(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2014-48797
(P2014-48797A)
(43)公開日　平成26年3月17日(2014.3.17)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 19/22 (2011.01)	G06F 19/22	

審査請求 未請求　請求項の数 18　ＯＬ　（全 27 頁）

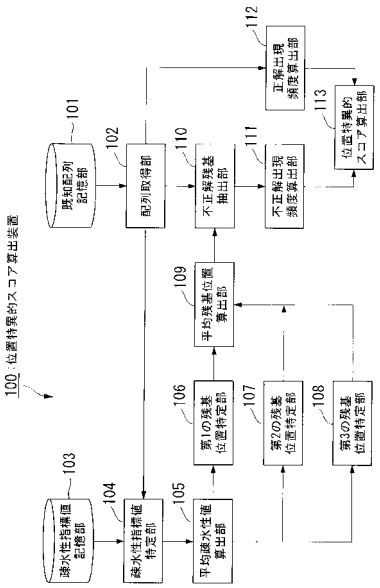
(21)出願番号　特願2012-189907 (P2012-189907)
(22)出願日　平成24年8月30日 (2012.8.30)

(71)出願人　801000027
学校法人明治大学
東京都千代田区神田駿河台１－１
(74)代理人　100064908
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)【要約】
【課題】ＧＰＩアンカー修飾部位（ωサイト）の特定に特化した位置特異的スコアを算出する。
【解決手段】正解出現頻度算出部１１２は、ωサイト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。不正解残基抽出部１１０は、位置特異的スコアの算出に用いるωサイト以外のアミノ酸残基を抽出する。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。位置特異的スコア算出部１１３は、アミノ酸残基の種類ごとに、正解出現頻度を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。
【選択図】図１



【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、
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、
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。
【請求項９】
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、
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ＧＰＩ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、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
ことを特徴とする請求項９に記載のＧＰＩアンカー修飾部位の特定装置。
【請求項１１】
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
ことを特徴とする請求項１０に記載のＧＰＩ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、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、
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、
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特定装置。

【請求項１６】
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特定する
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特定装置。

【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、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、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（ω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、ω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ω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、ω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ωサイトであるとは限らない。

【０００５】

また現在、既知のＧＰＩアンカー型タンパク質においても、ω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ω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、ω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 ω－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 Ｐｒｏ

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ω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ω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。
【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。

【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、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。
【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、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である。
【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。

【００２４】

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特定することを特徴とする。

【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、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、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。

【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できる。

【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】ω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ωサイトとの残基位置差を示す図である。

【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ω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ω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ω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ωサイト特定装置の構成を示す概略ブロック図である。
【図１４】本発明の一実施形態によるω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、ω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ωサイトである可能性を示す値であり、当該値が大きいほど、部分配列の中心のアミノ酸残基がωサイトである可能性が高いことを示す。これにより、ω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ω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。
【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ω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ω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。
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。

【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。

【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】

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ω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）ω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ＧＰＩアンカー型タンパク質における平均残基位置と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。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。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、ω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、ＰＳＳＭ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ＫＹＴＪ８２０１０１（Ｋｙｔｅ Ｊ．，Ｄｏｏｌｉｔｔｌｅ Ｒ．，「Ｊｏｕｒｎａｌ ｏｆ Ｍｏｌｅｃｕｌａｒ Ｂｉｏｌｏｇｙ」、１９８２年、ｖｏｌ．１５７、ｎｏ．１、ｐｐ．１０５－１３２）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ω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Ｓｗｉｓｓ Ｐｒｏｔ Ｒｅｌｅａｓｅ ２０１１０７を用いる。本データバンクのうち、ωサイトの位置が確定している実験的確証のあるＧＰＩ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、実験的確証のあるＧＰＩアンカー型タンパク質のアミノ酸配列情報に加えて、実験的確証のあるＧＰＩアンカー型タンパク質に類似する（Ｂｙ ｓｉｍｉｌａｒｉｔｙ）タンパク質のアミノ酸配列情報を用いる。実験的確証のあるＧＰＩ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、実験的確証のあるＧＰＩ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ＧＰＩ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、実験的確証のあるＧＰＩアンカー型タンパク質のアミノ酸配列情報とω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、ω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、実験的確証のあるＧＰＩアンカー型タンパク質のωサイトと揃った残基位置を、ω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。

【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、ω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、実験的確証のあるＧＰＩアンカー型タンパク質のωサイトとアラインメントを取ることで、ω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。

そして、当該１２２エントリのアミノ酸配列情報及びωサイトの残基位置を、既知のＧＰＩアンカー型タンパク質のアミノ酸配列情報及びω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、ω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、ω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ω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。

【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、ωサイトであるアミノ酸残基の数に対して、ω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、ω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、ω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、ω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。

