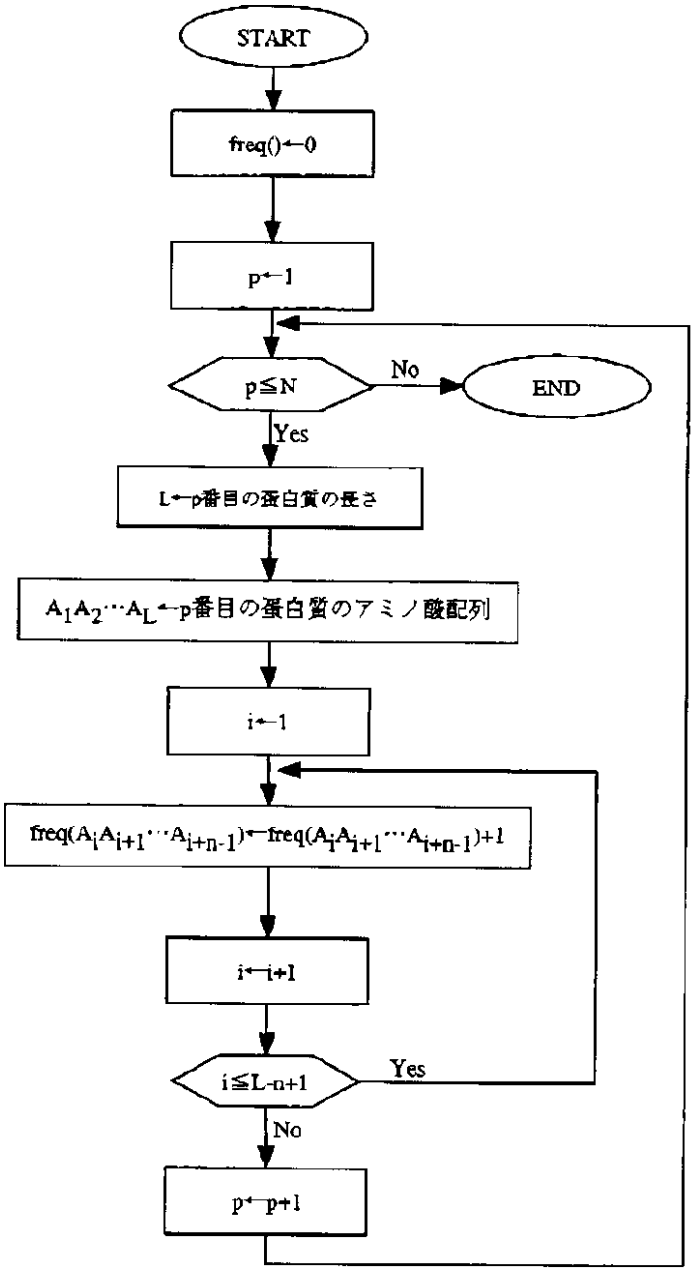
【図１】



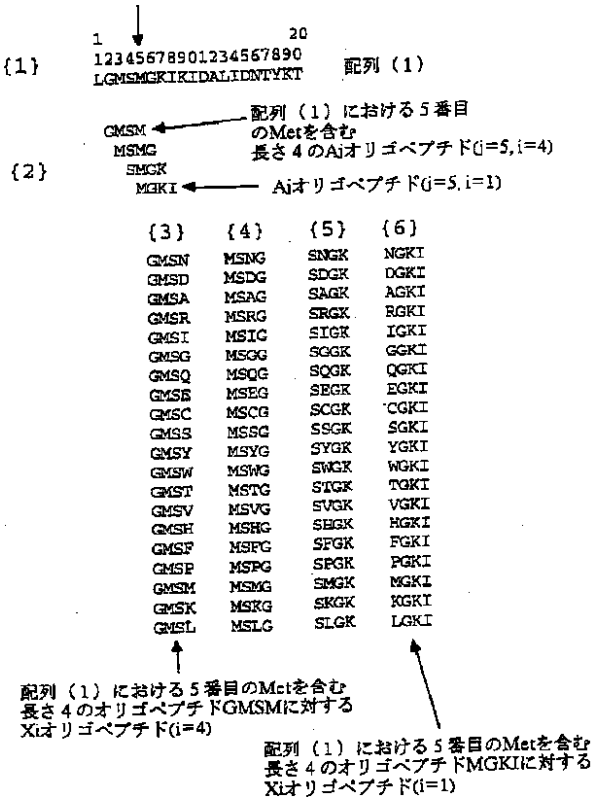
【図１１】



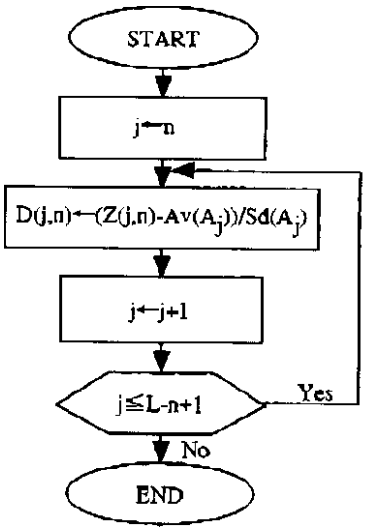
【図２】


START
freq()←0
p←1
p≦N
No
END
Yes
L←p番目の蛋白質の長さ
A1A2…AL←p番目の蛋白質のアミノ酸配列
i←1
freq(AiAi+1…Ai+n-1)←freq(AiAi+1…Ai+n-1)+1
i←i+1
i≦L-n+1
Yes
No
p←p+1

