(19)日本国特許庁(JP)

再公表特許(A1)

(11)国際公開番号
WO2004/051546

発行日　平成18年4月6日(2006.4.6)

(43)国際公開日　平成16年6月17日(2004.6.17)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 19/00	(2006.01)	G06F 19/00	600	5B075
G06F 17/30	(2006.01)	G06F 17/30	170F	
		G06F 17/30	210D	
		G06F 17/30	350C	

審査請求 未請求　予備審査請求 未請求　（全 51 頁）

出願番号	特願2004-556854 (P2004-556854)
(21)国際出願番号	PCT/JP2003/015245
(22)国際出願日	平成15年11月28日(2003.11.28)
(31)優先権主張番号	特願2002-348678 (P2002-348678)
(32)優先日	平成14年11月29日(2002.11.29)
(33)優先権主張国	日本国(JP)

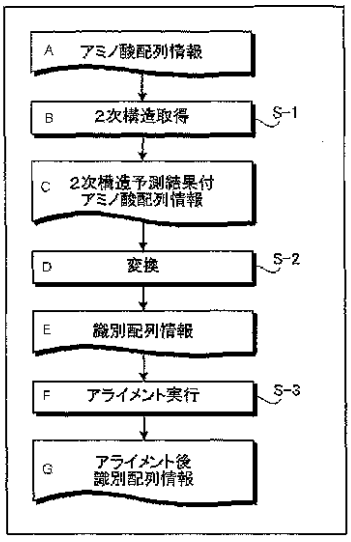
(71)出願人　502433955
株式会社インシリコサイエンス
東京都大田区東雪谷二丁目１５番９号
(74)代理人　100089118
弁理士　酒井　宏明
(74)代理人　100113103
弁理士　香島　拓也
(72)発明者　梅山　秀明
千葉県浦安市美浜１丁目７番１００２号
(72)発明者　岩舘　満雄
埼玉県羽生市大字尾崎２６番地５

Fターム(参考)　5B075 ND02　NR12　QM08　UU19

最終頁に続く

(54)【発明の名称】配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体

(57)【要約】
本発明は、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得し、取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、変換された複数の識別配列情報に対してアライメントを実行する。



A...AMINO ACID SEQUENCE DATA
B...ACQUISITION OF SECONDARY STRUCTURE
C...AMINO ACID SEQUENCE DATA WITH SECONDARY STRUCTURE ESTIMATION RESULT
D...CONVERSION
E...IDENTIFICATION SEQUENCE DATA
F...ALIGNMENT PERFORMANCE
G...POST-ALIGNMENT IDENTIFICATION SEQUENCE DATA

【特許請求の範囲】
【請求項１】
アミノ酸配列情報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得手段と、
上記二次構造取得手段にて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換手段と、
上記変換手段にて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行手段と、
を備えたことを特徴とする配列情報処理装置。
【請求項２】
上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納手段、
を備え、
上記変換手段は、
上記二次構造・類似性分類情報格納手段にて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換手段、
をさらに備えたことを特徴とする請求の範囲第１項に記載の配列情報処理装置。
【請求項３】
上記分類情報参照変換手段により変換された上記識別配列情報を格納する識別配列情報格納手段、
を備え、
上記変換手段は、
上記識別配列情報格納手段にて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索手段、
をさらに備えたことを特徴とする請求の範囲第１項または第２項に記載の配列情報処理装置。
【請求項４】
上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納手段、
および／または、
上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納手段、
を備え、
上記アライメント実行手段は、
上記二次構造置換マトリックス格納手段にて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納手段にて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行手段、
をさらに備えたことを特徴とする請求の範囲第１項から第３項のいずれか一つに記載の配列情報処理装置。
【請求項５】
上記置換マトリックス基準アライメント実行手段は、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行すること、
を特徴とする請求の範囲第４項に記載の配列情報処理装置。

【請求項６】
上記アライメント実行手段にてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換手段、
を備えたことを特徴とする請求の範囲第１項から第５項のいずれか一つに記載の配列情報処理装置。
【請求項７】
上記二次構造・類似性分類情報は、
上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類すること、
を特徴とする請求の範囲第１項から第６項のいずれか一つに記載の配列情報処理装置。
【請求項８】
上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であること、
を特徴とする請求の範囲第１項から第７項のいずれか一つに記載の配列情報処理装置。
【請求項９】
アミノ酸配列情報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得ステップと、
上記二次構造取得ステップにて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換ステップと、
上記変換ステップにて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行ステップと、
を含むことを特徴とする配列情報処理方法。
【請求項１０】
上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納ステップ、
を含み、
上記変換ステップは、
上記二次構造・類似性分類情報格納ステップにて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換ステップ、
をさらに含むことを特徴とする請求の範囲第９項に記載の配列情報処理方法。
【請求項１１】
上記分類情報参照変換ステップにより変換された上記識別配列情報を格納する識別配列情報格納ステップ、
を含み、
上記変換ステップは、
上記識別配列情報格納ステップにて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索ステップ、
をさらに含むことを特徴とする請求の範囲第９項または第１０項に記載の配列情報処理

方法。
【請求項１２】
上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納ステップ、
　および／または、
　上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納ステップ、
　を含み、
　上記アライメント実行ステップは、
　上記二次構造置換マトリックス格納ステップにて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納ステップにて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行ステップ、
　をさらに含むことを特徴とする請求の範囲第９項から第１１項のいずれか一つに記載の配列情報処理方法。
【請求項１３】
上記置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行すること、
　を特徴とする請求の範囲第１２項に記載の配列情報処理方法。
【請求項１４】
上記アライメント実行ステップにてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換ステップ、
　を含むことを特徴とする請求の範囲第９項から第１３項のいずれか一つに記載の配列情報処理方法。
【請求項１５】
上記二次構造・類似性分類情報は、
　上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類すること、
　を特徴とする請求の範囲第９項から第１４項のいずれか一つに記載の配列情報処理方法。
【請求項１６】
上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であること、
　を特徴とする請求の範囲第９項から第１５項のいずれか一つに記載の配列情報処理方法。
【請求項１７】
アミノ酸配列情報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得ステップと、
　上記二次構造取得ステップにて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記

アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換ステップと、
上記変換ステップにて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行ステップと、
を含む配列情報処理方法をコンピュータに実行させることを特徴とするプログラム。
【請求項１８】
上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納ステップ、
を含み、
上記変換ステップは、
上記二次構造・類似性分類情報格納ステップにて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換ステップ、
をさらに含むことを特徴とする請求の範囲第１７項に記載のプログラム。
【請求項１９】
上記分類情報参照変換ステップにより変換された上記識別配列情報を格納する識別配列情報格納ステップ、
を含み、
上記変換ステップは、
上記識別配列情報格納ステップにて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索ステップ、
をさらに含むことを特徴とする請求の範囲第１７項または第１８項に記載のプログラム。
【請求項２０】
上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納ステップ、
および／または、
上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納ステップ、
を含み、
上記アライメント実行ステップは、
上記二次構造置換マトリックス格納ステップにて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納ステップにて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行ステップ、
をさらに含むことを特徴とする請求の範囲第１７項から第１９項のいずれか一つに記載のプログラム。
【請求項２１】
上記置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行すること、
を特徴とする請求の範囲第２０項に記載のプログラム。
【請求項２２】
上記アライメント実行ステップにてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換ステップ、
を含むことを特徴とする請求の範囲第１７項から第２１項のいずれか一つに記載のプログラム。
【請求項２３】
上記二次構造・類似性分類情報は、
上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分

類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類すること、

を特徴とする請求の範囲第１７項から第２２項のいずれか一つに記載のプログラム。

【請求項２４】

上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であること、

を特徴とする請求の範囲第１７項から第２３項のいずれか一つに記載のプログラム。

【請求項２５】

上記請求の範囲第１７項から第２４項のいずれか一つに記載されたプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【技術分野】

本発明は、配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体に関するものであり、特に、アミノ酸配列の二次構造および／またはアミノ酸の類似性を加味したアライメントを得ることができる、配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体に関する。

【背景技術】

現在、タンパク質の二次構造をそのアミノ酸配列から予測できるようになっている。その代表的な二次構造予測ソフトであるＰＳＩＰＲＥＤ（ｐｒｏｔｅｉｎ ｓｔｒｕｃｔｕｒｅ ｐｒｅｄｉｃｔｉｏｎ）は高速で精度が高く、ＰＳＩ－ＢＬＡＳＴで作成されるマルチプルアライメントを用いて進化の過程で機能を維持するために保存された構造や、残基置換の制約といったことも考慮して二次構造を予測することができる。

また、ホモロジーモデリング法とは、立体構造既知のタンパク質に関する情報を利用し、目的タンパク質とのアライメントから立体構造を作成する方法である。このホモロジーモデリングを用いて作成されるモデルの精度は近年目覚しく向上しているが、未だ解決すべき問題点も多い。

まず、ホモロジーモデリングにおいて、精度の高いモデルを作成するためには、まず信頼性の高いアライメントを得ることが重要である。任意の配列が与えられたとき、ＰＤＢ（Ｐｒｏｔｅｉｎ Ｄａｔａ Ｂａｎｋ）やＳＣＯＰ（Ｓｔｒｕｃｔｕｒｅ Ｃｌａｓｓｉｆｉｃａｔｉｏｎ ｏｆ Ｐｒｏｔｅｉｎ）のような立体構造データベースから配列類似性の高いタンパク質を単数、または複数選び出し、アライメントを与えるホモロジー検索方法として、ＦＡＳＴＡ、ＰＳＩ－ＢＬＡＳＴ（Ｐｏｓｉｔｉｏｎ－Ｓｐｅｃｉｆｉｃ Ｉｔｅｒａｔｅｄ ＢＬＡＳＴ）、ＲＰＳ－ＢＬＡＳＴ（Ｒｅｖｅｒｓｅ ＰＳＩ－ＢＬＡＳＴ）、ＩＭＰＡＬＡ（Ｉｎｔｅｇｒａｔｉｎｇ Ｍａｔｒｉｘ Ｐｒｏｆｉｌｅｓ Ａｎｄ Ｌｏｃａｌ Ａｌｉｇｎｍｅｎｔｓ）などのアライメントソフトがある。

ここで、ホモロジー検索を行なうと、一般的に各アライメントについてホモロジーとｅ値を得ることができる。ここで、「ホモロジー」とは残基一致度（％）のことである。また、「ｅ値（Ｅｘｐｅｃｔｅｄ Ｖａｌｕｅ）」とは、データベースにおいて全く偶然に同じスコアになる配列の数の期待値、すなわちそのアライメントのスコアがどの程度まれであるのかを示す指標であり、小さければ小さいほど似た配列は他に見つかりにくく、偶然には見つかりにくいことを表わす。

また、ＦＡＳＴＡは２０種の天然アミノ酸のアルファベット文字配列マッチングを行なうプログラムであり、高ホモロジー（アミノ酸配列の一致度３０％以上、ｅ値０．０１以下に相当）の参照タンパク質とのアライメントを用いてモデルを作成すると信頼性の高い

モデルができるといわれている。一方ＰＳＩ－ＢＬＡＳＴでは、ＦＡＳＴＡと同じように文字列のマッチングを行なうが、文字が一致しているかどうかの情報ではなく、プロファイルと呼ばれる文字の一致の度合い、置換のしやすさなどを文字配列上の部位ごとに置換配列として算出し、さらに何回も繰り返して算出することによりアライメントを最適化する性質をもっている。なお、ＰＳＩ－ＢＬＡＳＴではホモロジーが低い場合であってもｅ値が低ければある程度信頼性のあるアライメントを得ることができるといわれている。

現在、タンパク質の二次構造をそのアミノ酸配列から予測できるようになっている。その代表的な二次構造予測ソフトであるＰＳＩＰＲＥＤ（ｐｒｏｔｅｉｎ ｓｔｒｕｃｔｕｒｅ ｐｒｅｄｉｃｔｉｏｎ）は高速で精度が高く、ＰＳＩ－ＢＬＡＳＴで作成されるマルチプルアライメントを用いて進化の過程で機能を維持するために保存された構造や、残基の置換の制約といったことも考慮して二次構造を予測することができる。

さらに、ホモロジーモデリング法を用いたＦＡＭＳ（Ｆｕｌｌ Ａｕｔｏｍａｔｉｃ Ｍｏｄｅｌｉｎｇ Ｓｙｓｔｅｍ）は、Ｘ線やＮＭＲなどにより実験的に構造が決定されたタンパク質の立体構造（参照タンパク質）をもとにして、立体構造未知のタンパク質（目的タンパク質）の構造をモデリングすることが可能である。ＦＡＭＳではホモロジー検索プログラムなどにより作成されるアライメント（目的タンパク質と参照タンパク質のアミノ酸配列を並置したもの）ファイルを入力としている。

しかしながら、ＦＡＳＴＡ、ＰＳＩ－ＢＬＡＳＴなどの従来のホモロジー検索プログラムにおいては、以下の２つの問題点が指摘されている。

（１）低ホモロジーで高ｅ値の場合、検索が困難であること

ＰＳＩ－ＢＬＡＳＴのように位置ごとに置換配列を作成（プロファイル）し、ホモロジー検索を行なうプログラムはＰＳＳＭ（Ｐｏｓｉｔｉｏｎ Ｓｐｅｃｉｆｉｃ Ｓｃｏｒｅ Ｍａｔｒｉｘ）と呼ばれ、ＦＡＳＴＡのように文字列の一致度のみを考慮してアライメントを行なう従来のプログラムよりも、より感受性が高く遠縁の配列も検索できるといわれている。しかしながら、低ホモロジーで高ｅ値の場合にはＰＳＳＭによるアライメントであっても信頼性は低下する、または検索が不可能という場合があり限界がある。ここで低ホモロジーの場合、より広く遠縁の配列を探索するホモロジー検索手法の開発が望まれていた。

（２）アライメントの信頼性

従来のアライメントプログラムでは、アミノ酸配列の文字情報のみを用いてアライメントを行なっていた。このアライメントでは、タンパク質の立体構造形成の単位であるαヘリックスやβシートなどの二次構造を分断してアライメントを行なってしまうことがある。また、このアライメントを用いてモデリングを行なった場合、疎水性アミノ酸が溶媒に露出してしまい、疎水性相互作用の減少から実際の形と比べて大きく異なったモデルを作成してしまうことがある。このように文字のみを考慮したアライメントでは、特に低ホモロジーの場合その信頼性が疑わしいことがある。この問題を解消するため、二次構造情報や残基の疎水性情報をも考慮したアライメントプログラムの開発が望まれていた。

