(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2003－14734
（Ｐ２００３－１４７３４Ａ）
(43)公開日　平成15年１月15日(2003.1.15)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ (参考)
G01N 33/48		G01N 33/48	Z 2G045

審査請求　未請求　請求項の数３　ＯＬ　（全10頁）

(21)出願番号　特願2001－181248(Ｐ2001－181248)

(22)出願日　平成13年６月15日(2001.6.15)

特許法第30条第１項適用申請有り　2000年12月18日～19日　日本バイオインフォマティクス学会開催の「ＧＩＷ2000（Ｔｈｅ　Ｅｌｅｖｅｎｔｈ　Ｗｏｒｋｓｈｏｐ　ｏｎ　Ｇｅｎｏｍｅ　Ｉｎｆｏｒｍａｔｉｃｓ）」において文書をもって発表

(71)出願人　396020800
科学技術振興事業団
埼玉県川口市本町４丁目１番８号
(72)発明者　美宅　成樹
国分寺市南町３－21－１－1108
(72)発明者　五味　雅裕
埼玉県狭山市北入曽402－17
(74)代理人　100087631
弁理士　滝田　清暉　（外１名）

Ｆターム(参考)　2G045 DA36 JA01

(54)【発明の名称】シグナルペプチドの判別方法、及びそのためのコンピュータプログラム

(57)【要約】　（修正有）
【課題】　アミノ酸配列が決定されたタンパク質に、シグナルペプチド又はシグナルアンカーの何れかが含まれるか否かを、高い精度で判別するための方法、及び、そのためのコンピュータプログラムの提供。
【解決手段】　タンパク質の構成要素である20種類のアミノ酸それぞれに対して、予め、疎水性指標H、負電荷残基指標NC、シグナルペプチド判別指標SP Index、及びシグナル配列判別指標SS Indexを割り当てたデータ、及び判別式を用い、シグナル配列の候補領域の二回平均疎水性値のピーク位置に近い正電荷残基の位置を判別基準位置Ｐｒとし、該判別基準位置Ｐｒからアミノ酸配列のＣ末端側へ下流10番目から27番目までの18残基の領域を判別対象領域Ｒとして設定し、１８個のアミノ酸残基それぞれに、ＳＰ Index、SS Indexを割り振り、判別式から該候補領域がシグナルペプチド、シグナルアンカー、シグナル配列無しの何れに該当するものかを判別する。

1

【特許請求の範囲】

【請求項１】　（ａ）タンパク質の構成要素である20種類のアミノ酸それぞれに対して、予め、疎水性指標H、負電荷残基指標NC、シグナルペプチド判別指標SP Index、及びシグナル配列判別指標SS Indexを割り当てておき、（ｂ）５～９個の何れかの数の一連の窓を有するウィンドウWを用いて被判別タンパク質のアミノ酸配列から一連のアミノ酸配列を抽出し、（ｃ）抽出された前記ウィンドウWと対応する全てのアミノ酸配列について、前記疎水性指標を用いて計算した二回平均疎水性値[[H]]を算出し、（ｄ）得られた[[H]]の値が一定の閾値ｋを越える領域の一連のアミノ酸残基の数が５～２５残基となり、かつその領域内に負電荷残基指標NCが１であるアミノ酸残基を含まないとき、その領域を構成するアミノ酸残基からなるセグメントをシグナル配列の候補領域Ｃとし、（ｅ）複数の候補領域が得られた場合には、被判別アミノ酸配列のＮ末端に最も近い候補領域を最終的な候補領域Ｃとして選択すると共に、その候補領域Ｃの始点及び終点のアミノ酸残基の位置を示す数(Xstart、Xend、並びに候補領域Ｃの長さXlength)を求め、（ｆ）前記候補領域Ｃが得られない場合には被判別タンパク質はシグナルペプチドを含まない水溶性タンパク質であると認定し、（ｇ）前記候補領域Ｃがある場合にはその平均疎水性値及び該候補領域Ｃ中の二回平均疎水性値[[H]]の最大ピーク位置ＰｐをＮ末端に向かって正電荷残基を検索し、最もピーク位置に近い正電荷残基の位置を判別基準位置Ｐｒとし、（ｉ）もし正電荷残基が見出されない場合には被判別アミノ酸配列のＮ末端をＰｒと定義し、（ｊ）該判別基準位置Ｐｒからアミノ酸配列のＣ末端側へ下流10番目から27番目までの18残基の領域を判別対象領域Ｒとして設定し、（ｋ）該判別対象領域Ｒを構成する１８個のアミノ酸残基それぞれに、前記

$$[H(i)] = \left[\sum_{j=i-W/2}^{i+W/2} H(j) \right] / W \qquad (1)$$

$$[[H(i)]] = \left[\sum_{j=i-W/2}^{i+W/2} [H(j)] \right] / W \qquad (2)$$

（ｄ）得られた[[H]]の値が一定の閾値ｋを越える領域の一連のアミノ酸残基の数が５～２５残基となり、かつその領域内に負電荷残基指標NCが１であるアミノ酸残基を含まないとき、その領域を構成するアミノ酸残基からなるセグメントをシグナル配列の候補領域Ｃとし、（ｅ－１）複数の候補領域Ｓ(i)が得られた場合には被判別アミノ酸配列のＮ末端に最も近い候補領域Ｓ(i)を最終的な候補領域Ｃとして選択すると共に、その候補領域Ｃの始点及び終点のアミノ酸残基の位置を示す数(Xstart及びXend、並びに候補領域Ｃの長さXlength)を求める。但し、（ｅ－２）候補領域Ｃが１００残基目より後にのみ現れるか否かを判別し、１００残基目より後にのみ現れる場合には、シグナル配列はないものとしてプログラムは終了する。（ｆ）候補領域Ｃが得られない場合には被判別タンパク質はシグナルペプチドを含まない水溶性タンパク質であると認定し、（ｇ）候補領域Ｃの平均疎水性値を下記式（３）によって求め、

$$[\overline{H}] = \left[\sum_{i=X_{start}}^{X_{end}} [H(i)] \right] / X_{length} \qquad (3)$$