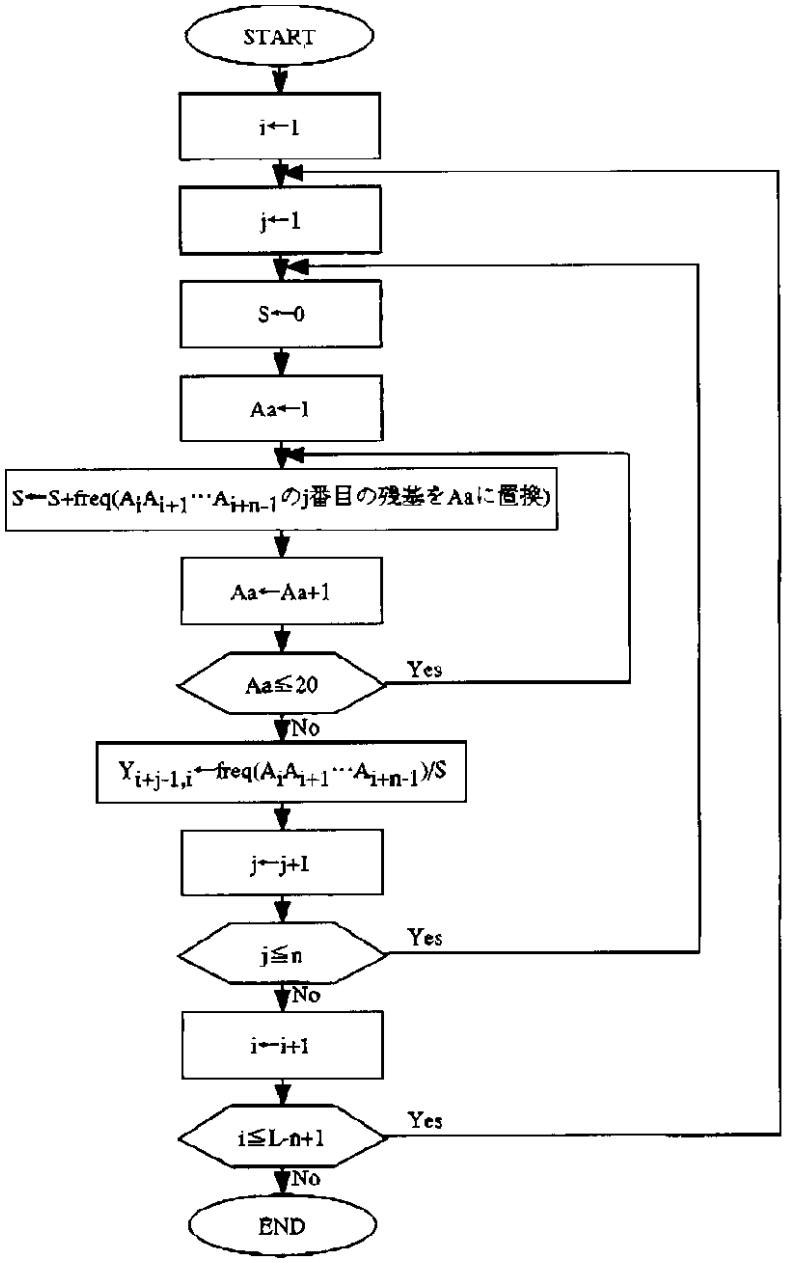
【図３】


1 20
12345678901234567890
{1} LGMSMGKIKIDALIDNTYKT 配列（1）
配列（1）における５番目のMetを含む長さ４のAiオリゴペプチド(j=5,i=4)
{2}
GMSM
MSMG
SMGK
MGKI
Aiオリゴペプチド(j=5,i=1)
{3} {4} {5} {6}
GMSN MSNG SNGK NGKI
GMSD MSDG SDGK DGKI
GMSA MSAG SAGK AGKI
GMSR MSRG SRGK RGKI
GMSI MSIG SIGK IGKI
GMSG MSGG SGGK GGKI
GMSQ MSQG SQGK QGKI
GMSE MSEG SEGK EGKI
GMSC MSCG SCGK CGKI
GMSS MSSG SSGK SGKI
GMSY MSYG SYGK YGKI
GMSW MSWG SWGK WGKI
GMST MSTG STGK TGKI
GMSV MSVG SVGK VGKI
GMSH MSHG SHGK HGKI
GMSF MSFG SFGK FGKI
GMSP MSPG SPGK PGKI
GMSM MSMG SMGK MGKI
GMSK MSKG SKGK KGKI
GMSL MSLG SLGK LGKI
配列（1）における５番目のMetを含む長さ４のオリゴペプチドGMSMに対するXiオリゴペプチド(i=4)
配列（1）における５番目のMetを含む長さ４のオリゴペプチドMGKIに対するXiオリゴペプチド(i=1)