また、上記で述べたホモロジー値、ｅ値はホモロジー検索プログラムが出力するものであり、アミノ酸配列の文字情報のみから算出したものであって、実際に作成されたモデル構造ついての評価は行なうことができない。そのため、ホモロジー値、ｅ値だけではなく作成したモデルの構造をも同時に考慮したモデルの選定方法の開発が望まれていた。

さらに、ＦＡＭＳでは、入力ファイルであるアライメント次第でモデリングできる領域、長さが決まってしまう。しかし、異なる参照タンパク質によるアライメントを用いると、より長くモデルを作成できる場合がある。このような場合、よい構造であるが領域が短いモデルに、長い領域をモデリングできたモデルの末端構造を貼り付けることによって、自動的にモデリング領域を伸長させる技術が望まれていた。

本発明は上記問題点に鑑みてなされたもので、上記問題点を解決することのできる、各種の方法等を提供することを目的としている。

【発明の開示】

上述した目的を達成するために、本発明にかかる配列情報処理装置は、アミノ酸配列情

報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得手段と、上記二次構造取得手段にて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換手段と、上記変換手段にて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行手段とを備えたことを特徴とする。

この装置によれば、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得（例えば、アミノ酸配列情報に対応する立体構造が未知の場合には、既存の二次構造予測プログラムなどを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラムなどを用いて取得）し、取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（例えば、各アミノ酸の疎水性情報等のアミノ酸毎の性質に関する類似性等）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、変換された複数の識別配列情報に対してアライメントを実行するので、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効果的に実行することができ、従来のホモロジー検索法と比べ、より遠縁の配列の探索が可能となる。

すなわち、本装置によれば、二次構造情報や疎水性情報なども考慮した探索・アライメントを行なうため、従来のアミノ酸の文字のみを考慮したアライメントと比較すると、立体構造も考慮したアライメントが可能となり、より高精度なアライメント作成が可能となる。

本装置のように二次構造情報や疎水性情報などを考慮して作成したアライメントを用いると、疎水エネルギーが安定化し、また二次構造が分断していないモデルを作成でき、モデルとしても真実構造に近いものができると考えられる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納手段を備え、上記変換手段は、上記二次構造・類似性分類情報格納手段にて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換手段をさらに備えたことを特徴とする。

これは二次構造・類似性分類情報格納手段および変換手段の一例を一層具体的に示すものである。この装置によれば、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納し、変換手段は、二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換するので、二次構造・類似性分類情報を予め作成して格納しておき、変換時に当該情報を参照することにより、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記分類情報参照変換手段により変換された上記識別配列情報を格納する識別配列情報格納手段を備え、上記変換手段は、上記識別配列情報格納手段にて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索手段をさらに備えたことを特徴とする。

これは識別配列情報格納手段および変換手段の一例を一層具体的に示すものである。この装置によれば、変換された識別配列情報を格納し、変換手段は、格納された識別配列情

報から、アミノ酸配列情報に対応する識別配列情報を検索するので、変換された識別配列情報をデータベースなどに予め格納しておき、変換時に当該データベースなどを参照することにより、変換処理を効率化、高速化することができるようになる。

また、これにより、例えば公共のデータベース（例えばＰＤＢなど）に登録されたアミノ酸配列情報の二次構造を既存の二次構造判定プログラムなどを用いて判定した後、当該アミノ酸配列情報を本装置により二次構造・類似性分類情報に基づいて変換して作成した識別配列情報を予めデータベースに格納して利用することができるようになり、取得したアミノ酸配列情報に対応する識別配列情報を効率よく検索することができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納手段、および／または、上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納手段を備え、上記アライメント実行手段は、上記二次構造置換マトリックス格納手段にて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納手段にて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行手段をさらに備えたことを特徴とする。

これは二次構造置換マトリックス格納手段、類似性置換マトリックス格納手段およびアライメント実行手段の一例を一層具体的に示すものである。この装置によれば、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックス、および／または、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納し、アライメント実行手段は、格納された二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、置換マトリックスを予め作成して格納しておき、アライメント実行時に当該置換マトリックスを参照して各スコア値に置換することにより、アミノ酸配列情報の二次構造および／またはアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したスコア値に置換することにより最適なアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記置換マトリックス基準アライメント実行手段は、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行することを特徴とする。

これは置換マトリックス基準アライメント実行手段の一例を一層具体的に示すものである。この装置によれば、置換マトリックス基準アライメント実行手段は、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、例えば生物学的な知見などに基づいて二次構造置換マトリックスおよび／または類似性置換マトリックスに対して適切な係数を設定して重み付けをすることによって、生物学的な知見などを反映した最適なアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記アライメント実行手段にてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換手段を備えたことを特徴とする。

この装置によれば、アライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換するので、二次構造および類似性を加味してアライメントが実行されたアミノ酸配列情報を効率よく得ることができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記二次構造・類似性分類情報は、上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）から

なる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類することを特徴とする。

これは二次構造・類似性分類情報の一例を一層具体的に示すものである。この装置によれば、二次構造・類似性分類情報は、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類するので、既知の置換マトリックスであるＢＬＯＳＵＭ６２に基づいてアミノ酸を７つの群に分類することによりアミノ酸情報の類似性を加味し、かつ、二次構造を３つに分類することによりアミノ酸配列情報の二次構造を加味して、アライメントが実行された識別配列情報を効率よく得ることができる。

つぎの発明にかかる配列情報処理装置は、上記に記載の配列情報処理装置において、上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であることを特徴とする。

これはアミノ酸の類似性の一例を一層具体的に示すものである。この装置によれば、アミノ酸の類似性は、アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であるので、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる。

また、本発明は配列情報処理方法に関するものであり、本発明にかかる配列情報処理方法は、アミノ酸配列情報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得ステップと、上記二次構造取得ステップにて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換ステップと、上記変換ステップにて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行ステップとを含むことを特徴とする。

この方法によれば、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得（例えば、アミノ酸配列情報に対応する立体構造が未知の場合には、既存の二次構造予測プログラムなどを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラムなどを用いて取得）し、取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（例えば、各アミノ酸の疎水性情報等のアミノ酸毎の性質に関する類似性等）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、変換された複数の識別配列情報に対してアライメントを実行するので、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効果的に実行することができ、従来のホモロジー検索法と比べ、より遠縁の配列の探索が可能となる。

すなわち、本方法によれば、二次構造情報や疎水性情報なども考慮した探索・アライメ

ントを行なうため、従来のアミノ酸の文字のみを考慮したアライメントと比較すると、立体構造も考慮したアライメントが可能となり、より高精度なアライメント作成が可能となる。

本方法のように二次構造情報や疎水性情報などを考慮して作成したアライメントを用いると、疎水エネルギーが安定化し、また二次構造が分断していないモデルを作成でき、モデルとしても真実構造に近いものができると考えられる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納ステップを含み、上記変換ステップは、上記二次構造・類似性分類情報格納ステップにて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換ステップをさらに含むことを特徴とする。

これは二次構造・類似性分類情報格納ステップおよび変換ステップの一例を一層具体的に示すものである。この方法によれば、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納し、変換ステップは、二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換するので、二次構造・類似性分類情報を予め作成して格納しておき、変換時に当該情報を参照することにより、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記分類情報参照変換ステップにより変換された上記識別配列情報を格納する識別配列情報格納ステップを含み、上記変換ステップは、上記識別配列情報格納ステップにて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索ステップをさらに含むことを特徴とする。

これは識別配列情報格納ステップおよび変換ステップの一例を一層具体的に示すものである。この方法によれば、変換された識別配列情報を格納し、変換ステップは、格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索するので、変換された識別配列情報をデータベースなどに予め格納しておき、変換時に当該データベースなどを参照することにより、変換処理を効率化、高速化することができるようになる。

また、これにより、例えば公共のデータベース（例えばＰＤＢなど）に登録されたアミノ酸配列情報の二次構造を既存の二次構造判定プログラムなどを用いて判定した後、当該アミノ酸配列情報を本方法により二次構造・類似性分類情報に基づいて変換して作成した識別配列情報を予めデータベースに格納して利用することができるようになり、取得したアミノ酸配列情報に対応する識別配列情報を効率よく検索することができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納ステップ、および／または、上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納ステップを含み、上記アライメント実行ステップは、上記二次構造置換マトリックス格納ステップにて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納ステップにて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行ステップをさらに含むことを特徴とする。

これは二次構造置換マトリックス格納ステップ、類似性置換マトリックス格納ステップおよびアライメント実行ステップの一例を一層具体的に示すものである。この方法によれば、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックス、および／または、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納し、アライメント実行ステップは、格納された二次構造置

換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、置換マトリックスを予め作成して格納しておき、アライメント実行時に当該置換マトリックスを参照して各スコア値に置換することにより、アミノ酸配列情報の二次構造および／またはアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したスコア値に置換することにより最適なアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行することを特徴とする。

これは置換マトリックス基準アライメント実行ステップの一例を一層具体的に示すものである。この方法によれば、置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、例えば生物学的な知見などに基づいて二次構造置換マトリックスおよび／または類似性置換マトリックスに対して適切な係数を設定して重み付けをすることによって、生物学的な知見などを反映した最適なアライメントを効率よく得ることができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記アライメント実行ステップにてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換ステップを含むことを特徴とする。

この方法によれば、アライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換するので、二次構造および類似性を加味してアライメントが実行されたアミノ酸配列情報を効率よく得ることができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記二次構造・類似性分類情報は、上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類することを特徴とする。

これは二次構造・類似性分類情報の一例を一層具体的に示すものである。この方法によれば、二次構造・類似性分類情報は、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類するので、既知の置換マトリックスであるＢＬＯＳＵＭ６２に基づいてアミノ酸を７つの群に分類することによりアミノ酸情報の類似性を加味し、かつ、二次構造を３つに分類することによりアミノ酸配列情報の二次構造を加味して、アライメントが実行された識別配列情報を効率よく得ることができる。

つぎの発明にかかる配列情報処理方法は、上記に記載の配列情報処理方法において、上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であることを特徴とする。

これはアミノ酸の類似性の一例を一層具体的に示すものである。この方法によれば、アミノ酸の類似性は、アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であるので、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる。

また、本発明はプログラムに関するものであり、本発明にかかる配列情報処理方法をコンピュータに実行させるためのプログラムは、アミノ酸配列情報を取得し、取得された上記アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得ステップと、上記二次構造取得ステップにて取得された上記二次構造および上記アミノ酸配列情報を構成する上記アミノ酸情報に対応するアミノ酸の類似性に基づいて、上記アミノ酸配列情報を構成する各アミノ酸情報を、同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の情報として識別するための識別情報に変換することにより、上記アミノ酸配列情報を上記識別情報からなる識別配列情報に変換する変換ステップと、上記変換ステップにて変換された複数の上記識別配列情報に対してアライメントを実行するアライメント実行ステップとを含むことを特徴とする。

このプログラムによれば、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得（例えば、アミノ酸配列情報に対応する立体構造が未知の場合には、既存の二次構造予測プログラムなどを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラムなどを用いて取得）し、取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（例えば、各アミノ酸の疎水性情報等のアミノ酸毎の性質に関する類似性等）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、変換された複数の識別配列情報に対してアライメントを実行するので、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効果的に実行することができ、従来のホモロジー検索法と比べ、より遠縁の配列の探索が可能となる。

すなわち、本プログラムによれば、二次構造情報や疎水性情報なども考慮した探索・アライメントを行なうため、従来のアミノ酸の文字のみを考慮したアライメントと比較すると、立体構造も考慮したアライメントが可能となり、より高精度なアライメント作成が可能となる。

本プログラムのように二次構造情報や疎水性情報などを考慮して作成したアライメントを用いると、疎水エネルギーが安定化し、また二次構造が分断していないモデルを作成でき、モデルとしても真実構造に近いものができると考えられる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記二次構造および上記類似性に基づいて同一の上記二次構造および同一の上記類似性を有する上記アミノ酸情報を同一の上記識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納ステップを含み、上記変換ステップは、上記二次構造・類似性分類情報格納ステップにて格納された上記二次構造・類似性分類情報に基づいて、上記アミノ酸配列情報を上記識別配列情報に変換する分類情報参照変換ステップをさらに含むことを特徴とする。

これは二次構造・類似性分類情報格納ステップおよび変換ステップの一例を一層具体的に示すものである。このプログラムによれば、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納し、変換ステップは、二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換するので、二次構造・類似性分類情報を予め作成して格納しておき、変換時に当該情報を参照することにより、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効率よく得ることができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記分類情報参照変換ステップにより変換された上記識別配列情報を格納する識別配列情報格納ステップを含み、上記変換ステップは、上記識別配列情報格納ステップにて格納された上記識別配列情報から、上記アミノ酸配列情報に対応する上記識別配列情報を検索する識別配列情報検索ステップをさらに含むことを特徴とする。

これは識別配列情報格納ステップおよび変換ステップの一例を一層具体的に示すものである。このプログラムによれば、変換された識別配列情報を格納し、変換ステップは、格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索するので、変換された識別配列情報をデータベースなどに予め格納しておき、変換時に当該データベースなどを参照することにより、変換処理を効率化、高速化することができるようになる。

また、これにより、例えば公共のデータベース（例えばＰＤＢなど）に登録されたアミノ酸配列情報の二次構造を既存の二次構造判定プログラムなどを用いて判定した後、当該アミノ酸配列情報を本プログラムにより二次構造・類似性分類情報に基づいて変換して作成した識別配列情報を予めデータベースに格納して利用することができるようになり、取得したアミノ酸配列情報に対応する識別配列情報を効率よく検索することができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納ステップ、および／または、上記識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納ステップを含み、上記アライメント実行ステップは、上記二次構造置換マトリックス格納ステップにて格納された上記二次構造置換マトリックスおよび／または上記類似性置換マトリックス格納ステップにて格納された上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行ステップをさらに含むことを特徴とする。

これは二次構造置換マトリックス格納ステップ、類似性置換マトリックス格納ステップおよびアライメント実行ステップの一例を一層具体的に示すものである。このプログラムによれば、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックス、および／または、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納し、アライメント実行ステップは、格納された二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、置換マトリックスを予め作成して格納しておき、アライメント実行時に当該置換マトリックスを参照して各スコア値に置換することにより、アミノ酸配列情報の二次構造および／またはアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したスコア値に置換することにより最適なアライメントを効率よく得ることができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた上記二次構造置換マトリックスおよび／または上記類似性置換マトリックスに基づいて、上記識別配列情報を置換してアライメントを実行することを特徴とする。