次いで前記候補領域Ｃ中の二回平均疎水性値[[H]]の最大ピーク位置ＰｐをＮ末端に向かって正電荷残基を検索し、最もピーク位置に近い正電荷残基の位置を判別基準位置Ｐｒとし、（ｉ）もし正電荷残基が見出されない場

2

したシグナルペプチド判別指標ＳＰ Index及びシグナル配列判別指標SS Indexを割り振り、（ｌ）前記判別対象領域ＲにおけるSP Index及びSS Indexの平均値を算出し、（ｍ）計算された各パラメータを元に、３つの判別式によって、該領域についてシグナルペプチド-シグナルアンカーの二群の判別、（ｎ）シグナルペプチド-シグナル配列無しの二群の判別、（ｏ）シグナルアンカー-シグナル配列無しの二群の判別、と二群の判別を３回行ってそれぞれで判別結果を得、（ｐ）得られた各々の判別結果の組み合わせから前記候補領域Ｃがシグナルペプチド、シグナルアンカー、シグナル配列無しの何れに該当するものかを判別することを特徴とするシグナルペプチドの判別方法。

【請求項２】　シグナルペプチド判別指標SP Index及びシグナル配列判別指標SS Indexを、真極生物由来のタンパク質を構成する場合と、原核生物由来のタンパク質を構成する場合とで別々に設定し、被判別タンパク質が何れの生物に由来するタンパク質であるかに従って、（１）の工程で割り振るシグナルペプチド判別指標SP Indexとシグナル配列判別指標SS Indexを選択する、請求項１に記載されたシグナルペプチドの判別方法。

【請求項３】　（ａ）真核生物由来のタンパク質と原核生物由来のタンパク質の各タンパク質別に、タンパク質の構成要素である20種類のアミノ酸それぞれについて、疎水性指標H、負電荷残基指標NC、シグナルペプチド判別指標SP Index、及びシグナル配列判別指標SS Indexを割り当てたデータ、及び、各種判別式を記憶部にあらかじめ格納させておき、（ｂ）５～９個の何れかの数の一連の窓を有するウィンドウWを用いて被判別タンパク質のアミノ酸配列から一連のアミノ酸配列を抽出し、（ｃ）抽出された前記ウィンドウWと対応する全てのアミノ酸配列について、下式（１）及び（２）に前記疎水性指標をあてはめて二回平均疎水性値[[H]]を算出し、

3

合には被判別アミノ酸配列のN末端をＰｒと定義し、
（ｊ）前記判別基準位置Ｐｒからアミノ酸配列のＣ末端側へ下流10番目から27番目までの18残基の領域を判別対象領域Ｒとして設定し、（ｋ）判別対象領域を構成する１８個のアミノ酸残基それぞれに、被判別タンパク質が真核生物由来か原核生物由来かに応じて、対応する前記シグナルペプチド判別指標ＳＰ Index及びシグナル配列判別指標SS Indexを割り振り、（ｌ）下式（４）及び（５）に基づいて判別対象領域におけるSP Index及びSS Indexの平均値を算出し、

$$\left[\overline{X}_{SP}\right]=\left[\sum_{i=Pr+10}^{Pr+27}\left[SP(i)\right]\right]/18 \qquad (4)$$

4

$$\left[\overline{X}_{SS}\right]=\left[\sum_{i=Pr+10}^{Pr+27}\left[SS(i)\right]\right]/18 \qquad (5)$$

計算された各パラメータを元に下記３つの判別式（６）、（７）、（８）によって、（ｍ）該領域についてシグナルペプチド-シグナルアンカーの二群の判別、（ｎ）シグナルペプチド-シグナル配列無しの二群の判別、（ｏ）シグナルアンカー-シグナル配列無しの二群の判別、と二群の判別を３回行ってそれぞれで判別結果を得、

$$F_{SP-SA}=b_1X_{start}+b_2X_{end}+b_3X_{length}+b_4\left[\overline{H}\right]+b_5\left[\overline{X}_{SP}\right] \qquad (6)$$

$$F_{SP-NoSignal}=b_6X_{start}+b_7X_{end}+b_8X_{length}+b_9\left[\overline{H}\right]+b_{10}\left[\overline{X}_{SS}\right] \qquad (7)$$

$$F_{SA-NoSignal}=b_{11}X_{start}+b_{12}X_{end}+b_{13}X_{length}+b_{14}\left[\overline{H}\right]+b_{15}\left[\overline{X}_{SS}\right] \qquad (8)$$

