

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3792039号  
(P3792039)

(45) 発行日 平成18年6月28日(2006.6.28)

(24) 登録日 平成18年4月14日(2006.4.14)

(51) Int. Cl.		F I		
<b>G06F 19/00</b>	<b>(2006.01)</b>	G06F 19/00	600	
<b>G06F 17/30</b>	<b>(2006.01)</b>	G06F 17/30	170F	
C07K 14/195	(2006.01)	C07K 14/195		
C12N 15/09	(2006.01)	C12N 15/00	A	

請求項の数 6 (全 16 頁)

(21) 出願番号	特願平10-18699	(73) 特許権者	503360115
(22) 出願日	平成10年1月30日(1998.1.30)		独立行政法人科学技術振興機構
(65) 公開番号	特開平11-213003		埼玉県川口市本町4丁目1番8号
(43) 公開日	平成11年8月6日(1999.8.6)	(73) 特許権者	597014682
審査請求日	平成16年11月12日(2004.11.12)		土居 洋文
			千葉県船橋市夏見5-29-4-515
		(73) 特許権者	500520628
			セレスター・レキシコ・サイエンス株式
			会社
			千葉県千葉市美浜区中瀬1丁目3番地 幕
			張テクノガーデンD17
		(74) 代理人	100093230
			弁理士 西澤 利夫
		(72) 発明者	土居 洋文
			千葉県船橋市夏見5-29-4-515
			最終頁に続く

(54) 【発明の名称】 蛋白質機能部位の予測方法と予測装置

(57) 【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特定する方法であって、

(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る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A<sub>j</sub>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A<sub>j</sub>(n<sub>j</sub> = 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A<sub>j</sub>オリゴペプチド ;

a<sub>j1</sub>a<sub>j2</sub>.....a<sub>ji</sub> .. a<sub>jn</sub>(1 ≤ i ≤ n + 1 ; A<sub>j</sub> = a<sub>ji</sub>でA<sub>j</sub>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、

A<sub>j</sub>オリゴペプチドに対応する長さnのX<sub>i</sub>オリゴペプチド ;

a<sub>j1</sub>a<sub>j2</sub>.....X<sub>i</sub>.....a<sub>jn</sub>(X<sub>i</sub>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) A<sub>j</sub>オリゴペプチドとX<sub>i</sub>オリゴペプチドの出現頻度の比Y<sub>ji</sub>を求め、

(4) Y<sub>ji</sub>の平均値Y(j,n) ;

Y(j,n) = Y<sub>ji</sub> / n(1 ≤ i ≤ n)

10

20

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  $A_j$  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  $A_a$  に対する  $Z(j, n)$  値の平均値  $Av(A_a)$  と標準偏差値  $Sd(A_a)$  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  $g$  ;

$$g = (Z(j, n), A_j) = \{Z(j, n) - Av(A_a)\} / Sd(A_a) \quad (\text{ただし } A_j = A_a)$$

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  $A_j$  についてステップ(7)で得られた関数  $g$  の値  $D(j, n)$  ;

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

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  $W_j$  ;

$$W_j = h(Z(j, 1), Z(j, 2), \dots, Z(j, M), D(j, 1), D(j, 2), \dots, D(j, M))$$

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  $W_j$  値の大きさを指標として特定すること特徴とする蛋白質の機能部位特定方法。

【請求項 2】

各アミノ酸残基の  $W_j$  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  $W_j$  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る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  $A_j$  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  $A_j$  ( $n - j - L - n + 1$ )を含む任意の長さ  $n$  ( $1 \leq n \leq M$ 、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  $A_j$  オリゴペプチド ;

$a_{j_1} a_{j_2} \dots a_{j_i} \dots a_{j_n}$  ( $1 \leq i \leq n + 1$  ;  $A_j = a_{j_i}$  で  $A_j$  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、

$A_j$  オリゴペプチドに対応する長さ  $n$  の  $X_i$  オリゴペプチド ;