これは置換マトリックス基準アライメント実行ステップの一例を一層具体的に示すものである。このプログラムによれば、置換マトリックス基準アライメント実行ステップは、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、例えば生物学的な知見などに基づいて二次構造置換マトリックスおよび／または類似性置換マトリックスに対して適切な係数を設定して重み付けをすることによって、生物学的な知見などを反映した最適なアライメントを効率よく得ることができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記アライメント実行ステップにてアライメントが実行された上記識別配列情報を構成する上記識別情報を、対応する上記アミノ酸情報に再変換する再変換ステップを含むことを特徴とする。

このプログラムによれば、アライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換するので、二次構造および類似性を加味してアライメントが実行されたアミノ酸配列情報を効率よく得ることができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記二次構造・類似性分類情報は、上記二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、上記類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類することを特徴とする。

これは二次構造・類似性分類情報の一例を一層具体的に示すものである。このプログラムによれば、二次構造・類似性分類情報は、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類するので、既知の置換マトリックスであるＢＬＯＳＵＭ６２に基づいてアミノ酸を７つの群に分類することによりアミノ酸情報の類似性を加味し、かつ、二次構造を３つに分類することによりアミノ酸配列情報の二次構造を加味して、アライメントが実行された識別配列情報を効率よく得ることができる。

つぎの発明にかかるプログラムは、上記に記載のプログラムにおいて、上記アミノ酸の上記類似性は、上記アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であることを特徴とする。

これはアミノ酸の類似性の一例を一層具体的に示すものである。このプログラムによれば、アミノ酸の類似性は、アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であるので、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる。

また、本発明は記録媒体に関するものであり、本発明にかかるコンピュータ読み取り可能な記録媒体は、上記に記載されたプログラムを記録したことを特徴とする。

この記録媒体によれば、上記に記載されたプログラムをコンピュータを利用して実現することができ、これら各方法と同様の効果を得ることができる。

（関数を用いたモデルの選定）

また、本発明は目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法に関するものであり、本発明にかかる方法は、同一の遺伝子のタンパク質について複数の立体構造が算出された場合、アライメントの適合の度合い（ｅ値等）ばかりでなくコンピュータで作成されたモデル構造のタンパク質らしさを表す構造の指標やホモロジーモデリング対象のタンパク質のアミノ酸シーケンスのみから予測する立体構造の指標やモデルを作成するとき参考にする実験で解明された立体構造の指標からなるある目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

また、つぎの発明にかかる方法は、ある目的関数が、アミノ酸シーケンスしかわかっていないタンパク質のアミノ酸残基とそれと文字並べで類似性ありと判定された立体構造が既知の参照タンパク質のアミノ酸残基のアライメントの適合の度合い（ｅ値等）、目的タ

ンパク質の予測二次構造と参照タンパク質の判定二次構造の適合率、モデルタンパク質の立体構造に発生する疎水エネルギー等の評価基準を組み合わせたものである上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

また、つぎの発明にかかる方法は、二次構造の適合率をクエリー配列（目的配列）から二次構造予測ソフトウェア（例えばＰＳＩ－ＰＲＥＤ等）を用いて予測した結果とモデル構造の二次構造判定ソフトウェア（例えばＤＳＳＰ、ＳＴＲＩＤＥ等）を用いて決定した結果とを見比べてモデルが構築された領域の残基数に対して二次構造が適合した残基数の割合で表した場合の上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

また、つぎの発明にかかる方法は、ホモロジーモデリングされたタンパク質の立体構造に対する疎水エネルギーの指標の評価は疎水性残基の側鎖同士が下記の表１の疎水性原子から構成される疎水性残基同士の距離がある閾値（例えば６．６Å）以下であったとき、疎水性相互作用していると見なし溶媒から遮蔽されており接触している状態よりもエネルギー的に安定であると考えて負のエネルギーの指標をカウントしそのエネルギーの指標をモデル立体構造の座標全体における合計で表した時の上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

（表１）

残基名	疎水性原子			
ＡＬＡ	ＣＢ			
ＶＡＬ	ＣＧ１	ＣＧ２		
ＰＨＥ	ＣＢ	ＣＤ１	ＣＤ２	ＣＺ
ＰＲＯ	ＣＧ			
ＭＥＴ	ＣＧ	ＣＥ		
ＩＬＥ	ＣＧ１	ＣＧ２	ＣＤ１	
ＬＥＵ	ＣＢ	ＣＤ１	ＣＤ２	
ＡＳＰ				
ＧＬＵ	ＣＢ			
ＬＹＳ	ＣＢ	ＣＤ		
ＡＲＧ	ＣＧ			
ＳＥＲ				
ＴＨＲ	ＣＧ２			
ＴＹＲ	ＣＢ	ＣＥ１	ＣＥ２	
ＨＩＳ	ＣＢ			
ＣＹＳ	ＣＢ			
ＡＳＮ	ＣＢ			
ＧＬＮ	ＣＢ	ＣＧ		
ＴＲＰ	ＣＢ	ＣＤ２	ＣＺ２	ＣＺ３
ＧＬＹ				

また、つぎの発明にかかる方法は、アライメントの適合率（ｅ値等）、二次構造の適合率、疎水エネルギーの指標等の評価基準を組み合わせる際に、複数種のアライメントソフト（ＦＡＳＴＡ、ＢＬＡＳＴ、ＰＳＩ－ＢＬＡＳＴ、ＲＰＳ－ＢＬＡＳＴ、ＩＭＰＡＬＡ等）で、ｅ値と二次構造適合率、疎水エネルギーの指標のどれを優先させるべきか決定するのに有用な特別な判定関数をもつ上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

また、つぎの発明にかかる方法は、二次構造適合率、疎水エネルギー値の積の符号を逆にした値が高い構造はタンパク質らしい構造を持つとして、得られた多数のモデル構造の評価関数とする上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルと

して算出する方法である。

また、つぎの発明にかかる方法は、複数アライメントそれぞれのｅ値の対数の符号を逆にしたものにある定数を加算した値に複数のアライメント中で最も低いｅ値の対数の符号を逆にしたものにある定数を加算した値で除した値と請求項６の評価関数との積を評価関数とする上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

また、つぎの発明にかかる方法は、ｅ値の対数が０で１００の値を持ち、－３で５０の値を、－１０で４０の値を持つようにして、その間では線形の関係があるような値に１を加算した値をある定数とする上記に記載の目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法である。

（本発明で得られたアライメントからホモロジーモデリングを用いて立体構造を得た後の埋め込み処理）

また、本発明は遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法に関するものであり、本発明にかかる方法は、同一の遺伝子のタンパク質についてある精密なモデル立体構造（以下局所構造１）とそのモデルのアミノ酸配列上の領域を含む別の領域をモデリングしているモデル立体構造（局所構造２）が存在する場合において、局所構造１の立体構造を保持しつつ局所構造２の領域をも包括し且つ局所構造１の精密さを保持した立体構造を算出し、遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法である。

また、つぎの発明にかかる方法は、精密な局所構造１を局所構造２が完全に包括する場合において、立体構造的にそのＲＭＳＤ（Ｒｏｏｔ　Ｍｅａｎ　Ｓｑｕａｒｅ　Ｄｉｓｔａｎｃｅ）を最小化するように重ね合わせて、局所構造１の末端部分と局所構造２における局所構造１の末端部分に相当するアミノ酸残基が距離的に（例えば、８Å以上程度）離れていたとき、精密な局所構造の末端部分１残基を切り短くして再度重ね合せを行う、という操作を前述の距離がある閾値（例えば、８Å）よりも小さくなるまで行い、局所構造１の精密さを最大限に保持し且つ局所構造２のモデル領域までも包括するモデル立体構造を算出する上記に記載の遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法である。

また、つぎの発明にかかる方法は、精密な局所構造１とその一部を含む別領域の局所構造２が存在しているとき、立体構造的にそのＲＭＳＤを最小化するように重ね合わせて、重ね合わせた部分の根平均自乗距離（ＲＭＳＤ）が例えば２Å以上のとき、局所構造２の重ね合わせの領域中、局所構造１と共通領域でもある末端部分をある残基数分（例えば、２残基）切り、短くして再度重ね合せを行う、という操作を前述のＲＭＳＤがある閾値（例えば、２Å）よりも小さくなるまで重ね合せ領域が短くなる操作を行う。ＲＭＳＤが閾値より小さくならずに局所構造２の残基数が４残基になった場合は閾値をある値だけ（例えば、１Å）上げて上記を繰り返し行う。その結果、局所構造１の精密さを最大限に保持し、局所構造２の領域をモデリングしたグローバル立体構造を複数の立体構造中の１つとする上記に記載の遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法である。

また、つぎの発明にかかる方法は、精密な局所構造１とその一部または全部を含む別領域の局所構造２が存在しているとき、また上記に記載の遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法の両方がうまく行かない場合においても、立体構造的にそのＲＭＳＤを最小化するように重ね合わせる長さを一定にして移動させ、重ね合わせた部分の根平均自乗距離（ＲＭＳＤ）がある閾値以下（例えば、２Å）以下で且つ重ねあわせ領域をなるべく大きくするように変化させる、という操作を行い、閾値以下の領域が見つからない場合は閾値を上げて操作を繰り替えし行う。閾値（例えば、４Å）以下にしても領域が見つからない場合は、重ね合わせる長さを二残基短くして、再度閾値を例えば２Åに戻し前述の操作を行う。その結果、局所構造１の精密さを最大限に保持し、局所構造２の領域を包括したモデル立体構造を得て、それを複数の立体構造中の１つとする上記に記載の遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法である。

また、つぎの発明にかかる方法は、ある遺伝子のある領域をモデリングした精密な立体構造モデル１が存在する場合に置いて、二次構造単位に細切れにした立体構造データベースを予め作成しておき、ＰＳＩ－ＢＬＡＳＴ等のアライメントによって、断片の検索を行い、その中で遺伝子のアミノ酸配列をＰＳＩ－ＰＲＥＤ等の二次構造予測によく一致する断片のみを取り出し、ＦＡＭＳ等のホモロジーモデリングによって断片をモデリングし、その断片をモデル１に対して重ね合わせて遺伝子が完全長でない場合は完全長になるまで重ね合わせつづける。その結果、最初のモデル１の精密さを損なわずに且つ遺伝子の完全長のモデルを得て、それを複数の立体構造中の１つとする上記に記載の遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法である。

また、本発明はプログラムに関するものであり、本発明にかかるプログラムは、上記に記載された目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法または遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法をコンピュータに実行させるためのプログラムであることを特徴とする。

このプログラムによれば、当該プログラムをコンピュータに読み取らせて実行することによって、上記に記載された目的関数に従ってもっとも高得点を得た立体構造をモデルとして算出する方法または遺伝子のアミノ酸配列と同じ長さを持つ完全長モデルを得る方法をコンピュータを利用して実現することができ、これら各方法と同様の効果を得ることができる。

また、本発明は記録媒体に関するものであり、本発明にかかる記録媒体は、上記に記載されたプログラムを記録したことを特徴とする。

この記録媒体によれば、当該記録媒体に記録されたプログラムをコンピュータに読み取らせて実行することによって、上記に記載されたプログラムをコンピュータを利用して実現することができ、これら各方法と同様の効果を得ることができる。

【図面の簡単な説明】

第１図は、本発明の基本原理を示す原理構成図であり、第２図は、本発明が適用される本システムの構成の一例を示すブロック図であり、第３図は、本実施形態におけるアミノ酸配列関連情報ファイル１０６ａに格納される情報の一例を示す図であり、第４図は、本実施形態における二次構造・類似性分類情報ファイル１０６ｂに格納される情報の一例を示す図であり、第５図は、本実施形態における識別配列情報データベース１０６ｃに格納される情報の一例を示す図であり、第６図は、本実施形態における二次構造置換マトリックス１０６ｄに格納される情報の一例を示す図であり、第７図は、本実施形態における類似性置換マトリックス１０６ｅに格納される情報の一例を示す図であり、第８図は、本実施形態におけるアライメント後識別配列情報ファイル１０６ｆに格納される情報の一例を示す図であり、第９図は、本実施形態における再変換アミノ酸配列情報ファイル１０６ｇに格納される情報の一例を示す図であり、第１０図は、本発明が適用される本システムの変換部１０２ｄの構成の一例を示すブロック図であり、第１１図は、本発明が適用される本システムのアライメント実行部１０２ｇの構成の一例を示すブロック図であり、第１２図は、本実施形態における本システムのメイン処理の一例を示すフローチャートであり、第１３図は、本発明を有効に適用するための処理を示すフローチャートであり、第１４図は、予測結果ファイルの一例を示す図であり、第１５図は、疎水性相互作用の指標（ｅｎｅｓｏｓｕｉ値）の一例を示す図であり、第１６図は、ＳＣＯＰドメインデータベース（９５％Ｎｏｎ－Ｒｅｄｕｎｄａｎｔ：ｔｏｔａｌ＿ｓｅｑ＿ｄｏｍ＿ａｌｌ＿９５（７４３３ドメイン））中のドメイン構造について、その残基数とｅｎｅｓｏｓｕｉ値との関係を調べた結果をまとめた表を示す図であり、第１７図は、本発明の入力ファイルの一例を示す図であり、第１８図は、本発明の出力ファイルの一例を示す図であり、第１９図は、本発明の処理を示すフローチャート（ペアワイズモードの場合）であり、第２０図は、本発明の処理を示すフローチャート（ホモロジー検索モードの場合）であり、第２１図は、２１種（二次構造３パターン×疎水性７グループ）のアミノ酸を分類する表の一例を示す図であり、第２２図は、二次構造によるスコアマトリックスの一例を示す図であり、第２３図は、アミノ酸類似性によるスコアマトリックスの一例を示す図であり、第２４図は、

高ホモロジーの場合における本発明の入力ファイルの一例を示す図であり、第２５図は、ＰＲＥＤ＿ＦＡＳＴＡによる１位の参照タンパク質についての出力ファイルを示す図であり、第２６図は、低ホモロジーの場合における本発明の入力ファイルの一例を示す図であり、第２７図は、ＰＲＥＤ＿ＦＡＳＴＡにより１位で検索された参照タンパク質についての出力ファイルを示す図であり、第２８図は、タンパク質（Ｔ０１７６：ＣＡＦＡＳＰ８８８０）のアミノ酸配列についてＰＳＩ－ＢＬＡＳＴ、ＰＲＥＤ＿ＦＡＳＴＡにより得られたアライメントを示す図であり、第２９図は、ＰＳＩ－ＢＬＡＳＴによるアライメントにＰＲＥＤ＿ＦＡＳＴＡによるアライメントの末端部分を付け加えることにより、長いアライメントを得ることができた場合を示す図であり、第３０図は、ペアワイズモードの場合の本発明の入力ファイルの一例を示す図であり、第３１図は、本発明の出力ファイルの一例を示す図であり、第３２図は、本発明で得られたアライメントからホモロジーモデリングを用いて立体構造を得た後の処理（埋め込み）を示すフローチャートであり、第３３図は、つなぎ目の残基の距離を判定して伸長する場合の概念図であり、第３４図は、埋め込みの処理によりつなぎ目の残基の距離を判定して伸長する場合の処理の一例を示すフローチャートであり、第３５図は、埋め込みの処理により、末端を優先させるフィッティングによる伸長を行う場合の処理の一例を示すフローチャートであり、第３６図は、ＲＭＳＤの閾値、フィッティング残基数の変化を示す図であり、第３７図は、埋め込みの処理によりフィッティング領域を移動させてフィッティングさせる場合の処理の一例を示すフローチャートであり、第３８図は、ＲＭＳＤの閾値、フィッティング残基数、フィッティング領域の変化の順序を示す図であり、第３９図は、埋め込みの処理による二次構造データベースの作成法を示すフローチャートであり、第４０図は、埋め込みの処理による二次構造データベースの作成法を示すフローチャートである。

