

(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2012-32163  
(P2012-32163A)

(43) 公開日 平成24年2月16日(2012.2.16)

(51) Int.Cl.  
G01N 33/48 (2006.01)

F I  
G O I N 33/48

テーマコード(参考)  
2 G O 4 5

審査請求 未請求 請求項の数 19 O L (全 33 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)

(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 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

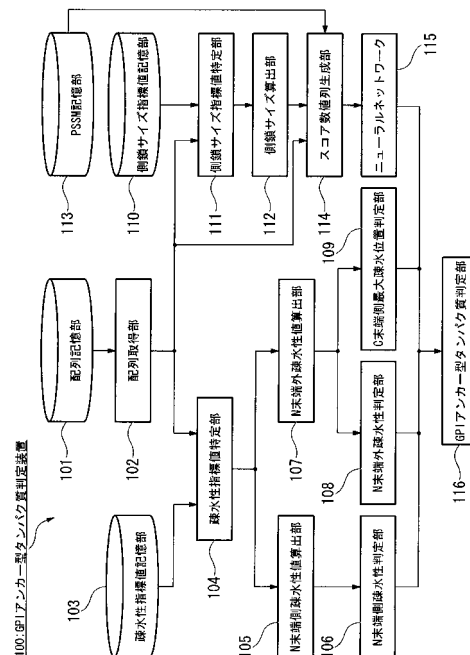
(54) 【発明の名称】 G P I アンカー型タンパク質の判定装置、判定方法及び判定プログラム

(57) 【要約】

【課題】高感度且つ高選択的に検査対象タンパク質が G P I アンカー型タンパク質であるか否かを判定する。

【解決手段】スコア数値列生成部 1 1 4 は、P S S M 記憶部 1 1 3 が記憶する P S S M 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。次に、ニューラルネットワーク 1 1 5 は、当該スコア数値列を入力し、G P I アンカー型タンパク質らしさを示す 0 以上 1 以下の期待値を出力する。そして、G P I アンカー型タンパク質判定部 1 1 6 は、当該期待値の値に基づいて、検査対象タンパク質が G P I アンカー型タンパク質であるか否かを判定する。

【選択図】図 1



**【特許請求の範囲】****【請求項 1】**

検査対象タンパク質が G P I アンカー型タンパク質であるか否かを判定する G P I 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 G P I 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、

10

記憶部に記憶されている既知の G P I 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前記記憶部から取得し、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、G P I アンカー型タンパク質らしさを示す 0 以上 1 以下の期待値を出力する分類部であって、既知の G P I アンカー型タンパク質の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 G P I 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、

20

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 G P I アンカー型タンパク質でないとして判定する G P I 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 G P I 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、

30

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 G P I アンカー型タンパク質の判定装置。

40

**【請求項 3】**

前記分類部は、既知の G P I 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 G P I 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 G P I 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 G P I アンカー型タンパク質の判定装置。

**【請求項 5】**

50

前記小側鎖サイズ判定領域は、

既知の G P I 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 1 から請求項 4 の何れか 1 項に記載の G P I アンカー型タンパク質の判定装置。

【請求項 6】

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 G P I 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 G P I 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 G P I 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

10

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 G P I 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 G P I 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 G P I 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、

20

【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 G P I アンカー型タンパク質の判定装置。

30

【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 G P I 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  $\epsilon$  と、アミノ酸残基の種類数  $s$  とを用いて、

【数 2】

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

40

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 G P I 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 G P I 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

50

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数を用いて、連続する当該N末端外疎水性特性抽出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

10

20

30

40

50

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 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 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 1 から請求項 15 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

10

20

30

40

50

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、記憶部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前記記憶部から取得し、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、

記憶部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前記記憶部から取得し、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ことを特徴とする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。

10

20

30

40

50

## 【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 ( 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、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、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」、

10

20

30

40

50

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、記憶部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前記記憶部から取得し、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して判定する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

10

20

30

40

50



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。

【0012】

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。

【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

10

20

30

40

50

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数を用いて、連続する当該N末端外疎水性特性抽出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

10

20

30

40

50

I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、記憶部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前記記憶部から取得し、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、記憶部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前記記憶部から取得し、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

10

20

30

40

50

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して判定する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。

10

【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。

20

【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。

30

【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。

40

【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 エントリー名を示す図である。

【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(-

50

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

10

20

30

40

50

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 ( 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。