$a_{j_1} a_{j_2} \dots X_i \dots a_{j_n}$  ( $X_i$  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)  $A_j$  オリゴペプチドと  $X_i$  オリゴペプチドの出現頻度の比  $Y_{j_i}$  を求める CPU と、 $Y_{j_i}$  を記憶する記憶装置とからなる計算/記憶装置、

(e)  $Y_{j_i}$  の平均値  $Y(j, n)$  ;

10

20

30

40

50

$$Y(j,n) = Y_{j,i} / n(1-i-n)$$

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 の全蛋白質のアミノ酸配列について、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  $A_a$  に対する  $Z(j,n)$  値の平均値  $Av(A_a)$  と標準偏差値  $Sd(A_a)$  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  $g$  ;

$$g = (Z(j,n), A_j) = \{ Z(j,n) - Av(A_a) \} / Sd(A_a) \quad (\text{ただし } A_j = A_a) \quad 10$$

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  $A_j$  について、装置(g)に記憶された関数  $g$  の値  $D(j,n)$  ;

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

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  $W_j$  ;

$$W_j = h(Z(j,1), Z(j,2), \dots, Z(j,M), D(j,1), D(j,2), \dots, D(j,M))$$

を求める計算装置と、 $W_j$  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特定装置。 20

#### 【請求項 5】

アミノ酸配列について、各アミノ酸残基の  $W_j$  値を 2 次元的な分布図として表示するディスプレイ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  $W_j$  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イ装置を備えた請求項 4 の装置。

#### 【発明の詳細な説明】

##### 【0001】 30

#### 【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 [*Haemophilus influenzae*] (Fleischmann et al., Science 269, 496-512, 1995)、メタノコッカス・ヤナシイ [*Methanococcus jannaschii*] (Bult et 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。 40

##### 【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た。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 50

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 $A_j$ 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 $A_j$ ( $n - j - L - n + 1$ )を含む任意の長さ $n$ ( $1 \leq n \leq M$ 、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 $A_j$ オリゴペプチド;

$a_{j_1} a_{j_2} \dots a_{j_i} \dots a_{j_n}$  ( $1 \leq i \leq n + 1$ ;  $A_j = a_{j_i}$ で $A_j$ 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、

$A_j$ オリゴペプチドに対応する長さnの $X_i$ オリゴペプチド;

$a_{j_1} a_{j_2} \dots X_i \dots a_{j_n}$  ( $X_i$ 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)  $A_j$ オリゴペプチドと $X_i$ オリゴペプチドの出現頻度の比 $Y_{j_i}$ を求め、

(4)  $Y_{j_i}$ の平均値 $Y(j, n)$ ;

$Y(j, n) = Y_{j_i} / n$  ( $1 \leq i \leq n$ )

を求め、

(5)  $Y(j, n)$ の関数値 $Z(j, n)$ ;

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

を求め、

10

20

30

40

50

(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A<sub>j</sub>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A<sub>a</sub>に対するZ(j,n)値の平均値Av(A<sub>a</sub>)と標準偏差値Sd(A<sub>a</sub>)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;

$$g = (Z(j,n), A_j) = \{ Z(j,n) - Ad(A_a) \} / Sd(A_a)$$

(ただしA<sub>j</sub> = A<sub>a</sub>)

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A<sub>j</sub>についてステップ(7)で得られた関数gの値D(j,n);

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

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W<sub>j</sub>;

$$W_j = h(Z(j,1), Z(j,2), \dots, Z(j,M), D(j,1), D(j,2), \dots, D(j,M))$$

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W<sub>j</sub>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る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A<sub>j</sub>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A<sub>j</sub>(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A<sub>j</sub>オリゴペプチド;

$$a_{j1} a_{j2} \dots a_{ji} \dots a_{jn} \quad (1 \leq i \leq n+1; A_j = a_{ji} \text{で} A_j \text{はこのオリゴペプチドの} i \text{番目の残基を示す})$$

の出現頻度と、

A<sub>j</sub>オリゴペプチドに対応する長さnのX<sub>i</sub>オリゴペプチド;

$$a_{j1} a_{j2} \dots X_i \dots a_{jn} \quad (X_i \text{は任意のアミノ酸残基を示す})$$

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) A<sub>j</sub>オリゴペプチドとX<sub>i</sub>オリゴペプチドの出現頻度の比Y<sub>ji</sub>を求めるCPUと、Y<sub>ji</sub>を記憶する記憶装置とからなる計算/記憶装置、