【発明を実施するための最良の形態】

以下に、本発明にかかる配列情報処理装置、配列情報処理方法、プログラム、記録媒体等の実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態によりこの発明が限定されるものではない。

（本発明の基本原理）

以下、本発明の概要について説明し、その後、本発明の構成および処理等について詳細に説明する。第１図は、本発明の基本原理を示す原理構成図である。

本発明は、概略的に、以下の基本的特徴を有する。まず、本発明は、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する（ステップＳ－１）。なお、二次構造の取得には、既存の二次構造予測プログラム（例えば、ＰＳＩ－ＰＲＥＤなど）などを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラム（例えば、ＤＳＳＰ、ＳＴＲＩＤＥなど）などを用いて取得してもよい。また、ステップＳ－１において、予めアライメントしたアミノ酸配列情報を取得してもよい。

ついで、ステップＳ－１にて取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（例えば、各アミノ酸の疎水性情報等のアミノ酸毎の性質に関する類似性等）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報（例えば、Ａ、Ｂ、Ｃなどの文字など）に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換する（ステップＳ－２）。

ここで、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を予め格納し、ステップＳ－２において、格納された二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換してもよい。これにより、変換時に当該情報を参照することにより、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効率よく得ることができる。なお、二次構造・類似性分類情報は、例えばアミノ酸間の置換の起こりやすさをスコア値で表したマトリックス形式の既知の情報であるＢＬＯＳＵＭ６２のスコア値などに基づいてクラスタリング（ク

ラスター解析）を行ったことにより得られた複数のグループを示す類似性情報と、複数のグループに分類された二次構造を示す二次構造情報とを相互に関連付けて構成され、類似性情報と二次構造情報との組み合わせ毎に、対応する識別情報が割り当てられた情報でもよい。具体的には、二次構造・類似性分類情報は、例えば、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類した情報でもよい。

また、二次構造・類似性分類情報に基づいて変換された識別配列情報を予め格納し、ステップＳ－２において、格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索してもよい。これにより、変換された識別配列情報をデータベースなどに予め格納しておき、変換時に当該データベースなどを参照することにより、変換処理を効率化、高速化することができるようになる。また、これにより、例えば公共のデータベース（例えばＰＤＢなど）に登録されたアミノ酸配列情報の二次構造を既存の二次構造判定プログラムなどを用いて判定した後、当該アミノ酸配列情報を本装置により二次構造・類似性分類情報に基づいて変換して作成した識別配列情報を予めデータベースに格納して利用することができるようになり、取得したアミノ酸配列情報に対応する識別配列情報を効率よく検索することができる。

ついで、ステップＳ－２にて変換された複数の識別配列情報に対してアライメントを実行する（ステップＳ－３）。

ここで、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスおよび／または識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを予め格納し、ステップＳ－３において、格納された二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行してもよい。これにより、置換マトリックスを予め作成して格納しておき、アライメント実行時に当該置換マトリックスを参照して各スコア値に置換することにより、アミノ酸配列情報の二次構造および／またはアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したスコア値に置換することにより最適なアライメントを効率よく得ることができる。

なお、ステップＳ－３において、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行してもよい。これにより、例えば生物学的な知見などに基づいて二次構造置換マトリックスおよび／または類似性置換マトリックスに対して適切な係数を設定して重み付けをすることによって、生物学的な知見などを反映した最適なアライメントを効率よく得ることができる。

ここで、ステップＳ－３にてアライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換してもよい。これにより、二次構造および／またはアミノ酸類似性を加味してアライメントが行われたアミノ酸配列情報を得ることができる。

なお、上述におけるアミノ酸の類似性は、アミノ酸の疎水性のほか、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性を用いてもよい。これにより、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる。

（システム構成）

まず、本システムの構成について説明する。第２図は、本発明が適用される本システムの構成の一例を示すブロック図であり、該構成のうち本発明に関係する部分のみを概念的に示している。本システムは、概略的に、第２図に示すように、配列情報処理装置１００と、塩基配列情報やアミノ酸配列情報などに関する外部データベース、配列情報に対するホモロジー検索やアミノ酸配列情報に対するアライメントの実行やタンパク質の立体構造（例えば二次構造、三次構造など）の解析などを行うための外部プログラム等を提供する外部システム２００とを、ネットワーク３００を介して通信可能に接続して構成されている。

第２図において、ネットワーク３００は、配列情報処理装置１００を相互に接続する機能を有し、例えば、インターネット等である。

第２図において、配列情報処理装置１００は、概略的に、配列情報処理装置１００の全体を統括的に制御するＣＰＵ等の制御部１０２、通信回線等に接続されるルータ等の通信装置（図示せず）に接続される通信制御インタフェース部１０４、入力装置１１２や出力装置１１４に接続される入出力制御インタフェース部１０８、および、各種のデータベースやテーブルなどを格納する記憶部１０６を備えて構成されており、これら各部は任意の通信路を介して通信可能に接続されている。さらに、この配列情報処理装置１００は、ルータ等の通信装置および専用線等の有線または無線の通信回線を介して、ネットワーク３００に通信可能に接続されている。

記憶部１０６に格納される各種のデータベースやテーブル（アミノ酸配列関連情報ファイル１０６ａ～再変換アミノ酸配列情報ファイル１０６ｇ）は、固定ディスク装置等のストレージ手段であり、各種処理に用いる各種のプログラムやテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

これら記憶部１０６の各構成要素のうち、アミノ酸配列関連情報ファイル１０６ａは、取得されたアミノ酸配列情報と、当該アミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造、類似性および識別情報と、を格納するアミノ酸配列関連情報格納手段である。ここで、アミノ酸配列関連情報ファイル１０６ａに格納される情報について第３図を参照して説明する。

第３図は、本実施形態におけるアミノ酸配列関連情報ファイル１０６ａに格納される情報の一例を示す図である。第３図に示すように、アミノ酸配列関連情報ファイル１０６ａに格納される情報は、アミノ酸配列情報を一意に識別するためのアミノ酸配列識別情報（第３図に示す「Ａ００１」）と、アミノ酸配列情報を構成する各アミノ酸情報（第３図に示す「「Ａｌａ」、「Ｐｈｅ」、「Ｔｒｐ」、・・・、「Ｌｙｓ」」）と、各アミノ酸情報に対応する二次構造情報（第３図に示す「「α」、「α」、「その他」、・・・、「β」」）と、アミノ酸の類似性に基づいて分類されたグループ（群）を示す類似性情報（第３図に示す「「１」、「４」、「４」、・・・、「６」」）と、変換後の識別配列情報を構成する識別情報（第３図に示す「「Ａ」、「Ｋ」、「Ｍ」、・・・、「Ｓ」」）と、を相互に関連付けて構成されている。

また、二次構造・類似性分類情報ファイル１０６ｂは、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納手段である。ここで、二次構造・類似性分類情報ファイル１０６ｂに格納される情報について第４図を参照して説明する。

第４図は、本実施形態における二次構造・類似性分類情報ファイル１０６ｂに格納される情報の一例を示す図である。第４図に示すように、二次構造・類似性分類情報ファイル１０６ｂに格納される情報は、アミノ酸の類似性に基づいて分類されたグループ（群）を示す類似性情報（第４図に示す「「１」、「２」、「３」、・・・、「７」」）と、類似性情報により分類されるグループに属するアミノ酸情報（第４図は、例えば、類似性情報「１」に属するアミノ酸情報が「Ａｌａ」および「Ｇｌｙ」であることを示している。）と、二次構造情報（第４図に示す「「α」、「β」および「その他」」）とを相互に関連付けて構成され、類似性情報および二次構造情報の組み合わせ毎に、対応する識別情報が

格納されている。例えば、第４図に示す例では、類似性情報が「１」、二次構造情報が「α」の組み合わせの場合には、識別情報が「Ａ」であることを示している。

また、識別配列情報データベース１０６ｃは、二次構造・類似性分類情報に基づいてアミノ酸配列情報を構成するアミノ酸情報を識別情報に変換した識別配列情報を格納する識別配列情報格納手段である。ここで、識別配列情報データベース１０６ｃに格納される情報について第５図を参照して説明する。

第５図は、本実施形態における識別配列情報データベース１０６ｃに格納される情報の一例を示す図である。第５図に示すように、識別配列情報データベース１０６ｃに格納される情報は、アミノ酸配列情報を一意に識別するためのアミノ酸配列識別情報（第５図に示す「Ｋ＿００１」）と、アミノ酸配列情報を構成する各アミノ酸情報（第５図に示す「「Ａｌａ」、「Ｐｈｅ」、「Ｔｒｐ」、・・・、「Ｌｙｓ」」）と、アミノ酸配列情報を構成する各アミノ酸情報が二次構造・類似性分類情報に基づいて変換された各識別情報（第５図に示す「「Ａ」、「Ｋ」、「Ｍ」、・・・、「Ｓ」」）と、を相互に関連付けて構成されている。

また、二次構造置換マトリックス１０６ｄは、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納手段である。ここで、二次構造置換マトリックス１０６ｄに格納される情報について第６図を参照して説明する。

第６図は、本実施形態における二次構造置換マトリックス１０６ｄに格納される情報の一例を示す図である。第６図に示すように、二次構造置換マトリックス１０６ｄに格納される情報は、二次構造情報（第６図に示す「「α」、「β」および「その他」」）の組み合わせに対応する構造スコア値（第６図は、例えば、二次構造情報が「α」と「β」の組み合わせの場合、構造スコア値が「－６」であることを示している。）で構成されている。

また、類似性置換マトリックス１０６ｅは、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納手段である。ここで、類似性置換マトリックス１０６ｅに格納される情報について第７図を参照して説明する。

第７図は、本実施形態における類似性置換マトリックス１０６ｅに格納される情報の一例を示す図である。第７図に示すように、類似性置換マトリックス１０６ｅに格納される情報は、識別情報（第７図に示す「「Ａ」、「Ｒ」、・・・、「Ｘ」」）の組み合わせに応じて割り当てられた類似性スコア値（第７図は、例えば、識別情報が「Ａ」と「Ｎ」の組み合わせの場合、類似性スコア値が「－１」であることを示している。）で構成されている。

また、アライメント後識別配列情報ファイル１０６ｆは、後述するアライメント実行部１０２ｇによりアライメントが実行された識別配列情報を格納するアライメント後識別配列情報格納手段である。ここで、アライメント後識別配列情報ファイル１０６ｆに格納される情報について第８図を参照して説明する。

第８図は、本実施形態におけるアライメント後識別配列情報ファイル１０６ｆに格納される情報の一例を示す図である。第８図に示すように、アライメント後識別配列情報ファイル１０６ｆに格納される情報は、識別配列情報に対応するアミノ酸配列識別情報（第８図に示す「Ａ００１」）と、アライメントが実行された識別配列情報（第８図に示す「「Ａ」、「Ｂ」、「Ｃ」、・・・、「Ｐ」」）と、を相互に関連付けて構成されている。

また、再変換アミノ酸配列情報ファイル１０６ｇは、後述する再変換部１０２ｈにより再変換されたアミノ酸配列情報を格納する再変換アミノ酸配列情報格納手段である。ここで、再変換アミノ酸配列情報ファイル１０６ｇに格納される情報について第９図を参照して説明する。

第９図は、本実施形態における再変換アミノ酸配列情報ファイル１０６ｇに格納される情報の一例を示す図である。第９図に示すように、再変換アミノ酸配列情報ファイル１０６ｇに格納される情報は、アミノ酸配列識別情報（第９図に示す「Ａ００１」）と、再変

換されたアミノ酸配列情報（第９図に示す「「Ａｌａ」、「Ｐｈｅ」、「Ｔｒｐ」、・・・、「Ｌｙｓ」」）と、を相互に関連付けて構成されている。

また、第２図において、通信制御インタフェース部１０４は、配列情報処理装置１００とネットワーク３００（またはルータ等の通信装置）との間における通信制御を行う。すなわち、通信制御インタフェース部１０４は、他の端末と通信回線を介してデータを通信する機能を有する。

また、第２図において、入出力制御インタフェース部１０８は、入力装置１１２や出力装置１１４の制御を行う。ここで、出力装置１１４としては、モニタ（家庭用テレビを含む）の他、スピーカを用いることができる（なお、以下においては出力装置１１４をモニタとして記載する場合がある）。また、入力装置１１２としては、キーボード、マウス、および、マイク等を用いることができる。また、モニタも、マウスと協働してポインティングデバイス機能を実現する。

また、第２図において、制御部１０２は、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）等の制御プログラム、各種の処理手順等を規定したプログラム、および所要データを格納するための内部メモリを有し、これらのプログラム等により、種々の処理を実行するための情報処理を行う。制御部１０２は、機能概念的に、二次構造取得部１０２ａ、二次構造・類似性分類情報格納部１０２ｂ、識別配列情報格納部１０２ｃ、変換部１０２ｄ、二次構造置換マトリックス格納部１０２ｅ、類似性置換マトリックス格納部１０２ｆ、アライメント実行部１０２ｇおよび再変換部１０２ｈを備えて構成される。

ここで、二次構造取得部１０２ａは、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得する二次構造取得手段である。

また、二次構造・類似性分類情報格納部１０２ｂは、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納する二次構造・類似性分類情報格納手段である。

また、識別配列情報格納部１０２ｃは、上記の二次構造・類似性分類情報に基づいてアミノ酸配列情報を構成するアミノ酸情報を識別情報に変換した識別配列情報を格納する識別配列情報格納手段である。

また、変換部１０２ｄは、二次構造取得部１０２ａにて取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換する変換手段である。ここで、変換部１０２ｄは、第１０図に示すように、分類情報参照変換部１０２ｉおよび識別配列情報検索部１０２ｊをさらに含んで構成される。