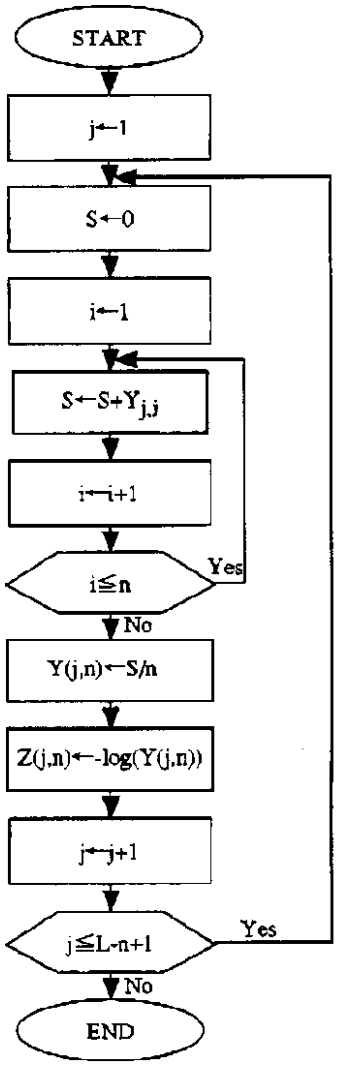
【図８】


START
j←n
D(j,n)←(Z(j,n)-Av(Aj))/Sd(Aj)
j←j+1
j≦L-n+1
Yes
No
END

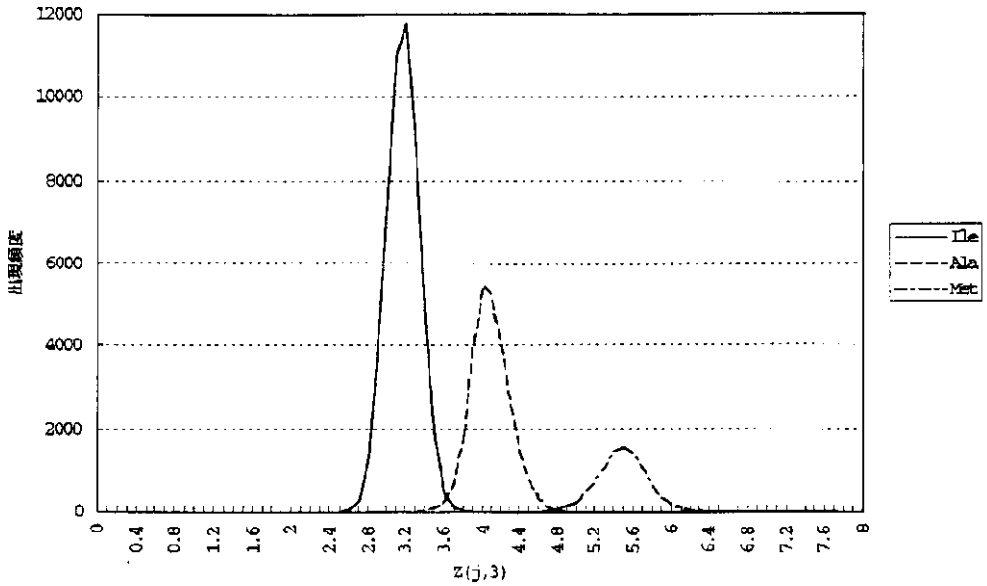
【図４】


START
i←1
j←1
S←0
Aa←1
S←S+freq(AiAi+1…Ai+n-1のj番目の残基をAaに置換)
Aa←Aa+1
Aa≦20
Yes
No
Yi+j-1,i←freq(AiAi+1…Ai+n-1)/S
j←j+1
j≦n
Yes
No
i←i+1
i≦L-n+1
Yes
No
END