(e) Y<sub>ji</sub>の平均値Y(j,n);

$$Y(j,n) = \sum_{i=1}^n Y_{ji} / n$$

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の全蛋白質のアミノ酸配列について、各アミノ酸残基のZ(j,n)を求め、ア

10

20

30

40

50

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  $A_a$  に対する  $Z(j,n)$  値の平均値  $Ad(A_a)$  と標準偏差値  $Sd(A_a)$  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  $g$  ;

$$g = (Z(j,n), A_j) = \{Z(j,n) - Ad(A_a)\} / Sd(A_a)$$

(ただし  $A_j = A_a$ )

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( $n = j \sim L - n + 1$ )の位置にある全アミノ酸残基  $A_j$  について、装置(g)に記憶された関数  $g$  の値  $D(j,n)$  ;

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

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  $W_j$  ;

$$W_j = h(Z(j,1), Z(j,2), \dots, Z(j,M), D(j,1),$$

$$D(j,2), \dots, D(j,M))$$

を求める計算装置と、 $W_j$  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で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られる。

#### 【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  $A_j$  オリゴペプチドと  $X_i$  オリゴペプチドの出現頻度の比  $Y_{ji}$  によって、アミノ酸残基  $A_j$  が  $A_j$  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  $A_j$  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  $A_j$  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  $A_j$  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  $W_j$  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。

10

20

30

40

50

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。

【 0 0 1 4 】

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。

【 0 0 1 5 】

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  $A_j$  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  $A_j$  ( $n = j, L - n + 1$ ) を含む任意の長さ  $n$  ( $1 \leq n \leq M$ 、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  $A_j$  オリゴペプチド;

$a_{j_1} a_{j_2} \dots a_{j_i} \dots a_{j_n}$  ( $1 \leq i \leq n$ ;  $A_j = a_{j_i}$  で  $A_j$  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、

$A_j$  オリゴペプチドに対応する長さ n の  $X_i$  オリゴペプチド;

$a_{j_1} a_{j_2} \dots X_i \dots a_{j_n}$  ( $X_i$  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。

【 0 0 1 6 】

このような  $A_j$  オリゴペプチドと  $X_i$  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、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(アミノ末端)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  $A_j$  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  $X_i$  オリゴペプチドの例を示している。

ステップ(3):

$A_j$  オリゴペプチドと  $X_i$  オリゴペプチドの出現頻度の比  $Y_{j_i}$  を求める。

【 0 0 1 7 】

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):

$Y_{j_i}$  の平均値  $Y(j, n)$  を以下のとおりに求める。

$$Y(j, n) = \sum_{i=1}^n Y_{j_i} / n$$

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。

【 0 0 1 8 】

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

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

10

20

30

40

50

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  $A_a$  に対する  $Z(j, n)$  値の平均値  $Ad(A_a)$  と標準偏差値  $Sd(A_a)$  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  $g$  ;

$$g = (Z(j, n), A_j) = \{ Z(j, n) - Ad(A_a) \} / Sd(A_a)$$

(ただし  $A_j = A_a$ )

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、 $A_j = Ile$  の場合の関数  $g$  が以下のとおり求められる。

#### 【0020】

$$g = (Z(j, n), A_j) = (Z(j, n) - 3.16) / 0.17$$

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  $A_j$  についてステップ(7)で得られた関数  $g$  の値 ;

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

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  $W_j$  を以下のとおりに求める。

$$W_j = W_j = h(Z(j, 1), Z(j, 2), \dots, Z(j, M), D(j, 1), D(j, 2), \dots, D(j, M))$$

そしてこの  $W_j$  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  $W_j$  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  $W_j$  値は、例えば、X軸にアミノ酸配列を、Y軸に  $W_j$  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 fifth international conference 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

10

20

30

40

50



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  $A_j$  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  $A_j$  ( $n = j, L - n + 1$ ) を含む任意の長さ n ( $1 \leq n \leq M$ 、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  $A_j$  オリゴペプチド ;