第１０図は、本発明が適用される本システムの変換部１０２ｄの構成の一例を示すブロック図であり、該構成のうち本発明に関する部分のみを概念的に示している。

第１０図において、分類情報参照変換部１０２ｉは、二次構造・類似性分類情報格納部１０２ｂにて格納された二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換する分類情報参照変換手段である。

また、識別配列情報検索部１０２ｊは、識別配列情報格納部１０２ｃにて格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索する識別配列情報検索手段である。

再び第２図に戻り、二次構造置換マトリックス格納部１０２ｅは、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを格納する二次構造置換マトリックス格納手段である。

また、類似性置換マトリックス格納部１０２ｆは、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納する類似性置換マトリックス格納手段である。

また、アライメント実行部１０２ｇは、変換部１０２ｄにて変換された複数の識別配列

情報に対してアライメントを実行するアライメント実行手段である。ここで、アライメント実行部１０２ｇは、第１１図に示すように、置換マトリックス基準アライメント実行部１０２ｋをさらに含んで構成されている。

第１１図は、本発明が適用される本システムのアライメント実行部１０２ｇの構成の一例を示すブロック図であり、該構成のうち本発明に関する部分のみを概念的に示している。

第１１図において、置換マトリックス基準アライメント実行部１０２ｋは、二次構造置換マトリックス格納部１０２ｅにて格納された二次構造置換マトリックスおよび／または類似性置換マトリックス格納部１０２ｆにて格納された類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行する置換マトリックス基準アライメント実行手段である。

再び第２図に戻り、再変換部１０２ｈは、アライメント実行部１０２ｇにてアライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換する再変換手段である。

なお、これら各部によって行なわれる処理の詳細については、後述する。

（システムの処理）

次に、このように構成された本実施の形態における本システムの処理の一例について、以下に第１２図等を参照して詳細に説明する。第１２図は、本実施形態における本システムのメイン処理の一例を示すフローチャートである。

まず、配列情報処理装置１００は、二次構造取得部１０２ａの処理により、アミノ酸配列情報を取得してアミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「アミノ酸情報」の項）に格納し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得し、取得された二次構造に関する情報をアミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「二次構造情報」の項）に格納する（ステップＳＡ－１）。なお、二次構造の取得には、既存の二次構造予測プログラム（例えば、ＰＳＩ－ＰＲＥＤなど）などを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラム（例えば、ＤＳＳＰ、ＳＴＲＩＤＥなど）などを用いて取得してもよい。また、ステップＳＡ－１において、予めアライメントしたアミノ酸配列情報を取得してもよい。

ついで、配列情報処理装置１００は、制御部１０２の処理により、アミノ酸配列情報を構成するアミノ酸情報を、同じ類似性を有するアミノ酸を同じグループに属させることにより２０種類の天然アミノ酸を複数のグループに分類した情報である予め定めたアミノ酸類似性分類情報に基づいて分類し、分類されたグループを示す類似性情報をアミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「類似性情報」の項）に格納する。なお、アミノ酸類似性分類情報は、例えば、アミノ酸間の置換の起こりやすさをスコア値で表したマトリックス形式の既知の情報であるＢＬＯＳＵＭ６２のスコア値などに基づいてクラスタリング（クラスター解析）を行ったことにより得られた複数のグループからなる情報でもよい。具体的には、アミノ酸類似性分類情報は、例えば、下記の表２に示すように、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類した情報でもよい。

(表２)

	アミノ酸（カッコは１文字コード）			
１	Ａｌａ（A）	Ｇｌｙ（G）		
２	Ａｓｐ（D）	Ｇｌｕ（E）	Ａｓｎ（N）	Ｇｌｎ（Q）
３	Ｃｙｓ（C）			
４	Ｐｈｅ（F）	Ｈｉｓ（H）	Ｔｒｐ（W）	Ｔｙｒ（Y）
５	Ｉｌｅ（I）	Ｌｅｕ（L）	Ｍｅｔ（M）	Ｖａｌ（V）
６	Ｌｙｓ（K）	Ａｒｇ（R）		
７	Ｐｒｏ（P）	Ｓｅｒ（S）	Ｔｈｒ（T）	

ついで、配列情報処理装置１００は、変換部１０２ｄの処理により、ステップＳＡ－１にて予測された二次構造（第３図に示す「二次構造情報」）およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（第３図に示す「類似性情報」）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造（第３図に示す「二次構造情報」）および同一の類似性（第３図に示す「類似性情報」）を有するアミノ酸情報を同一の情報として識別するための識別情報（例えば、Ａ、Ｂ、Ｃなどの文字など）に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、アミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「識別情報」の項）に格納する（ステップＳＡ－２）。

ここで、配列情報処理装置１００は、二次構造・類似性分類情報格納部１０２ｂの処理により、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を予め二次構造・類似性分類情報ファイル１０６ｂの所定の記憶領域（第４図参照）に格納し、ステップＳＡ－２において、変換部１０２ｄは、分類情報参照変換部１０２ｉの処理により、二次構造・類似性分類情報ファイル１０６ｂに格納された二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換し、アミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「識別情報」の項）に格納してもよい。なお、二次構造・類似性分類情報は、例えば、アミノ酸間の置換の起こりやすさをスコア値で表したマトリックス形式の既知の情報であるＢＬＯＳＵＭ６２のスコア値などに基づいてクラスタリング（クラスター解析）を行ったことにより得られた複数のグループを示す類似性情報と、複数に分類された二次構造を示す二次構造情報とを相互に関連付けて構成され、類似性情報と二次構造情報との組み合わせ毎に、対応する識別情報が割り当てられた情報でもよい。具体的には、二次構造・類似性分類情報は、例えば、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類した情報でもよい。

また、配列情報処理装置１００は、二次構造取得部１０２ａの処理により、例えば外部システム２００の外部データベースに格納されたアミノ酸配列情報（例えば、公共の立体構造データベースであるＰＤＢに登録されているアミノ酸配列情報など）をネットワーク３００を介して予め取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得し、分類情報参照変換部１０２ｉの処理により、取得したアミノ酸配列情報を構成するアミノ酸情報を二次構造・類似性分類情報に基づいて識別情報に予め

変換し、識別配列情報格納部１０２ｃの処理により、変換した識別情報からなる識別配列情報を識別配列情報データベース１０６ｃの所定の記憶領域に予め格納し、ステップＳＡ－２において、変換部１０２ｄは、識別配列情報検索部１０２ｊの処理により、識別配列情報データベース１０６ｃに格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索し、検索した識別配列情報をアミノ酸配列関連情報ファイル１０６ａの所定の記憶領域（第３図に示す「識別情報」の項）に格納してもよい。

ついで、配列情報処理装置１００は、アライメント実行部１０２ｇの処理により、ステップＳＡ－２にて変換された複数の識別配列情報に対してアライメントを実行し、アライメント後識別配列情報ファイル１０６ｆの所定の記憶領域に格納する（ステップＳＡ－３）。

ここで、配列情報処理装置１００は、二次構造置換マトリックス格納部１０２ｅの処理により、二次構造（例えば、αヘリックス、βシートおよびその他）の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックスを二次構造置換マトリックス１０６ｄの所定の記憶領域に予め格納し、および／または、類似性置換マトリックス格納部１０２ｆの処理により、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを類似性置換マトリックス１０６ｅの所定の記憶領域に予め格納し、ステップＳＡ－３において、アライメント実行部１０２ｇは、置換マトリックス基準アライメント実行部１０２ｋの処理により、二次構造置換マトリックス１０６ｄに格納された二次構造置換マトリックスおよび／または類似性置換マトリックス１０６ｅに格納された類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行し、アライメント後識別配列情報ファイル１０６ｆの所定の記憶領域に格納してもよい。

なお、ステップＳＡ－３において、アライメント実行部１０２ｇは、置換マトリックス基準アライメント実行部１０２ｋの処理により、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行し、アライメント後識別配列情報ファイル１０６ｆの所定の記憶領域に格納してもよい。例えば、二次構造置換マトリックスを重んじる場合は、例えば二次構造置換マトリックスの構造スコア値に例えば１以上の数値を掛ける、および／または、例えば類似性置換マトリックスの類似性スコア値に例えば１未満の数値を掛ける、などすることにより、類似性スコア値に対して構造スコア値が大きくなるように重み付けを行ってもよい。

ここで、配列情報処理装置１００は、再変換部１０２ｈの処理により、ステップＳＡ－３にてアライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換し、再変換アミノ酸配列情報ファイル１０６ｇの所定の記憶領域に格納してもよい。具体的には、配列情報処理装置１００は、再変換部１０２ｈの処理により、ステップＳＡ－３にてアライメントが実行された識別配列情報に対応するアミノ酸配列識別情報と同一のアミノ酸配列識別情報のアミノ酸配列情報を取得し、当該識別配列情報を構成する各識別情報を、取得したアミノ酸配列情報のアミノ酸情報と対応付けることにより再変換し、再変換アミノ酸配列情報ファイル１０６ｇの所定の記憶領域に格納してもよい。

なお、本実施形態におけるアミノ酸の類似性は、アミノ酸の疎水性のほか、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性を用いてもよい。これにより、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる。

これにて、メイン処理が終了する。

【実施例】

（関数を用いたモデルの選定）

第１３図は、本発明を有効に適用するための処理を示すフローチャートである。以下に、第１３図に示す本フローチャートを詳細に説明する。

（１）ホモロジー検索

目的タンパク質のアミノ酸配列を用いてホモロジー検索を行ない、アライメントを得る

。このとき本発明のｐｒｅｄ＿ｆａｓｔａのホモロジー検索モードの実行をする。さらに加えて複数のホモロジー検索プログラム（例えば、ＢＬＡＳＴ、ＦＡＳＴＡなど）を用いて探索するのが望ましい。これは同じ参照タンパク質を探索してもアライメント方法が異なる場合や、あるホモロジー検索プログラムでは探索できなかった参照タンパク質を他のホモロジー検索プログラムが探索できる場合など、モデル作成時に様々なアライメントや参照タンパク質を用いたほうがより精度の高いモデルを作成できる可能性があるためである。

（２）モデリング

（１）で得られたアライメントを用いて、ＦＡＭＳなどのホモロジーモデリングプログラムにより目的タンパク質のモデリングを行なう。ただし、（１）で得られた全てのアライメントについてモデルを作成するのは非常に非能率的であるので、ｅ値やホモロジーの基準でモデリングを行なうアライメントを絞った。

（３）目的タンパク質の二次構造を予測する

目的タンパク質のアミノ酸配列から、二次構造予測プログラムＰＳＩ－ＰＲＥＤを用いて目的タンパク質の二次構造を予測する。ＰＳＩ－ＰＲＥＤはアミノ酸配列から二次構造を予測するプログラムであり、かなり高い信頼度（予測率の指標であるQ_3値は約７８％といわれている）で二次構造の予測を行なうことが可能である。予測結果は第１４図に示す形式のファイルに出力される。

第１４図におけるＭＡ－１～ＭＡ－６について以下に説明する。

ＭＡ－１：予測する目的タンパク質アミノ酸配列の残基番号（１から始まる）

ＭＡ－２：アミノ酸の種類（一文字表記）

ＭＡ－３：この残基の二次構造判定結果（ＭＡ－４～ＭＡ－６の値のうち一番高いものの二次構造を表記）

以下の３文字で表現される。

Ｈ：αヘリックス

Ｅ：βシート

Ｃ：その他（コイル）

ＭＡ－４：この残基がコイルである確率

ＭＡ－５：この残基がαヘリックスである確率

ＭＡ－６：この残基がβシートである確率

（４）作成したモデルについて、二次構造を判定する

（２）で作成したモデルについて、その３次元座標から二次構造判定プログラムＳＴＲＩＤＥにより二次構造を判定した。これによりモデルの各残基について、どのような二次構造をとっているかを判定することができる。

（５）二次構造適合率の算出

モデルの二次構造判定結果を二次構造予測結果と比較することにより二次構造適合率を求める。この割合の算出方法は以下の通りである。

モデル中のある残基についての二次構造適合率を求めたい場合、（４）においてＳＴＲＩＤＥにより求めたモデルの残基ごとの二次構造判定結果から、その残基についての二次構造（αヘリックス、βシート、その他（コイル））を求める。次に（３）においてＰＳＩ－ＰＲＥＤにより求めた残基ごとの二次構造確率から当該の残基について、求めた二次構造についての確率を求める。この確率がその残基についての二次構造適合率であり、全ての残基について二次構造適合率を求めて和を算出する。この和（二次構造適合残基数）は二次構造の確率を足し合わせたもので、二次構造が適合した残基数を意味している。この二次構造適合残基数をモデル残基数で割って求めた平均を、このタンパク質全体についての二次構造適合率とした。

ＰＳＩ－ＰＲＥＤは高い予測率のため、ある程度信頼度の高い二次構造適合率を求めることが可能である、と考えられる。

（６）ホモロジー検索時のｅ値の算出

モデル作成に用いたアライメントに対応するホモロジー検索結果から、そのアライメン

トについてのｅ値を求めた。ただし、ホモロジー検索時に用いる立体構造データベースは異なる場合があり一概にｅ値を比較することはできない。ｅ値は以下の式で求められるため、異なるデータベースを使用した場合にはデータベース間の大きさの比をかけることによりｅ値を補正した。

ＤＢ－ｓｉｚｅ×ｐ－ｖａｌｕｅ（＝　Ｅ－ｖａｌｕｅ）

（７）疎水性相互作用の指標（ｅｎｅｓｏｓｕｉ値）の算出

タンパク質が立体構造を構成する因子として、疎水性相互作用がある。この疎水性相互作用を評価することにより、予測したモデルの立体構造を評価する指標として利用することが可能である。この指標を求めるために、「ｅｎｅｓｏｓｕｉ値」を用いた。ｅｎｅｓｏｓｕｉ値は疎水性相互作用の指標であり、第１５図に示す表で示したアミノ酸残基の側鎖中の疎水性を示す代表的な原子同士の距離が、ある閾値（例えば、６．６Å）より小さい場合、疎水性相互作用をしていると見なし、その残基にｅｎｅｓｏｓｕｉ値として「－１」を与える。この場合、溶媒から遮蔽されており溶媒と接触している状態よりもエネルギー的に安定であると考えられる。ｅｎｅｓｏｓｕｉ値の合計はタンパク質全体における疎水性相互作用の指標とすることができる。

