(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5773406号
(P5773406)
(45)発行日　平成27年9月2日(2015.9.2)　(24)登録日　平成27年7月10日(2015.7.10)

(51)Int.Cl.　F I
G06F 19/24 (2011.01)　G06F 19/24
G01N 33/48 (2006.01)　G01N 33/48 Z

請求項の数 19　(全 32 頁)

(21)出願番号　特願2010-169324(P2010-169324)
(22)出願日　平成22年7月28日(2010.7.28)
(65)公開番号　特開2012-32163(P2012-32163A)
(43)公開日　平成24年2月16日(2012.2.16)
審査請求日　平成25年6月3日(2013.6.3)

(73)特許権者　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　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　生田校舎内

最終頁に続く

(54)【発明の名称】ＧＰＩアンカー型タンパク質の判定装置、判定方法及び判定プログラム

(57)【特許請求の範囲】
【請求項１】
検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置であって、
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と、
前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出する側鎖サイズ算出部と、
既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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前記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、既知の非ＧＰＩアンカー型タンパク質の前記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た分類部と、
前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定するＧＰＩアンカー型タンパク質判定部と、
を備えることを特徴とするＧＰＩアンカー型タンパク質の判定装置。
【請求項２】
前記分類部は、ニューラルネットワークであり、
前記スコア数値列生成部が生成するスコア数値列の要素数と同数のノードで構成される入力層と、複数のノードで構成される隠れ層と、１つのノードで構成される出力層とを少なくとも含む階層型の構造を有し、
前記入力層の各ノードは、前記スコア数値列のうち自身に対応づけられた要素が示す値を前記隠れ層のノードのそれぞれに出力し、
前記隠れ層の各ノードは、前記入力層の各ノードが出力する値を所定の伝達関数に代入し、得られた値を前記出力層のノードに出力し、
前記出力層のノードは、前記隠れ層の各ノードが出力する値を所定の伝達関数に代入し、得られた値を期待値として出力する
ことを特徴とする請求項１に記載のＧＰＩアンカー型タンパク質の判定装置。
【請求項３】
前記分類部は、既知のＧＰＩアンカー型タンパク質の前記スコア数値列を入力した場合に、期待値として１を出力するように前記ノードの伝達関数の係数を変化させ、前記既知の非ＧＰＩアンカー型タンパク質の前記スコア数値列を入力した場合に、期待値として０を出力するように前記ノードの伝達関数の係数を変化させることで学習されたことを特徴とする請求項２に記載のＧＰＩアンカー型タンパク質の判定装置。
【請求項４】
前記ノードのそれぞれは伝達関数としてシグモイド関数を用いることを特徴とする請求項２または請求項３に記載のＧＰＩアンカー型タンパク質の判定装置。
【請求項５】
前記側鎖サイズ特性抽出必要数は、
当該側鎖サイズ特性抽出必要数を用いて、既知の複数のＧＰＩアンカー型タンパク質の小側鎖サイズ判定領域に対して平均側鎖サイズを算出した場合に、前記ＧＰＩアンカー型タンパク質から算出した平均側鎖サイズが最小となるアミノ酸残基のうち、当該アミノ酸残基のＣ末端側に隣接するアミノ酸残基がＧＰＩアンカー修飾部位であるものの個数が最大となるような値である
ことを特徴とする請求項１から請求項４の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。
【請求項６】
前記小側鎖サイズ判定領域は、
既知のＧＰＩアンカー型タンパク質の前記平均側鎖サイズが最小となる位置が含まれる領域である、
ことを特徴とする請求項５に記載のＧＰＩアンカー型タンパク質の判定装置。
【請求項７】
前記位置特異的スコアは、
既知の複数のＧＰＩアンカー型タンパク質の平均側鎖サイズが最小となる位置を基準位置とする前記所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を示す$f_{ip}^{positive}$、既知の複数の非ＧＰＩアンカー型タンパク質の平均側鎖サイズが

最小となる位置を基準位置とする前記所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を示すｆ$_{ip}^{negative}$を用いて、

【数１】

$$\ln \frac{f_{ip}^{positive}}{f_{ip}^{negative}}$$

から算出されたものであることを特徴とする請求項１から請求項６の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項８】

前記所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度は、

種類ｉのアミノ酸残基が位置ｐに存在する既知のＧＰＩアンカー型タンパク質の個数を示すｎ$_{ip}$と、当該出現頻度の調整値を示すεと、アミノ酸残基の種類数ｓとを用いて、

【数２】

$$\frac{n_{ip} + \varepsilon \frac{1}{s}}{\sum_{i=1}^{s} n_{ip} + \varepsilon}$$

から算出されたものであることを特徴とする請求項７に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項９】

前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＮ末端から予め定められた残基数の領域を特定し、当該Ｎ末端側の高疎水性領域に対応する領域のアミノ酸残基を抽出し、前記Ｎ末端側の高疎水性領域に対応する領域の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Ｎ末端側疎水性特性抽出必要数を用いて、連続する当該Ｎ末端側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Ｎ末端側平均疎水性値を、前記抽出した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複数算出するＮ末端側疎水性値算出部と、

前記Ｎ末端側疎水性値算出部が算出した複数のＮ末端側平均疎水性値のうちの最大値が、既知のＧＰＩアンカー型タンパク質におけるＮ末端側平均疎水性値の特性を示すＮ末端側疎水性閾値以上であるか否かを判定するＮ末端側疎水性判定部と

を備え、

前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｎ末端側疎水性判定部が、前記Ｎ末端側疎水性値算出部の算出したＮ末端側平均疎水性値の最大値が前記Ｎ末端側疎水性閾値以上であると判定したアミノ酸配列情報に対して処理を行う

ことを特徴とする請求項１から請求項８の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１０】

前記Ｎ末端側疎水性閾値は、

予め既知の複数のＧＰＩアンカー型タンパク質に対して前記Ｎ末端側平均疎水性値の算出を行い、当該算出されたＮ末端側平均疎水性値の最大値の集合における最小値である

ことを特徴とする請求項９に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１１】

前記Ｎ末端側疎水性特性抽出必要数は、

当該Ｎ末端側疎水性特性抽出必要数を用いて、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のアミノ酸残基のそれぞれに対してＮ末端側平均疎水性値を算出し、前記既知のＧＰＩアンカー型タンパク質から算出したＮ末端側平均疎水性値の最大値の集合における最小値を抽出し、前記Ｎ末端側疎水性特性抽出必要数を用いて、既知の複数の非ＧＰＩアンカー型タンパク質における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域のアミノ酸残基のそれぞれに対してＮ末端側平均疎水性値を算出した場合に、前記既知の非ＧＰＩアンカー型タンパク質から算出したＮ末端側平均疎水性値の最大値のうち、前記抽出した最小値より値が大きいものの個数が最小となるような値である

ことを特徴とする請求項９または請求項１０に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１２】

前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＮ末端から予め定められた残基数の領域を特定し、当該Ｎ末端側の高疎水性領域に対応する領域以外のアミノ酸残基を抽出し、前記Ｎ末端側の高疎水性領域に対応する領域以外の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Ｎ末端外疎水性特性抽出必要数を用いて、連続する当該Ｎ末端外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Ｎ末端外平均疎水性値を、前記抽出した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複数算出するＮ末端外疎水性値算出部と、

前記Ｎ末端外疎水性値算出部が算出した複数のＮ末端外平均疎水性値のうちの最大値が、既知のＧＰＩアンカー型タンパク質におけるＮ末端外平均疎水性値の特性を示すＮ末端外疎水性閾値以上であるか否かを判定するＮ末端外疎水性判定部と、

を備え、

前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｎ末端外疎水性判定部が、前記Ｎ末端外疎水性値算出部の算出したＮ末端外平均疎水性値の最大値が前記Ｎ末端外疎水性閾値以上であると判定したアミノ酸配列情報に対して処理を実行する

ことを特徴とする請求項１から請求項１１の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１３】

前記Ｎ末端外疎水性閾値は、

予め既知の複数のＧＰＩアンカー型タンパク質に対して前記Ｎ末端外平均疎水性値の算出を行い、当該算出されたＮ末端外平均疎水性値の最大値の集合における最小値である

ことを特徴とする請求項１２に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１４】

前記Ｎ末端外疎水性特性抽出必要数は、

当該Ｎ末端外疎水性特性抽出必要数を用いて、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以外の領域のアミノ酸残基のそれぞれに対してＮ末端外平均疎水性値を算出し、前記既知のＧＰＩアンカー型タンパク質から算出したＮ末端外平均疎水性値の最大値の集合における最小値を抽出し、前記Ｎ末端外疎水性特性抽出必要数を用いて、既知の複数の非ＧＰＩアンカー型タンパク質における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域以外の領域のアミノ酸残基のそれぞれに対してＮ末端外平均疎水性値を算出した場合に、前記既知の非ＧＰＩアンカー型タンパク質から算出したＮ末端外平均疎水性値の最大値のうち、前記抽出した最小値より値が大きいものの個数が最小となるような値である

ことを特徴とする請求項１２または請求項１３項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１５】

前記既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域は、

既知のＧＰＩアンカー型タンパク質において、前記Ｎ末端側平均疎水性値が最大となる位置が含まれる領域である、

ことを特徴とする請求項９から請求項１１の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１６】

既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数のアミノ酸残基を特定し、前記Ｎ末端外疎水性値算出部が算出したＮ末端外平均疎水性値が最大となるアミノ酸残基の位置が当該特定した領域内にあるか否かを判定するＣ末端側最大疎水位置判定部

を備え、

前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｃ末端側最大疎水位置判定部が、前記Ｎ末端外疎水性値算出部の算出したＮ末端外平均疎水性値が最大となるアミノ酸残基の位置が前記既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域内にあると判定したアミノ酸配列情報に対して処理を実行する

ことを特徴とする請求項１２から請求項１４の何れか１項に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１７】

前記既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域は、

既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域以外の領域において、前記Ｎ末端外平均疎水性値が最大となる位置が含まれる領域である、

ことを特徴とする請求項１６に記載のＧＰＩアンカー型タンパク質の判定装置。

【請求項１８】

検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置を用いた判定方法であって、

前記ＧＰＩアンカー型タンパク質の判定装置の配列取得部は、前記検査対象タンパク質のアミノ酸配列情報を取得し、

前記ＧＰＩアンカー型タンパク質の判定装置の側鎖サイズ算出部は、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出し、

前記ＧＰＩアンカー型タンパク質の判定装置のスコア数値列生成部は、既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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ＧＰＩアンカー型タンパク質の判定装置の分類部は、既知のＧＰＩアンカー型タンパク質の前記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、既知の非ＧＰＩアンカー型タンパク質の前記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、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であるか否かを示す０以上１以下の期待値を出力し、

前記ＧＰＩアンカー型タンパク質の判定装置のＧＰＩアンカー型タンパク質判定部は、前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定する
ことを特徴とする判定方法。
【請求項１９】
検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置を、
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、
前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出する側鎖サイズ算出部、
既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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であるか否かを示す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前記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、既知の非ＧＰＩアンカー型タンパク質の前記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た分類部、
前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定するＧＰＩアンカー型タンパク質判定部
として機能させるための判定プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、検査対象タンパク質がＧＰＩ（ｇｌｙｃｏｓｙｌｐｈｏｓｐｈａｔｉｄｙｌｉｎｏｓｉｔｏｌ）アンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置、判定方法及び判定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。

ＧＰＩアンカー型タンパク質の特徴としては、Ｎ末端のシグナルペプチド及びＣ末端のプロペプチドの疎水性が高く、ωサイトの近隣には残基サイズの小さいアミノ酸が存在することが知られている。

【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、非特許文献１には、真核生物のＧＰＩアンカー型タンパク質を学習のデータセットとして、隠れマルコフモデルとサポートベクターマシン（ＳＶＭ）とを組み合わせた判定手法を用いて、検査対象タンパク質のアミノ酸配列情報から、検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する方法が記載されている。