【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ω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）。

【００４７】

疎水性指標値特定部１０４は、疎水性指標値記憶部１０３を参照して、配列取得部１０２がステップＳ１０１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Ｓ１０２）。例えば、配列取得部１０２が取得したアミノ酸配列情報が、「ＭＬＬＥＰＧＲＧＣＣ……」という配列を示す場合、疎水性指標値特定部１０４は、疎水性指標値記憶部１０３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。

【００４８】

図４は、平均疎水性値の算出方法を示す図である。

次に、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（ステップＳ１０３）。具体的には、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、１７残基である。

【００４９】

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。

【００５０】

【数１】

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

【００５１】

但し、ｎは、平均化に用いる前後の残基数を示す。つまり、２ｎ＋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。

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ＮＵＬＬ値を代入しておくと良い。

【００５２】

図５は、ω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Ｓ１０３で、平均疎水性値算出部１０５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１００は、当該平均疎水性値を用いて、ω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。

【００５３】

まず、第１の残基位置特定部１０６は、Ｃ末端から１４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Ｓ１０４）。

【００５４】

また、第２の残基位置特定部１０７は、平均疎水性値算出部１０５が算出した平均疎水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）。

【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ω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ω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、ω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ω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）。

【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ω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。

【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、ω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、ω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ω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のことである。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ω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／に開示されている。）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ω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】

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

【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ωサイトを用いて、当該ω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ω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）。

【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】

$$\ln \frac{f_{ip}^{T\omega p}}{f_{ip}^{F\omega p}} \quad \cdots (3)$$

【００６６】

但し、$f_{ip}^{T\omega p}$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、$f_{ip}^{F\omega p}$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、ω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、ωサイトの特定に特化したＰＳＳＭを生成することができる。

【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、ω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】

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】

《ωサイト判定装置》

本実施形態によるω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ω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ω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ωサイト判定装置２００の構成を示す概略ブロック図である。

ω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）、ω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ω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。

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、ωサイトらしさを示す０以上１以下の期待値を出力する。

ω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ω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。

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～N_1－２５（以下、ノードN_1－１～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～N_2－２５（以下、ノードN_2－１～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。

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】

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

【数５】

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

【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。

【００７７】

【数６】

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

【００７８】

また、ニューラルネットワーク２０４は、既知のω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ω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。
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。
【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ω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ω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る。これにより、ニューラルネットワーク２０４は、既知のω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ω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ω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ω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。
【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ω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）。
【００８４】
ニューラルネットワーク２０４が期待値を出力すると、ω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、ω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、ωサイトであることを示す「１」とω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】
ω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、ωサイトであると判定する（ステップＳ２０６）。他方、ω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、ωサイトでないと判定する（ステップＳ２０７）。

【００８６】

上述した動作により、ω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ω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。

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Ｔωｐと１７２エントリのＦωｐ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ＴωｐとＦωｐ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ＴωｐとＦωｐ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、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ωサイト判定装置２００の判定精度を示す表である。

図１２では、ωサイト判定装置２００が、ω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ω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ωサイト及び非ω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。

【００８９】

図１２に示すように、４分割交差検定法によるω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。

【００９０】

また、図１２に示すように、４分割交差検定法による非ω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。

【００９１】
また、図１２に示すように、４分割交差検定法によるω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ω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。
【００９３】
また、図１２に示すように、４分割交差検定法によるω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ω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。
【００９５】
このように、本実施形態によるω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ωサイトであるか否かを判定することができる。
【００９６】
《ωサイト特定装置》
本実施形態によるω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ωサイトの残基位置を特定する。
【００９７】
図１３は、本発明の一実施形態によるωサイト特定装置３００の構成を示す概略ブロック図である。
ω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）、ω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。

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、ω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。

ω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ωサイトの位置を特定する。

【００９９】

次に、本実施形態によるω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ω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ω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）。

【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、ω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、ω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、ω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、ωサイト特定装置３００は、検査対象タンパク質のω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することが可能である。

例えば、上述したωサイト判定装置２００及びω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ω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ωサイトの判定・特定を行っても良い。この場合、ω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ωサイトであると判定する。また、ωサ

イト特定部３０５は、スコアが最も高いアミノ酸残基を、ωサイトと特定する。

【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、ωサイトの判定及びω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、ωサイト判定装置２００、及びω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。