また、ｅｎｅｓｏｓｕｉによりＳＣＯＰドメインデータベース（９５％Ｎｏｎ－Ｒｅｄｕｎｄａｎｔ化したもの）中のドメイン構造について、その残基数とｅｎｅｓｏｓｕｉ値との関係を調べた（第１６図参照）。ここで、第１６図は、ＳＣＯＰドメインデータベース（９５％Ｎｏｎ－Ｒｅｄｕｎｄａｎｔ：ｔｏｔａｌ＿ｓｅｑ＿ｄｏｍ＿ａｌｌ＿９５（７４３３ドメイン））中のドメイン構造について、その残基数とｅｎｅｓｏｓｕｉ値との関係を調べた結果をまとめた表を示す図である。その結果、「ｙ＝７．１１１８ｘ－１７．３４」という検量線（第１６図の表中の直線を参照）を得た。構築したモデルについてｅｎｅｓｏｓｕｉ値を計算したとき、この検量線付近の値をとることが望ましいといえる。この検量線を考慮した評価を行なうには、ｅｎｅｓｏｓｕｉ適合率を算出する必要がある。このｅｎｅｓｏｓｕｉ適合率は以下の数式１により求められる。

$$\text{enesosui適合率} = 1 - C\left(\frac{E_0 - E}{E_0}\right)^2 \quad \cdots \text{（数式１）}$$

Ｅ＝モデルのｅｎｅｓｏｓｕｉ値

E_0＝モデル残基数のときに検量線により求められるｅｎｅｓｏｓｕｉ値

Ｃ＝定数

（８）二次構造率を求める

二次構造率とは、モデル構造について二次構造判定プログラムＳＴＲＩＤＥにより二次構造判定を行ない、二次構造（αヘリックス、βシート）であると判定された残基の数を数え、その値をモデル残基数で割ったものである。二次構造が長くできているモデルほど、この値は大きくなる。

（９）スコア値の算出

二次構造、疎水性エネルギー、ｅ値、最低のｅ値をもとにモデルに対するスコアを計算する（二次構造率、ｅｎｅｓｏｓｕｉ適合率がなくてもスコアによる評価は可能である。）。具体的には、以下の数式２に示すように、二次構造適合率と疎水性相互作用の指標（ｅｎｅｓｏｓｕｉ値）の積にｅ値と最低のｅ値の対数比による重みをかけることによりスコアを計算する。このスコアはアライメント時に得られるｅ値だけではなくモデルの立体構造状態も考慮したものである。さらに（任意的に）二次構造率、ｅｎｅｓｏｓｕｉ適合率をかけることにより、モデルの構造をより厳しく判定することができる。

$$\text{スコア}=-(\text{二次構造適合率}\times\text{enesosui値})\times\frac{-\log_{10}(e)+\alpha}{-\log_{10}(e_{min})+\alpha}\times(\text{二次構造率}\times\text{enesosui適合率})$$

・・・（数式２）

e_{min}はモデル作成時に用いたアライメント中で、最も低いｅ値をもつもののｅ値である。αは定数であり、e_{min}値の対数が０で１００の値を持ち、－３で５０の値を、－１０で４０の値を持つようにして、その間では線形の関係があるような値に１を加算した値である。

ここで１を加算する理由を述べる。通常のＦＡＳＴＡ、ＰＳＩ－ＢＬＡＳＴなどのホモロジー検索プログラムでは、ｅ値の上限が１０で区切られて検索結果が出力される。このため、$-\log_{10}$（ｅ）の最低値は－１であり、最高値は＋∞である。このため、ｅ値、最低ｅ値の組み合わせによっては$-\log_{10}$（ｅ）と$-\log_{10}$（e_{min}）の符号が異なる場合がある。そこで、符号をそろえるために分子・分母の値に１を加えた。この値は、$\log_{10}$（ホモロジー検索プログラムでのｅ値の上限値）で表わされる。

この定数により、ｅ値が低い場合には二次構造適合率、疎水エネルギー値の重みが大きくなり、逆にｅ値が高い場合にはｅ値の重みが大きくなる。これはｅ値が大きい場合、目的タンパク質とその参照タンパク質とは類似性が薄く、構造もあまり類似していないと思われるため、二次構造や疎水エネルギーといったモデル構造に依存する値には重みを与えないようにしたためである。

このスコアをモデルごとに算出する。

（１０）スコア値の比較

（９）によりモデルごとにスコアが得られるが、これらのスコアを比較することによりモデルの精度の指標とした。スコア値が大きいほどアライメント的にも、立体構造的に見ても精度が高いモデルである、ということができると考えられる。

結果ファイルはスコア値順に出力される。

次に、上記プロセスの入出力画面について、第１７図および第１８図を参照して説明する。

第１７図は本発明の入力ファイルの一例を示す図である。入力ファイルとして目的タンパク質のアミノ酸配列を記したファイルを用いる。このファイルはＦＡＳＴＡフォーマット（１行目に“＞タンパク質名”、２行目にアミノ酸配列）でなければならない。

また、第１８図は、本発明の出力ファイルの一例を示す図である。このファイルはスコア値順に出力される。以下に第１８図に示すＭＢ－１～ＭＢ－１５について詳細に説明する。

ＭＢ－１：探索された参照タンパク質名
ＭＢ－２：参照タンパク質のアミノ酸配列の全長
ＭＢ－３：スコア値
ＭＢ－４：ｅ値
ＭＢ－５：ホモロジー値
ＭＢ－６：使用したホモロジー検索プログラムによる結果における順位
ＭＢ－７：目的タンパク質のモデリング領域（残基番号）
ＭＢ－８：参照タンパク質の対応領域（残基番号）
ＭＢ－９：使用したホモロジー検索プログラムの略名
ＭＢ－１０：使用したデータベース名

ＰＤＢ…立体構造をもつ鎖単位のタンパク質をアミノ酸配列に変換して作成したデータベース

ＳＣＯＰ…ＰＤＢをＳＣＯＰドメインにより分類したアミノ酸配列データベース

ＰＤＢ９５…ＰＤＢデータベースをホモロジー９５％以上によりクラスタリングし、配列冗長性を排除したデータベース

ＳＣＯＰ９５…ＳＣＯＰドメインデータベースをホモロジー９５％以上によりクラスタリングし、配列冗長性を排除したデータベース

ＭＢ－１１：モデルのｅｎｅｓｏｓｕｉ値
ＭＢ－１２：モデルと参照タンパク質のｅｎｅｓｏｓｕｉ値の比（モデルのｅｎｅｓｏｓｕｉ値／参照タンパク質の対応する領域のｅｎｅｓｏｓｕｉ値）
ＭＢ－１３：ＭＢ－１２の値をモデルの残基数で割ったもの

ＭＢ－１４：目的タンパク質全体の二次構造適合残基数
ＭＢ－１５：二次構造適合率（ＭＢ－１４の値をモデルの残基数で割ったもの）
（本発明のＰＲＥＤ＿ＦＡＳＴＡ）
第１９図および第２０図は、本発明の処理を示すフローチャートである。以下に、第１９図（ペアワイズモードの場合）および第２０図（ホモロジー検索モードの場合）に示す本フローチャートを詳細に説明する。
（１）アミノ酸類似性データベース（上述の実施形態におけるアミノ酸類似性分類情報に対応）の作成
通常、アライメントを行なう場合にはアミノ酸同士の相同性スコアを計算する。このときに使用される置換マトリックスとして代表的なものにＢＬＯＳＵＭ６２がある。このファイルには例えば次のようなことが書かれている。Ｆ（Ｐｈｅ）とＹ（Ｔｙｒ）との間のスコア値（置換のしやすさを示す）は３であり、Ｆ（Ｐｈｅ）とＰ（Ｐｒｏ）との間のスコア値は－４である。このことからＦとＰをアライメントするよりもＦとＹをアライメントしたときに、より高いスコアが付くようにしている。この置換マトリックスに記されている、あるアミノ酸同士の置換のしやすさはそれらのアミノ酸の間の類似性を示している。このため、ＢＬＯＳＵＭ６２に記されているスコア値をもとにクラスタリングを行ない、２０種の天然アミノ酸をその特質により７つのグループに分類した。
（２）二次構造・アミノ酸類似性データベース（第５図に示す識別配列情報データベース１０６ｃに対応）の準備
あらかじめＰＲＥＤ＿ＦＡＳＴＡで使用する二次構造・アミノ酸類似性データベース（第５図に示す識別配列情報データベース１０６ｃに対応）を準備しておく。
そして、タンパク質の立体構造から二次構造とその位置を判定するプログラム「ＳＴＲＩＤＥ」により、通常ホモロジー検索に用いる立体構造データベースに収載されているタンパク質について二次構造判定をする。これは通常のホモロジー検索を行なったときに、その検索されたタンパク質についての二次構造情報を必要とする場合があるためである。今回は立体構造データベース中のタンパク質配列について、ホモロジー９５％以上かつｅ値０．０１以下の基準でクラスタリングしたものを用いた。
ここで、残基ごとに二次構造の違い（αヘリックス、βシート、その他）３種類に分類できるが、この情報のほかに（１）で作成したアミノ酸類似性分類情報（２０種のアミノ酸を７つのグループに分類）を組み合わせて表現したアミノ酸を作成する。すなわち、第２１図に示すように、天然のアミノ酸をその二次構造・アミノ酸類似性の基準から便宜上二次構造３パターン×疎水性７グループ＝２１種のアミノ酸として新たな配列に置き換えるという作業を行なった（第４図に示す二次構造・類似性分類情報ファイル１０６ｂに対応）。なお、第２１図に示す例は一例であり、これ以外のグループに分類してもよい。
（３）アライメント、参照タンパク質の準備
まず、目的・参照タンパク質のアミノ酸配列間の関係を示したアライメントを作成する。アライメント用ソフトウエアとして、ＦＡＳＴＡ、ＰＳＩ－ＢＬＡＳＴなどがある。これらのプログラムを用いたときに、ある一定の閾値以下の条件で得られたアライメントは信頼性が高いといえる。今回の閾値として、ＰＳＩ－ＢＬＡＳＴを用いた際のｅ値が０．０１以下とした。このようにして得られたアライメント領域については、その対応するアミノ酸には大きな誤りはないといえる。しかし、このような通常のアライメントソフトはアミノ酸配列の文字情報のみからアライメントを得ているため、参照タンパク質側の二次構造を分断したアライメントを作成してしまう恐れがある。このため、ＦＡＳＴＡを用いて二次構造・アミノ酸類似性を崩さないアライメントを以下の手順により作成した。
（４）目的タンパク質の配列から二次構造を予測する（第２図に示す二次構造取得部１０２ａの処理に対応）
与えられた目的タンパク質の配列を用いて、ＰＳＩＰＲＥＤ（ｐｒｏｔｅｉｎ ｓｔｒｕｃｔｕｒｅ ｐｒｅｄｉｃｔｉｏｎ）を実行する。ＰＳＩＰＲＥＤは高速で精度の高い二次構造予測プログラムであり、ＰＳＩ－ＢＬＡＳＴで作成されるマルチプルアライメントを単純なニューラルネットで評価することにより二次構造を予測する。ＰＳＩ－ＰＲＥ

Ｄ中でマルチプルアライメントの結果を使用する理由は、進化の過程で機能を維持する上で構造の保存性は高く、構造を変えないように残基の置換も制約されるといったことも考慮して二次構造を予測することができるためである。

（５）アライメントを二次構造・アミノ酸類似性により置換した配列に変換（第２図に示す変換部１０２ｄ、分類情報参照変換部１０２ｉおよび識別配列情報検索部１０２ｊの処理に対応）

（３）で得られたアライメントの配列を（２）で定めた二次構造とアミノ酸の類似性の基準（上述の実施形態における二次構造・類似性分類情報に対応）によりアミノ酸一文字表記を変換し、新たなアミノ酸配列（上述の実施形態における識別配列情報に対応）を作成する。

目的タンパク質の配列に関しては、（４）で行なった二次構造予測結果とアミノ酸類似性（上述の実施形態におけるアミノ酸類似性分類情報に対応）に基づいて変換する。参照タンパク質の配列に関しては（２）で作成した二次構造・アミノ酸類似性データベース（第５図に示す識別配列情報データベース１０６ｃに対応）から当該参照タンパク質についての配列を抜き出した。このようにして目的・参照タンパク質共に二次構造・アミノ酸類似性により変換した配列を用意した。この新たに用意したアミノ酸配列は、（３）で得られたアライメント領域について作成した。

（６）ＦＡＳＴＡによる再アライメント（第２図に示すアライメント実行部１０２ｇおよび置換マトリックス基準アライメント実行部１０２ｋの処理に対応）

（５）で用意した目的・参照タンパク質配列（上述の実施形態における識別配列情報に対応）をＦＡＳＴＡにより再アライメントした。ここで用いた文字比較のための置換マトリックスはアミノ酸一文字表記の比較のためのＢＬＯＳＵＭ６２ではなく、二次構造とアミノ酸類似性を表わす文字（疎水性が同程度のアミノ酸残基であり、且つ同じ二次構造をもつ場合同じ文字として扱われる）の２つのマトリックス（上述の実施形態における二次構造置換マトリックスおよび類似性置換マトリックスに対応）から作成されたものである（第２２図および第２３図を参照）。デフォルトは１：１の均等の重み付けであるが、オプションによりその重み付けを変更することも可能である。ここで作成したマトリックスを用いることにより、二次構造とアミノ酸類似性情報を加味したアライメントを得ることができる。

ここで、第２２図は二次構造によるスコアマトリックスの一例を示す図である。第２２図に示すスコア（上述の実施形態における構造スコア値に対応）は、二次構造が一致したときに高くなるよう設定し、逆に相反する二次構造の場合にはアライメントしにくいように低く設定した。

また、第２３図はアミノ酸類似性によるスコアマトリックスの一例を示す図である。第２３図のスコアマトリックスは、（１）で定義したアミノ酸類似性により分類した識別情報のグループをもとに作成した。例えば、ある識別情報のグループ内のアミノ酸同士のスコア（上述の実施形態における類似性スコア値に対応）は、そのグループに属する全てのアミノ酸の組み合わせによるＢＬＯＳＵＭ６２スコアの平均（小数点以下は切り捨て、整数で表記）を割り当てた。同様に、あるグループに属するアミノ酸とあるグループに属するアミノ酸との間のスコアは、その二つのグループに属する全てのアミノ酸の組み合わせによるＢＬＯＳＵＭ６２スコアの平均を割り当てた。

（７）二次構造・アミノ酸類似性による配列をもとのアミノ酸配列に戻す（第２図に示す再変換部１０２ｈの処理に対応）

（６）で得られたアライメントは、二次構造とアミノ酸類似性により変換された配列で出力される。このため、もとのアミノ酸配列に再変換したアライメントを作成した。つまり、二次構造、アミノ酸類似性により変換された配列を２０種類のアミノ酸で表現される通常の配列に再変換する。