## 【 0 0 3 5 】

そして、G P Iアンカー型タンパク質判定装置 1 0 0において、配列取得部 1 0 2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 1 1 2は、配列取得部 1 0 2が取得したアミノ酸配列情報における既知のG P I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 1 1 4は、P S S M記憶部 1 1 3に記憶されている既知のG P I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P S S M記憶部 1 1 3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 1 1 2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 1 1 5は、スコア数値列生成部 1 1 4が生成したスコア数値列を入力し、G P Iアンカー型タンパク質らしさを示す0以上1以下の期待値を出力する。なお、ニューラルネットワーク 1 1 5は、既知のG P I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G P I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。

10

そして、G P Iアンカー型タンパク質判定部 1 1 6は、ニューラルネットワーク 1 1 5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G P Iアンカー型タンパク質でないと判定する。

20

これにより、G P Iアンカー型タンパク質判定装置 1 0 0は、高感度且つ高選択的に検査対象タンパク質がG P Iアンカー型タンパク質であるか否かを判定する。

## 【 0 0 3 6 】

次に、G P Iアンカー型タンパク質判定装置 1 0 0の動作を説明する。

図6は、G P Iアンカー型タンパク質判定装置 1 0 0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、G P Iアンカー型タンパク質判定装置 1 0 0が動作を開始すると、配列取得部 1 0 2は、配列記憶部 1 0 1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。

30

## 【 0 0 3 7 】

<ステップS2：疎水性指標値を特定>

配列取得部 1 0 2がアミノ酸配列情報を取得すると、疎水性指標値特定部 1 0 4は、疎水性指標値記憶部 1 0 3を参照して、配列取得部 1 0 2が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 1 0 2が取得したアミノ酸配列情報が、「M L L E P G R G C C ... ..」という配列を示す場合、疎水性指標値特定部 1 0 4は、疎水性指標値記憶部 1 0 3が記憶する図2に示す指標値より「1.9、3.8、3.8、-3.5、-1.6、-0.4、-4.5、-0.4、2.5、2.5... ..」という数値列を生成する。

40

## 【 0 0 3 8 】

<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 1 0 4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 1 0 5は、疎水性指標値特定部 1 0 4が生成した数値列から、G P I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、G P I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G P Iアンカー型タンパク質のアミノ酸残基のそれ

50

それぞれに対して、後述するステップ S 4 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。

【 0 0 3 9 】

図 7 は、G P I 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、S W I S S - P R O T ver 5 4 . 0 の B Y 5 5 \_ H U M A N ( 1 8 1 a a ) エントリーに対して、後述するステップ S 4 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。

10

図 7 に示すように、既知の G P I 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。

【 0 0 4 0 】

< ステップ S 4 : 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 S 3 で G P I 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。

20

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。

【 0 0 4 1 】

【 数 1 】

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

30

【 0 0 4 2 】

但し、n は、平均化に用いる前後の残基数を示す。つまり、2 n + 1 は、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 N U L L 値を代入しておくが良い。

40

【 0 0 4 3 】

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。

【 0 0 4 4 】

まず、既知の複数の G P I 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 G P I 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

50



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 ver 54.0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

10

20

30

40

50

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 N U L L 値を代入しておくが良い。

#### 【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。

10

#### 【0051】

まず、既知の複数の G P I 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 G P I 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 G P I アンカー型タンパク質における既知の複数の G P I 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 G P I 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 G P I 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 G P I 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。

20

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 G P I 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 G P I 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。

30

#### 【0052】

図 10 は、G P I 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、S W I S S - P R O T ver 5 4 . 0 の B Y 5 5 \_ H U M A N ( 1 8 1 a a ) エントリーに対して、ステップ S 7 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 G P I 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。

40

#### 【0053】

そして、本実施形態では、S W I S S - P R O T ver 5 4 . 0 より取得した既知の哺乳類 G P I 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 G P I 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 S 8 : N 末端外平均疎水性値の最大値の判定>

ステップ S 7 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

50

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 S 1 2 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。

【 0 0 5 9 】

< ステップ S 1 1 : 側鎖サイズ指標値を特定 >

側鎖サイズ指標値特定部 1 1 1 は、ステップ S 1 0 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 1 1 0 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 S 1 1)。例えば、配列取得部 1 0 2 が取得したアミノ酸配列情報が、「M L L E P G R G C C . . . . .」という配列を示す場合、側鎖サイズ指標値特定部 1 1 1 は、側鎖サイズ指標値記憶部 1 1 0 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。

10

【 0 0 6 0 】

< ステップ S 1 2 : 平均側鎖サイズを算出 >

ステップ S 1 1 で、側鎖サイズ指標値特定部 1 1 1 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 1 1 2 は、側鎖サイズ指標値特定部 1 1 1 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。