（ｐ）得られた各々の結果を下記①～⑧にあてはめて、候補領域を判定する（但し、各判別式中の定数は夫々±１０%の間で変動することがあるものとする。）；
①ＳＰ∩ＳＰ∩ＳＡ→ＳＰ
②ＳＰ∩ＳＰ∩ＮＳ→ＳＰ
③ＳＰ∩ＮＳ∩ＳＡ→ＮＳ
④ＳＰ∩ＮＳ∩ＮＳ→ＮＳ
⑤ＳＡ∩ＳＰ∩ＳＡ→保留
⑥ＳＡ∩ＳＰ∩ＮＳ→保留
⑦ＳＡ∩ＮＳ∩ＮＳ→ＮＳ
⑧ＳＡ∩ＮＳ∩ＳＡ→ＳＡ
（ｑ）また、上記判別結果が⑤又は⑥となって保留とした場合には、シグナルペプチド切除部位であるCleavage Siteを特徴付ける（－１、－３）ルールに即したパターン検索を行い、もし当てはまればＳＰ、当てはまらない場合はＮＳとすることを特徴とするコンピュータプログラム。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、アミノ酸配列が決定されたタンパク質について、シグナルペプチドの有無を判別する方法に関し、コンピュータを用いて大量のアミノ酸配列に対して迅速且つ高い精度で対象の配列がシグナルペプチドを含有するか否かを判別する方法、及びシグナルペプチドが存在すると判別される場合には、シグナルペプチド領域についての情報を出力する方法、及びそれを実施する為のコンピュータソフトウェアに関する。
【０００２】
【従来技術】シグナルペプチドは、分泌型水溶性タンパク質および一部の膜タンパク質のアミノ酸配列のN末端付近に存在する10～30残基長の機能性セグメントであり、細胞質内で生合成されたポリペプチド鎖の生体膜透過、及び膜への組み込みにおいて重要な役割を果たしている。シグナルペプチドの判別および領域予測は、これまで、シグナルペプチドが膜透過後に切除を受ける位置である、Cleavage Site付近のアミノ酸配列のパターンを認識することによって行われてきた。このようなアプローチによるシグナルペプチドの判別、及び領域予測には幾つかの方法が提唱されている。ひとつにはシグナルペプチドの配列パターンからウエイトマトリックスを作製し、これを用いてシグナルペプチドの判別・領域予測を行う統計的手法があり、またニューラルネットワークや隠れマルコフモデルのような機械学習的アルゴリズムを用いたパターン認識的手法、並びにこれらの手法を組み合わせた複合的手法などがある。
【０００３】典型的なシグナルペプチドは、疎水的な性質の側鎖を持つアミノ酸残基が比較的高頻度で現れる疎水性コア領域を有することが知られている。従って、シグナルペプチドを判別する初段階として、与えられた任意のアミノ酸配列からシグナルペプチドの候補領域を捕捉することを目的とする時、この疎水性コア領域をシグナルペプチドの候補とする手法が従来のシグナルペプチド判別技術でも用いられてきた。
【０００４】しかしながら、シグナルペプチドの疎水性コア領域を捉える際、ある疎水性インデックス値の閾値をもって決定するという単純な方法では、シグナルペプチドの疎水性コア領域だけでなく、膜タンパク質の膜貫通領域や、本来単なる水溶性タンパク質の一領域に過ぎないような配列までがシグナルペプチドの候補領域として誤って予測されることがあるという欠点があった。
【０００５】一方、典型的なシグナルペプチドを特徴付けるとされるものとして別の要素も知られている。その一つは、疎水性コア領域のN末端側に正電荷を有するア

ミノ酸残基が高頻度で現れるということである。またシグナルペプチドが、膜透過後に切除されるアミノ酸配列上の位置であるCleavage Siteには、明確な配列パターンは存在しないが、Cleavage SiteからN末端側へ１つ目と３つ目のアミノ酸残基に相当する位置（即ち（-1, -３）位）に、側鎖の体積が小さなアミノ酸残基が頻出することが知られている（（－１、－３）ルール；詳細はVon Heigne,Eur.J.Biochem.133:17-21(1981)に記載されている）。しかしながら、これら既存の典型的なシグナルペプチドを特徴付けるとされる要素だけでは、シグナルペプチドの判別に充分な精度を得ることはできなかった。

【０００６】一般に、アミノ酸配列上における機能予測には、ある機能に特徴的なアミノ酸残基の出現パターンを検索するモチーフ検索という手法が取られるが、このような方法ではシグナルペプチドの判別は不可能である。例えば、既存の技術では、シグナルペプチドに関する膨大な凡例を機械学習的アルゴリズムによって判別プログラムに学習させ、これによってシグナルペプチドを判別させることにより一定以上の精度でシグナルペプチドの判別を可能としてきた。しかしながら、隠れマルコフモデルやニューラルネットワークに代表される機械学習的アルゴリズムによる判別では、判別が可能なシグナルペプチドは学習に使用したデータセットに依存することになるという宿命的な問題が存在する。このことは、学習したパターンとは異なる判別対象を判別できないという欠陥につながる。また、精度を上げるために学習するデータを増やしていくと、本来シグナルペプチドではないものをシグナルペプチドと誤って判別する確率が増加していくという欠点も併せ持つ。

【０００７】既存のシグナルペプチド判別技術は、シグナルペプチドの持つ微弱なパターンをいかに捕捉して判別するかという点にのみ注力されている。従って、たとえ結果としてシグナルペプチドを精度よく判別できたとしても、シグナルペプチドの持つ物理化学的な性質やシグナルペプチドが関与するタンパク質の生体膜透過について、あるいは生体膜への組み込みシステムについての生物学的考察が欠如していると言える。

【０００８】一方、細胞質内において合成されたタンパク質が輸送されるプロセスを考えてみると、シグナルペプチドを形作るアミノ酸配列の持つべき性質が分かる。即ち、細胞質以外の場所で働くタンパク質は全て、最低一回は生体膜を透過するプロセスを経る必要がある。この場合の、生体内における膜透過を実現するシステムは複数存在することが知られている。最もよく用いられるのがシグナルペプチドが関与するタンパク質膜透過機構であり、多くの分泌タンパク質がこの経路を辿って生体膜を透過する。例えば、細胞質内で遊離リボソームにおけるタンパク質のポリペプチド鎖を合成する際には、このタンパク質がシグナルペプチドを持つ場合には、シグナルペプチドがシグナル認識粒子（SRP)によって認識を受ける。そして、SRPによる認識はシグナルペプチドの疎水性領域を認識すると言われている。

【０００９】上記SRPによる認識を受けるとポリペプチド鎖の伸張が停止する。一方、SRPの方は膜上のSRP受容体に認識されるとポリペプチド鎖の伸張が再開される。SRP受容体の傍にはタンパク質を膜透過させる機構であるトランスロコンと呼ばれるタンパク質の複合体があり、運ばれてきたポリペプチド鎖を膜透過させる。この時、シグナルペプチドはN末端側が細胞質側に向くトポロジーを形成しており、トランスロコンの中で丁度膜を貫通するような形になっている。ポリペプチド鎖の伸張が100～150残基まで進んだところで、膜表在性の酵素であるシグナルペプチダーゼによって、シグナルペプチドはCleavage Siteにおいて切除されるのである。