（８）モデリング

（７）で得られた修正アライメントを元に、適当なモデリングソフト、例えばＦＡＭＳを用いて立体構造を構築し、出来上がったモデルの二次構造が分断されていないかどうか

を検査する。

上記（１）～（８）に示した手順は、第１９図に示すペアワイズモード（１対１でアライメントすること）の場合の手順であり、これらの手順は、第２０図に示すホモロジー検索モードの場合には、目的タンパク質配列について、データベースを（２）で作成した二次構造・アミノ酸類似性データベース（第２図における識別配列情報データベース１０６ｃに対応）全体に対して検索することによって、通常行なわれるホモロジー検索としての利用も同様に可能である。

次に、本発明の入出力画面について、第２４図から第３１図を参照して説明する。

（１）ホモロジー検索モード

ホモロジー検索モードを使用する場合、目的タンパク質のアミノ酸配列を記したＦＡＳＴＡ形式のファイルを入力ファイルとする。ＦＡＳＴＡ形式とは、ファイルの一行目に「＞目的タンパク質名」、２行目に「目的タンパク質のアミノ酸配列」を記したファイルである。出力は通常のＦＡＳＴＡの出力と同じ形式である。

（例１；高ホモロジーの場合）

第２４図は、高ホモロジーの場合における本発明の入力ファイルの一例を示す図である。第２４図は、ＰＤＢに登録してあり、かつＰＲＥＤ＿ＦＡＳＴＡで用いられる立体構造データベースにも登録されている１Ａ４ＦのＡ鎖（１Ａ４ＦＡ）の配列を目的タンパク質ｔｅｓｔ１としてＰＲＥＤ＿ＦＡＳＴＡを実行した場合の入力ファイルである。

また、第２５図は、ＰＲＥＤ＿ＦＡＳＴＡによる１位の参照タンパク質についての出力ファイルを示す図である。これにより自身の配列である１Ａ４Ｆ＿Ａを探索できたことがわかる。第２５図のアミノ酸配列を見るとホモロジーが９１．４８９％となっているが、これは二次構造・アミノ酸類似性によりアミノ酸配列を変換してアライメントさせたときのホモロジーであるためである。

（例２；低ホモロジーの場合）

第２６図は、低ホモロジーの場合における本発明の入力ファイルの一例を示す図である。第２６図は、立体構造予測コンテストＣＡＳＰ５において出題された低ホモロジー配列（Ｔ０１９４：ＣＡＦＡＳＰ９０５８）について、ＰＲＥＤ＿ＦＡＳＴＡによりホモロジー検索を行なった際の入力ファイルである。

また、第２７図は、ＰＲＥＤ＿ＦＡＳＴＡにより１位で検索された参照タンパク質についての出力ファイルを示す図である。第２７図において、ホモロジー１８．７９７％と出力されているが、これは二次構造・アミノ酸類似性によりアライメントさせたときのホモロジーであり、実際のホモロジーは７％と非常に低い。このように非常にホモロジーが低く、他のホモロジー検索プログラムでは探索できない場合でもＰＲＥＤ＿ＦＡＳＴＡにより参照タンパク質を探索することが可能である。

（例３；アライメントを伸長させる）

ＰＲＥＤ＿ＦＡＳＴＡにより参照タンパク質を探索してアライメントを得ると、通常のホモロジー検索プログラムによるアライメントよりも長いアライメントが得られることがある。このような場合、通常のホモロジー検索プログラムによるアライメントとＰＲＥＤ＿ＦＡＳＴＡによるアライメントを組み合わせることにより、より長いアライメントを得ることができる。この例を以下に述べる。

第２８図は、立体構造予測コンテストＣＡＳＰ５において出題されたタンパク質（Ｔ０１７６：ＣＡＦＡＳＰ８８８０）のアミノ酸配列についてＰＳＩ－ＢＬＡＳＴ、ＰＲＥＤ＿ＦＡＳＴＡにより得られたアライメントを示す図である。ともに同じ参照タンパク質を探索しているがＰＲＥＤ＿ＦＡＳＴＡの方がアライメント領域を長くアライメントをしている。

また、第２９図は、ＰＳＩ－ＢＬＡＳＴによるアライメントにＰＲＥＤ＿ＦＡＳＴＡによるアライメントの末端部分を付け加えることにより、長いアライメントを得ることができた場合を示す図である。

（２）ペアワイズモード

ペアワイズモードで使用する場合、目的タンパク質と参照タンパク質とのアライメント

ファイルを入力とする。第３０図は、ペアワイズモードの場合の本発明の入力ファイルの一例を示す図である。この入力ファイルは１行目に「＞目的タンパク質名」、２行目に「アライメントされた目的タンパク質のアミノ酸配列」、３行目に「＞参照タンパク質名」、４行目に「アライメントされた参照タンパク質のアミノ酸配列」を記したものである。

ここで、第３０図に示す例では、ＰＤＢに登録してある１Ａ４ＦのＡ鎖（１Ａ４Ｆ＿Ａ）の一部を目的タンパク質ｔｅｓｔ３とした。このｔｅｓｔ３と１Ａ４Ｆ＿Ａとのアライメント（ホモロジーは１００％）について、１Ａ４Ｆ＿Ａの二次構造が壊れるようにわざとギャップを入れたアライメントファイルを入力としてＰＲＥＤ＿ＦＡＳＴＡによるアライメント修正を実行した。

また、第３１図は、本発明の出力ファイルの一例を示す図である。第３１図に示す出力ファイルの例では、二次構造上に入ったギャップは除かれてきちんとしたアライメントになっていることがわかる。

（埋め込み）

第３２図は、本発明でアライメントを行い立体構造モデルが得られた後、より現実に近い立体構造モデルを得るための処理を示すフローチャートである。以下に、第３２図に示す本フローチャートを詳細に説明する。

あるタンパク質Ａを参照タンパク質として作成したモデルＡ（第３２図におけるモデル１）より、あるタンパク質Ｂを参照タンパク質として作成したモデルＢ（第３２図におけるモデル２）の方がより長くモデリングできたとする。しかし領域の短いモデルＡの方が精度よい構造である場合（すなわち高優先順位である場合）、モデルＡの構造にモデルＢの末端の立体構造を貼り付けてモデリング領域を伸長させることができる場合がある。

このような場合に局所構造（短いモデル：ベースとなる）の末端にグローバル構造（長いモデル）の構造を貼り付けることによりモデルの伸長を行なう方法論を開発した。これらの伸長方法には以下の４つがあり、以下にその手順を述べる。

手順（１）：つなぎ目の残基の距離を判定して伸長（第３３図、第３４図参照）
手順（２）：末端を優先させるフィッティングによる伸長（第３５図参照）
手順（３）：領域を移動させるフィッティングによる伸長（第３７図参照）
手順（４）：二次構造等を主に考慮して断片を再検索して伸長（第３９図、第４０図参照）

以下に、これらの各手順の方法論などについて詳細に説明する。

手順（１）：つなぎ目の残基の距離を判定して伸長

第３４図は、本発明によりつなぎ目の残基の距離を判定して伸長する場合の処理の一例を示すフローチャートである。

局所構造（短いモデル：ベースとなる）の末端にグローバル構造（長いモデル）の構造を貼り付けることによりモデルの伸長を行なう方法論を開発した。この方法論では構造のつなぎ目の残基間の距離を考慮して、つなぎ目がなるべく滑らかになるように局所構造を削っていく。以下にその手順を述べる。

（１）アライメントにより共通領域を求める

グローバル構造（長いモデル）のアミノ酸配列と、局所構造（短いモデル：ベースとなる）のアミノ酸配列をアライメントし、共通領域を求めた。つまり、局所構造モデルとグローバル構造モデルのアミノ酸配列を用いてアライメントして対応する残基、共通領域を求めた。局所構造とグローバル構造が全く同一の領域をモデリングしている場合、つまりモデルの長さと領域が同じ場合はモデルの伸長させることができないので、ここで終了する。

（２）フィッティングを行なう

（１）で求めた共通領域を用いて最小二乗法などによりフィッティングを行なう。これは局所構造モデルとグローバル構造モデルの座標系を揃えるために行なう必要がある。局所構造とグローバル構造をフィッティングさせたときのＲＭＳＤが２Åより大きい場合は、この埋め込みは行なわない。これはあまりにも構造が異なる場合、埋め込みを行なってもよいモデルを作成することができないためである。

（３）局所構造モデルをグローバル構造モデルに埋め込む
　座標系を一致させた局所構造モデルの座標を、グローバル構造モデル座標上の対応する領域にはめ込む。
（４）境界の残基間距離を計算する
　（３）で埋め込んだ局所構造モデルとグローバル構造モデルとの境界残基同士のＣα原子間距離を計算する。具体的には、例えば、局所構造モデルとグローバル構造モデルの境界部分の隣接する残基のＣα原子間距離を求め、ある閾値（例えば、８．０Å）以下かどうかを調べる。この距離は次の（５）において、つなぎ目の構造を評価するときに必要である。
（５）つなぎ目の判定
　（４）で計算したつなぎ目残基のＣα原子間距離がある閾値（例えば、８．０Å：ＦＡＭＳにより処理可能な限界値など）以上離れている場合、局所構造モデルの末端を一残基カットし、再び（２）のフィッティングに戻る。つまり、閾値よりもＣα原子間距離が大きい場合は、局所構造モデルを末端から一残基カットして距離を調べる、という手順を繰り返す。例えば、グローバル構造モデルの残基２（第３４図の残基番号２の残基）と局所構造モデルの残基１’（第３４図の残基番号１’の残基）のＣα原子間距離が閾値よりも大きかった場合、局所構造モデルの残基１’（第３４図の残基番号１’の残基）をカットし、その残基については代わりにグローバル構造モデルの残基３（第３４図の残基番号３の残基）を用いる。もう一方の末端側について距離が閾値よりも小さい場合はこのカットは行なわない。この手順を距離が閾値以下になり、つなぎ目が滑らかになるまで繰り返して埋め込みモデルを作成する。残基を削っていっても最後までこの閾値を満たすことができない場合は埋め込みは不可能として終了する。
　この作業を行なう理由は、あまりにも接続残基部分の距離が離れているとＦＡＭＳでの再モデリング時に参照タンパク質座標異常として、参照タンパク質（局所構造モデルをグローバル構造モデルに埋め込んだもの）の座標を参照せずにＦＡＭＳ内部のデータベースを使用してしまうためである。
　このようにしてなるべく局所構造モデル領域を保存し、末端のみグローバル構造領域を使用するようにして座標ファイルをつなげる。
（６）ＦＡＭＳを用いて再モデリング（モデルを伸長させることができた場合のみ）
　（５）までの作業で埋め込んで作成したＰＤＢ形式の座標ファイルを参照タンパク質として、ＦＡＭＳにより同一領域の再モデリング（ホモロジー１００％のモデリング）を行なった。つまり、埋め込みを行った構造を参照タンパク質にして、ＦＡＭＳにより再モデリングを行った。再モデリングすることにより、つなぎ目部分の不自然な構造や、側鎖のぶつかりなどを解消して全体の形を整えることができる。
手順（２）：末端を優先させるフィッティングによる伸長
　第３５図は、本発明により、末端を優先させるフィッティングによる伸長を行う場合の処理の一例を示すフローチャートである。本発明者は、本方法により、局所構造（短いモデル：ベースとなる）の末端にグローバル構造（長いモデル）の構造を貼り付けることによりモデルの伸長を行なう方法論を開発した。この方法では伸長させようとする方の末端の構造を主に考慮してフィッティングさせる。この方法論について以下に述べる。なお、第３５図では、局所構造の末端にグローバル構造の一部を貼り付けることによりモデルを伸長させる。以下に局所構造モデルの１～１０残基目がグローバル構造モデルの６～１５残基目と対応しているときのフィッティングの手順を一例に解説する。
（１）ＲＭＳＤによる閾値の設定
　フィッティングの良し悪しの状態を評価するためにＲＭＳＤを用いた。ＲＭＳＤ（根平均二乗距離：Ｒｏｏｔ　Ｍｅａｎ　Ｓｑｕａｒｅ　Ｄｅｖｉａｔｉｏｎ）とは二つの構造を重ね合わせ、対応する残基間（Ｃα間）の距離の離れ具合を表わす指標である。ＲＭＳＤが小さいほどよくフィッティングしているといえる。
　フィッティング時のＲＭＳＤによる閾値の初期値を２Å（デフォルト）とした。このＲＭＳＤの閾値は、閾値以下のフィッティングができなかった場合に１Åごとに高くしてい

く。条件を徐々に甘くすることによって最適なフィッティングを探索する。このＲＭＳＤの閾値を上限（デフォルト４Å）まで変化させることにより繰り返してフィッティングを行なう。これらの基準値は変更することも可能である。

（２）共通領域全体についてのフィッティング

はじめに共通領域全体についてフィッティングを行なう。ＲＭＳＤの閾値以下であればそのフィッティングを採用して終了する。共通領域全体においてよくフィッティングすることが理想であるが、互いの構造があまり似ていないことも多い。このようにＲＭＳＤが閾値よりも大きい場合は（３）に進んだ。

（３）残基を削って繰り返し再フィッティング

先のフィッティングに用いた領域から２残基削って再びフィッティングを行なった。ここで削る２残基は、伸長させたい方ではない末端からの２残基とした。ＲＭＳＤが閾値以下であればそのフィッティングを採用して終了するが、ＲＭＳＤが閾値よりも大きい場合は再び２残基削ってフィッティングを行なう、ということを繰り返す。

（４）ＲＭＳＤの閾値の変更

（３）により繰り返してフィッティングを行なうと、フィッティング領域が削られていく。フィッティング領域が３残基以下になった場合、短い残基でフィッティングさせても意味がない。このような場合フィッティングをせずにＲＭＳＤの閾値を１Å増やし、基準を甘くして（２）～（３）を繰り返した。このＲＭＳＤの閾値の変更とフィッティングは、ＲＭＳＤの閾値が（１）で設定した上限（デフォルト４Å）に達し、かつフィッティング領域が３残基以下になるまで行なわれる。

最後まで閾値以下のフィッティングができなかった場合は、フィッティング不可能とみなして終了する。

また、（３）、（４）によるＲＭＳＤの閾値、フィッティング残基数の変化は、第３６図に示す図を参照する。

（５）構造の貼り付け

（４）までフィッティングがうまく行なうことができた場合、フィッティングで用いた領域をのりしろとして、グローバル構造の座標を局所構造モデルの末端に貼り付けることによりモデルを伸長させる。この貼り付けのときに貼り付けられるグローバル構造部分が局所構造にぶつからないかを確認する。ここでぶつかってしまうようであれば、ぶつからなくなるまで貼り付けられるグローバル構造部分をカットする。