また、非特許文献２には、原核生物及び真核生物のＧＰＩアンカー型タンパク質を学習のデータセットとして、ωサイト前後のアミノ酸配列におけるアミノ酸の性質及び出現頻度をスコア化し、ＧＰＩアンカー修飾部位を予測し、検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する方法が記載されている。

さらに、非特許文献３には、ニューラルネットワークの一種であるコホーネン自己組織化マップを用いて、検査対象の真核生物タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する方法が記載されている。

【先行技術文献】

【非特許文献】

【０００５】

【非特許文献１】Ｐｉｅｒｌｅｏｎｉら、「ＢＭＣ　Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ」、２００８年、ｖｏｌ．９、ｎｏ．３９２、ｐｐ．１－１１

【非特許文献２】Ｅｉｓｅｎｈａｂｅｒら、「Ｊｏｕｒｎａｌ　ｏｆ　Ｍｏｌｅｃｕｌａｒ　Ｂｉｏｌｏｇｙ」、１９９９年、ｖｏｌ．２９２、ｐｐ．７４１－７５８

【非特許文献３】Ｆｒａｎｋｈａｕｓｅｒら、「Ｂｉｏｉｎｆｏｒｍａｔｉｃｓ」、２００５年、ｖｏｌ．２１、ｎｏ．９、ｐｐ．１８４６－１８５２

【発明の概要】

【発明が解決しようとする課題】

【０００６】

上述したような従来のＧＰＩアンカー型タンパク質判定方法は、ＧＰＩアンカー型タンパク質のアミノ酸出現確率や疎水性値、分子量を解析手段（ニューラルネットワーク、ＳＶＭなど）への入力値として用いている。そのため、非ＧＰＩアンカー型タンパク質らしさについての判定がなされず、新規のＧＰＩアンカー型タンパク質を判定する感度及び選択性が十分ではない。そこで、より高い感度及び選択性で、検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定することへの要求がある。

本発明は、上記事情に鑑みてなされたものであって、高感度且つ高選択的に検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定することが可能なＧＰＩアンカー型タンパク質の判定装置、判定方法及び判定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】
【０００７】
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、検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置であって、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と、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出する側鎖サイズ算出部と、既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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前記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、既知の非ＧＰＩアンカー型タンパク質の前記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た分類部と、前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定するＧＰＩアンカー型タンパク質判定部と、を備えることを特徴とする。
【０００８】
また、本発明は、前記分類部は、ニューラルネットワークであり、前記スコア数値列生成部が生成するスコア数値列の要素数と同数のノードで構成される入力層と、複数のノードで構成される隠れ層と、１つのノードで構成される出力層とを少なくとも含む階層型の構造を有し、前記入力層の各ノードは、前記スコア数値列のうち自身に対応づけられた要素が示す値を前記隠れ層のノードのそれぞれに出力し、前記隠れ層の各ノードは、前記入力層の各ノードが出力する値を所定の伝達関数に代入し、得られた値を前記出力層のノードに出力し、前記出力層のノードは、前記隠れ層の各ノードが出力する値を所定の伝達関数に代入し、得られた値を期待値として出力することを特徴とする。
【０００９】
また、本発明において、前記分類部は、既知のＧＰＩアンカー型タンパク質の前記スコア数値列を入力した場合に、期待値として１を出力するように前記ノードの伝達関数の係数を変化させ、前記既知の非ＧＰＩアンカー型タンパク質の前記スコア数値列を入力した場合に、期待値として０を出力するように前記ノードの伝達関数の係数を変化させることで学習されたことを特徴とする。
【００１０】
また、本発明において、前記ノードのそれぞれは伝達関数としてシグモイド関数を用いることを特徴とする。
【００１１】
また、本発明において、前記側鎖サイズ特性抽出必要数は、当該側鎖サイズ特性抽出必要数を用いて、既知の複数のＧＰＩアンカー型タンパク質の小側鎖サイズ判定領域に対し

て平均側鎖サイズを算出した場合に、前記ＧＰＩアンカー型タンパク質から算出した平均側鎖サイズが最小となるアミノ酸残基のうち、当該アミノ酸残基のＣ末端側に隣接するアミノ酸残基がＧＰＩアンカー修飾部位であるものの個数が最大となるような値であることを特徴とする。

【００１２】

また、本発明において、前記小側鎖サイズ判定領域は、既知のＧＰＩアンカー型タンパク質の前記平均側鎖サイズが最小となる位置が含まれる領域である、ことを特徴とする。

【００１３】

また、本発明において、前記位置特異的スコアは、式（４）から算出されたものであることを特徴とする。

【００１４】

また、本発明において、前記所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度は、式（３）から算出されたものであることを特徴とする。

【００１５】

また、本発明は、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＮ末端から予め定められた残基数の領域を特定し、当該Ｎ末端側の高疎水性領域に対応する領域のアミノ酸残基を抽出し、前記Ｎ末端側の高疎水性領域に対応する領域の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Ｎ末端側疎水性特性抽出必要数を用いて、連続する当該Ｎ末端側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Ｎ末端側平均疎水性値を、前記抽出した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複数算出するＮ末端側疎水性値算出部と、前記Ｎ末端側疎水性値算出部が算出した複数のＮ末端側平均疎水性値のうちの最大値が、既知のＧＰＩアンカー型タンパク質におけるＮ末端側平均疎水性値の特性を示すＮ末端側疎水性閾値以上であるか否かを判定するＮ末端側疎水性判定部とを備え、前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｎ末端側疎水性判定部が、前記Ｎ末端側疎水性値算出部の算出したＮ末端側平均疎水性値の最大値が前記Ｎ末端側疎水性閾値以上であると判定したアミノ酸配列情報に対して処理を行うことを特徴とする。

【００１６】

また、本発明において、前記Ｎ末端側疎水性閾値は、予め既知の複数のＧＰＩアンカー型タンパク質に対して前記Ｎ末端側平均疎水性値の算出を行い、当該算出されたＮ末端側平均疎水性値の最大値の集合における最小値であることを特徴とする。

【００１７】

また、本発明において、前記Ｎ末端側疎水性特性抽出必要数は、当該Ｎ末端側疎水性特性抽出必要数を用いて、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のアミノ酸残基のそれぞれに対してＮ末端側平均疎水性値を算出し、前記既知のＧＰＩアンカー型タンパク質から算出したＮ末端側平均疎水性値の最大値の集合における最小値を抽出し、前記Ｎ末端側疎水性特性抽出必要数を用いて、既知の複数の非ＧＰＩアンカー型タンパク質における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域のアミノ酸残基のそれぞれに対してＮ末端側平均疎水性値を算出した場合に、前記既知の非ＧＰＩアンカー型タンパク質から算出したＮ末端側平均疎水性値の最大値のうち、前記抽出した最小値より値が大きいものの個数が最小となるような値であることを特徴とする。

【００１８】

また、本発明は、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＮ末端から予め定められた残基数の領域を特定し、当該Ｎ末端側の高疎水性領域に対応する領域以外のアミノ酸残基を抽出し、前記Ｎ末端側の高疎水性領域に対応する領域以外の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Ｎ末端外疎水性特性抽出必

要数を用いて、連続する当該Ｎ末端外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Ｎ末端外平均疎水性値を、前記抽出した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複数算出するＮ末端外疎水性値算出部と、前記Ｎ末端外疎水性値算出部が算出した複数のＮ末端外平均疎水性値のうちの最大値が、既知のＧＰＩアンカー型タンパク質におけるＮ末端外平均疎水性値の特性を示すＮ末端外疎水性閾値以上であるか否かを判定するＮ末端外疎水性判定部と、を備え、前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｎ末端外疎水性判定部が、前記Ｎ末端外疎水性値算出部の算出したＮ末端外平均疎水性値の最大値が前記Ｎ末端外疎水性閾値以上であると判定したアミノ酸配列情報に対して処理を実行することを特徴とする。

【００１９】

また、本発明において、前記Ｎ末端外疎水性閾値は、予め既知の複数のＧＰＩアンカー型タンパク質に対して前記Ｎ末端外平均疎水性値の算出を行い、当該算出されたＮ末端外平均疎水性値の最大値の集合における最小値であることを特徴とする。

【００２０】

また、本発明において、前記Ｎ末端外疎水性特性抽出必要数は、当該Ｎ末端外疎水性特性抽出必要数を用いて、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以外の領域のアミノ酸残基のそれぞれに対してＮ末端外平均疎水性値を算出し、前記既知のＧＰＩアンカー型タンパク質から算出したＮ末端外平均疎水性値の最大値の集合における最小値を抽出し、前記Ｎ末端外疎水性特性抽出必要数を用いて、既知の複数の非ＧＰＩアンカー型タンパク質における既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域以外の領域のアミノ酸残基のそれぞれに対してＮ末端外平均疎水性値を算出した場合に、前記既知の非ＧＰＩアンカー型タンパク質から算出したＮ末端外平均疎水性値の最大値のうち、前記抽出した最小値より値が大きいものの個数が最小となるような値であることを特徴とする。

【００２１】

また、本発明において、前記既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域は、既知のＧＰＩアンカー型タンパク質において、前記Ｎ末端側平均疎水性値が最大となる位置が含まれる領域である、ことを特徴とする。

【００２２】

また、本発明は、既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数のアミノ酸残基を特定し、前記Ｎ末端外疎水性値算出部が算出したＮ末端外平均疎水性値が最大となるアミノ酸残基の位置が当該特定した領域内にあるか否かを判定するＣ末端側最大疎水位置判定部を備え、前記側鎖サイズ算出部、前記スコア数値列生成部、前記分類部、前記ＧＰＩアンカー型タンパク質判定部は、前記Ｃ末端側最大疎水位置判定部が、前記Ｎ末端外疎水性値算出部の算出したＮ末端外平均疎水性値が最大となるアミノ酸残基の位置が前記既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域内にあると判定したアミノ酸配列情報に対して処理を実行することを特徴とする。

【００２３】

また、本発明において、前記既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域は、既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域以外の領域において、前記Ｎ末端外平均疎水性値が最大となる位置が含まれる領域である、ことを特徴とする。

【００２４】

また、本発明は、検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置を用いた判定方法であって、前記ＧＰＩアンカー型タンパク質の判定装置の配列取得部は、前記検査対象タンパク質のアミノ酸配列情報を取得し、前記ＧＰＩアンカー型タンパク質の判定装置の側鎖サイズ算出部は、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプ

ロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出し、前記ＧＰＩアンカー型タンパク質の判定装置のスコア数値列生成部は、既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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ＧＰＩアンカー型タンパク質の判定装置の分類部は、既知のＧＰＩアンカー型タンパク質の前記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、既知の非ＧＰＩアンカー型タンパク質の前記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、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であるか否かを示す０以上１以下の期待値を出力し、前記ＧＰＩアンカー型タンパク質の判定装置のＧＰＩアンカー型タンパク質判定部は、前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判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がＧＰＩアンカー型タンパク質であるか否かを判定するＧＰＩアンカー型タンパク質の判定装置を、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、前記配列取得部が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域として、前記アミノ酸配列情報のＣ末端から予め定められた残基数の領域を特定し、当該プロペプチド領域を含む領域のアミノ酸残基を抽出し、当該抽出したアミノ酸残基のそれぞれに対して、当該プロペプチド領域を含む領域のアミノ酸残基の側鎖サイズの平均化に用いる残基数である側鎖サイズ特性抽出必要数を用いて、連続する当該側鎖サイズ特性抽出必要数分のアミノ酸残基の各側鎖サイズ指標値の平均値である平均側鎖サイズを１残基ずつずらしながら複数算出する側鎖サイズ算出部、既知のＧＰＩアンカー型タンパク質の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と既知の非ＧＰＩアンカー型タンパク質の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とから求められる既知のＧＰＩアンカー型タンパク質のアミノ酸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取得し、当該位置特異的スコアに基づき、前記側鎖サイズ算出部が算出した平均側鎖サイズが最小となる位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所定の残基数のアミノ酸残基からなる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であるか否かを示す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前記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、既知の非ＧＰＩアンカー型タンパク質の前記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た分類部、前記分類部が出力した期待値が０．５未満であると判定した場合に、前記検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定するＧＰＩアンカー型タンパク質判定部として機能させるための判定プログラムである。

