

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5991524号
(P5991524)

(45) 発行日 平成28年9月14日(2016.9.14)

(24) 登録日 平成28年8月26日(2016.8.26)

(51) Int.Cl. F I
G O 6 F 19/22 (2011.01) G O 6 F 19/22

請求項の数 18 (全 27 頁)

(21) 出願番号	特願2012-189907 (P2012-189907)	(73) 特許権者	801000027
(22) 出願日	平成24年8月30日 (2012.8.30)		学校法人明治大学
(65) 公開番号	特開2014-48797 (P2014-48797A)		東京都千代田区神田駿河台1-1
(43) 公開日	平成26年3月17日 (2014.3.17)	(74) 代理人	100064908
審査請求日	平成27年3月23日 (2015.3.23)		弁理士 志賀 正武
		(74) 代理人	100106909
			弁理士 棚井 澄雄
		(74) 代理人	100108578
			弁理士 高橋 詔男
		(74) 代理人	100126882
			弁理士 五十嵐 光永
		(72) 発明者	田中 大貴
			神奈川県川崎市多摩区東三田1-1-1
			学校法人明治大学 生田キャンパス内

最終頁に続く

(54) 【発明の名称】 位置特異的スコアの算出装置、算出方法及びプログラム、GPIアンカー修飾部位の特定装置、特定方法及びプログラム、並びにGPIアンカー修飾部位の判定装置、判定方法及びプログラム

(57) 【特許請求の範囲】

【請求項1】

GPIアンカー型タンパク質のGPIアンカー修飾部位を基準位置としてN末端側及びC末端側に連続する所定の残基数の所定の領域の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアの算出装置であって、

複数のGPIアンカー型タンパク質のアミノ酸配列情報を取得する配列取得部と、

前記配列取得部が取得したアミノ酸配列情報のGPIアンカー修飾部位の残基位置を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出する正解出現頻度算出部と、

前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いるGPIアンカー修飾部位以外のアミノ酸残基を抽出する不正解残基抽出部と、

前記不正解残基抽出部が抽出した複数のアミノ酸残基を用いて、当該アミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出する不正解出現頻度算出部と、

アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出する位置特異的スコア算出部と

を備えることを特徴とする位置特異的スコアの算出装置。

【請求項2】

アミノ酸残基の疎水性値の平均化に用いる残基数である疎水性特性抽出必要数を用いて、連続する当該疎水性特性抽出必要数分のアミノ酸残基の各疎水性指標値の平均である平

10

20

均疎水性値を、前記配列取得部が取得したアミノ酸配列情報が示すアミノ酸残基のそれぞれに対して1残基ずつならしめながら算出する平均疎水性値算出部と、

前記配列取得部が取得したアミノ酸配列情報のC末端側の高疎水性領域におけるアミノ酸残基のうち、前記平均疎水性値が最も高いアミノ酸残基の残基位置である第1の残基位置を特定する第1の残基位置特定部と、

前記配列取得部が取得したアミノ酸配列情報のアミノ酸残基であって前記平均疎水性値が負数のアミノ酸残基のうち、最もC末端側に存在するアミノ酸残基の残基位置である第2の残基位置を特定する第2の残基位置特定部と、

前記配列取得部が取得したアミノ酸配列情報のアミノ酸残基であって前記平均疎水性値が前記第2の残基位置の平均疎水性値より低くかつ当該平均疎水性値がそれぞれ隣接するアミノ酸残基の平均疎水性値より低いアミノ酸残基のうち、最もC末端側に存在するアミノ酸残基の残基位置である第3の残基位置を特定する第3の残基位置特定部と、

前記第1の残基位置、前記第2の残基位置、及び前記第3の残基位置の平均値である平均残基位置を算出する平均残基位置算出部と、

を備え、

前記不正解残基抽出部は、前記平均残基位置算出部が算出した平均残基位置の近傍の所定の候補範囲内にあるアミノ酸残基を抽出する

ことを特徴とする請求項1に記載の位置特異的スコアの算出装置。

【請求項3】

前記候補範囲は、前記平均残基位置算出部が算出した平均残基位置とGPIアンカー修飾部位の残基位置との差の最小値から最大値までの範囲である

ことを特徴とする請求項2に記載の位置特異的スコアの算出装置。

【請求項4】

前記不正解残基抽出部は、残基位置が所定残基数以上C末端から離れているアミノ酸残基を抽出する

ことを特徴とする請求項1から請求項3の何れか1項に記載の位置特異的スコアの算出装置。

【請求項5】

前記所定残基数は、複数の既知のGPIアンカー型タンパク質のC末端からGPIアンカー修飾部位までの残基数の最小値である

ことを特徴とする請求項4に記載の位置特異的スコアの算出装置。

【請求項6】

前記不正解残基抽出部は、前記配列取得部が取得したアミノ酸配列情報のアミノ酸残基のうち、アラニン、システイン、アスパラギン酸、グリシン、アスパラギン、及びセリンを抽出する

ことを特徴とする請求項1から請求項5の何れか1項に記載の位置特異的スコアの算出装置。

【請求項7】

GPIアンカー型タンパク質のGPIアンカー修飾部位を基準位置としてN末端側及びC末端側に連続する所定の残基数の所定の領域の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアの算出装置を用いた位置特異的スコアの算出方法であって、

前記算出装置の配列取得部は、複数のGPIアンカー型タンパク質のアミノ酸配列情報を取得し、

前記算出装置の正解出現頻度算出部は、前記配列取得部が取得したアミノ酸配列情報のGPIアンカー修飾部位の残基位置を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出し、

前記算出装置の不正解残基抽出部は、前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いるGPIアンカー修飾部位以外のアミノ酸残基を抽出し、

、

10

20

30

40

50

前記算出装置の不正解出現頻度算出部は、前記不正解残基抽出部が抽出した複数のアミノ酸残基を用いて、当該アミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出し、

前記算出装置の位置特異的スコア算出部は、アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出することを特徴とする位置特異的スコアの算出方法。

【請求項 8】

コンピュータを、

複数の G P I アンカー型タンパク質のアミノ酸配列情報を取得する配列取得部、

前記配列取得部が取得したアミノ酸配列情報の G P I アンカー修飾部位の残基位置を基準位置とする所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出する正解出現頻度算出部、

前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いる G P I アンカー修飾部位以外のアミノ酸残基を抽出する不正解残基抽出部、

前記不正解残基抽出部が抽出したアミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出する不正解出現頻度算出部、

アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出する位置特異的スコア算出部

として機能させるためのプログラム。

【請求項 9】

検査対象タンパク質における G P I アンカー修飾部位の位置を特定する G P I アンカー修飾部位の特定装置であって、

請求項 1 から請求項 6 の何れか 1 項に記載の算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部と、

前記スコア数値列に基づいて G P I アンカー修飾部位を特定する G P I アンカー修飾部位特定部と

を備えることを特徴とする G P I アンカー修飾部位の特定装置。

【請求項 10】

前記スコア数値列生成部が生成したスコア数値列を入力し、G P I アンカー型タンパク質らしさを示す 0 以上 1 以下の期待値を出力する分類部であって、既知の G P I アンカー型タンパク質の G P I アンカー修飾部位を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として 1 を出力し、既知の G P I アンカー型タンパク質の G P I アンカー修飾部位でない残基位置を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として 0 を出力するように学習された分類部

を備え、

前記 G P I アンカー修飾部位特定部は、前記分類部が出力した期待値に基づいて G P I アンカー修飾部位を特定する

ことを特徴とする請求項 9 に記載の G P I アンカー修飾部位の特定装置。

【請求項 11】

前記 G P I アンカー修飾部位特定部は、前記分類部が出力した期待値が最も高いアミノ酸残基が G P I アンカー修飾部位であると特定する

ことを特徴とする請求項 10 に記載の G P I アンカー修飾部位の特定装置。

【請求項 12】

検査対象タンパク質における G P I アンカー修飾部位の位置を特定する G P I アンカー修飾部位の特定装置を用いた G P I アンカー修飾部位の特定方法であって、

前記特定装置のスコア数値列生成部は、請求項 1 から請求項 6 の何れか 1 項に記載の算

10

20

30

40

50

出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成し、

前記特定装置のGPIアンカー修飾部位特定部は、前記スコア数値列に基づいてGPIアンカー修飾部位を特定する

ことを特徴とするGPIアンカー修飾部位の特定方法。

【請求項13】

コンピュータを、

検査対象タンパク質のアミノ酸配列情報を取得する配列取得部、

請求項1から請求項6の何れか1項に記載の算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部、

前記スコア数値列に基づいてGPIアンカー修飾部位を特定するGPIアンカー修飾部位特定部

として機能させるためのプログラム。

【請求項14】

検査対象タンパク質を構成するアミノ酸残基である検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位の判定装置であって、

請求項1から請求項6の何れか1項に記載の算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸配列情報のうち、前記検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部と、

前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定部と

を備えることを特徴とするGPIアンカー修飾部位の判定装置。

【請求項15】

前記スコア数値列生成部が生成したスコア数値列を入力し、GPIアンカー型タンパク質らしさを示す0以上1以下の期待値を出力する分類部であって、既知のGPIアンカー型タンパク質のGPIアンカー修飾部位を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として1を出力し、既知のGPIアンカー型タンパク質のGPIアンカー修飾部位でない残基位置を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として0を出力するように学習された分類部

を備え、

前記GPIアンカー修飾部位判定部は、前記分類部が出力した期待値に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定する

ことを特徴とする請求項14に記載のGPIアンカー修飾部位の判定装置。

【請求項16】

前記GPIアンカー修飾部位判定部は、前記分類部が出力した期待値が0.5以上である場合に、前記検査対象残基がGPIアンカー修飾部位であると判定する

ことを特徴とする請求項15に記載のGPIアンカー修飾部位の判定装置。

【請求項17】

検査対象タンパク質を構成するアミノ酸残基である検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位の判定装置を用いたGPIアンカー修飾部位判定方法であって、

前記判定装置のスコア数値列生成部は、請求項1から請求項6の何れか1項に記載の算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸配列

10

20

30

40

50

のうち、前記検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定して当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成し、

前記判定装置のGPIアンカー修飾部位判定部は、前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定する

ことを特徴とするGPIアンカー修飾部位の判定方法。

【請求項18】

コンピュータを、

請求項1から請求項6の何れか1項に記載の算出装置が算出した位置特異的スコアに基づいて、検査対象タンパク質のアミノ酸配列のうち、検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部、