【００１０】もしCleavage Siteを持たず、それ以外はシグナルペプチドと同様の配列がタンパク質のN末端に存在する場合には、シグナルペプチドの場合と同様にSRPによる認識とトランスロコン通過というプロセスを踏むものの、Cleavage Siteが無いためにシグナルペプチダーゼによる切除を受けない。その結果としてこのセグメントは膜に組み込まれ、N末端が膜の内側を向いた膜貫通領域を形成する。このような膜貫通領域を形成するセグメントを本明細書ではシグナルアンカー(SA)と呼び、特にシグナルペプチドと同様の経路で膜を透過するシグナルアンカーをTypeII型シグナルアンカー（SA-II）と呼ぶ。また、本明細書では、SRPにより認識を受けてトランスロコンを通過して膜を透過した後最終的に切除されるセグメントをシグナルペプチド（SP）、シグナルペプチドと同様の経路を通って膜を透過するが切除を受けないセグメントをTypeII型シグナルアンカー（SA-II）、そしてシグナルペプチドとシグナルアンカーを含む機能性セグメント全体を総称してシグナル配列と定義する。

【００１１】シグナルペプチドを保有する水溶性タンパク質の場合には、シグナルペプチドが切除された後、タンパク質本体が膜を透過し、生体膜を挟んで反対側の空間へと分泌される。一方、シグナルペプチドを有する膜タンパク質の場合には、シグナルペプチド領域が切除された後、下流の膜貫通領域のN末端が外側向きで生体膜に挿入されるために、N末端が膜外を向いたトポロジーを有する膜タンパク質となる。

【００１２】このように、シグナルペプチドの役割には次の三つの別々の段階が存在する。

１）細胞質側においてSRPによる認識を受ける

２）トランスロコンによって認識を受ける

３）生体膜透過後、シグナルペプチダーゼによる認識を受け切除される

これらのうちの１）と２）は、シグナルペプチド及びTypeII型シグナルアンカー両者に共通する特徴的機能と考

えられ、３）が特にシグナルペプチドに特有の特徴的機能である。SRPが認識するシグナルペプチドの部位は、シグナルペプチドの持つ疎水性の高いアミノ酸配列の領域であると考えられる。

【００１３】SRPによるシグナルペプチドの認識と同様に、トランスロコンによる認識やシグナルペプチダーゼによるシグナルペプチドの認識の場合にも、配列特異性というよりもむしろこの領域を構成するアミノ酸の側鎖の疎水性、極性といった物理化学的性質によって認識されると考えられる。既存の方法ではシグナルペプチドを特徴付けるはっきりとしたアミノ酸配列パターンが見出せないのも、このような要因に起因すると考えられる。

【００１４】そこで本発明者等は、シグナルペプチドをより高い精度で予測することについて鋭意検討した結果、Kyte-Doolittleの疎水性指標（以下、疎水性指標とする）と、新しく定義した負電荷残基指標によってシグナルペプチド候補領域を抽出し、列挙された候補領域に新しく定義したシグナルペプチド判別指標およびシグナルアンカー判別指標を適用し、併せて候補領域の位置や長さを用いてシグナルペプチドを予測した場合には、従来技術に比して格段に高精度の予測が可能となることを見出し、本発明に到達した。

【００１５】

【発明が解決しようとする課題】従って本発明の第一の目的は、アミノ酸配列が決定されたタンパク質に、シグナルペプチド又はシグナルアンカーの何れかが含まれるか否かを、高い精度で判別するための方法を提供することである。更に本発明の第二の目的は、与えられた任意のアミノ酸配列がシグナルペプチドを含むかどうか、含む場合にはその領域およびシグナルペプチドの切除部位であるCleavage Siteを高い精度で判別するための、コンピュータプログラムを提供することにある。

【００１６】

【課題を解決するための手段】本発明の上記の諸目的は、アミノ酸配列の決定されたタンパク質の判別方法であって、該方法が、

a）タンパク質の構成要素である20種類のアミノ酸それぞれに対して、予め、疎水性指標H、負電荷残基指標NC、シグナルペプチド判別指標SP Index、及びシグナル配列判別指標SS Indexを割り当て、（ｂ）５～９個の何れかの数の一連の窓を有するウィンドウWを用いて被判別タンパク質のアミノ酸配列から一連のアミノ酸配列を抽出し、（ｃ）抽出された前記ウィンドウWと対応する全てのアミノ酸配列について、前記疎水性指標を用いて計算した二回平均疎水性値[[H]]を算出し、（ｄ）得られた[[H]]の値が一定の閾値ｋを越える領域の一連のアミノ酸残基の数が５～２５残基となり、かつその領域内に負電荷残基指標NCが１であるアミノ酸残基を含まないとき、その領域を構成するアミノ酸残基からなるセグメントをシグナル配列の候補領域Ｃとし、（ｅ）複数の候補領域S(i)が得られた場合には、被判別アミノ酸配列のN末端に最も近い候補領域S(i)を最終的な候補領域Ｃとして選択し、その候補領域Ｃの始点及び終点のアミノ酸残基の位置を示す数Xstart、及び、Xend並びに候補領域Ｃを構成するアミノ酸の数Xlengthを求め、（ｆ）候補領域Ｃが得られない場合には被判別タンパク質はシグナルペプチドを含まない水溶性タンパク質であると認定し、（ｇ）候補領域Ｃがある場合にはその平均疎水性値及び該候補領域中の二回平均疎水性値[[H]]の最大ピーク位置Ｐｐを求め、（ｈ）該ピーク位置Ｐｐから被判別アミノ酸配列のN末端に向かって正電荷残基を検索し、最もピーク位置に近い正電荷残基の位置を判別基準位置Ｐｒとし、（ｉ）もし正電荷残基が見出されない場合には被判別アミノ酸配列のN末端をＰｒと定義し、（ｊ）該判別基準位置Ｐｒからアミノ酸配列のＣ末端側へ下流10番目から27番目までの18残基の領域を判別対象領域Ｒとして設定し、（ｋ）判別対象領域Ｒを構成する１８個のアミノ酸残基それぞれに、前記したシグナルペプチド判別指標ＳＰ Index、シグナル配列判別指標SS Indexを割り振り、（ｌ）判別対象領域ＲにおけるSP IndexおよびSS Indexの平均値を算出し、（ｍ）計算された各パラメータを元に、３つの判別式によって、該領域についてシグナルペプチド-シグナルアンカーの二群の判別、（ｎ）シグナルペプチド-シグナル配列無しの二群の判別、（ｏ）シグナルアンカー-シグナル配列無しの二群の判別、と二群の判別を３回行ってそれぞれで判別結果を得、（ｐ）得られた各々の判別結果の組み合わせから該候補領域Ｃがシグナルペプチド、シグナルアンカー、シグナル配列無しの何れに該当するものかを判別することを特徴とするシグナルペプチドの判別方法、及びそのためのコンピュータプログラムによって達成された。