【発明の効果】

【００２６】
本発明によれば、ＰＳＳＭ（ｐｏｓｉｔｉｏｎ ｓｐｅｃｉｆｉｃ ｓｃｏｒｉｎｇ ｍａｔｒｉｘ；位置特異的スコアリングマトリックス）によって検査対象タンパク質のアミノ酸配列の各アミノ酸残基の位置特異的スコアを示すスコア数値列を生成する。そして、機械学習された分類部が当該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ことで検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する。これにより、本発明によるＧＰＩアンカー型タンパク質の判定装置は、高感度且つ高選択的に検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定することができる。

【図面の簡単な説明】

【００２７】
【図１】本発明の一実施形態によるＧＰＩアンカー型タンパク質判定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】側鎖サイズ指標値記憶部が記憶する情報を示す図である。
【図４】ＰＳＳＭ記憶部が記憶するＰＳＳＭを示す図である。
【図５】ＰＳＳＭ記憶部が記憶するＰＳＳＭを示す図である。
【図６】ＧＰＩアンカー型タンパク質判定装置１００の動作を示すフローチャートである。
【図７】ＧＰＩアンカー型タンパク質の疎水性プロファイルを示す第１のグラフである。
【図８】Ｎ末端側平均疎水性値の算出方法を示す図である。
【図９】既知のＧＰＩアンカー型タンパク質のＮ末端から３０残基以内におけるＮ末端側平均疎水性値の最大値の分布を示すグラフである。
【図１０】ＧＰＩアンカー型タンパク質の疎水性プロファイルを示す第２のグラフである。
【図１１】既知のＧＰＩアンカー型タンパク質及び既知の非ＧＰＩアンカー型タンパク質のＮ末端外平均疎水性値の最大値を示すグラフである。
【図１２】ＧＰＩアンカー型タンパク質の側鎖サイズのプロファイルを示すグラフである。
【図１３】アミノ酸配列の抽出方法を示す図である。
【図１４】位置特異的スコアの割り当て方法を示す図である。
【図１５】冗長性を排除したＧＰＩアンカー型タンパク質データセットに含まれる１１３のＳＷＩＳＳ－ＰＲＯＴ　エントリーネームを示す図である。
【図１６】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１７】本実施形態によるＧＰＩアンカー型タンパク質判定装置の判定精度を示す第１の表である。
【図１８】本実施形態によるＧＰＩアンカー型タンパク質判定装置の判定精度を示す第２の表である。
【図１９】基準位置を含む所定の範囲を基準位置から（－１２残基～＋１２残基）を（－１０残基～＋１２残基）に変更した場合の判定精度を示す表である。
【図２０】基準位置を含む所定の範囲を基準位置から（－１２残基～＋１２残基）を（－１２残基～＋９残基）に変更した場合の判定精度を示す表である。

【発明を実施するための形態】

【００２８】
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。
図１は、本発明の一実施形態によるＧＰＩアンカー型タンパク質判定装置の構成を示す概略ブロック図である。
ＧＰＩアンカー型タンパク質判定装置１００は、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、Ｎ末端側疎水性値算出部１０５、Ｎ末端側疎水性判定部１０６、Ｎ末端外疎水性値算出部１０７、Ｎ末端外疎水性判定

部１０８、Ｃ末端側最大疎水位置判定部１０９、側鎖サイズ指標値記憶部１１０、側鎖サイズ指標値特定部１１１、側鎖サイズ算出部１１２、ＰＳＳＭ記憶部１１３、スコア数値列生成部１１４、ニューラルネットワーク１１５（分類部）、ＧＰＩアンカー型タンパク質判定部１１６を備える。

【００２９】

配列記憶部１０１は、機能未知の哺乳類のタンパク質の完全長アミノ酸配列情報を記憶する。

配列取得部１０２は、配列記憶部１０１から検査対象となるタンパク質のアミノ酸配列情報を取得する。

疎水性指標値記憶部１０３は、アミノ酸残基に対応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の各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が記憶する疎水性指標値から特定し、アミノ酸残基毎の疎水性指標値を示す連続する数値列を生成する。

Ｎ末端側疎水性値算出部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、配列取得部１０２が取得したアミノ酸配列情報が示すＮ末端側の連続するアミノ酸残基の平均疎水性値（Ｎ末端側平均疎水性値）を算出する。

Ｎ末端側疎水性判定部１０６は、Ｎ末端側疎水性値算出部１０５が算出した平均疎水性値の最大値がＮ末端側疎水性閾値以上であるか否かを判定する。ここで、Ｎ末端側疎水性閾値とは、既知のＧＰＩアンカータンパク質におけるＮ末端側平均疎水性値の特性を示す閾値である。

【００３０】

Ｎ末端外疎水性値算出部１０７は、疎水性指標値特定部１０４が生成した数値列に基づいて、配列取得部１０２が取得したアミノ酸配列情報のうち、Ｎ末端側疎水性値算出部１０５が平均疎水性値を算出した範囲以外の連続するアミノ酸残基の平均疎水性値（Ｎ末端外平均疎水性値）を算出する。

Ｎ末端外疎水性判定部１０８は、Ｎ末端外疎水性値算出部１０７が算出した平均疎水性値の最大値がＮ末端外疎水性閾値以上であるか否かを判定する。ここで、Ｎ末端外疎水性閾値とは、既知のＧＰＩアンカー型タンパク質におけるＮ末端外平均疎水性値の特性を示す閾値である。

Ｃ末端側最大疎水位置判定部１０９は、Ｎ末端外疎水性値算出部１０７が算出した平均疎水性値が最大となるアミノ酸残基の位置が既知のＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域内にあるか否かを判定する。

【００３１】

側鎖サイズ指標値記憶部１１０は、アミノ酸残基に対応付けて当該アミノ酸残基の側鎖サイズ指標値を記憶する。

側鎖サイズ指標値特定部１１１は、配列取得部１０２が取得したアミノ酸配列の各アミノ酸残基それぞれの側鎖サイズ指標値を、側鎖サイズ指標値記憶部１１０が記憶する側鎖サイズ指標値から特定し、アミノ酸残基毎の側鎖サイズ指標値を示す連続する数値列を生成する。

側鎖サイズ算出部１１２は、側鎖サイズ指標値特定部１１１が生成した数値列に基づいて、配列取得部１０２が取得したアミノ酸配列情報が示すＣ末端側のアミノ酸残基の平均残基サイズを算出する。

ＰＳＳＭ記憶部１１３は、ＧＰＩアンカー型タンパク質のアミノ酸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保持するＰＳＳＭを記憶する。ここで、位置特異的スコアとは、ＧＰＩアンカー型タンパク質である可能性を示す値であり、当該値が大きいほどＧＰＩアンカー型タンパク質である可能性が高いことを表す。

スコア数値列生成部１１４は、ＰＳＳＭ記憶部１１３が記憶するＰＳＳＭに基づいて、側鎖サイズ算出部１１２が算出した側鎖のサイズの平均が最小となるアミノ酸残基の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、配列取得部１０２が取得した検査対象となるタンパク質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１１５は、スコア数値列生成部１１４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。なお、ニューラルネットワーク１１５は、予め、既知のＧＰＩアンカー型タンパク質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、既知の非ＧＰＩアンカー型タンパク質の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。

ＧＰＩアンカー型タンパク質判定部１１６は、配列取得部１０２が取得した検査対象となるタンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する。

【００３２】

図２は、疎水性指標値記憶部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。

【００３３】

図３は、側鎖サイズ指標値記憶部が記憶する情報を示す図である。

側鎖サイズ指標値記憶部１１０は、図３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側鎖のサイズを示す指標値を記憶している。なお、本実施形態では、側鎖サイズ指標値としてＤＡＷＤ７２０１０１（Ｄａｗｓｏｎ　Ｄ．Ｍ．，「Ｔｈｅ　Ｂｉｏｌｏｇｉｃａｌ　Ｇｅｎｅｔｉｃｓ　ｏｆ　Ｍａｎ」、Ａｃａｄｅｍｉｃ　Ｐｒｅｓｓ、１９７２年、ｐｐ．１－３８）で示される側鎖サイズ指標値を用いている。

【００３４】

図４及び図５は、ＰＳＳＭ記憶部が記憶するＰＳＳＭを示す図である。

ＰＳＳＭ記憶部１１３は、図４及び図５に示すように、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ＰＳＳＭを記憶している。図４及び図５では、アミノ酸残基位置の基準位置を０とし、負数側をＮ末端側、正数側をＣ末端側としている。なお、ＰＳＳＭの作成方法については、後述する。ここで、基準位置とは、ＧＰＩアンカー型タンパク質のＧＰＩアンカー修飾部位（ωサイト）のＣ末端側に隣接するアミノ酸残基の位置を示す。

【００３５】

そして、ＧＰＩアンカー型タンパク質判定装置１００において、配列取得部１０２は、検査対象タンパク質のアミノ酸配列情報を取得し、側鎖サイズ算出部１１２は、配列取得部１０２が取得したアミノ酸配列情報における既知のＧＰＩアンカー型タンパク質のプロペプチド領域に対応する領域のアミノ酸残基のそれぞれに対して、連続する側鎖サイズ特性抽出必要数分のアミノ酸残基の側鎖サイズ指標値の平均値である平均側鎖サイズを算出する。具体的には、スコア数値列生成部１１４は、ＰＳＳＭ記憶部１１３に記憶されている既知のＧＰＩアンカー型タンパク質のアミノ酸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ＰＳＳＭ記憶部１１３から取得し、当該位置特異的スコアに基づいて、側鎖サイズ算出部１１２が算出した平均側鎖サイズが最小となるアミノ酸残基の位置を基準位置とする、当該基準位置からＮ末端側及びＣ末端側に連続する所

定の残基数のアミノ酸残基からなる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。次に、ニューラルネットワーク１１５は、スコア数値列生成部１１４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。なお、ニューラルネットワーク１１５は、既知のＧＰＩアンカー型タンパク質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、既知の非ＧＰＩアンカー型タンパク質の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。

そして、ＧＰＩアンカー型タンパク質判定部１１６は、ニューラルネットワーク１１５が出力した期待値が０．５未満であると判定した場合に、検査対象タンパク質がＧＰＩアンカー型タンパク質でないと判定する。

これにより、ＧＰＩアンカー型タンパク質判定装置１００は、高感度且つ高選択的に検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する。

【００３６】

次に、ＧＰＩアンカー型タンパク質判定装置１００の動作を説明する。

図６は、ＧＰＩアンカー型タンパク質判定装置１００の動作を示すフローチャートである。

＜ステップＳ１：配列を取得＞

まず、使用者による動作開始指示により、ＧＰＩアンカー型タンパク質判定装置１００が動作を開始すると、配列取得部１０２は、配列記憶部１０１から検査対象となるタンパク質のアミノ酸配列情報を取得する。

【００３７】

＜ステップＳ２：疎水性指標値を特定＞

配列取得部１０２がアミノ酸配列情報を取得すると、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。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。

【００３８】

＜ステップＳ３：Ｎ末端側の疎水性指標値を抽出＞

ステップＳ２で、疎水性指標値特定部１０４が疎水性指標値を示す数値列を生成すると、Ｎ末端側疎水性値算出部１０５は、疎水性指標値特定部１０４が生成した数値列から、ＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域のアミノ酸残基を示す部分数値列を抽出する。

本実施形態では、ＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域として、Ｎ末端から３０残基以内のアミノ酸残基を用いる。Ｎ末端から３０残基以内のアミノ酸残基の領域は、既知の複数のＧＰＩアンカー型タンパク質のアミノ酸残基のそれぞれに対して、後述するステップＳ４と同様の処理によって平均疎水性値（Ｎ末端側平均疎水性値）を算出した場合に、当該算出した平均疎水性値が最大となるアミノ酸残基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。