前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定部

として機能させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、GPI (glycosylphosphatidylinositol) アンカー型タンパク質のGPIアンカー修飾部位を基準位置としてN末端側及びC末端側に連続する所定の残基数の所定の領域の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアの算出装置、算出方法及びプログラムに関する。また、本発明は、検査対象タンパク質からGPIアンカー修飾部位の位置を特定するGPIアンカー修飾部位特定装置、特定方法及びプログラムに関する。また、本発明は、検査対象タンパク質の検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定装置、判定方法及びプログラムに関する。

【背景技術】

【0002】

生体内の多くのタンパク質は、糖鎖、脂質、糖脂質等により翻訳後修飾を受けており、これらの修飾がタンパク質の機能や細胞内局在に影響することが知られている。これらの翻訳後修飾の中でも、脂質と糖鎖とからなる糖脂質であるGPIアンカーによる修飾は、非常に重要な意味を有するとされている。このことは、GPIアンカーが真核生物や古細菌において広く保存されていること、GPIアンカーを欠損した酵母や原虫は生存できず、GPIアンカーを欠損したヒトは造血幹細胞に異常を生じること等からも明らかである。

GPIにより修飾を受けるタンパク質は、GPIアンカー型タンパク質と呼ばれる。GPIアンカー型タンパク質は、そのアミノ酸配列のN末端に小胞体輸送のシグナルペプチドを有するため、小胞体内に輸送された後に翻訳を完了する。その後、GPIアンカー修飾部位(サイト)のC末端側に存在するプロペプチドが、トランスアミダーゼにより切断及び除去され、GPIアンカー型タンパク質は小胞体内で生合成されたGPIアンカーと結合する。GPIアンカーと結合したGPIアンカー型タンパク質は、ゴルジ体を経て細胞膜表面に輸送され、GPIアンカーにより細胞膜に繋ぎ止められる。

【0003】

GPIアンカー型タンパク質としては、CD14、CD16b等の受容体、5'-ヌクレオチダーゼ、アルカリフォスファターゼ等の酵素等の生体反応に極めて重要なタンパク質が多く発見されている。また、狂牛病関連のプリオンタンパク質や、癌関連のヒト癌胎児性抗原(CEA)等、重篤な疾患に関わるタンパク質も見出されている。しかしながら、現在までに真核生物で知られているGPIアンカー型タンパク質は100種類程度であり、未だ発見されていないGPIアンカー型タンパク質が多く存在すると考えられている

。そこで、近年では、コンピュータを用いたバイオインフォマティクス手法により、アミノ酸配列から GPI アンカー型タンパク質を新たに見つける試みがなされている。

【0004】

例えば、特許文献1には、 サイトを含むアミノ酸残基の部分配列について誤差逆伝播型ニューラルネットワークを使用して GPI アンカー型タンパク質を判別する発明が開示されている。特許文献1によれば、平均側鎖サイズが最小となるアミノ酸残基の C 末端側に隣接するアミノ酸残基を サイトと推定している。

しかしながら、平均側鎖サイズが最小となるアミノ酸残基の C 末端側に隣接するアミノ酸残基の多くは、 サイトであることが分かっているが、当該アミノ酸残基は、必ずしも サイトであるとは限らない。

10

【0005】

また現在、既知の GPI アンカー型タンパク質においても、 サイトの位置が未知のものが存在する。これらの GPI アンカー型タンパク質の サイトの位置を正確に予測することで、 GPI アンカー型タンパク質についての詳しい情報を明らかにすることができる。

そのため、近年、 サイトの位置を特定する手法が研究されている（例えば、非特許文献1 - 5を参照）。

【先行技術文献】

【特許文献】

【0006】

【特許文献1】特開2012-32163号公報

【非特許文献】

【0007】

【非特許文献1】Birgit Eisenhaberら、Sequence properties of GPI-anchored proteins near the -site: constraints for the polypeptide binding site of the putative transamidase、「Protein Engineering」、1998年

【非特許文献2】Birgit Eisenhaberら、Prediction of Potential GPI-modification Sites in Protein Sequences、「J Mol Biol」、1999年9月

【非特許文献3】Niklaus Fankhauserら、Identification of GPI anchor attachment signals by a Kohonen self-organizing map、「BMC Bioinformatics」、2005年5月

【非特許文献4】Guylaine Poissonら、FragAnchor: A Large-Scale Predictor of Glycosylphosphatidylinositol Anchors in Eukaryote Protein Sequences by Qualitative Scoring、「Genomics Proteomics Bioinformatics」、2007年5月

【非特許文献5】Andrea Pierleoniら、PredGPI: a GPI-anchor predictor、「BMC Bioinformatics」、2008年9月

【発明の概要】

【発明が解決しようとする課題】

【0008】

上述した非特許文献1 - 5に挙げた サイトの位置の予測方法においては、いずれもその選択性の評価がなされていない。また、感度の評価も58% ~ 88%程度であり、より正確に サイトの位置を予測することが望まれている。

本発明は、上記事情に鑑みてなされたものであって、高感度かつ高選択的に検査対象タ

50

ンパク質のGPIアンカー修飾部位を判定するための位置特異的スコアの算出装置、算出方法及びプログラム、高感度かつ高選択的に検査対象タンパク質のGPIアンカー修飾部位を特定することが可能なGPIアンカー修飾部位特定装置、特定方法及びプログラム、並びに高感度かつ高選択的に検査対象残基がGPIアンカー修飾部位であるか否かを判定することが可能なGPIアンカー修飾部位判定装置、判定方法及びプログラムを提供することを目的とする。

【課題を解決するための手段】

【0009】

本発明は上記の課題を解決するためになされたものであり、GPIアンカー型タンパク質のGPIアンカー修飾部位を基準位置としてN末端側及びC末端側に連続する所定の残基数の所定の領域の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアの算出装置であって、複数のGPIアンカー型タンパク質のアミノ酸配列情報を取得する配列取得部と、前記配列取得部が取得したアミノ酸配列情報のGPIアンカー修飾部位の残基位置を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出する正解出現頻度算出部と、前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いるGPIアンカー修飾部位以外のアミノ酸残基を抽出する不正解残基抽出部と、前記不正解残基抽出部が抽出した複数のアミノ酸残基を用いて、当該アミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出する不正解出現頻度算出部と、アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出する位置特異的スコア算出部とを備えることを特徴とする。

10

20

【0010】

また、本発明は、アミノ酸残基の疎水性値の平均化に用いる残基数である疎水性特性抽出必要数を用いて、連続する当該疎水性特性抽出必要数分のアミノ酸残基の各疎水性指標値の平均である平均疎水性値を、前記配列取得部が取得したアミノ酸配列情報が示すアミノ酸残基のそれぞれに対して1残基ずつずらしながら算出する平均疎水性値算出部と、前記配列取得部が取得したアミノ酸配列情報のC末端側の高疎水性領域におけるアミノ酸残基のうち、前記平均疎水性値が最も高いアミノ酸残基の残基位置である第1の残基位置を特定する第1の残基位置特定部と、前記配列取得部が取得したアミノ酸配列情報のアミノ酸残基であって前記平均疎水性値が負数のアミノ酸残基のうち、最もC末端側に存在するアミノ酸残基の残基位置である第2の残基位置を特定する第2の残基位置特定部と、前記配列取得部が取得したアミノ酸配列情報のアミノ酸残基であって前記平均疎水性値が前記第2の残基位置の平均疎水性値より低くかつ当該平均疎水性値がそれぞれ隣接するアミノ酸残基の平均疎水性値より低いアミノ酸残基のうち、最もC末端側に存在するアミノ酸残基の残基位置である第3の残基位置を特定する第3の残基位置特定部と、前記第1の残基位置、前記第2の残基位置、及び前記第3の残基位置の平均値である平均残基位置を算出する平均残基位置算出部と、を備え、前記不正解残基抽出部は、前記平均残基位置算出部が算出した平均残基位置の近傍の所定の候補範囲内にあるアミノ酸残基を抽出することを特徴とする。

30

40

【0011】

また、本発明において前記候補範囲は、前記平均残基位置算出部が算出した平均残基位置とGPIアンカー修飾部位の残基位置との差の最小値から最大値までの範囲であることを特徴とする。

【0012】

また、本発明において前記不正解残基抽出部は、前記平均残基位置算出部が算出した平均残基位置の近傍の所定の候補範囲内にあり、かつ残基位置が所定残基数以上C末端から離れているアミノ酸残基を抽出することを特徴とする。

【0013】

また、本発明において前記所定残基数は、複数のGPIアンカー型タンパク質のC末端

50

から GPI アンカー修飾部位までの残基数の最小値であることを特徴とする。

【0014】

また、本発明において前記不正解残基抽出部は、前記平均残基位置算出部が算出した平均残基位置の近傍の所定の候補範囲内にあるアミノ酸残基のうち、アラニン、システイン、アスパラギン酸、グリシン、アスパラギン、及びセリンを抽出することを特徴とする。

【0015】

また、本発明は、GPI アンカー型タンパク質の GPI アンカー修飾部位を基準位置として N 末端側及び C 末端側に連続する所定の残基数の所定の領域の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアの算出装置を用いた位置特異的スコアの算出方法であって、前記算出装置の配列取得部は、複数の GPI アンカー型タンパク質のアミノ酸配列情報を取得し、前記算出装置の正解出現頻度算出部は、前記配列取得部が取得したアミノ酸配列情報の GPI アンカー修飾部位の残基位置を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出し、前記算出装置の不正解残基抽出部は、前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いる GPI アンカー修飾部位以外のアミノ酸残基を抽出し、前記算出装置の不正解出現頻度算出部は、前記不正解残基抽出部が抽出した複数のアミノ酸残基を用いて、当該アミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出し、前記算出装置の位置特異的スコア算出部は、アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出することを特徴とする。