20

ここで、平均側鎖サイズ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平均側鎖サイズは、式 ( 2 ) を用いて算出できる。

【 0 0 6 1 】

【 数 2 】

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

30

【 0 0 6 2 】

但し、n は、平均化に用いる前後の残基数を示す。つまり、2 n + 1 は、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 N U L L 値を代入しておくが良い。

40

【 0 0 6 3 】

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。

【 0 0 6 4 】

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。

50

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（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 ver 54.0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。

10

#### 【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 ver 54.0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。

20

#### 【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。

30

#### 【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。

40

#### 【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 ver 54.0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。

50

## 【 0 0 7 0 】

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。

10

## 【 0 0 7 1 】

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。

20

## 【 0 0 7 2 】

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。

30

## 【 0 0 7 3 】

## 【 数 3 】

$$\frac{n_{ip} + \epsilon \frac{1}{s}}{\sum_{i=1}^s n_{ip} + \epsilon} \quad \dots (3)$$

40

## 【 0 0 7 4 】

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、 $\epsilon$ 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。

50

同様に、データセットのうち非 G P I 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 G P I 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 G P I 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 G P I 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】

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

【0077】

但し、 $f_{ip}^{positive}$  は、既知の G P I 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 G P I 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、G P I 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 P S S M として生成し、P S S M 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 P S S M を生成することができる。

【0078】

<ステップ S 15 : ニューラルネットワークによる期待値出力>

ステップ S 14 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、G P I アンカー型タンパク質らしさを示す 0 以上 1 以下の期待値を出力する。なお、P S S M 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<sub>1</sub>、隠れ層 S<sub>2</sub>、出力層 S<sub>3</sub> の3段の階層構造を有する。

入力層 S<sub>1</sub> 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<sub>1</sub>-1 ~ N<sub>1</sub>-25 (以下、ノード N<sub>1</sub>-1 ~ N<sub>1</sub>-25 を総称する場合は、ノード N<sub>1</sub> と記載する) で構成される。

隠れ層 S<sub>2</sub> は、入力層 S<sub>1</sub> のノード数と同数のノード N<sub>2</sub>-1 ~ N<sub>2</sub>-25 (以下、ノード N<sub>2</sub>-1 ~ N<sub>2</sub>-25 を総称する場合は、ノード N<sub>2</sub> と記載する) で構成される。

出力層 S<sub>3</sub> は、1つのノード N<sub>3</sub> で構成される。

【0080】

ノード N<sub>1</sub> 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<sub>2</sub> のそれぞれに出力する。ノード N<sub>2</sub>

10

20

30

40

50

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】

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

10

【数6】

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

20

【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、 $\theta$  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。

30

【0083】

【数7】

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

【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。

40

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

50



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

10

20

30

40

50

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。

10

#### 【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。

20

30

#### 【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】

40

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。

50

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

10

20

30

40

50

ることによって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115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。

10

#### 【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行うことが可能である。

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。

20

#### 【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。

30

#### 【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。

40

#### 【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。

50

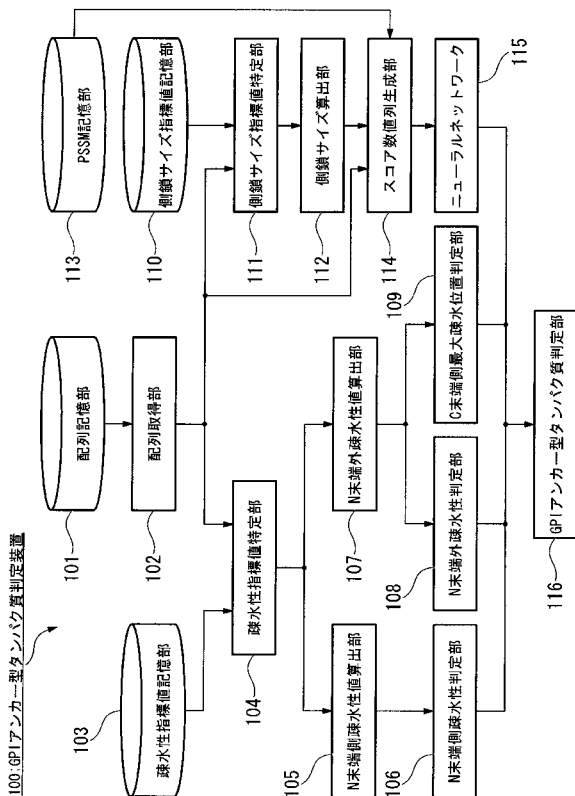
さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