【００１７】

【発明の実施の形態】一般的に、シグナルペプチドは典型的な膜貫通領域と同様の特徴をもっており、特に疎水性が高いアミノ酸残基が頻出するという点で両者は類似している。このため、疎水性だけではシグナルペプチドとシグナルアンカー等の膜貫通セグメントを区別することは困難であるといえる。一方、その疎水性領域の長さが膜貫通セグメントのそれと比較して短い場合や、極性のアミノ酸残基を多く含み全体として比較的親水的なセグメントであるものも多い。このような特徴をもつシグナルペプチドの場合には、水溶性タンパク質の配列中に散在して見られる比較的短めの疎水性セグメントと区別することが困難である。

【００１８】しかしながら、シグナルペプチドが細胞質側で生合成されてから膜を透過して切除されるまでの一連の流れを考慮したとき、その第一段階となるSRP（SignalRecognition Particle：シグナル認識粒子）によるシグナル配列認識に伴うシグナル配列とSRPとの相互作用は疎水性相互作用に基づいており、この段階では、シ

グナルペプチドと膜貫通セグメントであるシグナルアンカーは区別されていない。従って、配列の疎水性を基に、シグナルペプチドとシグナルアンカーを含む候補領域を列挙することは妥当であると考えられる。しかしながら、SRPはシグナル配列認識の段階で水溶性タンパク質の配列を捕捉しないことから明らかなように、シグナル配列（シグナルペプチド＋シグナルアンカー）と水溶性タンパク質上の疎水性セグメントを分ける要素が存在する。

【００１９】上記の要素は、本発明者等の解析の結果、シグナル配列のもつ疎水性領域には、極性の、特に負電荷を有するアミノ酸残基が現れない連続した領域があるのに対して、シグナル配列を有さない水溶性タンパク質の場合には、たとえN末端近傍に疎水性領域があったとしても、そこには負電荷残基が無秩序に分布するという相違点のあることが明らかとなった。言い換えれば、負電荷残基が存在しないことがSRPの認識を受けるひとつの条件であると言える。そこでこの負電荷残基の効果を、候補領域を列挙する際に取り入れるために、新たに負電荷残基指標NCを作成した。表１にこのパラメータを示す。尚、三文字のアルファベット表記は20種類のアミノ酸の3文字表記、カッコ内のアルファベットは三文字表記のアミノ酸を一文字表記する場合の記号である。

【表１】

	疎水性指標Ｈ	負電荷残基指標ＮＣ
Ala(A)	1.8	0
Cys(C)	2.5	0
Asp(D)	-3.5	1
Glu(E)	-3.5	1
Phe(F)	2.8	0
Gly(G)	-0.4	0
His(H)	-3.2	0
Ile(I)	4.5	0
Lys(K)	-3.9	0
Leu(L)	3.8	0
Met(M)	1.9	0
Asn(N)	-3.5	0
Pro(P)	-1.6	0
Gln(Q)	-3.5	0
Arg(R)	-4.5	0
Ser(S)	-0.8	0
Thr(T)	-0.7	0
Val(V)	4.2	0
Trp(W)	-0.9	0
Tyr(Y)	-1.3	0

【００２０】上記のアルゴリズムによって候補領域を列挙した次の段階は、シグナルペプチドと膜貫通セグメント（シグナルアンカー）の区別である。今、仮に任意のアミノ酸配列が与えられたと仮定すると、列挙された候補領域には、シグナルペプチド、膜貫通セグメント、シグナル配列をもたない水溶性タンパク質の疎水性セグメントが含まれると考えられるが、これらをどのように判別するかの問題である。シグナルペプチドとシグナルアンカーの最大の違いは膜透過後の切除プロセスの有無である。

【００２１】シグナルペプチドの切除部位であるCleavage Siteには、それを特徴付ける配列モチーフのような明確なパターンは存在せず、Cleavage SiteのN末端側にシグナルペプチド及びシグナルアンカーに共通的に見られる疎水性セグメントが存在することから、シグナルペプチドとシグナルアンカーを特徴付ける要素は、疎水性セグメントのC末端からCleavage Siteを跨いで全アミノ酸配列のC末端側に至る領域に存在すると考えることができる。そこで①シグナルペプチドのCleavageSiteを挟んだ前後10残基ずつ計20残基の領域、②シグナルアンカーの膜貫通領域のC末端側境界を挟んだ前後10残基ずつ計20残基の領域、③シグナル配列をもたない水溶性タンパク質の疎水性領域であるC末端側境界を挟む前後10残基ずつ計20残基；という３つの領域について、それぞれアミノ酸残基の出現傾向を解析し、どのようなアミノ酸残基が頻出するのかを調べ、これをもとにシグナルペプチド判別指標SP Index、およびシグナル配列判別指標SS-Indexを作成した。元となったアミノ酸残基の出現傾向には、タンパク質の由来生物種について真核生物（Eukaryote)と原核生物(Prokaryote)という大きな区分で差異が認められたため、SP-Index及び SS-Indexについて、真核生物由来のアミノ酸配列に適用するための指標と、原核生物由来のアミノ酸配列に適用するための指標とを別々に作成した。それぞれの指標の値については表２に示した。SP-Index、SS-Index両者を総称して以後SSインデックスと呼称する。