【0016】

また、本発明は、コンピュータを、複数の GPI アンカー型タンパク質のアミノ酸配列情報を取得する配列取得部、前記配列取得部が取得したアミノ酸配列情報の GPI アンカー修飾部位の残基位置を基準位置とする所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出する正解出現頻度算出部、前記配列取得部が取得したアミノ酸配列情報から、位置特異的スコアの算出に用いる GPI アンカー修飾部位以外のアミノ酸残基を抽出する不正解残基抽出部、前記不正解残基抽出部が抽出したアミノ酸残基を基準位置とする前記所定の領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出する不正解出現頻度算出部、アミノ酸残基の種類ごとに、前記正解出現頻度を前記不正解出現頻度で除算した値に基づいて位置特異的スコアを算出する位置特異的スコア算出部として機能させるためのプログラムである。

【0017】

また、本発明は、検査対象タンパク質における GPI アンカー修飾部位の位置を特定する GPI アンカー修飾部位の特定装置であって、上記算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部と、前記スコア数値列に基づいて GPI アンカー修飾部位を特定する GPI アンカー修飾部位特定部とを備えることを特徴とする。

【0018】

また、本発明は、前記スコア数値列生成部が生成したスコア数値列を入力し、GPI アンカー型タンパク質らしさを示す 0 以上 1 以下の期待値を出力する分類部であって、既知の GPI アンカー型タンパク質の GPI アンカー修飾部位を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として 1 を出力し、既知の GPI アンカー型タンパク質の GPI アンカー修飾部位でない残基位置を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として 0 を出力するように学習された分類部を備え、前記 GPI アンカー修飾部位特定部は、前記分類部が出力した期待値に基づいて GPI アンカー修飾部位を特定することを特徴とする。

【0019】

また、本発明において前記GPIアンカー修飾部位特定部は、前記分類部が出力した期待値が最も高いアミノ酸残基がGPIアンカー修飾部位であると特定することを特徴とする。

【0020】

また、本発明は、検査対象タンパク質におけるGPIアンカー修飾部位の位置を特定するGPIアンカー修飾部位の特定装置を用いたGPIアンカー修飾部位の特定方法であって、前記特定装置のスコア数値列生成部は、上記算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成し、前記特定装置のGPIアンカー修飾部位特定部は、前記スコア数値列に基づいてGPIアンカー修飾部位を特定することを特徴とする。

10

【0021】

また、本発明は、コンピュータを、前記検査対象タンパク質のアミノ酸配列情報を取得する配列取得部、上記算出装置が算出した位置特異的スコアに基づいて、検査対象タンパク質のアミノ酸残基のそれぞれについて、当該アミノ酸残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部、前記スコア数値列に基づいてGPIアンカー修飾部位を特定するGPIアンカー修飾部位特定部として機能させるためのプログラムである。

20

【0022】

また、本発明は、検査対象タンパク質を構成するアミノ酸残基である検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位の判定装置であって、上記算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸配列情報のうち、前記検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部と、前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定部とを備えることを特徴とする。

【0023】

また、本発明は、前記スコア数値列生成部が生成したスコア数値列を入力し、GPIアンカー型タンパク質らしさを示す0以上1以下の期待値を出力する分類部であって、既知のGPIアンカー型タンパク質のGPIアンカー修飾部位を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として1を出力し、既知のGPIアンカー型タンパク質のGPIアンカー修飾部位でない残基位置を基準位置とした部分配列のスコア数値列を入力とした場合に、期待値として0を出力するように学習された分類部を備え、前記GPIアンカー修飾部位判定部は、前記分類部が出力した期待値に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定することを特徴とする。

30

【0024】

また、本発明において前記GPIアンカー修飾部位判定部は、前記分類部が出力した期待値が0.5以上である場合に、前記検査対象残基がGPIアンカー修飾部位であると判定することを特徴とする。

40

【0025】

また、本発明は、検査対象タンパク質を構成するアミノ酸残基である検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定装置を用いたGPIアンカー修飾部位判定方法であって、前記判定装置のスコア数値列生成部は、上記算出装置が算出した位置特異的スコアに基づいて、前記検査対象タンパク質のアミノ酸配列のうち、前記検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定して当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成し、前記判定装置のGPIアンカー修飾部

50

位判定部は、前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定することを特徴とする。

【0026】

また、本発明は、コンピュータを、上記算出装置が算出した位置特異的スコアに基づいて、検査対象タンパク質のアミノ酸配列のうち、検査対象残基を基準位置とした前記所定の領域におけるアミノ酸残基の部分配列の各アミノ酸残基の位置特異的スコアを特定し、当該各アミノ酸残基の位置特異的スコアを示す数値列であるスコア数値列を生成するスコア数値列生成部、前記スコア数値列に基づいて前記検査対象残基がGPIアンカー修飾部位であるか否かを判定するGPIアンカー修飾部位判定部として機能させるためのプログラムである。

10

【発明の効果】

【0027】

本発明によれば、位置特異的スコアの算出装置は、GPIアンカー修飾部位の特定に特化したPSSM(Position Specific Scoring Matrix; 位置特異的スコアリングマトリックス)を生成することができる。そして、本発明によるGPIアンカー修飾部位の特定装置は、当該PSSMを用いることにより、高感度かつ高選択的に検査対象タンパク質のGPIアンカー修飾部位を特定することができる。また、本発明によるGPIアンカー修飾部位判定装置は、当該PSSMを用いることにより、高感度かつ高選択的に検査対象残基がGPIアンカー修飾部位であるか否かを判定することができる。

20

【図面の簡単な説明】

【0028】

【図1】本発明の一実施形態による位置特異的スコア算出装置の構成を示す概略ブロック図である。

【図2】疎水性指標値記憶部が記憶する情報を示す図である。

【図3】本実施形態による位置特異的スコア算出装置の動作を示すフローチャートである。

【図4】平均疎水性値の算出方法を示す図である。

【図5】 サイトの近傍に存在する特徴的なアミノ酸残基の位置を示す図である。

【図6】平均残基位置と サイトとの残基位置差を示す図である。

30

【図7】位置特異的スコア算出装置が算出した位置特異的スコアを用いて生成したPSSMの一例を示す図である。

【図8】本発明の一実施形態による サイト判定装置の構成を示す概略ブロック図である。

【図9】本実施形態で用いるニューラルネットワークの構成を示す図である。

【図10】位置特異的スコアの割り当て方法を示す図である。

【図11】本発明の一実施形態による サイト判定装置の動作を示すフローチャートである。

【図12】本実施形態による サイト判定装置の判定精度を示す表である。

【図13】本発明の一実施形態による サイト特定装置の構成を示す概略ブロック図である。

40

【図14】本発明の一実施形態による サイト特定装置の動作を示すフローチャートである。

【発明を実施するための形態】

【0029】

以下、図面を参照しながら本発明の実施形態について詳しく説明する。

《位置特異的スコア算出装置》

本実施形態に係る位置特異的スコア算出装置は、GPIアンカー修飾部位(以下、 サイトという)の特定に特化した位置特異的スコアを算出する。ここで、位置特異的スコアとは、アミノ酸残基の部分配列の中心のアミノ酸残基が サイトである可能性を示す値で

50

あり、当該値が大きいほど、部分配列の中心のアミノ酸残基が サイトである可能性が高いことを示す。これにより、 サイトの特定に特化したPSSMを生成することができる。PSSMとは、GPIアンカー型タンパク質の サイトを中心とした所定の部分配列の各残基位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアを格納する行列 (Matrix) である。

【0030】

図1は、本発明の一実施形態による位置特異的スコア算出装置100の構成を示す概略ブロック図である。

位置特異的スコア算出装置100は、既知配列記憶部101、配列取得部102、疎水性指標値記憶部103、疎水性指標値特定部104、平均疎水性値算出部105、第1の残基位置特定部106、第2の残基位置特定部107、第3の残基位置特定部108、平均残基位置算出部109、不正解残基抽出部110、不正解出現頻度算出部111、正解出現頻度算出部112、位置特異的スコア算出部113を備える。

10

【0031】

既知配列記憶部101は、複数の既知のGPIアンカー型タンパク質のアミノ酸配列情報と、当該GPIアンカー型タンパク質の サイトの残基位置を記憶する。

配列取得部102は、既知配列記憶部101からアミノ酸配列情報と サイトの残基位置を読み出す。

【0032】

疎水性指標値記憶部103は、アミノ酸残基に関連付けて当該アミノ酸残基の疎水性指標値を記憶する。

20

疎水性指標値特定部104は、配列取得部102が取得したアミノ酸配列情報のアミノ酸残基それぞれの疎水性指標値を疎水性指標値記憶部103から特定し、アミノ酸残基ごとの疎水性指標値を示す連続する数値列を生成する。

平均疎水性値算出部105は、疎水性指標値特定部104が生成した数値列に基づいて、連続するアミノ酸残基の平均疎水性値を算出する。なお、アミノ酸残基の平均疎水性値とは、算出対象となるアミノ酸残基の前後に連続する所定数のアミノ酸残基の疎水性指標値の平均値のことである。

【0033】

第1の残基位置特定部106は、平均疎水性値算出部105が算出した各アミノ酸残基の平均疎水性値に基づいて、GPIアンカー型タンパク質のC末端側の高疎水性領域におけるアミノ酸残基のうち、平均疎水性値が最も高いアミノ酸残基の残基位置である第1の残基位置を特定する。GPIアンカー型タンパク質のC末端側の高疎水性領域とは、既知の複数のGPIアンカー型タンパク質のN末端側の高疎水性領域 (N末端から30残基) を除くアミノ酸残基のそれぞれに対して平均疎水性値を算出した場合に、当該平均疎水性値が最大となるアミノ酸残基の部分配列の中央に位置するアミノ酸残基が含まれる領域である。なお、本実施形態におけるGPIアンカー型タンパク質のC末端側の高疎水性領域は、C末端から14残基以内の領域を示す。

30

【0034】

第2の残基位置特定部107は、平均疎水性値算出部105が算出した各アミノ酸残基の平均疎水性値に基づいて、平均疎水性値が負数のアミノ酸残基のうち最もC末端側に存在するアミノ酸残基の残基位置である第2の残基位置を特定する。

40

【0035】

第3の残基位置特定部108は、平均疎水性値算出部105が算出した各アミノ酸残基の平均疎水性値に基づいて、以下の条件を満たすアミノ酸残基のうち最もC末端側に存在するアミノ酸残基の残基位置である第3の残基位置を特定する。第3の残基位置の条件は、(1)平均疎水性値が第2の残基位置の平均疎水性値より低いこと、(2)平均疎水性値がN末端側およびC末端側にそれぞれ隣接する各アミノ酸残基の平均疎水性値より低いこと、である。