$a_{j_1} a_{j_2} \dots a_{j_i} \dots a_{j_n}$  ( $1 \leq i \leq n+1$  ;  $A_j = a_{j_i}$  で  $A_j$  はこのオリゴペプチドの i 番目の残基を示す)

10

の出現頻度と、

$A_j$  オリゴペプチドに対応する長さ n の  $X_i$  オリゴペプチド ;

$a_{j_1} a_{j_2} \dots X_i \dots a_{j_n}$  ( $X_i$  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) :

$A_j$  オリゴペプチドと  $X_i$  オリゴペプチドの出現頻度の比  $Y_{j_i}$  を求める CPU と、 $Y_{j_i}$  を記憶する記憶装置とからなる装置。

計算/記憶装置(e) :

$Y_{j_i}$  の平均値  $Y(j, n)$  ;

20

$Y(j, n) = \sum Y_{j_i} / n$  ( $1 \leq i \leq n$ )

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 の全蛋白質のアミノ酸配列について、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  $A_a$  に対する  $Z(j, n)$  値の平均値  $Ad(A_a)$  と標準偏差値  $Sd(A_a)$  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 ;

30

$g = (Z(j, n), A_j) = \{ Z(j, n) - Ad(A_a) \} / Sd(A_a)$

(ただし  $A_j = A_a$ )

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  $A_j$  について、装置(g) に記憶された関数 g の値  $D(j, n)$  ;

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

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) :

40

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  $W_j$  ;

$W_j = h(Z(j, 1), Z(j, 2), \dots, Z(j, M), D(j, 1),$

$D(j, 2), \dots, D(j, M))$

を求める計算装置と、 $W_j$  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るようによることができる。

ディスプレイ装置(j) :

アミノ酸配列について、各アミノ酸残基の  $W_j$  値を分布図として表示する装置。

計算/記憶装置(k) :

50

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。

ディスプレイ装置(l) :

アミノ酸配列について、各アミノ酸残基の $W_j$ 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ボード(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、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 $Z(j,1) = -\log Y(j,1)$ 、 $Z(j,3) = -\log Y(j,3)$ 、 $Z(j,4) = -\log Y(j,4)$ 、 $Z(j,5) = -\log Y(j,5)$ を算出し、 $W_j = Z(j,3) - Z(j,1)$ を算出した( $h = Z(j,3) - Z(j,1)$ )。同様に、 $W_j = Z(j,4) - Z(j,3)$  ( $h = Z(j,4) - Z(j,3)$ )、 $W_j = Z(j,5) - Z(j,3)$  ( $h = Z(j,5) - Z(j,3)$ )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 $W_j = Z(j,5) - Z(j,3)$ 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、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 $W_j$ 値を図12にプロットした。この図12に示したように、 $W_j$ 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、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 $W_j = D(j,3)$ 値と $W_j = D(j,5)$ 値をプロットしたものである( $h = D(j,3)$ と $h = D(j,5)$ )。  $W_j = D(j,n)$ 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、 $W_j = D(j,3)$ 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

10

20

30

40

50

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A<sub>j</sub>オリゴペプチド、およびX<sub>i</sub>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 $W_j = Z(j,3) - Z(j,1)$ 値(実線)、 $W_j = Z(j,4) - Z(j,3)$ 値(破線)、 $W_j = Z(j,5) - Z(j,3)$ 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、 $W_j = Z(j,5) - Z(j,3)$ 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、 $W_j = D(j,3)$ 値(濃色)と $W_j = D(j,5)$ 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、 $W_j = D(j,3)$ 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。