【表２】

11　　12

	SP Index-Eukaryote	SP Index-Prokaryote	SS Index-Eukaryote	SS Index-Prokaryote
Ala(A)	1.35	2.40	1.50	2.29
Cys©	1.73	0.58	2.07	1.12
Asp(D)	2.02	2.40	0.65	0.72
Glu(E)	2.59	2.27	0.71	0.61
Phe(F)	0.55	0.62	0.96	1.10
Gly(G)	1.23	1.47	1.02	1.04
His(H)	0.99	0.71	0.87	0.48
Ile(I)	0.42	0.40	0.67	0.61
Lys(K)	1.01	1.01	0.47	0.63
Leu(L)	0.79	0.52	1.59	0.92
Met(M)	0.52	0.74	0.63	0.92
Asn(N)	1.03	1.19	0.75	0.89
Pro(P)	1.94	1.31	1.25	0.94
Gln(Q)	1.25	1.06	1.19	0.92
Arg (R)	1.06	0.39	0.65	0.24
Ser(S)	1.40	1.24	1.41	1.38
Thr(T)	0.98	1.29	1.18	1.59
Val(V)	0.73	0.85	0.97	1.24
Trp(W)	0.98	0.48	1.53	0.58
Tyr(Y)	0.60	0.26	0.72	0.31

【００２２】以下、本発明の判別方法を具体的手順に従って説明する。本発明においては、先ずシグナルペプチドを有するか否か判別しようとするタンパク質について、アミノ酸配列と、そのタンパク質が真核生物由来のものか、原核生物由来のものであるかの情報を与える。尚、被判別タンパク質が真核生物由来のものか原核生物由来のものかの情報は、後述する如く、SP Index及びSS Indexを割り振る前に与えれば良く、必ず初めに与えなくてはならないというものではない。そこで、先ず与えられたアミノ酸配列を構成する各アミノ酸に対し、表１によって予め設定されている疎水性指標Hを割り当てる。次に、例えば7残基の、連続するアミノ酸残基に当てはめることのできるウィンドウWを用いて、下記（１）式によって一回平均疎水性値[H]を計算する。

$$[H(i)] = \left[\sum_{j=i-W/2}^{i+W/2} H(j) \right] / W \qquad (1)$$

但し、ｉはウィンドウの中心アミノ酸の位置を示す。上記の計算は、タンパク質を構成するアミノ酸鎖のＮ末端側から例えば７残基ウィンドウを１残基毎にずらしながら全ての単位について計算する。一通りＣ末端側まで適用し終えた後、下記（２）式によって二回平均疎水性値[[H]]を計算する。

$$[[H(i)]] = \left[\sum_{j=i-W/2}^{i+W/2} [H(j)] \right] / W \qquad (2)$$

【００２３】得られた[[H]]を用いて疎水性プロファイルを作成し、該プロファイルにおいて、[[H]]が連続して閾値kを超えるアミノ酸残基の数がＬよりも大である部分を、シグナル配列の候補領域セグメントS(i)とする。但しＬは５～２５から選択される何れかのアミノ酸残基の数、即ち長さであり、好ましくは８～１０から選択される整数である。尚、Ｌを９とした場合には、前記ｋとして０を設定することが好ましい。

【００２４】上記の如くして抽出された候補領域セグメントS(i)の各セグメントについて、領域を構成する各アミノ酸に負電荷残基指標NCを割り振る。NCが連続して0であるアミノ酸残基を候補領域として残し、NCが１の残基によって分割されたS(ｉ）の各分割領域の長さがLを超えない場合には、その領域を候補から排除する。

【００２５】以上の操作をしても１つも候補領域が列挙されなかった場合には、被判別タンパク質はシグナル配列を持たない水溶性タンパク質と判定され、判定操作（プログラム）は終了する。また、列挙された候補領域がアミノ酸配列のＮ末端から100残基目よりも後にのみ現れる場合には、シグナル配列無しとして判定操作は終了する。従ってこれらのタンパク質については以降の演算はされず、別の被判別タンパク質について、初めから本判定操作（プログラム）が実行される。一方、複数の候補領域S(i)が列挙された場合には、最もＮ末端側に現れたS(i)を候補領域Ｃとして採用する。

【００２６】次に、シグナルペプチドの判別に必要なパラメータを、次のようにして候補領域より抽出する。下記式（３）によって、Ｃの平均疎水性値を求めると共に、候補領域Ｃの始点（Xstart）および終点（Xend）を求める。

$$[\overline{H}] = \left[\sum_{i=X_{start}}^{X_{end}} [H(i)] \right] / X_{length} \qquad (3)$$

また候補領域Ｃの長さ即ち候補領域Ｃを構成するアミノ酸の数をXlengthとする。また候補領域Ｃ中で、二回平均疎水性値［［Ｈ］］が最も大きい位置を疎水性ピーク位置Ｐｐとする。次いで、Ｐｐから被判別アミノ酸配列のＮ末端側に遡って最初に現れた正電荷残基の位置を基準位置Ｐｒとする。もし正電荷残基が見つからなかった場合には、被判別アミノ酸配列のＮ末端を基準位置Ｐｒとする。

【００２７】求められた基準位置Ｐｒからアミノ酸配列のＣ末端側へ、下流10番目から27番目までの18残基の領

13

域を判別対象領域Ｒとし、判別対象領域Ｒを構成する１８残基のアミノ酸残基それぞれに、シグナルペプチド判別指標SP Index及びシグナル配列判別指標SS-Indexを割り振り、下記式（４）及び（５）によって判別対象領域におけるSP IndexおよびSS Indexの平均値を算出する。

$$\left[\overline{X}_{SP}\right]=\left[\sum_{i=Pt+10}^{Pt+27}\left[SP(i)\right]\right]/18 \qquad (4)$$

$$\left[\overline{X}_{SS}\right]=\left[\sum_{i=Pt+10}^{Pt+27}\left[SS(i)\right]\right]/18 \qquad (5)$$