【0036】

50

平均残基位置算出部 109 は、前記第 1 の残基位置、前記第 2 の残基位置、及び前記第 3 の残基位置の平均値である平均残基位置を算出する。

【0037】

不正解残基抽出部 110 は、配列取得部 102 が取得したアミノ酸配列情報に含まれるアミノ酸残基のうち、以下の条件を満たすものを、位置特異的スコアの算出に用いる サイト以外のアミノ酸残基である不正解残基として抽出する。不正解残基の条件は、(1) サイトでないこと、(2) 平均残基位置を中心とした所定の候補範囲内に存在すること、(3) C 末端から所定残基数以降に存在すること、(3) アラニン、システイン、アスパラギン酸、グリシン、アスパラギン、セリンのいずれかであること、である。

【0038】

不正解残基の条件の(2)の所定の候補範囲とは、既知の複数の G P I アンカー型タンパク質における平均残基位置と G P I アンカー修飾部位の残基位置との差の最小値から最大値までの範囲である。なお、本実施形態における候補範囲は、平均残基位置から N 末端側に 21 残基、平均残基位置から C 末端側に 14 残基の範囲である。また、不正解残基の条件の(3)の所定残基数とは、既知の複数の G P I アンカー型タンパク質の C 末端から G P I アンカー修飾部位までの残基数の最小値である。なお、本実施形態における所定残基数は 17 残基である。

【0039】

不正解出現頻度算出部 111 は、不正解残基抽出部 110 が抽出した複数の不正解残基を用いて、当該不正解残基を中心とする所定領域内の位置に存在するアミノ酸残基の種類の出現頻度である不正解出現頻度を算出する。

正解出現頻度算出部 112 は、配列取得部 102 が取得した複数のアミノ酸配列情報を用いて、 サイトを中心とする所定領域内の位置に存在するアミノ酸残基の種類の出現頻度である正解出現頻度を算出する。

位置特異的スコア算出部 113 は、正解出現頻度を不正解出現頻度で除算した値の対数をとることで、アミノ酸残基の種類ごと、中心残基からの位置ごとの位置特異的スコアを算出する。位置特異的スコア算出部 113 が算出した位置特異的スコアを行列形式にすることで、P S S M を生成することができる。

【0040】

図 2 は、疎水性指標値記憶部 103 が記憶する情報を示す図である。

疎水性指標値記憶部 103 は、図 2 に示すように、アミノ酸残基の各々に対して、当該アミノ酸残基の疎水性を示す指標値を記憶している。なお、本実施形態では、疎水性指標値として K Y T J 8 2 0 1 0 1 (K y t e J . , D o o l i t t l e R . , 「J o u r n a l o f M o l e c u l a r B i o l o g y」, 1982 年, v o l . 1 5 7 , n o . 1 , p p . 1 0 5 - 1 3 2) で示される疎水性指標値を用いている。図 2 において、アミノ酸残基の「A」はアラニンを示し、「R」はアルギニンを示し、「N」はアスパラギンを示し、「D」はアスパラギン酸を示し、「C」はシステインを示し、「Q」はグルタミンを示し、「E」はグルタミン酸を示し、「G」はグリシンを示し、「H」はヒスチジンを示し、「I」はイソロイシンを示し、「L」はロイシンを示し、「K」はリシンを示し、「M」はメチオニンを示し、「F」はフェニルアラニンを示し、「P」はプロリンを示し、「S」はセリンを示し、「T」はトレオニンを示し、「W」はトリプトファンを示し、「Y」はチロシンを示し、「V」はバリンを示す。

【0041】

ここで、既知配列記憶部 101 に記憶させるアミノ酸配列情報及び サイトの残基位置について説明する。

本実施形態では、データセットを生成するためのデータバンクとして S w i s s P r o t R e l e a s e 2 0 1 1 0 7 を用いる。本データバンクのうち、 サイトの位置が確定している実験的確認のある G P I アンカー型タンパク質は 20 エントリである。しかしながら 20 エントリというデータ数は、データセットとして用いるには不足であるため、実験的確認のある G P I アンカー型タンパク質のアミノ酸配列情報に加えて、実験的

10

20

30

40

50

確証のある G P I アンカー型タンパク質に類似する (B y s i m i l a r i t y) タンパク質のアミノ酸配列情報を用いる。実験的確認のある G P I アンカー型タンパク質のアミノ酸配列情報とそれに類似するタンパク質のアミノ酸配列情報を合わせたエントリ数は、101エントリである。

【0042】

ここで、実験的確認のある G P I アンカー型タンパク質に類似するタンパク質としてデータバンクに格納されているアミノ酸配列情報には、アラインメントがそろっていないものや、どの G P I アンカー型タンパク質に類似するかが不明のものが存在するため、これらをデータセットから除外する。具体的には、各エントリをクラスタリングしてクラスタごとにアラインメントをし、実験的確認のある G P I アンカー型タンパク質のアミノ酸配列情報と サイトの残基位置が揃っていないタンパク質のアミノ酸配列情報を、データセットから除外する。本実施形態では、配列類似性を40%に設定してクラスタリングを行った。このとき、 サイトの残基位置が確定していないタンパク質については、実験的確認のある G P I アンカー型タンパク質の サイトと揃った残基位置を、 サイトの残基位置としてアミノ酸配列情報に関連付ける。この時点で、データセットのエントリ数は85エントリである。

10

【0043】

次に、データバンクから当該85エントリに類似するアミノ酸配列情報を完全長で検索し、当該類似するアミノ酸配列情報を、データセットに加える。このとき、 サイトの残基位置が確定していないタンパク質については、実験的確認のある G P I アンカー型タンパク質の サイトとアラインメントを取ることで、 サイトの残基位置を特定し、アミノ酸配列情報に関連付ける。この時点で、データセットのエントリ数は122エントリである。

20

そして、当該122エントリのアミノ酸配列情報及び サイトの残基位置を、既知の G P I アンカー型タンパク質のアミノ酸配列情報及び サイトの残基位置として、既知配列記憶部101に記憶させる。

【0044】

次に、本実施形態による位置特異的スコア算出装置100の動作について説明する。

位置特異的スコア算出装置100は、 サイトを中心としたアミノ酸残基の部分配列におけるアミノ酸残基の種類の出現頻度と、 サイトでないアミノ酸残基を中心としたアミノ酸残基の部分配列におけるアミノ酸残基の種類の出現頻度とを算出し、その比の対数を位置特異的スコアとして算出する。これにより、位置特異的スコア算出装置100は、部分配列の中心のアミノ酸残基が サイトである可能性を示す位置特異的スコアを算出することができる。

30

【0045】

ここで、タンパク質を構成するアミノ酸残基の数は、平均300残基程度であるため、 サイトであるアミノ酸残基の数に対して、 サイトでないアミノ酸残基の数はその300倍程度存在することになる。このうち、 サイトになり得ないことが明らかなアミノ酸残基がほとんどであるため、 サイトでないアミノ酸残基を全て用いて位置特異的スコアを算出した場合、その精度が低くなることが考えられる。そこで、本実施形態では、 サイトになり得ないことが明らかなアミノ酸残基を除いて位置特異的スコアを算出することで、位置特異的スコアの精度を高めている。

40

【0046】

以下、位置特異的スコア算出装置100の具体的な動作について説明する。

図3は、本実施形態による位置特異的スコア算出装置100の動作を示すフローチャートである。

まず、位置特異的スコア算出装置100の配列取得部102は、既知配列記憶部101が記憶するアミノ酸配列情報と サイトの残基位置の組み合わせを1つずつ取得し、当該組み合わせごとに、以下に示すステップS102～ステップS111の処理を実行する(ステップS101)。

50

【 0 0 4 7 】

疎水性指標値特定部 1 0 4 は、疎水性指標値記憶部 1 0 3 を参照して、配列取得部 1 0 2 がステップ S 1 0 1 で取得したアミノ酸配列情報の各アミノ酸残基の疎水性指標値を特定し、当該疎水性指標値を示す数値列を生成する（ステップ S 1 0 2）。例えば、配列取得部 1 0 2 が取得したアミノ酸配列情報が、「M L L E P G R G C C」という配列を示す場合、疎水性指標値特定部 1 0 4 は、疎水性指標値記憶部 1 0 3 が記憶する図 2 に示す指標値より「1.9、3.8、3.8、-3.5、-1.6、-0.4、-4.5、-0.4、2.5、2.5」という数値列を生成する。

【 0 0 4 8 】

図 4 は、平均疎水性値の算出方法を示す図である。

10

次に、平均疎水性値算出部 1 0 5 は、疎水性指標値特定部 1 0 4 が生成した数値列に基づいて、連続するアミノ酸残基の平均疎水性値を算出する（ステップ S 1 0 3）。具体的には、平均疎水性値算出部 1 0 5 は、疎水性指標値特定部 1 0 4 が生成した数値列の連続する疎水性特性抽出必要数分の各疎水性指標値の平均である平均疎水性値を、図 4 に示すように、1 残基ずつずらしながら算出する。なお、本実施形態における疎水性特性抽出必要数は、1 7 残基である。

【 0 0 4 9 】

ここで、疎水性特性抽出必要数の連続するアミノ酸残基の部分配列における中央のアミノ酸残基の位置が C 末端から r 残基目であるときの平均疎水性値は、式 (1) を用いて算出できる。

20

【 0 0 5 0 】

【 数 1 】

$$\frac{1}{2n+1} \sum_{i=r-n}^{r+n} H(i) \quad \dots (1)$$

【 0 0 5 1 】

但し、n は、平均化に用いる前後の残基数を示す。つまり、2 n + 1 は、疎水性特性抽出必要数を示す。また、H (i) は、疎水性特性抽出必要数の連続するアミノ酸残基の部分配列における中央のアミノ酸残基の位置が N 末端から i 残基目である場合のアミノ酸残基の疎水性指標値を示す。

30

つまり、N 末端から r 残基目のアミノ酸残基が中央に位置する部分配列の平均疎水性値は、N 末端から r - n 残基目のアミノ酸残基から、N 末端から r + n 残基目のアミノ酸残基までの疎水性指標値の平均となる。なお、このとき、N 末端から n 残基以内のアミノ酸残基及び C 末端から n 残基以内のアミノ酸残基については、前後 n 残基の平均値を算出できないため、平均疎水性値として例えば N U L L 値を代入しておくが良い。