【００３９】

図７は、ＧＰＩアンカー型タンパク質の疎水性プロファイルを示す第１のグラフである。

図７は、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０のＢＹ５５＿ＨＵＭＡＮ（１８１ａａ）エントリーに対して、後述するステップＳ４と同様の処理によって算出したＮ末端側平均疎水性値（１１残基平均の場合）を示すグラフである。ここで、横軸は、Ｎ末端側疎水性特性抽出必要数の連続するアミノ酸残基列の中央に位置するアミノ酸残基のＮ末端からの残基位置を示し、縦軸はＮ末端側平均疎水性値の値を示す。

図７に示すように、既知のＧＰＩアンカー型タンパク質のＮ末端側の領域は疎水性が高く、Ｎ末端から３０残基以内にＮ末端側平均疎水性値が最大となる位置が存在する。

【００４０】

＜ステップＳ４：Ｎ末端側平均疎水性値を算出＞

図８は、Ｎ末端側平均疎水性値の算出方法を示す図である。

Ｎ末端側疎水性値算出部１０５は、ステップＳ３でＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域のアミノ酸残基を示す部分数値列を抽出すると、当該部分数値列の連続するＮ末端側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Ｎ末端側平均疎水性値を、図８に示すように、１残基ずつずらしながら算出する。

ここで、Ｎ末端側疎水性特性抽出必要数の連続するアミノ酸残基列の中央のアミノ酸残基の位置がＮ末端からｒ残基目であるときのＮ末端側平均疎水性値は、式（１）を用いて算出できる。

【００４１】

【数１】

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

【００４２】

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、Ｎ末端側疎水性特性抽出必要数を示す。また、Ｈ（ｉ）は、Ｎ末端側疎水性特性抽出必要数の連続するアミノ酸残基列の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アミノ酸残基列のＮ末端側平均疎水性値は、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は、前後ｎ残基の平均値を算出できないため、Ｎ末端側平均疎水性値として例えばＮＵＬＬ値を代入しておくと良い。

【００４３】

本実施形態では、Ｎ末端側疎水性特性抽出必要数として１１残基を用いる。つまり、Ｎ末端側平均疎水性値として、Ｎ末端からｒ残基目のアミノ酸残基の前後５残基のアミノ酸残基の疎水性指標値の平均を算出する。ここで、Ｎ末端側疎水性特性抽出必要数を１１残基と決定する方法を説明する。

【００４４】

まず、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、すなわちＮ末端から３０残基以内のアミノ酸残基から、Ｎ末端側疎水性特性抽出必要数の候補となる範囲の平均疎水性値を、１残基ずつずらしながら算出する。次に、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値を抽出する。そして、抽出した最大値の集合における最小値を抽出する。

次に、既知の複数の非ＧＰＩアンカー型タンパク質における、既知のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域、すなわち既知の複数の非ＧＰＩアンカー型タンパク質のＮ末端から３０残基以内のアミノ酸残基から、Ｎ末端側疎水性特性抽出必要数の候補となる個数の連続するアミノ酸残基列の平均疎水性値を、１残基ずつずらしながら算出する。そして、非ＧＰＩアンカー型タンパク質から算出した平均疎水性値の最大値のうち、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値の集合から抽出した最小値より値が大きいものの個数を計数する。

この処理をＮ末端側疎水性特性抽出必要数の候補となる値を変えて実行し、非ＧＰＩアンカー型タンパク質から算出した平均疎水性値の最大値のうち、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値の集合から抽出した最小値より値が

大きいものの個数が最小となるようなＮ末端側疎水性特性抽出必要数の候補を、Ｎ末端側疎水性特性抽出必要数として決定する。

【００４５】

そして、本実施形態では、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０より取得した既知の哺乳類ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセット、及び既知の哺乳類非ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセットを用いて上述した方法を実行した結果、Ｎ末端側疎水性特性抽出必要数を１１残基として決定した。

【００４６】

＜ステップＳ５：Ｎ末端側平均疎水性値の最大値の判定＞

ステップＳ４で、Ｎ末端側疎水性値算出部１０５が、部分数値列の各疎水性指標値のＮ末端側平均疎水性値を算出すると、Ｎ末端側疎水性判定部１０６は、算出したＮ末端側平均疎水性値の最大値がＮ末端側疎水性閾値以上であるか否かを判定する。なお、Ｎ末端側疎水性閾値は、ＧＰＩアンカー型タンパク質におけるＮ末端側平均疎水性値の特性を示す閾値であり、本実施形態では、Ｎ末端側疎水性閾値として１．５０を用いる。１．５０という値は、予め既知の複数のＧＰＩアンカー型タンパク質に対してＮ末端側平均疎水性値の算出を行い、当該算出されたＮ末端側平均疎水性値の最大値の集合における最小値として算出された値である。

【００４７】

図９は、既知のＧＰＩアンカー型タンパク質のＮ末端から３０残基以内におけるＮ末端側平均疎水性値の最大値の分布を示すグラフである。ここで、横軸はＮ末端側平均疎水性値の最大値を示し、縦軸はＧＰＩアンカー型タンパク質が当該最大値をとる頻度を示す。

図９に示すように、既知のＧＰＩアンカー型タンパク質のＮ末端から３０残基以内のアミノ酸残基から算出されたＮ末端側平均疎水性値の最大値は、Ｎ末端側疎水性閾値である１．５０以上の値となる。従って、検査対象タンパク質のＮ末端から３０残基以内のアミノ酸残基から算出されたＮ末端側平均疎水性値の最大値が１．５０以上であれば、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が高く、当該最大値が１．５０未満であれば、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が低いと判定できる。

【００４８】

＜ステップＳ６：Ｎ末端外の疎水性指標値を抽出＞

ステップＳ５でＮ末端側疎水性判定部１０６が、算出したＮ末端側平均疎水性値の最大値がＮ末端側疎水性閾値以上であると判定した場合（ステップＳ５：ＹＥＳ）、Ｎ末端外疎水性値算出部１０７は、ステップＳ２で疎水性指標値特定部１０４が生成した数値列から、ステップＳ３でＮ末端側疎水性値算出部１０５が抽出した部分数値列以外の残りの部分数値列を抽出する。すなわち、疎水性指標値特定部１０４が生成した数値列から、Ｎ末端から３０残基以降のアミノ酸残基を示す部分数値列を抽出する。

【００４９】

＜ステップＳ７：Ｎ末端外平均疎水性値を算出＞

次に、Ｎ末端外疎水性値算出部１０７は当該部分数値列の連続するＮ末端外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Ｎ末端外平均疎水性値を、１残基ずつずらしながら算出する。

ここで、Ｎ末端外疎水性特性抽出必要数の連続するアミノ酸残基列の中央のアミノ酸残基の位置がＮ末端からｒ残基目であるときのＮ末端側平均疎水性値は、Ｎ末端側平均疎水性値と同様に、式（１）を用いて算出できる。なお、このとき、Ｃ末端からｎ残基以内のアミノ酸残基は、前後ｎ残基の平均値を算出できないため、Ｎ末端外平均疎水性値として例えばＮＵＬＬ値を代入しておくと良い。

【００５０】

本実施形態では、Ｎ末端外疎水性特性抽出必要数として１７残基を用いる。つまり、Ｎ末端外平均疎水性値として、Ｎ末端からｒ残基目のアミノ酸残基を中心とする前後８残基のアミノ酸残基の疎水性指標値の平均を算出する。ここで、Ｎ末端外疎水性特性抽出必要

数を１７残基と決定する方法を説明する。

【００５１】

まず、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以外の領域、すなわちＮ末端から３０残基以降のアミノ酸残基から、Ｎ末端外疎水性特性抽出必要数の候補となる個数の連続するアミノ酸残基列の平均疎水性値を、１残基ずつずらしながら算出する。次に、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値を抽出する。そして、抽出した最大値の集合における最小値を抽出する。

次に、既知の複数の非ＧＰＩアンカー型タンパク質における既知の複数のＧＰＩアンカー型タンパク質のＮ末端側の高疎水性領域に対応する領域以外の領域、すなわち既知の複数の非ＧＰＩアンカー型タンパク質のＮ末端から３０残基以降のアミノ酸残基から、Ｎ末端外疎水性特性抽出必要数の候補となる範囲の平均疎水性値を、１残基ずつずらしながら算出する。そして、非ＧＰＩアンカー型タンパク質から算出した平均疎水性値の最大値のうち、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値の集合から抽出した最小値より値が大きいものの個数を計数する。

この処理をＮ末端外疎水性特性抽出必要数の候補となる値を変えて実行し、非ＧＰＩアンカー型タンパク質から算出した平均疎水性値の最大値のうち、既知の複数のＧＰＩアンカー型タンパク質のそれぞれの平均疎水性値の最大値の集合から抽出した最小値より値が大きいものの個数が最小となるＮ末端外疎水性特性抽出必要数の候補を、Ｎ末端外疎水性特性抽出必要数として決定する。

【００５２】

図１０は、ＧＰＩアンカー型タンパク質の疎水性プロファイルを示す第２のグラフである。

図１０は、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０のＢＹ５５＿ＨＵＭＡＮ（１８１ａ）エントリーに対して、ステップＳ７と同様の処理によって算出したＮ末端外平均疎水性値（１７残基平均の場合）を示すグラフである。ここで、横軸は、Ｎ末端外疎水性特性抽出必要数の連続するアミノ酸残基列の中央に位置するアミノ酸残基のＮ末端からの残基位置を示し、縦軸はＮ末端外平均疎水性値の値を示す。

図１０に示すように、既知のＧＰＩアンカー型タンパク質のＣ末端側の領域は、Ｎ末端からの３０残基に次いで疎水性が高い。

【００５３】

そして、本実施形態では、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０より取得した既知の哺乳類ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセット、及び既知の哺乳類非ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセットを用いて上述した方法を実行した結果、Ｎ末端外疎水性特性抽出必要数を１７残基として決定した。

【００５４】

＜ステップＳ８：Ｎ末端外平均疎水性値の最大値の判定＞

ステップＳ７で、Ｎ末端外疎水性値算出部１０７が、部分数値列の連続するＮ末端外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Ｎ末端外平均疎水性値を、１残基ずつずらしながら算出すると、Ｎ末端外疎水性判定部１０８は、算出したＮ末端外平均疎水性値の最大値がＮ末端外疎水性閾値以上であるか否かを判定する。なお、Ｎ末端外疎水性閾値は、既知のＧＰＩアンカー型タンパク質のＮ末端外平均疎水性値の特性を示す閾値であり、本実施形態では、Ｎ末端外疎水性閾値として１．３８を用いている。

１．３８という値は、予め既知の複数のＧＰＩアンカー型タンパク質に対してＮ末端外平均疎水性値の算出を行い、当該算出されたＮ末端側平均疎水性値の最大値の集合における最小値として算出された値である。

【００５５】

図１１は、既知のＧＰＩアンカー型タンパク質及び既知の非ＧＰＩアンカー型タンパク質のＮ末端外平均疎水性値の最大値を示すグラフである。ここで、横軸は、Ｎ末端外疎水性特性抽出必要数の連続するアミノ酸残基列の中央に位置するアミノ酸残基のＣ末端からの残基位置を示し、縦軸はＮ末端外平均疎水性値の値を示す。

図１１に示すように、既知のＧＰＩアンカー型タンパク質のＮ末端から３０残基以降のアミノ酸残基から算出されたＮ末端外平均疎水性値の最大値は、Ｎ末端外疎水性閾値である１．３８以上の値となる。従って、検査対象タンパク質のＮ末端から３０残基以降のアミノ酸残基から算出されたＮ末端外平均疎水性値の最大値が１．３８以上であれば、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が高く、当該最大値が１．３８未満であれば、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が低いと判定できる。

【００５６】

＜ステップＳ９：Ｎ末端外平均疎水性値が最大となるアミノ酸残基位置の判定＞

Ｎ末端外疎水性判定部１０８が、算出したＮ末端外平均疎水性値の最大値がＮ末端外疎水性閾値以上であると判定した場合（ステップＳ８：ＹＥＳ）、Ｃ末端側最大疎水位置判定部１０９は、ステップＳ７で算出したＮ末端外平均疎水性値が最大となるアミノ酸残基の位置が、ＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域内にあるか否かを判定する。