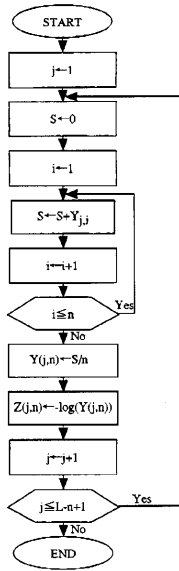
10

20

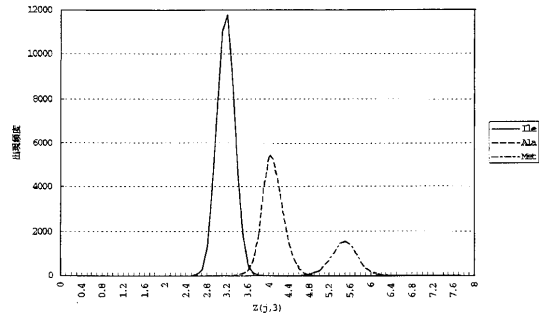
30



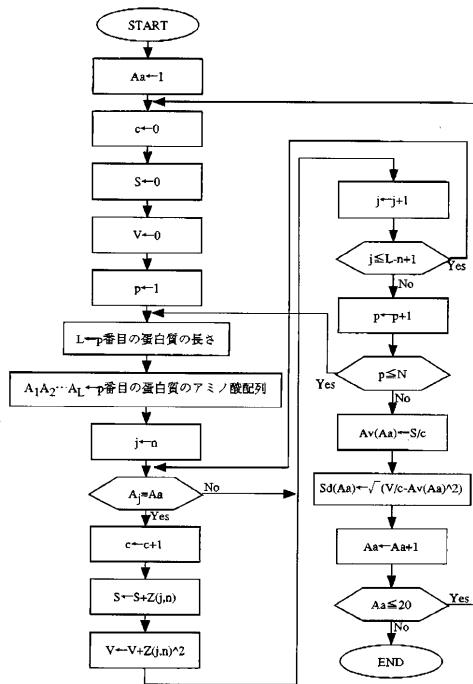
【 図 5 】



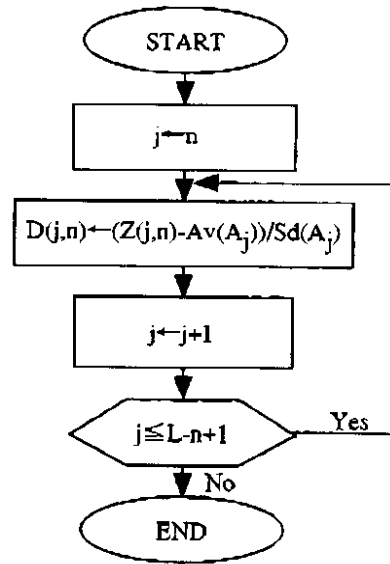
【 図 6 】



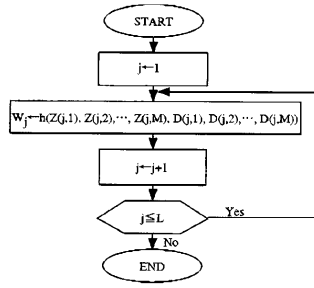
【 図 7 】



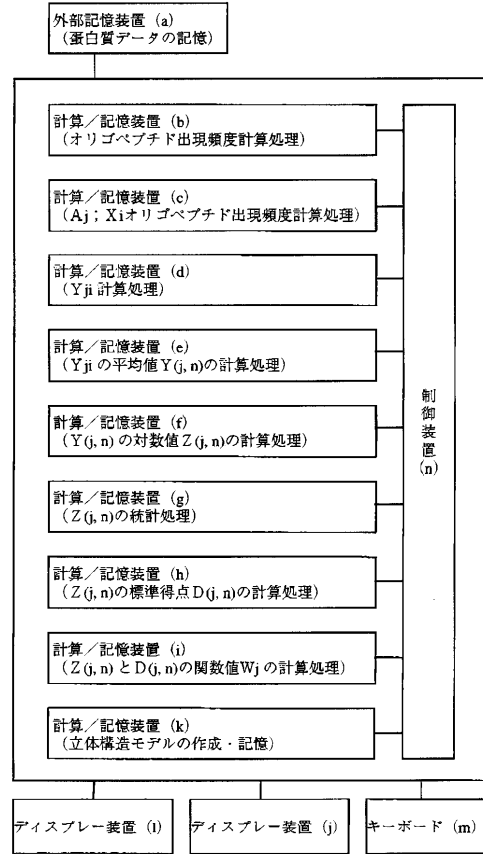
【 図 8 】



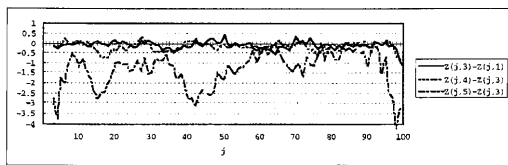
【 図 9 】



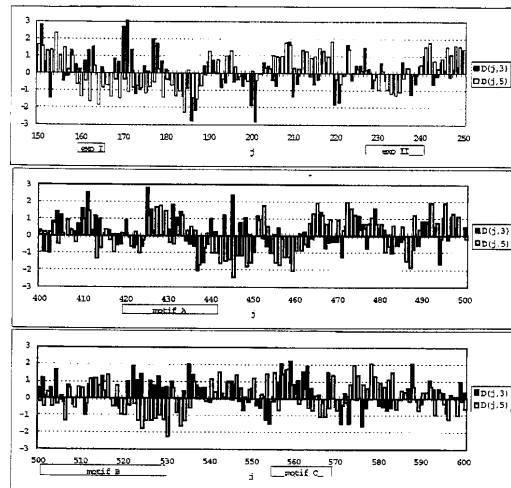
【 図 10 】



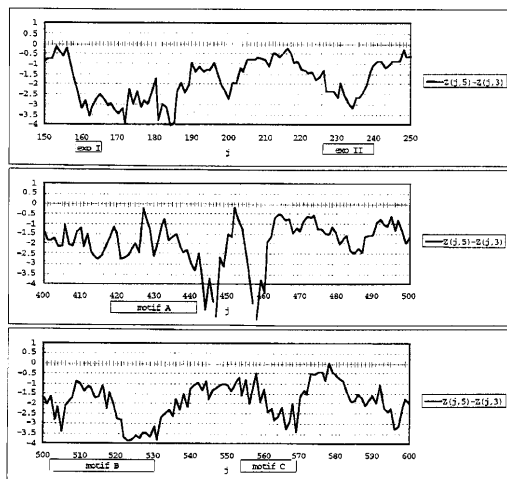
【 図 11 】



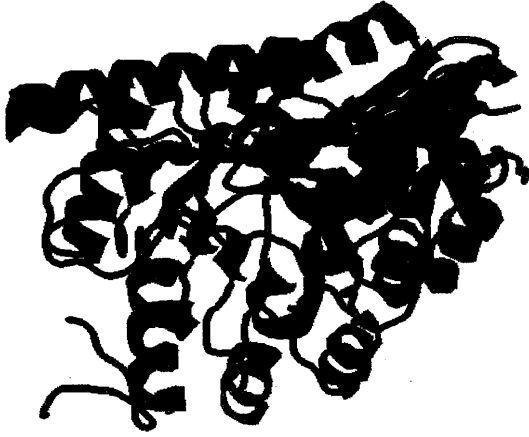
【 図 13 】



【 図 12 】



【 図 1 4 】



## フロントページの続き

(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

審査官 岩間 直純

(56)参考文献 特開平10-222486(JP,A)

特開平10-197529(JP,A)

特開平10-045795(JP,A)

DOI H, et al., Proc Natl Acad Sci USA, 米国, 1995年 3月28日, 92(7), 2879-83

金井昭夫, ゲノム情報工学を用いた古細菌DNA合成酵素のエンジニアリング, 創造科学技術推進事業 1997 創造科学技術研究報告会(東京)第3部講演要旨集, 日本, 1997年, 85-92

平木秀明, オリゴ配列の頻度によるゲノム情報工学, 創造科学技術推進事業 1997 創造科学技術研究報告会(東京)第3部講演要旨集, 日本, 1997年, 78-84

SOLOVYEV V. V., MAKAROVA K. S., Comput Appl Biosci, 1993年, vol. 9, no. 1, pp. 17-24

(58)調査した分野(Int.Cl., DB名)

G06F 19/00

G06F 17/30

C07K 14/195

C12N 15/00

JICSTファイル(JOIS)