【 0 0 5 2 】

図 5 は、サイトの近傍に存在する特徴的なアミノ酸残基の位置を示す図である。

ステップ S 1 0 3 で、平均疎水性値算出部 1 0 5 が平均疎水性値を算出すると、位置特異的スコア算出装置 1 0 0 は、当該平均疎水性値を用いて、サイトの近傍に存在する特徴的なアミノ酸残基の位置を特定する。特徴的なアミノ酸残基とは、具体的には、C 末端にある最大平均疎水性位置のアミノ酸残基（図 5 (A) : 第 1 の残基位置）、C 末端から見て初めに平均疎水性値が負数となるアミノ酸残基（図 5 (B) : 第 2 の残基位置）、及び、第 2 の残基位置より平均疎水性値が低く、隣接する前後 2 残基の平均疎水性値よりも低いアミノ酸残基（図 5 (C) : 第 3 の残基位置）である。

40

【 0 0 5 3 】

まず、第 1 の残基位置特定部 1 0 6 は、C 末端から 1 4 残基以内のアミノ酸残基のうち、平均疎水性値が最も高いアミノ酸残基の残基位置を、第 1 の残基位置として特定する（ステップ S 1 0 4）。

【 0 0 5 4 】

また、第 2 の残基位置特定部 1 0 7 は、平均疎水性値算出部 1 0 5 が算出した平均疎水

50

性値が負数のアミノ酸残基を抽出する。次に、第2の残基位置特定部107は、抽出したアミノ酸残基のうち、最もC末端側に存在するもののアミノ酸残基の残基位置を、第2の残基位置として特定する(ステップS105)。

【0055】

また、第3の残基位置特定部108は、平均疎水性値が第2の残基位置の平均疎水性値より低いアミノ酸残基を抽出する。次に、第3の残基位置特定部108は、抽出したアミノ酸残基のうち、平均疎水性値が、N末端側及びC末端側にそれぞれ隣接するアミノ酸残基の平均疎水性値より低いものを抽出する。そして、第3の残基位置特定部108は、抽出したアミノ酸残基のうち、最もC末端側に存在するもののアミノ酸残基の残基位置を、第3の残基位置として特定する(ステップS106)。

10

【0056】

次に、平均残基位置算出部109は、第1の残基位置、第2の残基位置、及び第3の残基位置の平均値である平均残基位置を算出する(ステップS107)。

図6は、平均残基位置と サイトとの残基位置差を示す図である。

平均残基位置は、上述した サイトの近傍に存在する特徴的なアミノ酸残基の位置の平均値であるため、図6に示すように、 サイトの近傍の残基位置となる。本実施形態においては、平均残基位置からC末端側に14残基、平均残基位置からN末端側に21残基の範囲内に、必ず サイトが存在することが分かる。

【0057】

次に、不正解残基抽出部110は、ステップS101で配列取得部102が取得したアミノ酸配列から、平均残基位置算出部109が算出した平均残基位置からC末端側に14残基、平均残基位置からN末端側に21残基の範囲内のアミノ酸残基を抽出する(ステップS108)。

20

【0058】

次に、不正解残基抽出部110は、抽出したアミノ酸残基のうち、C末端側から17残基以上離れているものを抽出する(ステップS109)。これは、GPIアンカー型タンパク質のアタッチメントシグナルの最小残基数が17残基であり、これらの残基は サイトになり得ないため、不正解残基から除外している。なお、アタッチメントシグナルの最小残基数は、複数の既知のGPIアンカー型タンパク質のC末端からGPIアンカー修飾部位までの残基数の最小値を算出することで求められる。

30

【0059】

次に、不正解残基抽出部110は、抽出したアミノ酸残基から、アラニン、システイン、アスパラギン酸、グリシン、アスパラギン、及びセリンを抽出する(ステップS110)。これは、 サイトとなり得るアミノ酸残基がアラニン、システイン、アスパラギン酸、グリシン、アスパラギン、セリンの何れかのみであるからである。そして、不正解残基抽出部110は、抽出したアミノ酸残基から、 サイトのアミノ酸残基を除外したアミノ酸残基を位置特異的スコアの算出に用いる不正解残基として、不正解出現頻度算出部111に出力する(ステップS111)。

【0060】

上述したステップS101~ステップS102の処理を、既知配列記憶部101が記憶するアミノ酸配列情報と サイトの残基位置の組み合わせの全てについて実行すると、不正解残基として、1007エントリが抽出される。次に、不正解出現頻度算出部111は、不正解残基抽出部110から受け付けた複数の不正解残基のうち、冗長性が高いものを除去する(ステップS112)。冗長性が高いアミノ酸残基とは、例えば、当該アミノ酸残基を基準位置とする所定の範囲のアミノ酸残基の部分配列が同一または酷似しているものである。同様に、正解出現頻度算出部112は、配列取得部102が取得した各 サイトのうち、冗長性が高いものを除去する(ステップS112)。なお、冗長性の除去は、CD-HIT(<http://weizhong-lab.ucsd.edu/cd-hit/>)を開示されている。)を用いて80%以上の相同性(アミノ酸配列の同一性)を有する配列ごとにクラスタリングし、各クラスタから無作為に代表配列を決定して行

40

50

った。これにより、位置特異的スコアの算出に用いる不正解残基の数は172エントリとなり、位置特異的スコアの算出に用いる サイトの数は45エントリとなる。

【0061】

次に、不正解出現頻度算出部111は、冗長性の排除を行った複数の不正解残基を用いて、当該不正解残基を基準位置とする所定の範囲（基準位置のアミノ酸残基と基準位置からN末端側に連続する12残基のアミノ酸残基とC末端側に連続する12残基のアミノ酸残基とからなる範囲とすることが好ましい）に存在するアミノ酸残基から、式(2)を用いて不正解残基を基準位置とする所定範囲内の位置pに存在するアミノ酸残基の種類iの出現頻度である不正解出現頻度を算出する（ステップS113）。

【0062】

【数2】

$$\frac{n_{ip} + 0.05}{\sum_{i=1}^{20} n_{ip} + 1} \dots (2)$$

10

【0063】

但し、 n_{ip} は、種類iのアミノ酸残基が位置pに存在するタンパク質の個数を示す。これにより、データセットの全てのエントリにおいて位置pに種類iが存在しない場合にも、ゼロで除算を行うことを防ぐことができる。同様に、正解出現頻度算出部112は、冗長性の排除を行った複数の サイトを用いて、当該 サイトを基準位置とする所定の範囲におけるアミノ酸残基から、式(2)を用いて サイトを基準位置とする所定範囲内の位置pに存在するアミノ酸残基の種類iの出現頻度である正解出現頻度を算出する（ステップS113）。

20

【0064】

そして、位置特異的スコア算出部113は、不正解出現頻度算出部111が算出した不正解出現頻度と正解出現頻度算出部112が算出した正解出現頻度とを用いて、式(3)を用いて位置特異的スコアを算出する（ステップS114）。

【0065】

【数3】

$$\ln \frac{f_{ip}^{Top}}{f_{ip}^{Fop}} \dots (3)$$

30

【0066】

但し、 f_{ip}^T は、位置pに存在するアミノ酸残基の種類iの正解出現頻度を示す。また、 f_{ip}^F は、位置pに存在するアミノ酸残基の種類iの不正解出現頻度を示す。

【0067】

このように、位置特異的スコア算出装置100は、 サイトになり得ない残基位置のアミノ酸残基を除いた不正解残基を用いて位置特異的スコアを算出する。これにより、当該位置特異的スコアを用いてPSSMを生成することで、 サイトの特定に特化したPSSMを生成することができる。

40

【0068】

図7は、位置特異的スコア算出装置100が算出した位置特異的スコアを用いて生成したPSSMの一例を示す図である。

図7に示すように、PSSMは、アミノ酸残基の位置におけるアミノ酸残基の種類の出現度合いを示す位置特異的スコアを要素とする。図7では、 サイトの残基位置を0とし、負数側をN末端側、正数側をC末端側としている。

【0069】

50

以下、上述した手順により算出された位置特異的スコアから生成したPSSMの使用方法について説明する。

【0070】

《 サイト判定装置 》

本実施形態による サイト判定装置は、検査対象となるタンパク質のアミノ酸配列情報と サイトであるか否かの判定対象となる残基位置の入力を受け付け、当該残基位置が サイトであるか否かを判定する。

【0071】

図8は、本発明の一実施形態による サイト判定装置200の構成を示す概略ブロック図である。

10

サイト判定装置200は、入力部201、PSSM記憶部202、スコア数値列生成部203、ニューラルネットワーク204（分類部）、 サイト判定部205（GPIアンカー修飾部位判定部）を備える。

【0072】

入力部201は、検査対象となるタンパク質のアミノ酸配列情報と サイトであるか否かの判定対象となる残基位置（検査対象残基位置）の入力を受け付ける。

PSSM記憶部202は、位置特異的スコア算出装置100が算出した位置特異的スコアを用いて生成されたPSSMを記憶する。

スコア数値列生成部203は、PSSM記憶部202が記憶するPSSMに基づいて、入力部201が受け付けた残基位置を基準位置とする所定の領域におけるスコア数値列を生成する。ここで生成するスコア数値列とは、入力部201が受け付けたアミノ酸配列情報の所定の領域のそれぞれのアミノ酸残基の位置特異的スコアを要素とする配列である。

20

ニューラルネットワーク204は、スコア数値列生成部203が生成したスコア数値列を入力し、 サイトらしさを示す0以上1以下の期待値を出力する。

サイト判定部205は、入力部201が受け付けた検査対象残基位置が サイトであるか否かを判定する。

【0073】

ここで、ニューラルネットワーク204の挙動について説明する。

図9は、本実施形態で用いるニューラルネットワークの構成を示す図である。

ニューラルネットワーク204は、入力層 S_1 、隠れ層 S_2 、出力層 S_3 の3段の階層構造を有する。

30

入力層 S_1 は、スコア数値列生成部203が生成するスコア数値列の要素数と同数のノード $N_1 - 1 \sim N_1 - 25$ （以下、ノード $N_1 - 1 \sim N_1 - 25$ を総称する場合は、ノード N_1 と記載する）で構成される。