14

【００２８】次いで、下記判別式（６）によってシグナルペプチド(SP)-シグナルアンカー(SA)間の二群の判別を行い、同様にして下記判別式（７）によってシグナルペプチド(SP)-非シグナル配列(NS)の二群の判別を、下記判別式（８）によってシグナルアンカー(SA)-非シグナル配列(NS)の二群の判別を行う。下記式中の係数等は下記表３に示した。

$$F_{SP-SA}=b_1X_{start}+b_2X_{end}+b_3X_{length}+b_4\left[\overline{H}\right]+b_5\left[\overline{X}_{SP}\right] \qquad (6)$$

$$F_{SP-NoSignal}=b_6X_{start}+b_7X_{end}+b_8X_{length}+b_9\left[\overline{H}\right]+b_{10}\left[\overline{X}_{SS}\right] \qquad (7)$$

$$F_{SA-NoSignal}=b_{11}X_{start}+b_{12}X_{end}+b_{13}X_{length}+b_{14}\left[\overline{H}\right]+b_{15}\left[\overline{X}_{SS}\right] \qquad (8)$$

【表３】

Prokaryote	Eukaryote
$b_1=-10.82$	$b_1=-1.33$
$b_2=10.81$	$b_2=1.32$
$b_3=-10.82$	$b_3=-1.33$
$b_4=-0.50$	$b_4=-0.18$
$b_5=1.65$	$b_5=1.40$
$b_6=-0.52$	$b_6=0.040$
$b_7=0.52$	$b_7=-0.043$
$b_8=-0.48$	$b_8=0.083$
$b_9=1.00$	$b_9=1.04$
$b_{10}=1.46$	$b_{10}=1.09$
$b_{11}=0.15$	$b_{11}=0.13$
$b_{12}=-0.15$	$b_{12}=-0.13$
$b_{13}=0.19$	$b_{13}=0.16$
$b_{14}=0.97$	$b_{14}=0.87$
$b_{15}=-0.37$	$b_{15}=0.24$

【００２９】上記判別式（６）～（８）を用いた三回の２群の判別の結果の組み合わせにより、最終的な判別結果は下記の通りとなる。

① SP ∩ SP ∩ SA → SP
② SP ∩ SP ∩ NS → SP
③ SP ∩ NS ∩ SA → NS
④ SP ∩ NS ∩ NS → NS
⑤ SA ∩ SP ∩ SA → 保留
⑥ SA ∩ SP ∩ NS → 保留
⑦ SA ∩ NS ∩ NS → NS
⑧ SA ∩ NS ∩ SA → SA

【００３０】上記⑤および⑥では三群の判別で矛盾が生じるため例外処理を行う。当てはまらない場合も多いが、シグナルペプチド切除部位であるCleavage Siteを特徴づけるルールとして広く認知されている、前述した(-1,-3)ルールに即したパターン検索を行い、もし当てはまれば、これを手がかりにしてSPであると最終判断を下し、ない場合にはNSとする。

【００３１】本発明のコンピュータプログラムは、被判別タンパク質のアミノ酸配列について以上のデータ入力と演算を行わせ、必要に応じて、得られた結果をモニター及び/又はプリンターによって出力する。上記の判別を実施する為のコンピュータプログラムのフローチャートは、図１および図２に示される通りである。本発明のコンピュータプログラムは、C言語等を用いて記載することができる。以下に更に詳述する。

【００３２】プログラムを起動し、被判別タンパク質を形成するアミノ酸配列を入力し（STP101）、被判別タンパク質の由来生物種が真核生物（Eukaryote）と原核生物（Prokaryote）のどちらに属するのかを選択して入力する（STP102）。一方、前記表１～３のデータ及び（１）～（８）の数式等を予め記憶部に格納しておく。次に、入力された全てのアミノ酸残基に、予め記憶部に格納されている前記表１のデータのうち、該当する疎水性指標Hの値を割り当てる（STP103）。尚、被判別タンパク質の由来生物種についての情報入力は、ＳＴＰ１０２に限定されるものではなく、後述するＳＴＰ１１２の前であればどの段階であっても良い。

【００３３】所定のウィンドウを、アミノ酸配列の端から１残基ごとにずらしながら、各ウィンドウに対応するアミノ酸列を抽出し、抽出された全てのアミノ酸配列に対して前記（１）式に従って一回平均疎水性値[H]を求め、ついで前記（２）式に従って二回平均疎水性値[[H]]を求めて疎水性プロファイルを作成する（STP104）。

【００３４】[[H]]が連続して閾値kを超えるアミノ酸残基の数がLよりも大であるものを、シグナル配列の候補領域セグメントS(i)とする。但しLは5～25から選択された何れかのアミノ酸残基の数、即ち長さであり、好ましくは８～１０から選択される整数である。尚、Lを９とした場合には、前記kとして０を設定することが好まし

い。抽出された候補S(i)の各セグメントについて、領域を構成する各アミノ酸に負電荷残基指標NCを割り振る。NCが連続して0である領域の長さがＬ以上であれば、候補領域Ｃとして残し、NCが１の残基によって分割されたS(ｉ）の各分割領域の長さがLを超えない場合にはその領域を候補から排除する（STP105)。

【００３５】ここで、１つも候補領域Ｃが列挙されなかった場合には被判別タンパク質はシグナル配列を持たない水溶性タンパク質と判定され、判定操作（プログラム）は終了する。次に、列挙された候補領域Ｃが100残基目よりも後に現れたか否かを判定し、100残基よりも後にのみ現れる場合にはシグナル配列無しとして、判定操作は終了する（STP106）。

【００３６】一方、S（ｉ）が一つに絞られた場合はそれを候補領域Ｃとして、また複数の候補領域が列挙された場合には最もＮ末端側に現れたS(i)を候補領域Ｃとして採用する（STP107）。従って候補領域Ｃは、最終的に１個に絞られる。得られた候補領域Ｃにおける領域の平均疎水性値を前記（３）式によって求めると共に、候補領域Ｃ中における疎水性値の最大ピーク位置Pp、候補領域始点Xstart、候補領域終点Xend及び候補領域長さXlengthの各パラメータを求める（STP108）。