【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

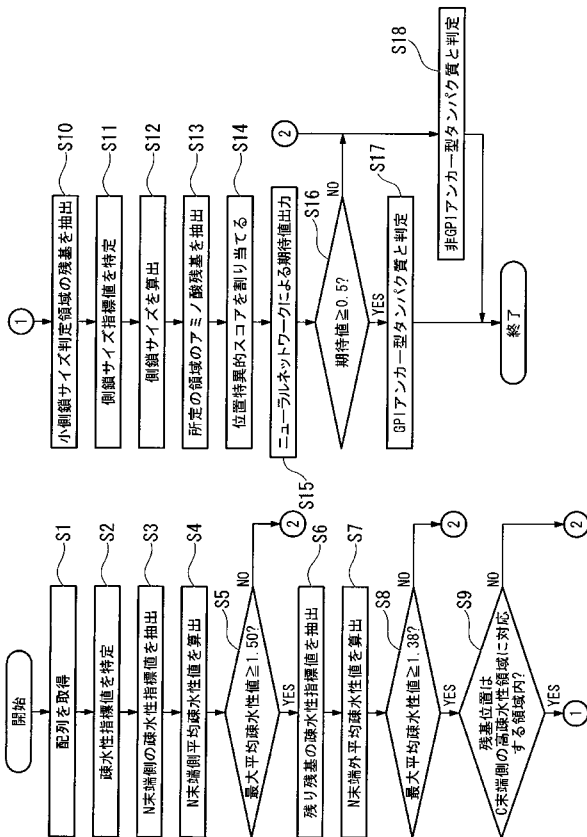
【 図 4 】

	-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	-0.06	-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