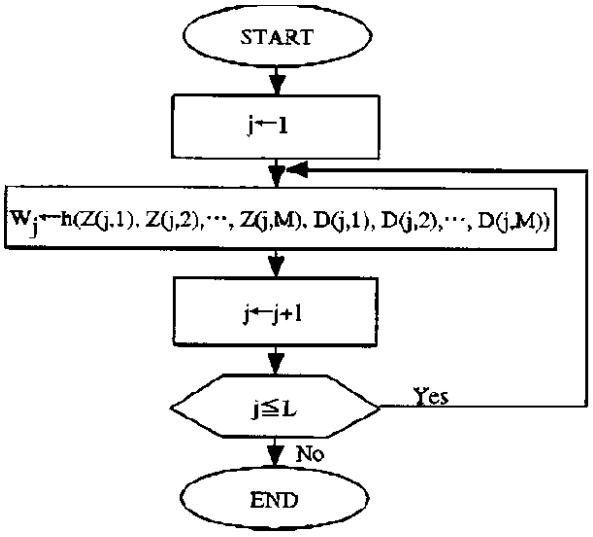
【図５】


START
j←1
S←0
i←1
S←S+Yj,j
i←i+1
i≦n
Yes
No
Y(j,n)←S/n
Z(j,n)←-log(Y(j,n))
j←j+1
j≦L-n+1
Yes
No
END

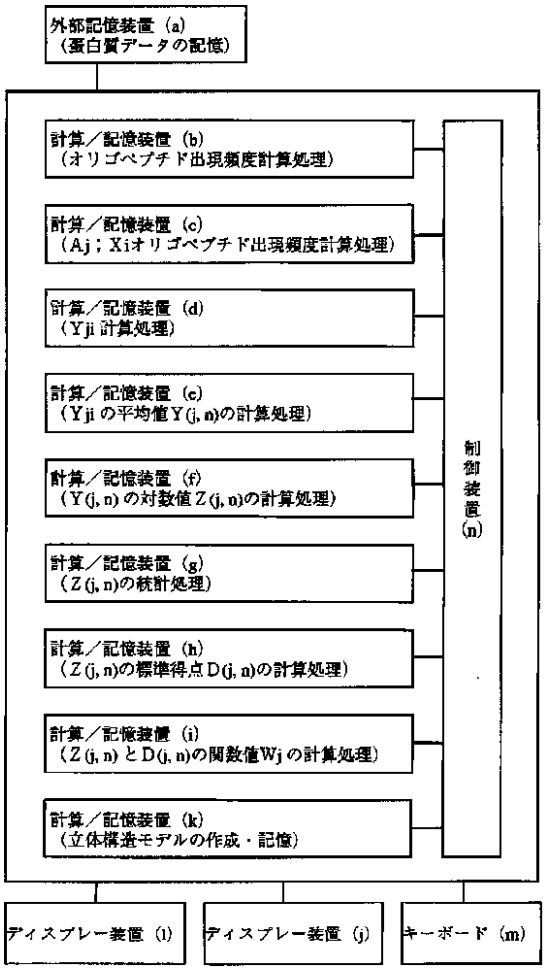
【図６】


12000
10000
8000
6000
4000
2000
0
出現頻度
0
0.4
0.8
1.2
1.6
2
2.4
2.8
3.2
3.6
4
4.4
4.8
5.2
5.6
6
6.4
6.8
7.2
7.6
8
Z(j,3)
Ile
Ala
Met