【００３７】候補領域Cの最大ピーク位置からN末端側へ遡って正電荷残基を検索し（STP109）、もっともPpに近いところで見つかった正電荷残基の位置をPrとし、正電荷残基が見つからなかった場合にはN末端をPrとする（STP110）。次に、PrからC末端側へ下流10番目から27番目までの18残基の領域を判別対象領域Ｒと決定する（STP111）。判別対象領域Ｒを構成する１８残基のアミノ酸残基それぞれに、予め記憶部に格納してある表２のシグナルペプチド判別指標SP-Index及びシグナル配列判別指標SS-Indexを割り振って、前記式（４）及び（５）によって判別対象領域におけるSP IndexおよびSS Indexの平均値を算出する（STP112）。次に、上記候補領域に(-1,-3）ルールに即したパターンが見出されるかどうかを検索する。もし見つかった場合にはフラグMotifを立てる。ここでの結果は、シグナルペプチド判別における、後の例外処理でのみ使われる（STP113）

【００３８】前記判別式（６）によってSP-SA間の二群の判別を行い、同様にして、前記判別式（７）によってSP-NS間の二群の判別を行う。更に、前記判別式（８）によってSA-NS間の二群の判別を行う（STP114）。これらの式中における係数等は予め記憶部に記録されている前記表３から読み出して使用される。上記判別結果は、シグナルペプチド（SP）、シグナルアンカー（SA）、又はシグナル配列なし(NS)として得られる。具体的には、前式（６）の計算結果F_{SP-SA}がシグナルペプチド－シグナルアンカー（ＳＰ－ＳＡ）を判別する閾値T_{SP-SA}以上である場合には、ＳＰ－ＳＡ判別の結果をＳＰとし、T_{SP-SA}未満である場合にはＳＰ－ＳＡ判別の結果をＳＡとする。この場合のT_{SP-SA}には、真核生物の場合には－０．３８２を、原核生物の場合には－９．９８を設定することが好ましい。前式（７）の計算結果F_{SP-NS}が、シグナルペプチド－シグナル配列なし（ＳＰ－ＮＳ）を判別する閾値T_{SP-NS}以上である場合には、ＳＰ－ＮＳ判別の結果をＳＰとし、T_{SP-NS}未満である場合にはＳＰ－ＮＳ判別の結果をＮＳ（シグナル配列なし）とする。この場合のT_{SP-NS}には、真核生物の場合には３．００を、原核生物の場合には２．７４を設定する。前式（８）の計算結果F_{SA-NS}がシグナルアンカー－シグナル配列なし（ＳＡ－ＮＳ）を判別する閾値T_{SA-NS}以上である場合には、ＳＡ－ＮＳ判別の結果をＳＡとし、T_{SA-NS}未満である場合にはＳＡ－ＮＳ判別の結果をＮＳ（シグナル配列なし）とする。この場合のT_{SA-NS}には、真核生物の場合には２．００を、原核生物の場合には２．３０を設定することが好ましい。

【００３９】上記の判別結果を用い、①SP-SAの判別がSPでSP-NSの判別がSP、かつSA-NSの判別がSAの場合には、判別結果をSPとし（STP115)、②SP-SAの判別がSPでSP-NSの判別がSP、かつSA-NSの判別がNSの場合には、判別結果をSPとする（STP116)。また、③SP-SAの判別がSPでSP-NSの判別がNS、かつSA-NSの判別がSAの場合には判別結果をNSとし（STP117)、④SP-SAの判別がSPでSP-NSの判別がNS、かつSA-NSの判別がNSの場合ににおける判別結果をNSとする（STP118)。⑤SP-SAの判別がSAでSP-NSの判別がSP、かつSA-NSの判別がSAの場合には、既に記憶部に格納されているMotifフラグをチェックする例外処理を行い（STP119）、その結果がtrueであればSPとし、falseの場合にはSAとする（STP120)。同様に、⑥SP-SAの判別がSAでSP-NSの判別がSP、かつSA-NSの判別がNSである場合にも、Motifフラグをチェックする例外処理を行い（STP121）、その結果がtrueであればSPとし、falseの場合にはNSとする（STP122)。⑦SP-SAの判別がSAで、SP-NSの判別がNS、且つSA-NSの判別がNSの場合には判別結果はNSとし(STP123)、⑧SP-SAの判別がSAで、SP-NSの判別がNS且つSA-NSの判別がSAの場合には判別結果はSAとする(STP124)。これら全ての演算が完了すると、判別結果と候補領域の終点を出力して(STP125)、プログラムは終了する。

【００４０】

【発明の効果】本発明によれば、一次配列が解明されたタンパク質について、94%以上の正答率で、迅速に、被判別タンパク質がシグナルペプチドを持つか否かを判定することができる。

【図面の簡単な説明】

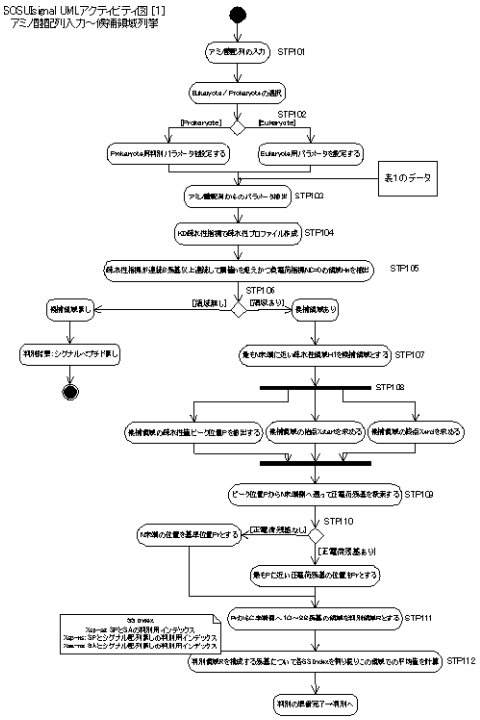
【図１】本発明のコンピュータプログラムのフローチャートＳＴＰ101からＳＴＰ112までの一例である。

【図２】図1に続くＳＴＰ１１３以降のフローチャート

17 18

である。

【図１】



【図２】