本実施形態では、ＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域として、Ｃ末端から１４残基以内のアミノ酸残基を用いる。Ｃ末端から１４残基以内のアミノ酸残基という領域は、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以外の領域、すなわちＮ末端から３０残基以降のアミノ酸残基のそれぞれに対してＮ末端外平均疎水性値を算出した場合に、当該算出したＮ末端外平均疎水性値が最大となる連続するアミノ酸残基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。

【００５７】

図１１に示すように、既知のＧＰＩアンカー型タンパク質のＮ末端から３０残基以降のアミノ酸残基から算出されたＮ末端外平均疎水性値が最大となるアミノ酸残基列の中央に位置するアミノ酸残基は、Ｃ末端側の高疎水性領域内に存在する。従って、検査対象タンパク質のＮ末端から３０残基以降のアミノ酸残基から算出されたＮ末端外平均疎水性値が最大となるアミノ酸残基列の中央に位置するアミノ酸残基がＧＰＩアンカー型タンパク質のＣ末端側の高疎水性領域に対応する領域内に存在すれば、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が高く、当該領域内に存在しなければ、検査対象タンパク質がＧＰＩアンカー型タンパク質である可能性が低いと判定できる。

つまり、図１１における網掛け矩形の範囲が、Ｎ末端外疎水性閾値及びＣ末端側の高疎水性領域の条件を満たす範囲を示し、当該範囲内に含まれる非ＧＰＩアンカー型タンパク質の個数が最小となるよう、Ｎ末端外疎水性閾値及びＣ末端側の高疎水性領域に対応する領域とを決定している。

【００５８】

＜ステップＳ１０：小側鎖サイズ判定領域の残基を抽出＞

Ｃ末端側最大疎水位置判定部１０９が、Ｎ末端外平均疎水性値が最大となるアミノ酸残基の位置がＣ末端から１４残基以内の位置であると判定した場合（ステップＳ９：ＹＥＳ）側鎖サイズ指標値特定部１１１は、ステップＳ１で配列取得部１０２が取得したアミノ酸配列情報から、小側鎖サイズ判定領域のアミノ酸残基に相当する部分配列を抽出する。ここで、小側鎖サイズ判定領域とは、既知のＧＰＩアンカー型タンパク質のプロペプチド領域を含む領域であり、本実施形態では、Ｃ末端から３０残基以内のアミノ酸残基を用いる。Ｃ末端から３０残基以内のアミノ酸残基という領域は、既知のＧＰＩアンカー型タンパク質において、後述するステップＳ１２と同様の処理によって平均側鎖サイズを算出した場合に、当該算出した平均側鎖サイズが最小となるアミノ酸残基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。

【００５９】

＜ステップＳ１１：側鎖サイズ指標値を特定＞

側鎖サイズ指標値特定部１１１は、ステップＳ１０で小側鎖サイズ判定領域のアミノ酸残基に相当する部分配列を抽出すると、側鎖サイズ指標値記憶部１１０を参照して、抽出した部分配列が示す各アミノ酸残基に側鎖サイズ指標値を割り当てた数値列を生成する（

ステップＳ１１）。例えば、配列取得部１０２が取得したアミノ酸配列情報が、「ＭＬＥＰＧＲＧＣＣ……」という配列を示す場合、側鎖サイズ指標値特定部１１１は、側鎖サイズ指標値記憶部１１０が記憶する図３に示す指標値より「6、5.5、5.5、5、5.5、0.5、7.5、0.5、3、3……」という数値列を生成する。

【００６０】

＜ステップＳ１２：平均側鎖サイズを算出＞

ステップＳ１１で、側鎖サイズ指標値特定部１１１が側鎖サイズ指標値を示す数値列を生成すると、側鎖サイズ算出部１１２は、側鎖サイズ指標値特定部１１１が生成した数値列の連続する側鎖サイズ特性抽出必要数分の各側鎖サイズ指標値の平均である平均側鎖サイズを、１残基ずつずらしながら算出する。

ここで、平均側鎖サイズ特性抽出必要分の連続するアミノ酸残基列の中央のアミノ酸残基の位置がＮ末端からｒ残基目であるときの平均側鎖サイズは、式（２）を用いて算出できる。

【００６１】

【数２】

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

【００６２】

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、側鎖サイズ特性抽出必要数を示す。また、Ｖ（ｉ）はＮ末端からｉ残基目に存在するアミノ酸残基の側鎖サイズ指標値を示す。

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アミノ酸残基列の平均側鎖サイズは、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側鎖サイズ指標値の平均となる。なお、このとき、Ｃ末端からｎ残基以内のアミノ酸残基は、前後ｎ残基の平均値を算出できないため、平均側鎖サイズとして例えばＮＵＬＬ値を代入しておくと良い。

【００６３】

本実施形態では、側鎖サイズ特性抽出必要数として３残基を用いる。つまり、Ｎ末端側平均疎水性値として、Ｎ末端からｒ残基目のアミノ酸残基に隣接するアミノ酸残基の疎水性指標値の平均を算出する。ここで、側鎖サイズ特性抽出必要数を３残基と決定する方法を説明する。

【００６４】

まず、既知の複数のＧＰＩアンカー型タンパク質の小側鎖サイズ判定領域、すなわちＣ末端から３０残基以内のアミノ酸残基から、側鎖サイズ特性抽出必要数の候補となる範囲の平均疎水性値を、１残基ずつずらしながら算出する。次に、既知の複数のＧＰＩアンカー型タンパク質のそれぞれから、平均側鎖サイズが最小となるアミノ酸残基を特定する。そして、当該抽出したアミノ酸残基のＣ末端側に隣接するアミノ酸残基がＧＰＩアンカー修飾部位（ωサイト）であるものの個数を計数する。

この処理をＮ末端側疎水性特性抽出必要数の候補となる値を変えて実行し、全ＧＰＩアンカー型タンパク質のうち、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がＧＰＩアンカー修飾部位であるものの個数が最大となる側鎖サイズ特性抽出必要数の候補を、側鎖サイズ特性抽出必要数として決定する。

【００６５】

そして、本実施形態では、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０より取得した既知の哺乳類ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセット、及び既知の哺乳類非ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセットを用いて上述した方法を実行した結果、Ｎ末端側疎水性特性抽出必要数を３残基として決定した。

【００６６】

図１２は、ＧＰＩアンカー型タンパク質の側鎖サイズのプロファイルを示すグラフである。

図１２は、ＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０のＢＹ５５＿ＨＵＭＡＮ（１８１ａａ）エントリーに対して、ステップＳ１２と同様の処理によって算出した平均側鎖サイズを示すグラフである。ここで、横軸は、平均側鎖サイズのアミノ酸残基列の中央に位置するアミノ酸残基のＣ末端からの残基位置を示し、縦軸は平均側鎖サイズの値を示す。

図１２に示すように、既知のＧＰＩアンカー型タンパク質のＧＰＩアンカー修飾部位は、平均側鎖サイズが最小となるアミノ酸残基のＣ末端側に隣接している。

【００６７】

＜ステップＳ１３：所定の領域のアミノ酸残基を抽出＞

図１３は、アミノ酸配列の抽出方法を示す図である。

ステップＳ１２で、側鎖サイズ算出部１１２が平均側鎖サイズを算出すると、スコア数値列生成部１１４は、図１３（１）に示すように、側鎖サイズ算出部１１２が算出した平均側鎖サイズが最小となるアミノ酸残基の位置を基準位置として決定する。次に、スコア数値列生成部１１４は、図１３（２）に示すように、当該基準位置を含む所定の領域におけるアミノ酸残基を、ステップＳ１で配列取得部１０２が取得したアミノ酸配列情報から抽出する。

本実施形態では、当該所定の領域として、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。

【００６８】

＜ステップＳ１４：位置特異的スコアを割り当てる＞

図１４は、位置特異的スコアの割り当て方法を示す図である。

次に、スコア数値列生成部１１４は、ＰＳＳＭ記憶部１１３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。例えば、抽出した所定の範囲のアミノ酸残基が、図１４に示すように「ＣＱＮＡ……Ｓ」という配列を示す場合、スコア数値列生成部１１４は、図４及び図５に示すＰＳＳＭを参照して、「0.21、-0.54、2.69、-0.77、……、1.13」という数値列を生成する。

【００６９】

ここで、ステップＳ１４で用いるＰＳＳＭの作成方法を説明する。

まず、既知の哺乳類ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセット、及び既知の哺乳類非ＧＰＩアンカー型タンパク質の完全長アミノ酸配列データセットを、取得する。本実施形態では、これらのデータセットをＳＷＩＳＳ－ＰＲＯＴ　ｖｅｒ５４．０より取得した。また、ＧＰＩアンカー型タンパク質のデータセットについては、当該アミノ酸配列から翻訳されるＧＰＩアンカー型タンパク質としての特性が実証されていないもの、明らかに完全長ではないもの等を除外した。その結果、ＧＰＩアンカー型タンパク質のエントリー数は３９１であり、非ＧＰＩアンカー型タンパク質のエントリー数は４８９８３であった。

【００７０】

データセットを取得すると、次に、データセットの各エントリーについて、疎水性のスクリーニングを行う。

まず、上述した式（１）及び図２に示す疎水性指標値を用いて、Ｎ末端側疎水性特性抽出必要数を１１残基に設定して（すなわち、式（１）においてｎ＝５に設定して）各エントリーのＮ末端平均疎水性値を算出し、Ｎ末端から３０残基以内の領域における最大のＮ末端側平均疎水性値が１．５０以上のものを抽出する。次に、抽出されたデータセット中の各エントリーの平均疎水性値を、前記式（１）及び図２に示す疎水性指標値を用いて、Ｎ末端外疎水性特性抽出必要数を１７残基に設定して（すなわち、式（１）においてｎ＝８に設定して）算出し、Ｎ末端から３０残基を除く全領域における最大のＮ末端外平均疎水性値が１．３８であり、且つ、該最大のＮ末端外平均疎水性値を示す残基位置がＣ末端から１４残基以内であるものを抽出する。この結果、実際は完全長でないエントリーや、

タンパク質としての発現が推定であるエントリーは排除されることとなる。本実施形態では、疎水性スクリーニング後のＧＰＩアンカー型タンパク質データセットのエントリー数は１２１であり、非ＧＰＩアンカー型タンパク質データセットのエントリー数は２１８であった。

【００７１】

次いで、疎水性スクリーニングにより抽出されたデータセットに含まれる同一アミノ酸配列を有するエントリーを除き、冗長性を排除する。この結果、本実施形態では、ＧＰＩアンカー型タンパク質データセットのエントリー数は１１３であり、非ＧＰＩアンカー型タンパク質データセットのエントリー数は２１０であった。冗長性を排除したＧＰＩアンカー型タンパク質データセットに含まれる１１３のＳＷＩＳＳ－ＰＲＯＴ　エントリーネームを図１５に示す。

【００７２】

上記により得られた各データセット中の各エントリーのＣ末端から３０アミノ酸残基までの平均側鎖サイズを、上述した式（２）及び図３に示す側鎖サイズ指標値を用いて、側鎖サイズ特性抽出必要数を３に設定して（すなわち、式（２）においてｎ＝１に設定して）算出する。

そして、データセットのうちＧＰＩアンカー型タンパク質の各エントリーの、平均側鎖サイズが最小となるアミノ酸残基の位置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）におけるアミノ酸残基から、式（３）を用いて既知のＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を算出する。

【００７３】

【数３】

$$\frac{n_{ip}+\varepsilon^{\frac{1}{s}}}{\sum_{i=1}^{s} n_{ip}+\varepsilon} \qquad \cdots(3)$$

【００７４】

但し、n_{ip}は、種類ｉのアミノ酸残基が位置ｐに存在する既知のＧＰＩアンカー型タンパク質の個数を示す。また、εは算出する出現頻度の調整値を示し、本実施形態では１を用いている。また、ｓは、アミノ酸残基の種類数を示す。