隠れ層 S_2 は、入力層 S_1 のノード数と同数のノード $N_2 - 1 \sim N_2 - 25$ （以下、ノード $N_2 - 1 \sim N_2 - 25$ を総称する場合は、ノード N_2 と記載する）で構成される。

出力層 S_3 は、1つのノード N_3 で構成される。

【0074】

ノード N_1 のそれぞれは、スコア数値列生成部203が生成するスコア数値列のうち、自身に対応づけられた要素の値を入力し、ノード N_2 のそれぞれに出力する。ノード N_2 は、ノード N_1 のそれぞれが出力する値を入力し、当該入力した値を所定の記憶領域に記憶した伝達関数に代入し、得られた値をノード N_3 に出力する。ノード N_3 は、ノード N_2 のそれぞれが出力する値を入力し、当該入力した値を所定の記憶領域に記憶した伝達関数に代入し、得られた値を期待値として出力する。

40

なお、ノード N_2 、 N_3 が用いる伝達関数とは、前段のノードから入力したそれぞれの値と入力元のノードに対応する結合加重との積を総和し、得られる値が所定の閾値を超えた場合にのみ値を発火（出力）する関数である。ここで、ノード N_2 の伝達関数を式（4）に、ノード N_3 の伝達関数を式（5）に示す。

【0075】

【数 4】

$$f\left(\sum_{i=1}^n w_i x_i - \theta\right) \quad \dots (4)$$

【数 5】

$$f\left(\sum_{j=1}^m w_j x_j - \theta\right) \quad \dots (5)$$

【0076】

但し、 n は、ノード N_1 の総数を示す値であり、本実施形態では25となる。また、 w_i は、ノード $N_1 - i$ に対応する結合加重を示す。また、 x_i は、ノード $N_1 - i$ から入力した値を示す。また、 m は、ノード N_2 の総数を示す値であり、本実施形態では25となる。また、 w_j は、ノード $N_2 - j$ に対応する結合加重を示す。また、 x_j は、ノード $N_2 - j$ から入力した値を示す。また、 θ は、発火のための閾値を示す。また、関数 f は、0以上1以下の値を出力するシグモイド関数である。なお、シグモイド関数は、式(6)に示す関数である。

10

【0077】

【数 6】

$$f(x) = \frac{1}{1 + e^{-x}} \quad \dots (6)$$

20

【0078】

また、ニューラルネットワーク204は、既知のサイトを基準位置としたスコア数値列を入力とした場合に、期待値として1を出力し、既知のサイトでないアミノ酸残基を基準位置としたスコア数値列を入力した場合に、期待値として0を出力するように学習されている。

ここで、ニューラルネットワーク204の学習方法を説明する。

【0079】

まず、位置特異的スコア算出装置100が出現頻度の算出に用いたアミノ酸残基の部分配列におけるアミノ酸残基のそれぞれに対して、PSSM記憶部202が記憶する位置特異的スコアを割り当て、スコア数値列を生成する。

30

図10は、位置特異的スコアの割り当て方法を示す図である。例えば、抽出した所定の範囲のアミノ酸残基が、図10に示すように「VLY.....FSA.....SLI」という配列を示す場合、図7に示すPSSMを参照して、「-0.40、1.61、0.92、.....、0.09、0.78、1.25、.....、-1.22、0.86、-0.45」という数値列を生成する。

【0080】

次に、生成したスコア数値列をニューラルネットワーク204の入力層 S_1 の各ノード N_1 に入力する。ノード N_1 のそれぞれは、入力した値をノード N_2 のそれぞれに出力する。ノード N_2 は、ノード N_1 のそれぞれが出力する値を伝達関数に代入し、得られた値をノード N_3 に出力する。ノード N_3 は、ノード N_2 のそれぞれが出力する値を伝達関数に代入し、得られる値を期待値として出力する。

40

【0081】

他方、ニューラルネットワーク204のノード N_3 は、教師データの入力を受け付ける。教師データとは、入力したデータに対して期待される出力値を示すデータのことである。本実施形態においては、既知のサイトを基準位置としたスコア数値列を入力した場合、教師データは1であり、既知のサイトでないアミノ酸残基を基準位置としたスコア数値列を入力した場合、教師データは0である。次に、ニューラルネットワーク204の各ノードは、教師データと出力した期待値との誤差を最小にするように、自身が用いる伝達関数の結合加重 w_i 、閾値 θ を変化させる。

この処理をPSSMの作成に用いたそれぞれのアミノ酸残基の部分配列に対して実行す

50

る。これにより、ニューラルネットワーク204は、既知の サイトを基準位置としたスコア数値列を入力とした場合に、期待値として1を出力し、既知の サイトでないアミノ酸残基を基準位置としたスコア数値列を入力した場合に、期待値として0を出力することとなる。

【0082】

次に、本実施形態による サイト判定装置200の動作について説明する。

図11は、本発明の一実施形態による サイト判定装置200の動作を示すフローチャートである。

まず、入力部201は、検査対象タンパク質のアミノ酸配列情報と検査対象残基位置の入力を受け付ける(ステップS201)。次に、スコア数値列生成部203は、入力部201が受け付けた検査対象残基位置を含む所定の領域における複数のアミノ酸残基を、入力部201が受け付けたアミノ酸配列情報から抽出する(ステップS202)。なお、本実施形態では、所定の領域として、基準位置からN末端側に連続する12残基のアミノ酸残基とC末端側に連続する12残基のアミノ酸残基とを用いる。

10

【0083】

次に、スコア数値列生成部203は、PSSM記憶部202が記憶するPSSMに基づいて、抽出した所定の範囲の各アミノ酸残基の位置特異的スコアを特定し、当該疎水性指標値を示す数値列を生成する(ステップS203)。次に、ニューラルネットワーク204は、当該スコア数値列を入力し、検査対象残基位置の サイトらしさを示す0以上1以下の期待値を出力する(ステップS204)。

20

【0084】

ニューラルネットワーク204が期待値を出力すると、 サイト判定部205は、出力した期待値が0.5以上であるか否かを判定する(ステップS205)。つまり、 サイト判定部205は、ニューラルネットワーク204が出力した期待値が、 サイトであることを示す「1」と サイトでないことを示す「0」との何れに近いかを判定する。

【0085】

サイト判定部205は、ニューラルネットワーク204が出力した期待値が0.5以上であると判定した場合(ステップS205: YES)、ステップS201で入力部201が受け付けた検査対象残基位置が、 サイトであると判定する(ステップS206)。他方、 サイト判定部205は、ニューラルネットワーク204が出力した期待値が0.5未満であると判定した場合(ステップS205: NO)、ステップS201で入力部201が受け付けた検査対象残基位置が、 サイトでないと判定する(ステップS207)。

30

【0086】

上述した動作により、 サイト判定装置200は、高感度且つ高選択的に検査対象残基位置が サイトであるか否かを判定することができる。

なお、GPIアンカー型タンパク質及び非GPIアンカー型タンパク質それぞれの判定精度を求める方法としては、n-fold cross validation法(n分割交差検定法)、bootstrap法、jackknife法、Self-consistency(自己無撞着)な手法などを挙げることができる。ここで、判定精度とは、判定の感度、選択性、及び成功率のことを言う。

40

以下に、4分割交差検定法について詳述する。

【0087】

本実施形態では、以下の処理により、4分割交差検定法による判定精度を算出した。

まず、上述した45エントリのT_pと172エントリのF_pの位置特異的アミノ酸出現頻度を用いて、PSSMを生成する。次に、生成したPSSMに基づくスコアをT_pとF_pのアミノ酸配列データに割り当てる。次に、スコアを割り当てたT_pとF_pのアミノ酸配列データを4分割し、そのうちの3つの部分データセットを用いてニューラルネットワーク204の学習を行う。次に、PSSMに基づいて、他の1つの部分データセットの各エントリのスコア数値列を生成する。次に、当該算出したスコアに基づいて

50

、感度、選択性、成功率を算出する。そして、ニューラルネットワーク204を学習する部分データセットとスコアを算出する部分データセットとの全ての組み合わせに対して判定精度を算出し、それぞれの平均値をデータセット全体に対する判定精度として算出する。

【0088】

4分割交差検定法について、図12を用いて、さらに具体的に説明する。

図12は、本実施形態による サイト判定装置200の判定精度を示す表である。

図12では、 サイト判定装置200が、 サイトであると判定した検査対象残基位置の判定精度、及び サイトでないと判定した検査対象残基位置の判定精度を示している。また、図12に示す サイト及び非 サイトそれぞれの判定精度を求めるにあたり、4分割交差検定法を100回実行した。

10

【0089】

図12に示すように、4分割交差検定法による サイトの判定精度は、冗長性94%の場合、4分割交差検定法を100回実行した平均の感度が92.99%、選択性が92.98%、成功率が0.93であった。なお、ここで冗長性の百分率(ここでは「94%」)は、基準位置のアミノ酸残基と基準位置からN末端側に連続する12残基のアミノ酸残基とC末端側に連続する12残基のアミノ酸残基とからなる範囲の全アミノ酸残基のうち、一致又は類似しているアミノ酸残基の割合を示す。また、4分割交差検定法を100回実行した場合における上位10回の成功率の平均の感度が94.83%、選択性が95.84%、成功率が0.95であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が94.08%、選択性が98.33%、成功率が0.96であった。

20

【0090】

また、図12に示すように、4分割交差検定法による非 サイトの判定精度は、冗長性94%の場合、4分割交差検定法を100回実行した平均の感度が98.96%、選択性が98.99%、成功率が0.99であった。また、4分割交差検定法を100回実行した場合における上位10回の成功率の平均の感度が99.30%、選択性が99.21%、成功率が0.99であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が99.78%、選択性が99.11%、成功率が0.99であった。

30

【0091】

また、図12に示すように、4分割交差検定法による サイトの判定精度は、冗長性90%の場合、4分割交差検定法を100回実行した平均の感度が95.04%、選択性が95.99%、成功率が0.95であった。また、4分割交差検定法を100回実行した場合における上位10回の成功率の平均の感度が97.96%、選択性が98.71%、成功率が0.98であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が98.33%、選択性が100.00%、成功率が0.99であった。

【0092】

また、図12に示すように、4分割交差検定法による非 サイトの判定精度は、冗長性90%の場合、4分割交差検定法を100回実行した平均の感度が99.27%、選択性が99.11%、成功率が0.99であった。また、4分割交差検定法を100回実行した場合における上位10回の成功率の平均の感度が99.75%、選択性が99.63%、成功率が1.00であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が100.00%、選択性が99.69%、成功率が1.00であった。