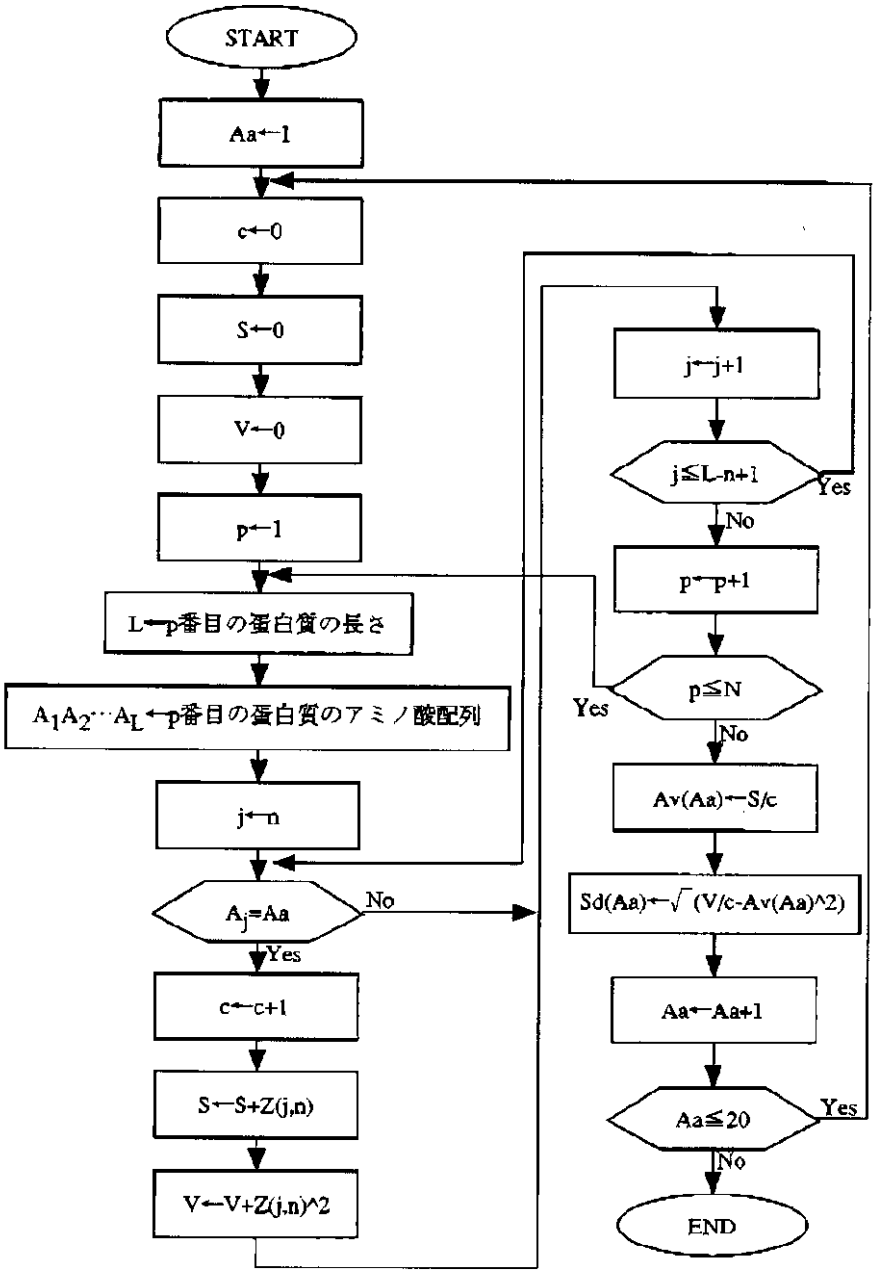
【図９】


START
j←1
Wj←h(Z(j,1), Z(j,2),…, Z(j,M), D(j,1), D(j,2),…, D(j,M))
j←j+1
j≦L
Yes
No
END