【０１０７】

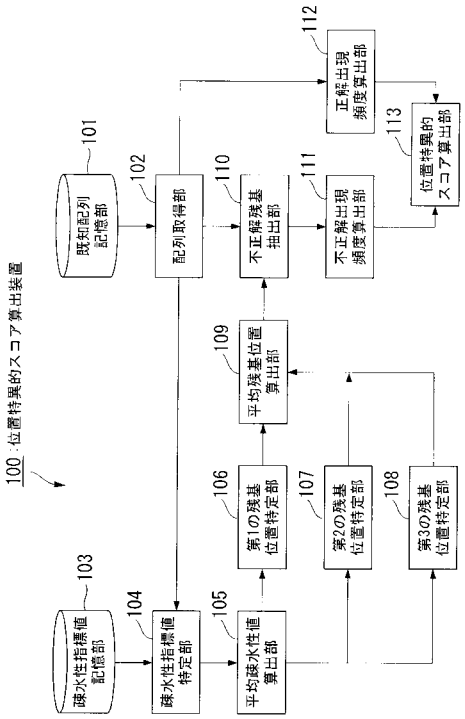
また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

【符号の説明】

【０１０８】

１００…位置特異的スコア算出装置　１０１…既知配列記憶部　１０２…配列取得部　１０３…疎水性指標値記憶部　１０４…疎水性指標値特定部　１０５…平均疎水性値算出部　１０６…第１の残基位置特定部　１０７…第２の残基位置特定部　１０８…第３の残基位置特定部　１０９…平均残基位置算出部　１１０…不正解残基抽出部　１１１…不正解出現頻度算出部　１１２…正解出現頻度算出部　１１３…位置特異的スコア算出部　２００…ωサイト判定装置　２０１…入力部　２０２…ＰＳＳＭ記憶部　２０３…スコア数値列生成部　２０４…ニューラルネットワーク　２０５…ωサイト判定部　３００…ωサイト特定装置　３０１…入力部　３０２…ＰＳＳＭ記憶部　３０３…スコア数値列生成部　３０４…ニューラルネットワーク　３０５…ωサイト特定部

【図１】

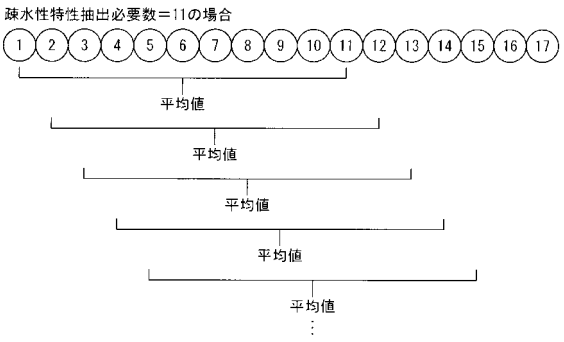


【図２】

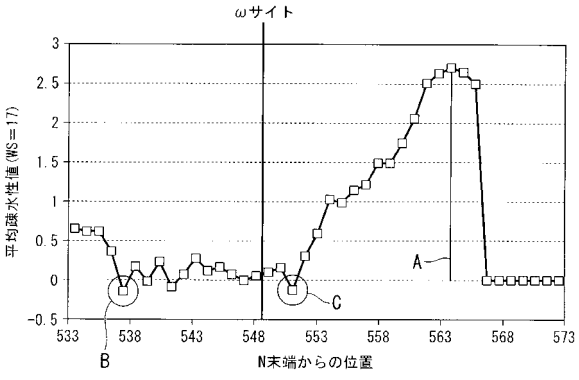
アミノ酸	指標	アミノ酸	指標
A	1.8	L	3.8
R	-4.5	K	-3.9
N	-3.5	M	1.9
D	-3.5	F	2.8
C	2.5	P	-1.6
Q	-3.5	S	-0.8
E	-3.5	T	-0.7
G	-0.4	W	-0.9
H	-3.2	Y	-1.3
I	4.5	V	4.2