40

【0093】

また、図12に示すように、4分割交差検定法による サイトの判定精度は、冗長性80%の場合、4分割交差検定法を100回実行した平均の感度が89.45%、選択性が90.88%、成功率が0.90であった。また、4分割交差検定法を100回実行した

50

場合における上位10回の成功率の平均の感度が95.64%、選択性が96.97%、成功率が0.96であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が100.00%、選択性が93.75%、成功率が0.97であった。

【0094】

また、図12に示すように、4分割交差検定法による非 サイトの判定精度は、冗長性80%の場合、4分割交差検定法を100回実行した平均の感度が98.83%、選択性が98.57%、成功率が0.99であった。また、4分割交差検定法を100回実行した場合における上位10回の成功率の平均の感度が99.59%、選択性が99.38%、成功率が0.99であった。また、4分割交差検定法を100回実行した場合における成功率が最高値のときの感度が98.94%、選択性が100.00%、成功率が0.99であった。

10

【0095】

このように、本実施形態による サイト判定装置200によれば、非特許文献1～非特許文献5に係る方法(感度58%～88%)と比較して、高感度且つ高選択的に検査対象残基位置が サイトであるか否かを判定することができる。

【0096】

《 サイト特定装置》

本実施形態による サイト特定装置は、検査対象となるタンパク質のアミノ酸配列情報の入力を受け付け、当該タンパク質における サイトの残基位置を特定する。

20

【0097】

図13は、本発明の一実施形態による サイト特定装置300の構成を示す概略ブロック図である。

サイト特定装置300は、入力部301、PSSM記憶部302、スコア数値列生成部303、ニューラルネットワーク304(分類部)、 サイト特定部305(GPIアンカー修飾部位特定部)を備える。

【0098】

入力部301は、検査対象となるタンパク質のアミノ酸配列情報の入力を受け付ける。

PSSM記憶部302は、位置特異的スコア算出装置100が算出した位置特異的スコアを用いて生成されたPSSMを記憶する。

30

スコア数値列生成部303は、PSSM記憶部302が記憶するPSSMに基づいて、入力部301が受け付けたアミノ酸配列情報が示す各アミノ酸残基の残基位置を基準位置とする所定の領域におけるスコア数値列を生成する。ここで生成するスコア数値列とは、入力部301が受け付けたアミノ酸配列情報の所定の領域のそれぞれのアミノ酸残基の位置特異的スコアを要素とする配列である。

ニューラルネットワーク304は、スコア数値列生成部303が生成したスコア数値列を入力し、 サイトらしさを示す0以上1以下の期待値を出力する。なお、ニューラルネットワーク304は、ニューラルネットワーク204と同様の学習がなされている。

サイト特定部305は、入力部301が受け付けたアミノ酸配列情報における サイトの位置を特定する。

40

【0099】

次に、本実施形態による サイト特定装置300の動作について説明する。

図14は、本発明の一実施形態による サイト特定装置300の動作を示すフローチャートである。

まず、入力部301は、検査対象タンパク質のアミノ酸配列情報の入力を受け付ける(ステップS301)。次に、スコア数値列生成部303は、入力部301が受け付けたアミノ酸配列情報が示すアミノ酸残基を1つずつ選択し、当該アミノ酸残基ごとに、以下に示すステップS303～ステップS305の処理を実行する(ステップS302)。

【0100】

まず、スコア数値列生成部303は、ステップS302で選択したアミノ酸残基を含む

50

所定の領域における複数のアミノ酸残基を、入力部 301 が受け付けたアミノ酸配列情報から抽出する（ステップ S303）。なお、本実施形態では、所定の領域として、基準位置から N 末端側に連続する 12 残基のアミノ酸残基と C 末端側に連続する 12 残基のアミノ酸残基とを用いる。

【0101】

次に、スコア数値列生成部 303 は、PSSM 記憶部 302 が記憶する PSSM に基づいて、抽出した所定の範囲の各アミノ酸残基の位置特異的スコアを特定し、当該疎水性指標値を示す数値列を生成する（ステップ S304）。次に、ニューラルネットワーク 304 は、当該スコア数値列を入力し、検査対象残基位置の サイトらしさを示す 0 以上 1 以下の期待値を出力する（ステップ S305）。

10

【0102】

入力部 301 が受け付けたアミノ酸配列情報が示す全てのアミノ酸残基について、ニューラルネットワーク 304 が期待値を出力すると、サイト特定部 305 は、最も期待値が大きい値を示すアミノ酸残基を、サイトと特定する（ステップ S306）。これにより、期待値が 0.5 以上のアミノ酸残基が複数出現した場合や、全ての期待値が 0.5 未満であった場合にも、サイトの位置を特定することができる。

【0103】

上述した動作により、サイト特定装置 300 は、検査対象タンパク質のサイトの位置を精度良く特定することができる。

【0104】

以上、図面を参照してこの発明の一実施形態について詳しく説明してきたが、具体的な構成は上述のものに限られることはなく、この発明の要旨を逸脱しない範囲内において様々な設計変更等を行うことが可能である。

20

例えば、上述した サイト判定装置 200 及び サイト特定装置 300 は、それぞれニューラルネットワーク 204、ニューラルネットワーク 304 により期待値を算出し、当該期待値に基づいて サイトの判定・特定を行う場合について説明したが、これに限られない。例えば、ニューラルネットワーク 204、ニューラルネットワーク 304 による期待値の算出に代えて、スコア数値列の平均値をスコアとして算出し、当該スコアを用いて サイトの判定・特定を行っても良い。この場合、サイト判定部 205 は、スコアが所定の閾値以上である場合に、検査対象残基位置が サイトであると判定する。また、サイト特定部 305 は、スコアが最も高いアミノ酸残基を、サイトと特定する。

30

【0105】

また、本実施形態では、タンパク質の完全長アミノ酸配列情報に基づいて位置特異的スコアの算出、サイトの判定及びサイトの特定を行ったが、これに限られず、完全長塩基配列情報を用いても良い。ただし、この場合、常法によるイントロ配列の除去処理及びアミノ酸配列情報への翻訳処理を行ってから、各処理を行うこととなる。

【0106】

なお、上述の位置特異的スコア算出装置 100、サイト判定装置 200、及び サイト特定装置 300 は、内部にコンピュータシステムを有している。そして、上述した各処理部の動作は、プログラムの形式でコンピュータ読み取り可能な記録媒体に記憶されており、このプログラムをコンピュータが読み出して実行することによって、上記処理が行われる。ここでコンピュータ読み取り可能な記録媒体とは、磁気ディスク、光磁気ディスク、CD-ROM、DVD-ROM、半導体メモリ等をいう。また、このコンピュータプログラムを通信回線によってコンピュータに配信し、この配信を受けたコンピュータが当該プログラムを実行するようにしても良い。

40

【0107】

また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

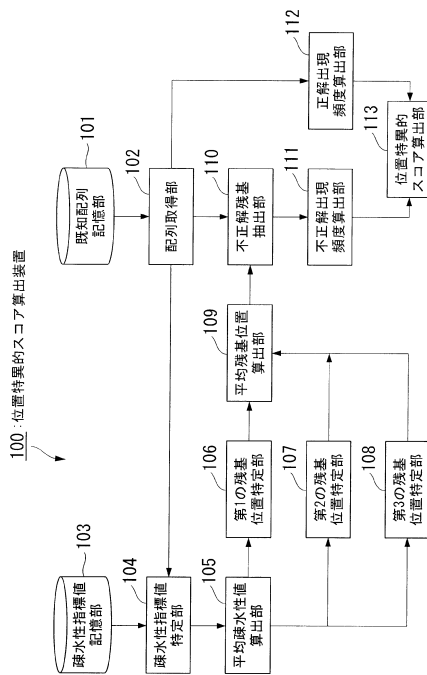
【符号の説明】

50

【 0 1 0 8 】

1 0 0 ... 位置特異的スコア算出装置 1 0 1 ... 既知配列記憶部 1 0 2 ... 配列取得部
 1 0 3 ... 疎水性指標値記憶部 1 0 4 ... 疎水性指標値特定部 1 0 5 ... 平均疎水性値算出部
 1 0 6 ... 第1の残基位置特定部 1 0 7 ... 第2の残基位置特定部 1 0 8 ... 第3の残基位置特定部
 1 0 9 ... 平均残基位置算出部 1 1 0 ... 不正解残基抽出部 1 1 1 ... 不正解出現頻度算出部
 1 1 2 ... 正解出現頻度算出部 1 1 3 ... 位置特異的スコア算出部 2
 0 0 ... サイト判定装置 2 0 1 ... 入力部 2 0 2 ... P S S M 記憶部 2 0 3 ... スコア数値列生成部
 2 0 4 ... ニューラルネットワーク 2 0 5 ... サイト判定部 3 0 0 ... サイト特定装置
 3 0 1 ... 入力部 3 0 2 ... P S S M 記憶部 3 0 3 ... スコア数値列生成部
 3 0 4 ... ニューラルネットワーク 3 0 5 ... サイト特定部

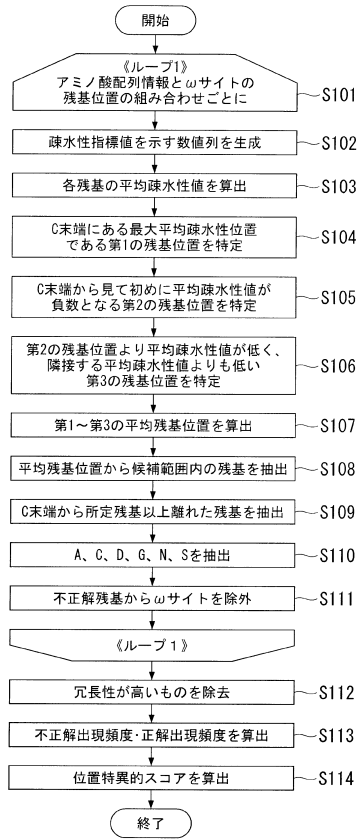
【 図 1 】



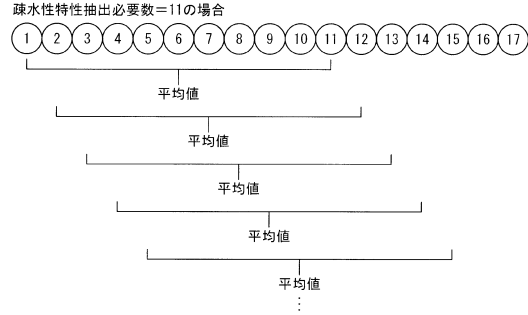
【 図 2 】

アミノ酸	指標	アミノ酸	指標
A	1.8	L	3.8
R	-4.5	K	-3.9
N	-3.5	M	1.9
D	-3.5	F	2.8
C	2.5	P	-1.6
Q	-3.5	S	-0.8
E	-3.5	T	-0.7
G	-0.4	W	-0.9
H	-3.2	Y	-1.3
I	4.5	V	4.2