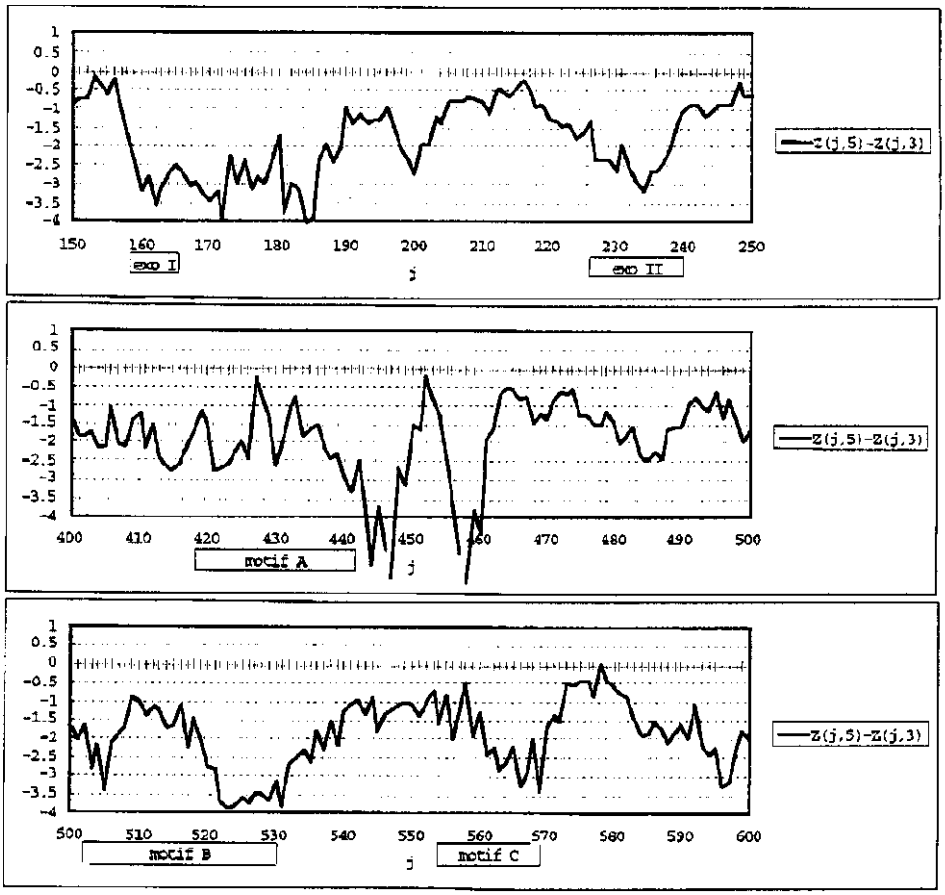
【図１０】


外部記憶装置（a）
（蛋白質データの記憶）
計算／記憶装置（b）
（オリゴペプチド出現頻度計算処理）
計算／記憶装置（c）
（Aj；Xiオリゴペプチド出現頻度計算処理）
計算／記憶装置（d）
（Yji 計算処理）
計算／記憶装置（e）
（Yji の平均値Y(j, n)の計算処理）
計算／記憶装置（f）
（Y(j, n) の対数値Z(j, n)の計算処理）
計算／記憶装置（g）
（Z(j, n)の統計処理）
計算／記憶装置（h）
（Z(j, n)の標準得点D(j, n)の計算処理）
計算／記憶装置（i）
（Z(j, n) とD(j, n)の関数値Wj の計算処理）
計算／記憶装置（k）
（立体構造モデルの作成・記憶）
制御装置（n）
ディスプレー装置（l）
ディスプレー装置（j）
キーボード（m）