これにより、データセットの全てのエントリーにおいて位置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、データセットのうち非ＧＰＩアンカー型タンパク質の各エントリーの、平均側鎖サイズが最小となるアミノ酸残基の位置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（３）を用いて既知の非ＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を算出する。

【００７５】

既知のＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度、及び既知の非ＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を算出すると、次に、式（４）を用いて、アミノ酸残基の位置ｐにおけるアミノ酸残基の種類ｉの位置特異的スコアを算出する。

【００７６】

【数４】

$$\ln \frac{f_{ip}^{positive}}{f_{ip}^{negative}} \qquad \cdots (4)$$

【００７７】

但し、$f_{ip}^{positive}$は、既知のＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を示す。また、$f_{ip}^{negative}$は、既知の非ＧＰＩアンカー型タンパク質の所定の領域内の位置ｐに存在するアミノ酸残基の種類ｉの出現頻度を示す。つまり、位置特異的スコアは、所定の範囲におけるあるアミノ酸残基の位置におけるアミノ酸残基の種類の、ＧＰＩアンカー型タンパク質における出現度合いを示している。

このように算出された位置特異的スコアを要素とする２５（所定の領域内のアミノ酸残基数）×２０（アミノ酸残基の種類数）の行列をＰＳＳＭとして生成し、ＰＳＳＭ記憶部１１３に格納しておく。これにより、図４及び図５に示すＰＳＳＭを生成することができる。

【００７８】

＜ステップＳ１５：ニューラルネットワークによる期待値出力＞

ステップＳ１４でスコア数値列生成部１１４がスコア数値列を生成すると、ニューラルネットワーク１１５は、当該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。なお、ＰＳＳＭから得られた複数の位置特異的スコアは、従来、その平均値の高低によって検査対象タンパク質が目的タンパク質であるか否かを判定するために用いられている。本発明の骨子は、スコアの算出に用いられていた複数の位置特異的スコアをニューラルネットワーク１１５の入力値とした点にある。

【００７９】

ここで、ニューラルネットワーク１１５の処理について詳細に説明する。

図１６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１１５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１１４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。

【００８０】

ノード$Ｎ_1$のそれぞれは、スコア数値列生成部１１４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（５）に、ノード$Ｎ_3$の伝達関数を式（６）に示す。

【００８１】

【数５】

$$f\left(\sum_{i=1}^{n} w_i x_i - \theta\right) \qquad \cdots (5)$$

【数６】

$$f\left(\sum_{j=1}^{m} w_j x_j - \theta\right) \qquad \cdots (6)$$

【００８２】
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（７）に示す関数である。

【００８３】
【数７】

$$f(x) = \frac{1}{1+e^{-x}} \qquad \cdots (7)$$

【００８４】
また、ニューラルネットワーク１１５は、既知のＧＰＩアンカー型タンパク質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、既知の非ＧＰＩアンカー型タンパク質の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１１５の学習方法を説明する。
【００８５】
まず、ＰＳＳＭの作成に用いたＧＰＩアンカー型タンパク質データセット及び非ＧＰＩアンカー型タンパク質データセットを読み出す。次に、当該データセットの各エントリーから、平均側鎖サイズが最小となるアミノ酸残基の位置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）におけるアミノ酸残基のそれぞれに対して、ＰＳＳＭ記憶部１１３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。
【００８６】
次に、生成したスコア数値列をニューラルネットワーク１１５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。
【００８７】
他方、ニューラルネットワーク１１５のノードＮ$_3$は、教師データを入力する。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、ＧＰＩアンカー型タンパク質のスコア数値列を入力した場合、教師データは１であり、非ＧＰＩアンカー型タンパク質の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１１５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ＧＰＩアンカー型タンパク質データセット及び非ＧＰＩアンカー型タンパク質データセットのそれぞれのエントリーに対して実行する。これにより、ニューラルネットワーク１１５は、既知のＧＰＩアンカー型タンパク質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、既知の非ＧＰＩアンカー型タンパク質の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。

【００８８】

＜ステップＳ１６：スコアの判定＞

ステップＳ１５でニューラルネットワーク１１５が期待値を出力すると、ＧＰＩアンカー型タンパク質判定部１１６は、出力した期待値が０．５以上であるか否かを判定する。つまり、ＧＰＩアンカー型タンパク質判定部１１６は、ニューラルネットワーク１１５が出力した期待値が、ＧＰＩアンカー型タンパク質を示す１と非ＧＰＩアンカー型タンパク質を示す０との何れに近いかを判定する。

【００８９】

＜ステップＳ１７：ＧＰＩアンカー型タンパク質と判定＞

ＧＰＩアンカー型タンパク質判定部１１６は、ステップＳ１６でニューラルネットワーク１１５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Ｓ１６：ＹＥＳ）、ステップＳ１で配列取得部１０２が取得したアミノ酸配列情報が、ＧＰＩアンカー型タンパク質のものであると判定する。

【００９０】

＜ステップＳ１８：非ＧＰＩアンカー型タンパク質と判定＞

他方、ステップＳ５でＮ末端側疎水性判定部１０６が、算出したＮ末端側平均疎水性値の最大値がＮ末端側疎水性閾値未満であると判定した場合（ステップＳ５：ＮＯ）、ステップＳ８でＮ末端外疎水性判定部１０８が、算出したＮ末端外平均疎水性値の最大値がＮ末端外疎水性閾値未満であると判定した場合（ステップＳ８：ＮＯ）、ステップＳ９でＣ末端側最大疎水位置判定部１０９が、Ｎ末端外平均疎水性値が最大となるアミノ酸残基の位置がＣ末端側の高疎水性領域に対応する領域内にないと判定した場合（ステップＳ９：ＮＯ）、またはステップＳ１６でニューラルネットワーク１１５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Ｓ１６：ＮＯ）、ＧＰＩアンカー型タンパク質判定部１１６は、ステップＳ１で配列取得部１０２が取得したアミノ酸配列情報が、非ＧＰＩアンカー型タンパク質のものであると判定する。

【００９１】

上述した動作により、ＧＰＩアンカー型タンパク質判定装置１００は、高感度且つ高選択的に検査対象タンパク質がＧＰＩアンカー型タンパク質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分割交差検定法及び自己無撞着な手法について詳述する。

【００９２】

４分割交差検定法による判定精度とは、以下の処理により算出した判定精度である。

まず、既知のＧＰＩアンカー型タンパク質及び既知の非ＧＰＩアンカー型タンパク質のデータセットを４等分する。次に、分割したデータセットのうち３つの部分データセットを用いてＰＳＳＭを生成する。また、分割したデータセットのうち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１１５の学習を行う。次に、３つの部分データセットを用いて生成した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ー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、ＰＳＳＭを生成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。

【００９３】

自己無撞着な手法による判定精度とは、以下の処理により算出した判定精度である。

まず、上述したスコア判定閾値の決定方法と同様に、既知のＧＰＩアンカー型タンパク質及び既知の非ＧＰＩアンカー型タンパク質のデータセットを用いてＰＳＳＭを生成する。また、当該データセットを用いてニューラルネットワーク１１５の学習を行う。次に、当該ＰＳＳＭを用いて、ＰＳＳＭの生成に用いたデータセットの各エントリーのスコアを算出する。そして、当該算出したスコアに基づいてデータセット全体に対する判定精度を算出する。但し、本実施形態では、ニューラルネットワーク１１５が、既知のＧＰＩアンカー型タンパク質のスコア数値列を入力とした場合に、期待値として必ず１を出力し、既知の非ＧＰＩアンカー型タンパク質のスコア数値列を入力した場合に、期待値として必ず０を出力するように学習されている。そのため、自己無撞着な手法によって算出された感度、選択性、成功率は、すべて１００％となる。

【００９４】

４分割交差検定法について、図１７～図２０を用いて、さらに具体的に説明する。

図１７は、本実施形態によるＧＰＩアンカー型タンパク質判定装置の判定精度を示す第１の表である。

図１７では、ＧＰＩアンカー型タンパク質判定装置１００がＧＰＩアンカー型タンパク質であると判定した検査対象タンパク質の判定精度、及び非ＧＰＩアンカー型タンパク質であると判定した検査対象タンパク質の判定精度を示している。また、図１７に示す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にあたり、４分割交差検定法を用いた。

【００９５】

図１７に示すように、本実施形態による、ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、感度が９１．５％、選択性が９１．５％、成功率が０．９１５であった。また、非ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、感度が９８．２％、選択性が９３．１％、成功率が０．９５６であった。なお、図１７に示す判定制度は、１００回試行のうち、成功率が最高値のときのものである。

【００９６】

図１８は、本実施形態によるＧＰＩアンカー型タンパク質判定装置の判定精度を示す第２の表である。

図１７では、１００回試行のうち、成功率が最高値のときの判別精度を示したが、図１８では、１００回試行のうち、成功率上位１０％の平均精度を示す。

図１８に示すように、本実施形態による、ＧＰＩアンカー型タンパク質の４分割交差検定法による成功率上位１０％の平均精度は、感度が９１．４％、選択性が９０．２％、成功率が０．９０７であった。また、非ＧＰＩアンカー型タンパク質の４分割交差検定法による成功率上位１０％の平均精度は、感度が９４．８％、選択性が９１．３％、成功率が０．９４９であった。このように、本実施形態によれば、成功率が最高の場合に限らず、平均的に高い判定精度を得ることができることが分かる。

【００９７】

以下に、基準位置を含む所定の範囲を変化させてＧＰＩアンカー型タンパク質の判定を行った場合の判定精度を示す。

【００９８】

図１９は、基準位置を含む所定の範囲を基準位置から（－１２残基～＋１２残基）を（－１０残基～＋１２残基）に変更した場合の判定精度を示す表である。

図１９に示すように、所定の範囲を、基準位置からＮ末端側に１０残基、Ｃ末端側に１２残基の範囲とした場合の、ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、成功率が最高の場合、感度が９０．０％、選択性が９２．３％、成功率が０．９１１であった。また、１００回試行のうち成功率上位１０％の平均精度は、感度が９０．５％、選択性が９０．０％、成功率が０．９０１であった。

他方、非ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、成功率が最高の場合、感度が９５．５％、選択性が９４．７％、成功率が０．９５１であった。また、１００回試行のうち成功率上位１０％の平均精度は、感度が９４．５％、選択性が９４．９％、成功率が０．９４７であった。

図１９に示す本実施形態による判定精度（基準位置を含む所定の範囲を、基準位置からＮ末端側に１０残基、Ｃ末端側に１２残基の範囲とした場合の判定精度）を、図１７に示す本実施形態による判定精度（基準位置を含む所定の範囲を、基準位置からＮ末端側に１２残基、Ｃ末端側に１２残基の範囲とした場合の判定精度）と比較すると、ＧＰＩアンカー型タンパク質と非ＧＰＩ型タンパク質とで図１７に示す本実施形態による判定精度の方が感度と成功率が高いことが分かる。

【００９９】

図２０は、基準位置を含む所定の範囲を基準位置から（－１２残基～＋１２残基）を（－１２残基～＋９残基）に変更した場合の判定精度を示す表である。

図２０に示すように、所定の範囲を、基準位置からＮ末端側に１２残基、Ｃ末端側に９残基の範囲とした場合の、ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、成功率が最高の場合、感度が９２．９％、選択性が９０．５％、成功率が０．９１６であった。また、１００回試行のうち成功率上位１０％の平均精度は、感度が９０．８％、選択性が８９．４％、成功率が０．９００であった。

他方、非ＧＰＩアンカー型タンパク質の４分割交差検定法による判定精度は、成功率が最高の場合、感度が９４．９％、選択性が９６．２％、成功率が０．９５５であった。また、１００回試行のうち成功率上位１０％の平均精度は、感度が９４．２％、選択性が９５．０％、成功率が０．９４６であった。