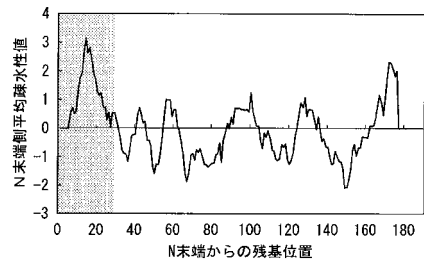
【 図 5 】

	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

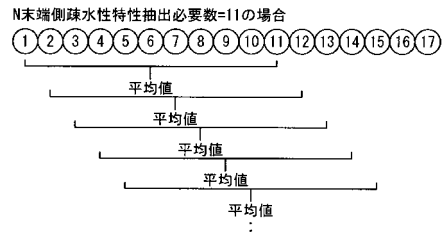
【 図 6 】



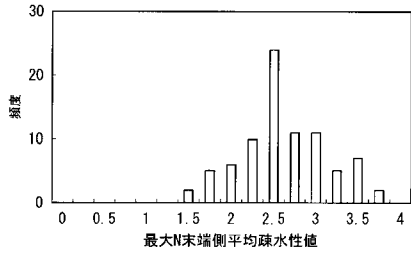
【 図 7 】



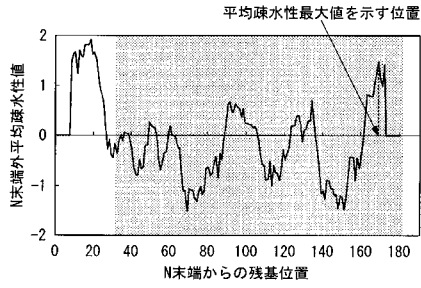
【 図 8 】



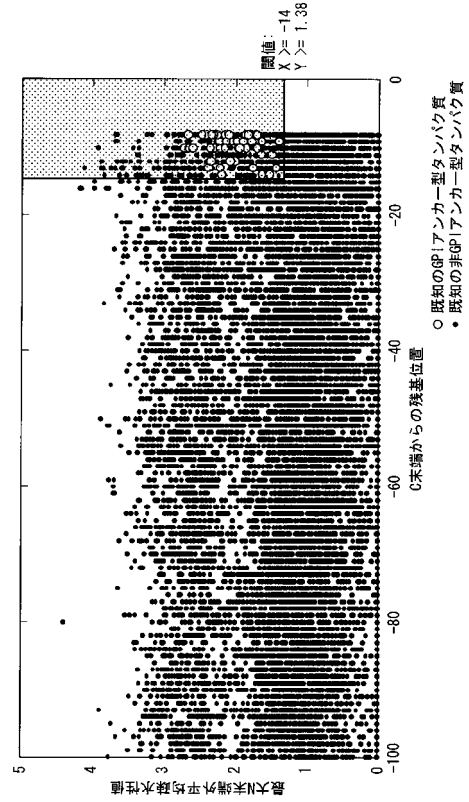
【 図 9 】



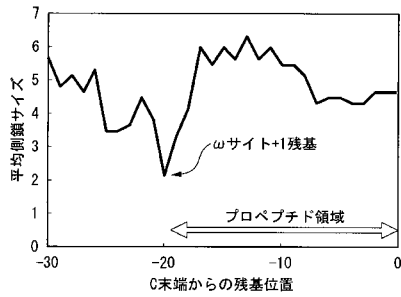
【 図 10 】



【 図 11 】



【 図 12 】

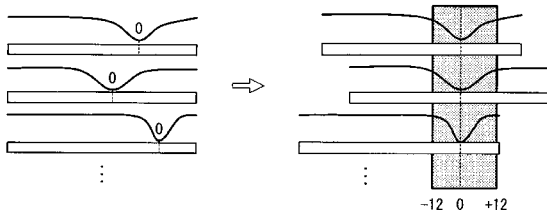


【 図 14 】

-10   -9   -8   -7   .....   12  
 C - Q - N - A - ..... - S  
 スコア数値列 = { 0.21, -0.54, 2.69, -0.77, ..., 1.13 }

【 図 13 】

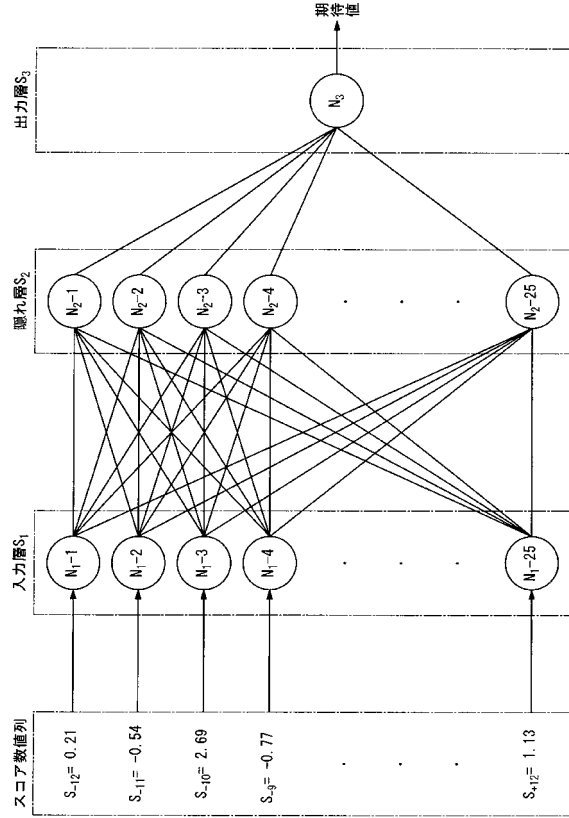
- (1) 平均側鎖サイズが最小値を示す残基位置を基準位置 (0) とする。
- (2) 基準位置を含む所定の領域におけるアミノ酸残基の配列を抽出する。



【 図 1 5 】

ENTD_BOVIN	PPB1_RAT
ENTD_HUMAN	PPBN_HUMAN
ENTD_MOUSE	PPBT_BOVIN
ENTD_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_BISB1
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_CR1GR
CD59_CALSQ	PRIO_CRIMI
CD59_CERAE	PRIO_FELCA
CD59_HUMAN	PRIO_GORGO
CD59_PAFSP	PRIO_HUMAN
CD59_PIG	PRIO_HYLLA
CD59_PONPY	PRIO_MACAR
CD59_RABIT	PRIO_MESAU
CD59_RAT	PRIO_MOSCH
CD59_SAI3C	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_SAI3C
DPEP1_RAT	PRIO_SHEEP
DPEP1_SHEEP	PRIO_SIGHI
FOLR1_HUMAN	PRIO_TRAIM
FOLR1_MOUSE	PRIO_TRIVU
FOLR2_HUMAN	PRIO1_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
PPB1_HUMAN	THY1_RAT
PPB1_BOVIN	XPP2_HUMAN
PPB1_HUMAN	XPP2_PIG
PPB1_MOUSE	

【 図 1 6 】



【 図 1 7 】

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 1 9 】

(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 1 8 】

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 2 0 】

(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
- Fターム(参考) 2G045 DA30 DA36 DA60 JA01