【図３】

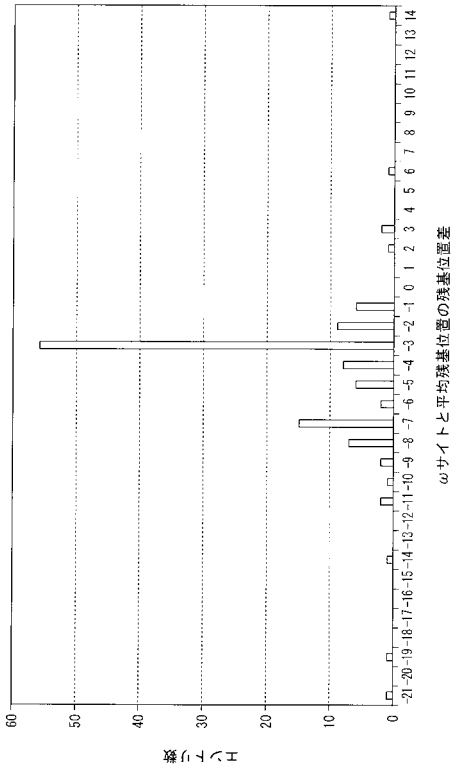
【図４】



【図５】



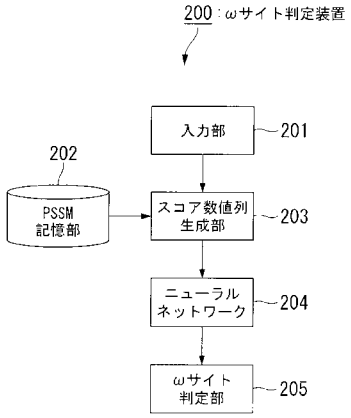
【図６】



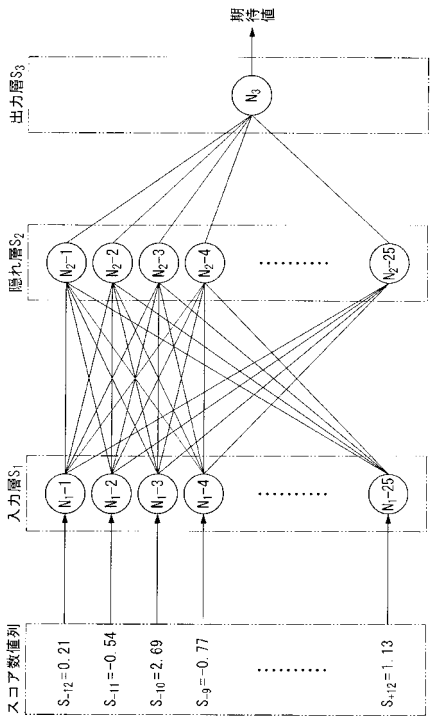
【図７】

	-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	…	-0.36	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	-3.98	…	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

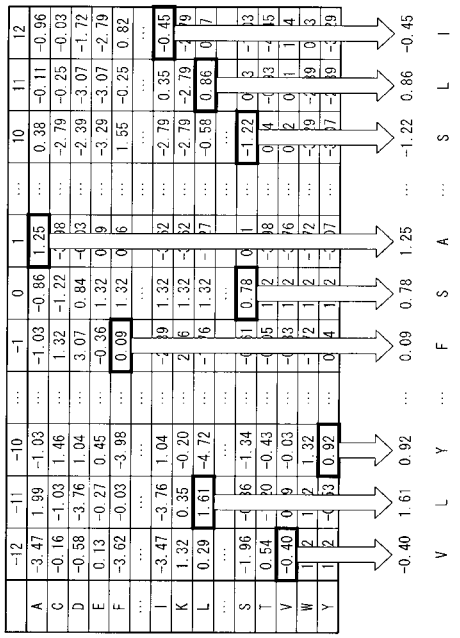
【図８】



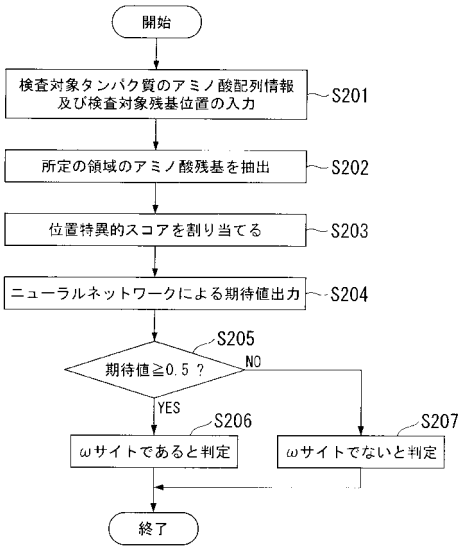
【図９】



【図１０】



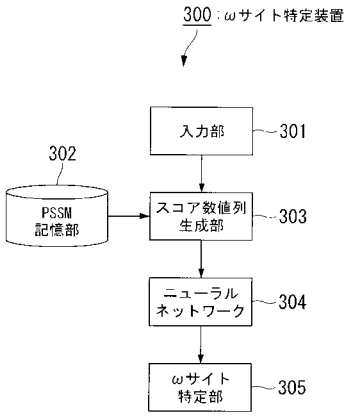
【図１１】



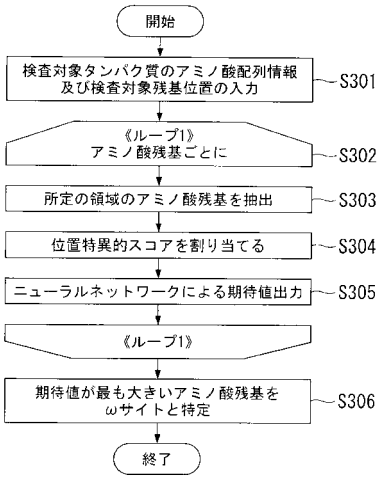
【図１２】

		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