図２０に示す本実施形態による判定精度（基準位置を含む所定の範囲を、基準位置からＮ末端側に１２残基、Ｃ末端側に９残基の範囲とした場合の判定精度）を、図１７に示す本実施形態による判定精度（基準位置を含む所定の範囲を、基準位置からＮ末端側に１２残基、Ｃ末端側に１２残基の範囲とした場合の判定精度）と比較すると、ＧＰＩアンカー型タンパク質では図２０に示す本実施形態による判定精度の方が感度と成功率が高いことが分かる。

【０１００】

このように、本実施形態によれば、ＧＰＩアンカー型タンパク質判定装置１００は、ＰＳＳＭによって検査対象タンパク質のアミノ酸配列の各アミノ酸残基の位置特異的スコアを示すスコア数値列を生成する。そして、ニューラルネットワーク１１５が当該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ことで検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定する。ＰＳＳＭは、既知のＧＰＩアンカー型タンパク質のアミノ酸出現頻度と既知の非ＧＰＩアンカー型タンパク質のアミノ酸出現頻度とを用いて生成されるため、ＰＳＳＭから生成されたスコア数値列は、ＧＰＩアンカー型タンパク質らしさのみならず非ＧＰＩアンカー型タンパク質らしさをも示すこととなる。これにより、ＧＰＩアンカー型タンパク質判定装置１００は、高感度且つ高選択的に検査対象タンパク質がＧＰＩアンカー型タンパク質であるか否かを判定することができる。

【０１０１】

また、本実施形態によれば、Ｎ末端側疎水性判定部１０６、Ｎ末端外疎水性判定部１０８、及びＣ末端側最大疎水位置判定部１０９による判定処理をした後に、ニューラルネットワーク１１５による期待値の算出を行う。これにより、ニューラルネットワーク１１５の処理対象となるアミノ酸配列情報の量を減らすことができ、ニューラルネットワーク１

１５による期待値算出処理の計算量が多い場合にも、処理の高速化を図ることができる。

【０１０２】

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、本実施形態では、タンパク質の完全長アミノ酸配列情報を検査対象として判定を行ったが、これに限られず、完全長塩基配列情報を検査対象として判定を行っても良い。但し、この場合、ステップＳ１で配列取得部１０２が完全長塩基配列情報を取得した後、図示しない翻訳処理部が、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い、当該アミノ酸配列情報を用いてステップＳ２以降の処理を行う。

【０１０３】

また、本実施形態では、期待値を算出する分類部としてニューラルネットワーク１１５を用いる場合を説明したが、これに限られず、例えば、サポートベクターマシンや、ベイジアンネットワークなど、分類部として他の解析手法を用いても良い。

【０１０４】

また、本実施形態では、ニューラルネットワーク１１５が入力層S_1、隠れ層S_2、出力層S_3の３層構造である場合を説明したが、これに限られず、ニューラルネットワーク１１５が複数の隠れ層を有する４層以上の構造を有していても良い。但し、隠れ層の数が増えると、学習時に、最適解（期待値と教師データとの誤差が最小値となる値）に到達せずに、局所解（期待値と教師データとの誤差が極小値となる値）に陥り、最適な学習がなされない可能性がある。

【０１０５】

また、本実施形態では、隠れ層のノード数と入力層のノード数とを同数とする場合を説明したが、これに限られず、隠れ層のノード数を入力層のノード数より多くしても良いし、隠れ層のノード数を入力層のノード数より少なくしても良い。但し、隠れ層のノード数を多くした場合、本実施形態と比較して、学習時に、局所解に陥る可能性が高くなり、また計算量が増える。また、隠れ層のノード数を少なくした場合、本実施形態と比較して計算量が減る一方、判別精度が低くなる。

【０１０６】

上述のＧＰＩアンカー型タンパク質判定装置１００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】

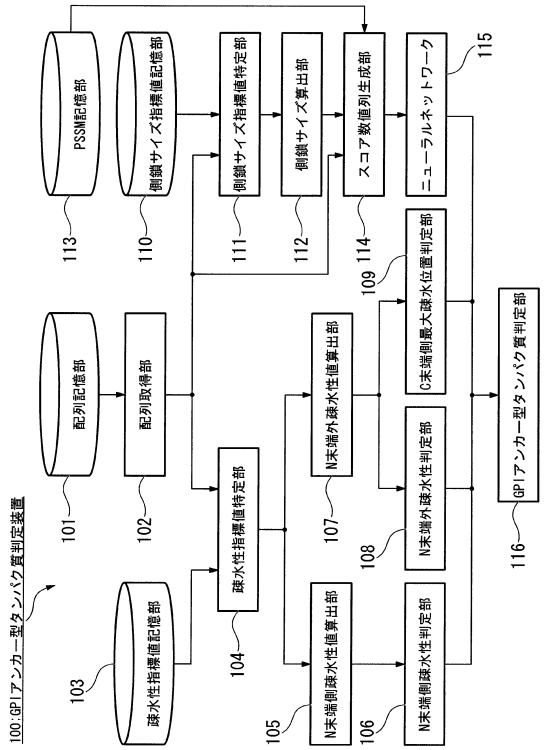
また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

【符号の説明】

【０１０８】

１００…ＧＰＩアンカー型タンパク質判定装置　１０１…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…Ｎ末端側疎水性値算出部　１０６…Ｎ末端側疎水性判定部　１０７…Ｎ末端外疎水性値算出部　１０８…Ｎ末端外疎水性判定部　１０９…Ｃ末端側最大疎水位置判定部　１１０…側鎖サイズ指標値記憶部　１１１…側鎖サイズ指標値特定部　１１２…側鎖サイズ算出部　１１３…ＰＳＳＭ記憶部　１１４…スコア数値列生成部　１１５…ニューラルネットワーク　１１６…ＧＰＩアンカー型タンパク質判定部

【図１】



【図２】

アミノ酸残基	指標	アミノ酸残基	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】

アミノ酸残基	指標	アミノ酸残基	指標
A	2.5	L	5.5
R	7.5	K	7.0
N	5.0	M	6.0
D	2.5	F	6.5
C	3.0	P	5.5
Q	6.0	S	3.0
E	5.0	T	5.0
G	0.5	W	7.0
H	6.0	Y	7.0
I	5.5	V	5.0

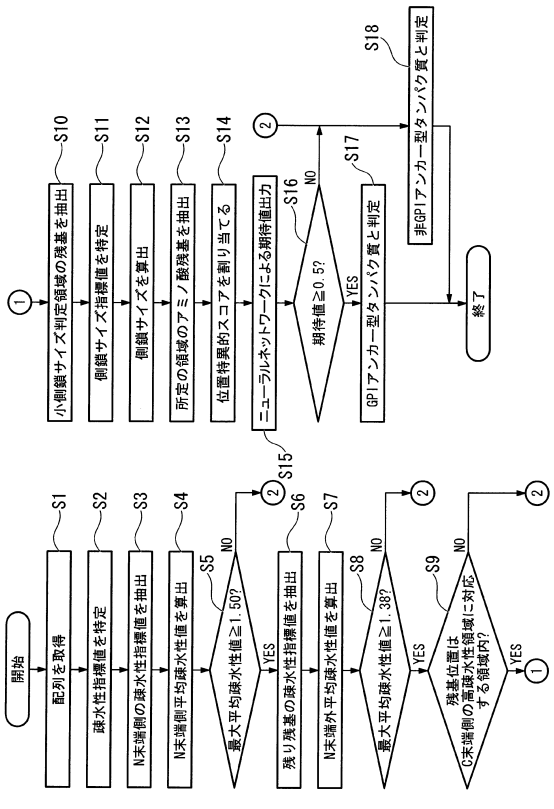
【図４】

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0
A	1.71	0.15	0.84	-0.77	2.46	1.31	-0.86	0.17	-0.07	-0.52	1.11	0.12	0.01
C	0.21	1.91	1.02	-0.07	1.91	-0.07	0.33	-0.99	1.71	-5.21	-5.50	-5.21	-4.12
D	-0.30	0.43	0.28	-0.59	0.62	0.87	-6.19	-6.06	-5.21	-0.38	-0.08	-2.37	1.24
E	-0.17	2.32	-1.33	0.95	-1.94	-0.08	-0.99	-1.94	0.43	-6.31	-4.12	-6.19	-0.48
F	-0.77	0.62	-0.77	-0.43	-5.50	0.33	-1.17	0.62	0.39	-5.50	-1.68	-4.12	-0.77
G	-0.23	-1.52	0.11	-1.57	1.02	-0.36	1.15	0.90	-5.72	0.71	-0.21	0.47	-1.77
H	-6.75	-1.32	-1.86	-0.08	-0.08	-0.36	-1.32	-0.48	-1.09	-4.12	1.31	-4.12	-0.08
I	-0.63	-0.08	-6.60	-0.99	-0.77	-1.17	-0.08	-1.40	-0.99	-5.90	-1.32	6.66	-0.59
K	-0.08	-0.39	-2.21	-0.34	-0.66	-0.56	-0.17	0.03	-0.19	-4.12	-5.72	-6.31	-1.46
L	0.37	-0.56	-1.36	-0.99	-2.08	-0.25	-1.24	-0.48	-1.89	-0.99	-0.48	-4.81	-1.43
M	1.71	-4.81	-0.99	-5.21	-1.94	-5.50	-0.99	2.00	-0.99	1.24	-5.50	1.24	2.41
N	1.13	1.02	2.69	0.21	2.12	0.21	-0.56	0.03	-0.48	0.70	-5.72	-6.68	-6.88
P	-0.08	-2.59	0.16	-0.08	-0.30	-6.68	0.21	-0.68	0.00	-5.21	-0.99	-1.17	-0.18
Q	-0.72	-0.54	-0.07	2.44	-0.23	-0.48	-0.08	1.84	-0.68	-4.81	-6.06	-5.50	0.39
R	1.61	0.62	0.50	-1.33	-1.12	-0.48	-1.33	-0.80	1.02	-5.21	-6.06	-5.21	-0.48
S	0.28	-0.48	1.61	0.00	-1.46	-1.32	0.90	0.11	0.28	0.59	0.40	0.66	0.75
T	-0.48	-0.38	0.01	-0.41	-0.72	-0.33	-0.41	0.03	-0.17	-0.59	-1.09	-0.07	0.93
V	-1.58	-1.33	-0.84	-1.94	-1.23	-1.86	-0.41	-0.26	-0.99	-7.29	-6.19	-6.19	1.35
W	-4.12	0.62	-4.12	-6.19	-6.19	-4.81	-4.81	-0.99	-1.32	-4.12	-4.12	1.24	-5.72
Y	-5.50	-0.48	0.43	1.53	0.71	1.07	4.17	0.06	0.62	1.24	0.62	1.24	-6.41