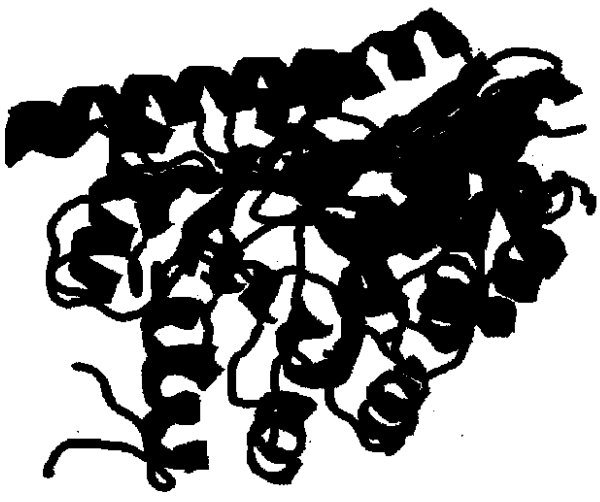
【図７】


START
Aa←1
c←0
S←0
V←0
p←1
L←p番目の蛋白質の長さ
A1A2…AL←p番目の蛋白質のアミノ酸配列
j←n
Aj=Aa
No
Yes
c←c+1
S←S+Z(j,n)
V←V+Z(j,n)^2
j←j+1
j≦L-n+1
Yes
No
p←p+1
p≦N
Yes
No
Av(Aa)←S/c
Sd(Aa)←√(V/c-Av(Aa)^2)
Aa←Aa+1
Aa≦20
Yes
No
END

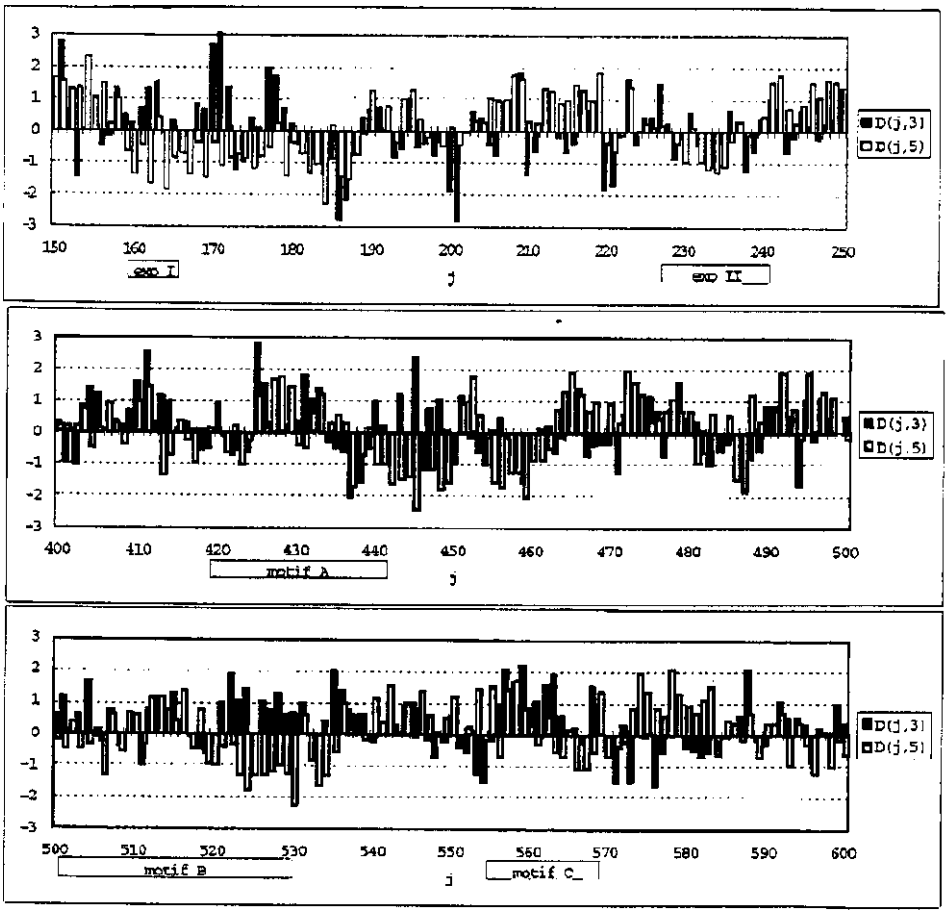
【図１２】



【図１４】



【図１３】


3
2
1
0
-1
-2
-3
150 160 170 180 190 200 210 220 230 240 250
exp I
j
exp II
D(j,3)
D(j,5)
400 410 420 430 440 450 460 470 480 490 500
motif A
500 510 520 530 540 550 560 570 580 590 600
motif B
motif C