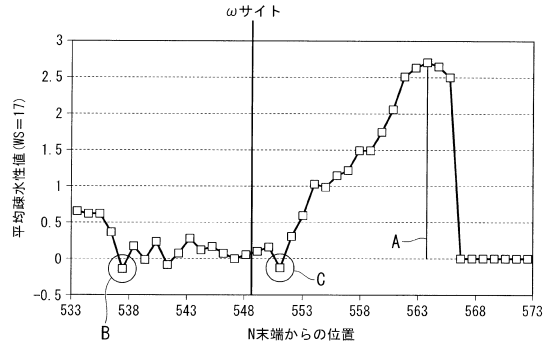
【図3】



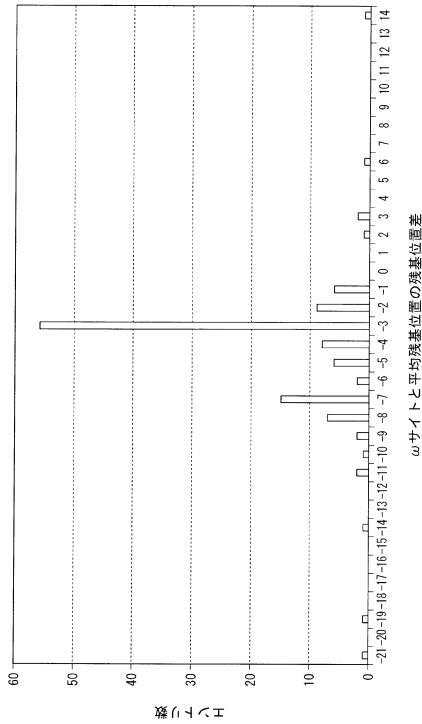
【図4】



【図5】



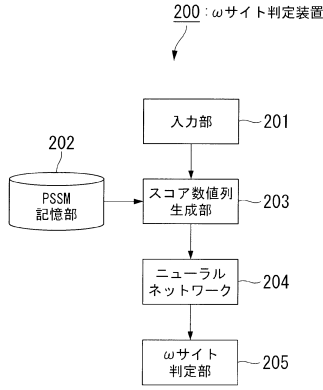
【図6】



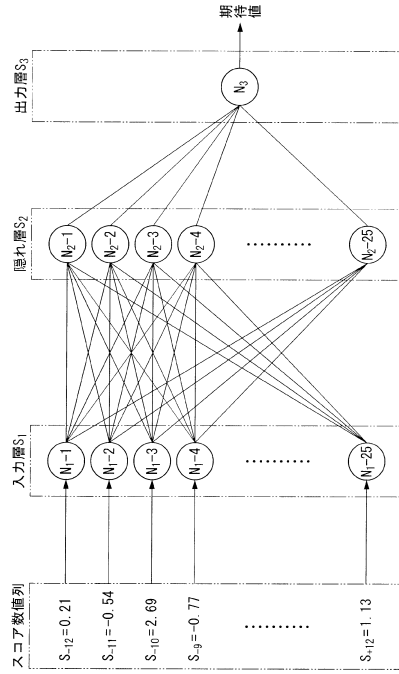
【図7】

	-12	-11	-10	...	-1	0	1	...	10	11	12
A	-3.47	1.99	-1.03	...	-1.03	-0.86	1.25	...	0.38	-0.11	-0.96
C	-0.16	-1.03	1.46	...	1.32	-1.22	-3.98	...	-2.79	-0.25	-0.03
D	-0.58	-3.76	1.04	...	-3.07	0.84	-0.03	...	-2.39	-3.07	-1.72
E	0.13	-0.27	0.45	...	-0.35	1.32	0.09	...	-3.29	-3.07	-2.79
F	-3.62	-0.03	-3.98	...	0.09	1.32	0.26	...	1.55	-0.25	0.82
...
I	-3.47	-3.76	1.04	...	-2.39	1.32	-3.62	...	-2.79	0.35	-0.45
K	1.32	0.35	-0.20	...	2.16	1.32	-3.62	...	-2.79	-2.79	-2.79
L	0.29	1.61	-4.72	...	-1.76	1.32	-1.27	...	-0.58	0.86	0.37
...
S	-1.96	-0.36	-1.34	...	-0.61	0.78	0.11	...	-1.22	0.03	-1.03
T	0.54	-1.20	-0.43	...	-0.05	1.32	-3.88	...	0.04	-0.93	-4.45
V	-0.40	0.99	-0.03	...	-0.83	1.32	-3.76	...	0.72	0.11	1.04
W	1.32	1.32	1.32	...	-1.72	1.32	-1.72	...	-3.29	-2.39	0.93
Y	1.32	-0.53	0.92	...	0.24	1.32	-3.07	...	-3.07	-2.39	-3.29

【図8】



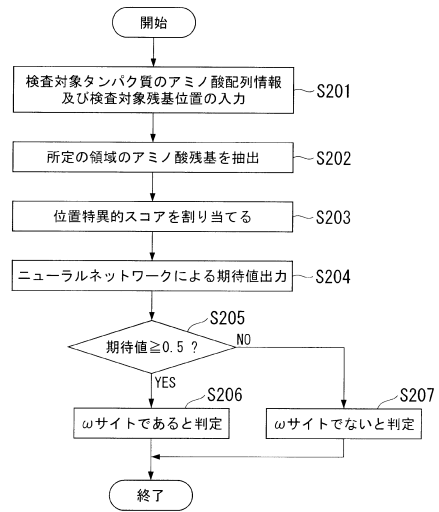
【図9】



【図10】

	-12	-11	-10	-1	0	1	10	11	12	
A	-3.47	1.99	-1.03	-1.03	-0.86	1.25	0.38	-0.11	-0.96	
C	-0.16	-1.03	1.46	1.32	-1.22	0.88	-2.79	-0.25	-0.03	
D	-0.58	-3.76	1.04	3.07	0.84	0.33	-3.29	-3.07	-1.72	
E	0.13	-0.27	0.45	-0.36	1.32	0.6	1.55	-3.07	-2.79	
F	-3.62	-0.03	-3.98	0.09	1.32	0.6	1.55	-0.25	0.82	
I	-3.47	-3.76	1.04	1.9	1.32	0.32	-2.79	0.35	-0.45	
K	1.32	0.35	-0.20	2.6	1.32	0.32	-2.79	-2.79	0.79	
L	0.29	1.61	-4.72	1.6	1.32	0.17	-0.58	0.86	0.7	
S	-1.96	-0.36	-1.34	1.1	0.78	0.1	-1.22	0.3	-0.33	
T	0.54	-0.20	-0.43	0.5	1.2	0.8	0.4	-0.33	-0.15	
V	-0.40	0.9	-0.03	0.3	1.2	0.16	0.2	0.1	0.4	
W	0.2	1.32	0.2	0.2	1.2	0.2	-0.79	-0.79	0.3	
Y	1.2	0.33	0.92	0.4	1.2	0.17	-0.79	-0.79	0.19	
	-0.40	1.61	0.92	0.09	0.78	1.25	-1.22	0.86	-0.45	

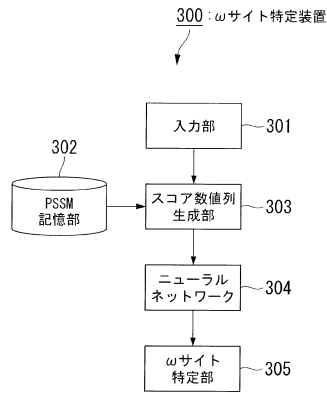
【図11】



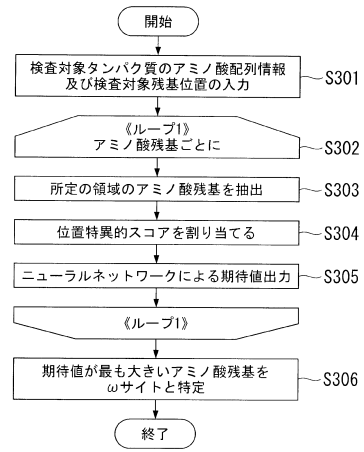
【図12】

		ωサイト			非ωサイト		
		感度	選択性	成功率	感度	選択性	成功率
100エントリー平均	冗長性94%	92.99	92.98	0.93	98.96	98.99	0.99
	冗長性90%	95.04	95.99	0.95	99.27	99.11	0.99
	冗長性80%	89.45	90.88	0.90	98.83	98.57	0.99
Top10平均	冗長性94%	94.83	95.84	0.95	99.30	99.21	0.99
	冗長性90%	97.96	98.71	0.98	99.75	99.63	1.00
	冗長性80%	95.64	96.97	0.96	99.59	99.38	0.99
BEST1	冗長性94%	94.08	98.33	0.96	99.78	99.11	0.99
	冗長性90%	98.33	100.00	0.99	100.00	99.69	1.00
	冗長性80%	100.00	93.75	0.97	98.94	100.00	0.99

【図13】



【図14】



フロントページの続き

(72)発明者 池田 有理

神奈川県川崎市多摩区東三田 1 - 1 - 1 学校法人明治大学 生田キャンパス内

(72)発明者 佐々木 貴規

神奈川県川崎市多摩区東三田 1 - 1 - 1 学校法人明治大学 生田キャンパス内

審査官 田付 徳雄

(56)参考文献 特開 2012 - 032163 (JP, A)

特開 2011 - 060004 (JP, A)

(58)調査した分野(Int.Cl., DB名)

G06F 19/10 - 19/28