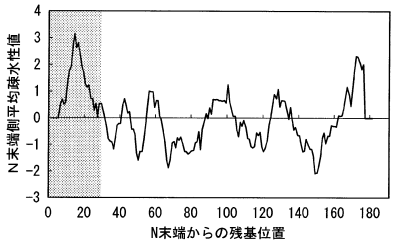
【図５】

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.31	-0.34	-1.33	0.21	-0.54	-0.77	-0.54	-0.70	-1.37	-0.15	0.21	-1.07
C	-1.17	6.66	-0.08	-0.99	-4.81	-6.60	-1.32	-1.17	-1.17	-6.19	-4.81	-6.68
D	5.97	-4.12	-0.07	1.31	-4.12	-4.81	-4.81	-6.19	1.24	-4.81	-4.12	1.24
E	0.33	-1.63	0.62	-0.30	-4.81	-5.90	-5.50	-4.81	-4.12	1.24	1.24	1.24
F	-7.33	0.50	1.91	-4.81	-7.20	-0.54	-0.40	-0.99	0.28	0.39	-6.41	-0.88
G	-0.36	0.06	-2.01	1.13	-0.83	0.37	0.17	-1.52	-0.54	-1.73	-2.59	-7.16
H	0.80	0.62	1.31	-5.50	0.15	-4.81	-0.76	-5.90	-5.72	-4.81	-4.81	1.24
I	0.90	-0.30	1.71	-0.77	-7.00	-1.68	-2.51	-0.19	1.02	-0.48	0.42	1.34
K	-6.41	-2.21	0.15	0.62	1.71	-6.41	-5.72	-6.19	-4.81	-5.21	-5.90	-4.81
L	0.26	1.95	-0.77	-0.55	-2.25	-0.22	-0.58	0.46	0.24	0.82	0.46	-0.13
M	-0.99	-0.07	-5.72	-6.82	-0.48	-5.21	-0.48	0.11	-1.32	-7.16	-0.23	-6.31
N	0.21	-5.90	0.62	0.62	6.66	-5.50	-4.81	-5.50	-4.81	-5.21	-5.72	-4.81
P	-0.08	-0.38	-1.06	-0.73	1.67	1.38	2.07	-1.17	-0.07	0.62	0.90	0.48
Q	-6.51	-0.77	-0.48	-0.77	0.62	-0.48	6.66	-4.81	1.24	5.97	-6.19	-0.99
R	-0.48	-0.88	-1.68	-1.09	-0.63	-0.07	-0.76	-4.12	-4.81	-4.12	-5.21	1.24
S	-0.30	-0.87	1.24	1.13	2.30	0.15	-0.70	-1.12	-0.19	-0.88	-0.36	1.13
T	0.26	-0.70	-0.42	-0.77	-0.39	0.95	0.33	1.09	-0.88	1.13	0.46	1.82
V	0.82	-2.59	-1.28	-1.25	-0.44	-1.12	0.69	1.36	-0.08	-0.43	0.25	-0.83
W	5.97	1.02	-5.21	-0.30	-4.12	0.95	6.66	6.66	-5.72	-4.81	-4.12	-4.81
Y	0.62	1.31	0.62	-5.50	-4.12	-4.81	-5.21	-4.81	1.24	-5.21	-4.12	-4.81

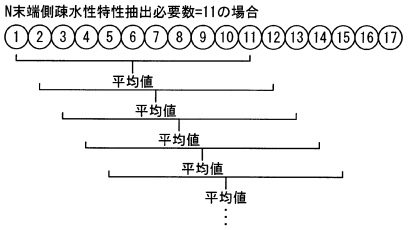
【図６】



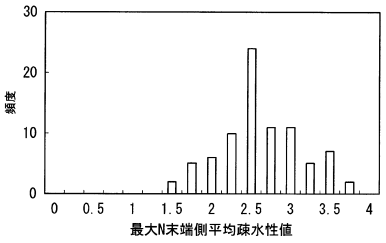
【図７】



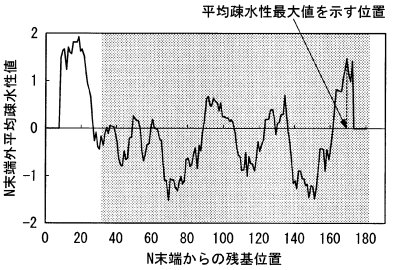
【図８】



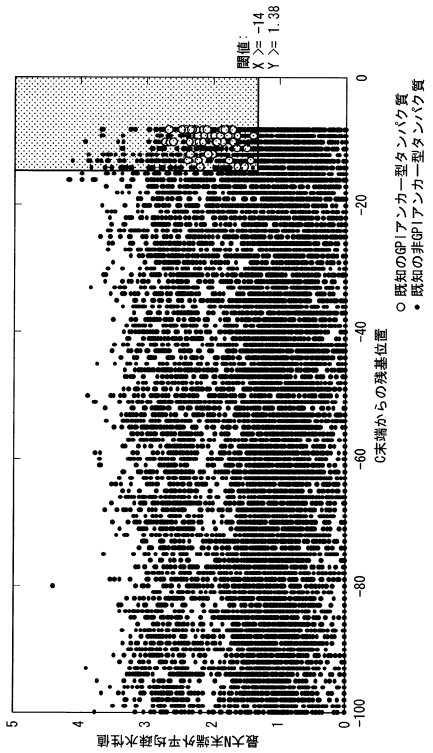
【図９】



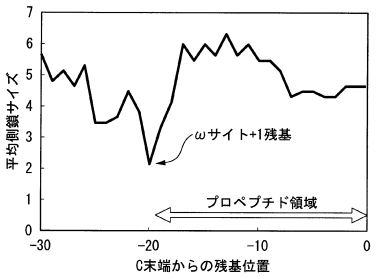
【図１０】



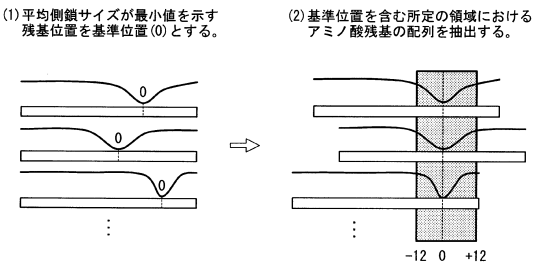
【図１１】



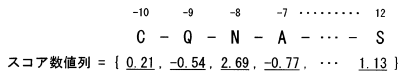
【図１２】



【図１３】



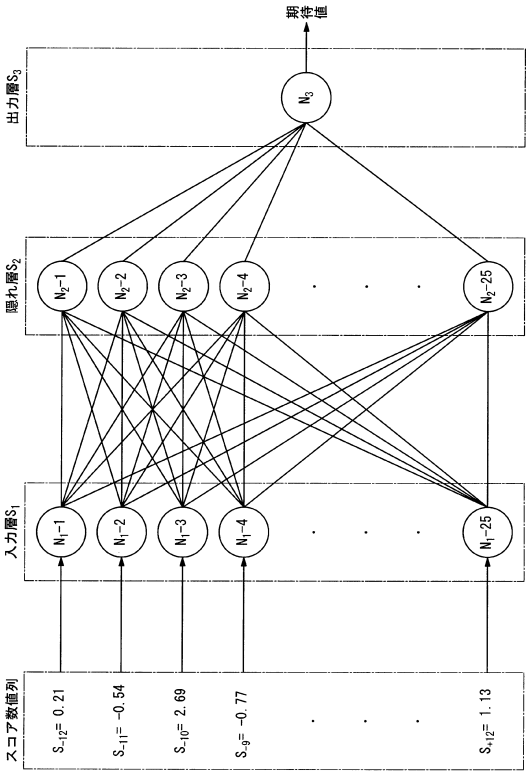
【図１４】



【図１５】

5NTD_BOVIN	PPBI_RAT
5NTD_HUMAN	PPBN_HUMAN
5NTD_MOUSE	PPBT_BOVIN
5NTD_RAT	PPBT_FELCA
BY55_HUMAN	PPBT_HUMAN
BY55_MOUSE	PPBT_MOUSE
CAH4_BOVIN	PPBT_RAT
CAH4_HUMAN	PRIO_AILME
CAH4_MOUSE	PRIO_ANTCE
CAH4_RABIT	PRIO_ATEPA
CAH4_RAT	PRIO_BISBI
CBPM_HUMAN	PRIO_BOSTR
CBPM_MOUSE	PRIO_BUBBU
CBPM_PONPY	PRIO_BUDTA
CD48_HUMAN	PRIO_CALJA
CD48_MOUSE	PRIO_CAMDR
CD48_RAT	PRIO_CANFA
CD52_CANFA	PRIO_CAPHI
CD52_HUMAN	PRIO_CEBAP
CD52_MACFA	PRIO_CERAE
CD52_MOUSE	PRIO_CEREL
CD52_RAT	PRIO_COLGU
CD59_AOTTR	PRIO_CRIGR
CD59_CALSQ	PRIO_CRIMI
CD59_CERAE	PRIO_FELCA
CD59_HUMAN	PRIO_GORGO
CD59_PAPSP	PRIO_HUMAN
CD59_PIG	PRIO_HYLLA
CD59_PONPY	PRIO_MACAR
CD59_RABIT	PRIO_MESAU
CD59_RAT	PRIO_MOSCH
CD59_SAISC	PRIO_MOUSE
CD59A_MOUSE	PRIO_MUSPF
CD59B_MOUSE	PRIO_MUSVI
CEAM6_HUMAN	PRIO_ODOHE
CEAM8_HUMAN	PRIO_OVICA
DAF_HUMAN	PRIO_PIG
DAF1_MOUSE	PRIO_PONPY
DPEP1_BOVIN	PRIO_PREFR
DPEP1_HUMAN	PRIO_RABIT
DPEP1_MOUSE	PRIO_RAT
DPEP1_RABIT	PRIO_SAISC
DPEP1_RAT	PRIO_SHEEP
DPEP1_SHEEP	PRIO_SIGHI
FOLR1_HUMAN	PRIO_TRAIM
FOLR1_MOUSE	PRIO_TRIVU
FOLR2_HUMAN	PRIOT_TRAST
FOLR2_MOUSE	PRIO2_TRAST
NAR2A_MOUSE	PSCA_MOUSE
NAR2A_RAT	SACA4_BOVIN
NAR2B_MOUSE	SACA4_MOUSE
NAR2B_RAT	THY1_HUMAN
NAR3_HUMAN	THY1_MOUSE
PPBI_HUMAN	THY1_RAT
PPBI_BOVIN	XPP2_HUMAN
PPBI_HUMAN	XPP2_PIG
PPBI_MOUSE	

【図１６】



【図１７】

	GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	91.5	91.5	0.915

	非GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	98.2	93.1	0.956

100回試行のうち、成功率が最高値のときの平均精度

【図１８】

	GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	91.4	90.2	0.907

	非GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	94.8	91.3	0.949

100回試行のうち、成功率上位10%の平均精度

【図１９】

(1)位置特異的スコア: -10→+12

	GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	90.0(90.5)	92.3(90.0)	0.911(0.901)

	非GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	95.5(94.5)	94.7(94.9)	0.951(0.947)

(カッコ内は、100回試行のうち成功率上位10%の平均精度)

【図２０】

(2)位置特異的スコア: -12→+9

	GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	92.9(90.8)	90.5(89.4)	0.916(0.900)

	非GPIアンカー型タンパク質		
	感度(%)	選択性(%)	成功率
4-fold	94.9(94.2)	96.2(95.0)	0.955(0.946)

(カッコ内は、100回試行のうち成功率上位10%の平均精度)

フロントページの続き

(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　生田校舎内
(72)発明者　吉澤　昌朗
神奈川県川崎市多摩区東三田１－１－１　学校法人明治大学　生田校舎内
(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　生田校舎内
(72)発明者　池田　修己
神奈川県川崎市多摩区東三田１－１－１　学校法人明治大学　生田校舎内

審査官　松野　広一

(56)参考文献　特開２００４－１２５６２３（ＪＰ，Ａ）
特開２００３－０１４７３４（ＪＰ，Ａ）
特開２００２－２１５６３４（ＪＰ，Ａ）
特開２００６－００３９７０（ＪＰ，Ａ）
Birgit Eisenhaber et al, Prediction of Potential GPI-modification Sites inProprotein Sequences, Joournal of Molecular Biology, １９９９年, Vol.292, pp.741-758
Niklaus fankhauser, Pascal Maser, Identification of GPI anchor attachment signals by a Kohonenself-organizing map, Bioinformatics, ２００５年, Vol.21 No.9, pp.1846-1852, ＵＲＬ, http://bioinformatics.oxfordjournals.org/content/21/9/1846.full.pdf+html
Andrea Pierleoni et al, PredGPI: a GPI-anchor predictor, BMC Bioinformatics, ２００８年, Vol.9 No.392, pp.1-11, ＵＲＬ, http://www.biomedcentral.com/content/pdf/1471-2105-9-392.pdf
新島 耕一, ゲノムデータ解析用の高速ニューラルネットワークの研究, ゲノム解析に伴う大量知識情報処理の研究　平成５年度研究成果報告書, 日本, 京都大学化学研究所, １９９４年　３月３１日, ｐｐ．１５０－１５５

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ　１９／１０－１９／２８
Ｇ０１Ｎ　３３／４８