（６）ＦＡＭＳを用いて再モデリング（モデルを伸長させることができた場合のみ）

（５）までの作業で埋め込んで作成したＰＤＢ形式の座標ファイルを参照タンパク質として、ＦＡＭＳにより同一領域を再モデリング（ホモロジー１００％のモデリング）を行なった。再モデリングすることにより、つなぎ目部分の不自然な構造や、側鎖のぶつかりなどを解消して全体の形を整えることができる。

手順（３）：領域を移動させるフィッティングによる伸長

第３７図は、本発明によりフィッティング領域を移動させてフィッティングさせる場合の処理の一例を示すフローチャートである。本発明者は、本方法により、局所構造（短いモデル：ベースとなる）の末端にグローバル構造（長いモデル）の構造を貼り付けることによりモデルの伸長を行なう方法論を開発した。この方法では貼り付けようとする方の末端から反対の末端に向けてフィッティングさせる領域を移動させていく。これにより伸長させようとする末端の構造が少しくらい異なっても、フィッティング領域を内部に移動させることによりフィッティングを行なうことができる。この方法論について以下に述べる。なお、第３７図では、局所構造の末端にグローバル構造の一部を貼り付けることによりモデルを伸長させるものであり、局所構造モデルの１～２４残基目がグローバル構造モデルの６～２９残基目と対応しているときのフィッティングの手順を一例に解説する。

（１）ＲＭＳＤによる閾値、フィッティング残基数の設定

フィッティング時の基準であるＲＭＳＤによる閾値（デフォルト２Å）を設定する。この閾値はフィッティングができない場合に上限（デフォルト４Å）まで１Åずつ高くしていく。

次にフィッティング残基数（フィッティングさせる領域の残基数）の初期値（デフォルト２０残基）を設定する。このフィッティング残基数はフィッティングができない場合に下限（デフォルト１２残基）まで２残基ずつ削って繰り返しフィッティングさせる。

このＲＭＳＤの閾値、フィッティング残基数の二つの条件を変化させることにより、フィッティング時の基準を徐々に甘くして繰り返しフィッティングを行なう。これらの基準値は変更することも可能である。

（２）共通領域全体についてのフィッティング

はじめに共通領域全体についてフィッティングを行なった。ＲＭＳＤの閾値以下であればそのフィッティングを採用して（６）へ移る。ＲＭＳＤの閾値は２Å（初期値）から４Å（上限値）まで１Åずつ変化し、フィッティングの具合を調べる。共通領域全体においてよくフィッティングすることが理想であるが、互いの構造があまり似ていないことも多い。このようにＲＭＳＤが閾値よりも大きい場合は（３）に進んだ。

（３）フィッティング領域を移動させて繰り返しフィッティング

伸長させたい末端側からフィッティング残基数分の領域についてフィッティングを行なった。ＲＭＳＤが閾値以下であればそのフィッティングを採用するが、ＲＭＳＤが閾値よりも大きい場合はこのフィッティング領域を伸長させたい方ではない末端側に２残基ずらし、再びフィッティングを行なう。これをフィッティング領域が反対側の末端まで移動するまで繰り返した。

（４）ＲＭＳＤの閾値の変更

（３）により閾値以下のフィッティンができない場合、ＲＭＳＤの閾値を１Åずつ増やし条件を甘くして（３）を行なう。これはＲＭＳＤの閾値が２Å（初期値）から１Åずつ変化させ４Å（上限値）になるまで（４Åを含む）行なわれる。

（５）フィッティングさせる残基数を削る

（１）～（４）を行なっても（ＲＭＳＤの閾値が上限に達しても）フィッティングがうまくいかない場合、フィッティング残基数を２残基減らして（３）に戻ってフィッティングを繰り返す。この場合ＲＭＳＤの閾値も（１）で設定した初期値（デフォルト２Å）に戻り、その上限（デフォルト４Å）になるまで変化する。

（５）により繰り返してフィッティングを行なうと、フィッティング領域が削られていく。削ることによってフィッティング残基数が下限以下になった場合は、フィッティングが不可能とみなして終了する。

（２）から（５）の手順において、ＲＭＳＤの閾値、フィッティング残基数、フィッティング領域が複雑に変化する。この変化の順序について第３８図に示す表にまとめた。

（６）構造の貼り付け

（５）までででフィッティングがうまく行なうことができた場合、フィッティングで用いた領域をのりしろとして、グローバル構造の座標を局所構造モデルの末端に貼り付けることによりモデルを伸長させる。この貼り付けのときに貼り付けられるグローバル構造部分が局所構造にぶつからないかを確認する。ここでぶつかってしまうようであれば、ぶつからなくなるまで貼り付けられるグローバル構造部分をカットする。

（７）ＦＡＭＳを用いて再モデリング（モデルを伸長させることができた場合のみ）

（６）までの作業で埋め込んで作成したＰＤＢ形式の座標ファイルを参照タンパク質として、ＦＡＭＳにより同一領域の再モデリング（ホモロジー１００％のモデリング）を行なった。再モデリングすることにより、つなぎ目部分の不自然な構造や、側鎖のぶつかりなどを解消して全体の形を整えることができる。

手順（４）：二次構造等を主に考慮して断片を再検索して伸長

このプログラムでは、手順（１）、手順（２）、手順（３）の結果、遺伝子が完全長ではない場合にその座標をさらに伸長する方法を提案する。本方法は、以下の３段階に大きく分かれる。

１）二次構造データベース上で検索

２）二次構造のマッチの度合いで１）の結果をふるいにかける

３）二次構造データベースから選ばれた断片を用いて、伸長を行う

以下に、１）～３）について詳細に説明する。
１）　二次構造データベースの検索はプログラムＰＳＩ－ＢＬＡＳＴを用いている。二次構造データベースは、２、３、４、５個の二次構造単位を持ったデータベースがそれぞれ用意される。そのそれぞれに対して検索が行われる。検索が行われる順番は、５、４、３、２個の二次構造単位を持ったデータベースの順である。二次構造データベースの作成方法について後述する「二次構造データベースの作成法」を参照（第３９図参照）。
２）　ＰＳＩ－ＢＬＡＳＴで得られた多数のアライメントを、ｐｓｉ－ｐｒｅｄによる二次構造予測とのアライメントで得られたシークエンスの二次構造情報の一致度に基づいて振るいにかける。
３）　２）で残った複数のアライメントを用いて、ＦＡＭＳによるモデリングを行う。できた複数モデル（フラグメントモデル）を使って、伸長を行う。
（ｉ）　手順（１）、手順（２）、手順（３）で得られた座標データ（ベースモデル）とアミノ酸残基が一致する部分（のりしろ）を二残基以上持つフラグメントモデルを選び出す。
（ｉｉ）　のりしろ部分で、ベースモデルの末端から二残基を最小自乗フィットさせる。
（ｉｉｉ）　この段階で全長でない場合は、他のフラグメントモデルから更に伸長する。即ち（ｉ）（ｉｉ）を繰り返す。
ここで、第３９図および第４０図は、本発明による二次構造データベースの作成法を示すフローチャートである。二次構造データベースとは、クラスタリングされたＰＤＢデータベースを加工して作られたものである。具体的な方法として、ＰＤＢシークエンスとその二次構造の情報をもとに、２、３、４、５個の二次構造単位でシークエンスをフラグメント化したものである。このときの二次構造単位とは、ループ構造を含めないαヘリックス、βシート等を指す。
次に、本発明の入出力画面について説明する。
プログラムの入力として、引数に局所構造モデルの座標ファイル名とグローバル構造モデルの座標ファイル名を入力する。
手順（２）、手順（３）の方法では、これらの引数の他にどちらの末端がベースとなる局所構造であるのかを入力する。伸長させた構造は標準出力により出力される。
また手順（１）では目的タンパク質名を引数で入力することにより、伸長させた構造がこの名前のファイルに出力される。
手順（４）の方法では、手順（１）、手順（２）、手順（３）の方法で伸長されたＰＤＢファイル名を指定する。出力は単数および複数の座標ファイルとして出力される。
これらの出力は全て座標ファイル（ＰＤＢ形式）である。
（他の実施の形態）
さて、これまで本発明の実施の形態について説明したが、本発明は、上述した実施の形態以外にも、請求の範囲に記載した技術的思想の範囲内において種々の異なる実施の形態にて実施されてよいものである。
例えば、装置の分散・統合の具体的形態は明細書および図面に示すものに限られず、その全部または一部を、各種の負荷等に応じた任意の単位で、機能的または物理的に分散・統合して構成することができる（例えば、グリッド・コンピューティングなど）。
例えば、配列情報処理装置１００は、配列情報処理装置１００とは別筐体で構成されるクライアント端末からの要求に応じて処理を行い、その処理結果を当該クライアント端末に返却するように構成してもよい。
また、実施形態において説明した各処理のうち、自動的に行なわれるものとして説明した処理の全部または一部を手動的に行うこともでき、あるいは、手動的に行なわれるものとして説明した処理の全部または一部を公知の方法で自動的に行うこともできる。
この他、上記文書中や図面中で示した処理手順、制御手順、具体的名称、各種の登録データや検索条件等のパラメータを含む情報、画面例、データベース構成については、特記する場合を除いて任意に変更することができる。
また、配列情報処理装置１００に関して、図示の各構成要素は機能概念的なものであり

、必ずしも物理的に図示の如く構成されていることを要しない。

例えば、配列情報処理装置１００の各部または各装置が備える処理機能、特に制御部１０２にて行なわれる各処理機能については、その全部または任意の一部を、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）および当該ＣＰＵにて解釈実行されるプログラムにて実現することができ、あるいは、ワイヤードロジックによるハードウェアとして実現することも可能である。なお、プログラムは、後述する記録媒体に記録されており、必要に応じて配列情報処理装置１００に機械的に読み取られる。

すなわち、ＲＯＭまたはＨＤなどの記憶部１０６などには、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）と協働してＣＰＵに命令を与え、各種処理を行うためのコンピュータプログラムが記録されている。このコンピュータプログラムは、ＲＡＭ等にロードされることによって実行され、ＣＰＵと協働して制御部１０２を構成する。また、このコンピュータプログラムは、配列情報処理装置１００に対して任意のネットワーク３００を介して接続されたアプリケーションプログラムサーバに記録されてもよく、必要に応じてその全部または一部をダウンロードすることも可能である。

また、本発明にかかるプログラムを、コンピュータ読み取り可能な記録媒体に格納することもできる。ここで、この「記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＣＤ－ＲＯＭ、ＭＯ、ＤＶＤ等の任意の「可搬用の物理媒体」や、各種コンピュータシステムに内蔵されるＲＯＭ、ＲＡＭ、ＨＤ等の任意の「固定用の物理媒体」、あるいは、ＬＡＮ、ＷＡＮ、インターネットに代表されるネットワークを介してプログラムを送信する場合の通信回線や搬送波のように、短期にプログラムを保持する「通信媒体」を含むものとする。

また、「プログラム」とは、任意の言語や記述方法にて記述されたデータ処理方法であり、ソースコードやバイナリコード等の形式を問わない。なお、「プログラム」は必ずしも単一的に構成されるものに限られず、複数のモジュールやライブラリとして分散構成されるものや、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）に代表される別個のプログラムと協働してその機能を達成するものをも含む。なお、実施の形態に示した各装置において記録媒体を読み取るための具体的な構成、読み取り手順、あるいは、読み取り後のインストール手順等については、周知の構成や手順を用いることができる。

記憶部１０６に格納される各種のデータベース等（アミノ酸配列関連情報ファイル１０６ａ～再変換アミノ酸配列情報ファイル１０６ｇ）は、ＲＡＭ、ＲＯＭ等のメモリ装置、ハードディスク等の固定ディスク装置、フレキシブルディスク、光ディスク等のストレージ手段であり、各種処理やウェブサイト提供に用いる各種のプログラムやテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

また、配列情報処理装置１００は、既知のパーソナルコンピュータ、ワークステーション等の情報処理端末等の情報処理装置にプリンタやモニタやイメージスキャナ等の周辺装置を接続し、該情報処理装置に本発明の方法を実現させるソフトウェア（プログラム、データ等を含む）を実装することにより実現してもよい。

さらに、配列情報処理装置１００の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷等に応じた任意の単位で、機能的または物理的に分散・統合して構成することができる。例えば、各データベースを独立したデータベース装置として独立に構成してもよく、また、処理の一部をＣＧＩ（Ｃｏｍｍｏｎ　Ｇａｔｅｗａｙ　Ｉｎｔｅｒｆａｃｅ）を用いて実現してもよい。

また、ネットワーク３００は、複数の配列情報処理装置１００を相互に接続する機能を有し、例えば、インターネットや、イントラネットや、ＬＡＮ（有線／無線の双方を含む）や、ＶＡＮや、パソコン通信網や、公衆電話網（アナログ／デジタルの双方を含む）や、専用回線網（アナログ／デジタルの双方を含む）や、ＣＡＴＶ網や、ＩＭＴ２０００方式、ＧＳＭ方式またはＰＤＣ／ＰＤＣ－Ｐ方式等の携帯回線交換網／携帯パケット交換網や、無線呼出網や、Ｂｌｕｅｔｏｏｔｈ等の局所無線網や、ＰＨＳ網や、ＣＳ、ＢＳまたはＩＳＤＢ等の衛星通信網等のうちいずれかを含んでもよい。すなわち、本システムは、有線・無線を問わず任意のネットワークを介して、各種データを送受信することができる

。

以上詳細に説明したように、本発明によれば、アミノ酸配列情報を取得し、取得されたアミノ酸配列情報を構成する各アミノ酸情報に対応する二次構造を取得（例えば、アミノ酸配列情報に対応する立体構造が未知の場合には、既存の二次構造予測プログラムなどを用いて予測して取得、また、対応する立体構造が既知の場合には既存の二次構造判定プログラムなどを用いて取得）し、取得された二次構造およびアミノ酸配列情報を構成するアミノ酸情報に対応するアミノ酸の類似性（例えば、各アミノ酸の疎水性情報等のアミノ酸毎の性質に関する類似性等）に基づいて、アミノ酸配列情報を構成する各アミノ酸情報を、同一の二次構造および同一の類似性を有するアミノ酸情報を同一の情報として識別するための識別情報に変換することにより、アミノ酸配列情報を識別情報からなる識別配列情報に変換し、変換された複数の識別配列情報に対してアライメントを実行するので、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効果的に実行することができ、従来のホモロジー検索法と比べ、より遠縁の配列の探索が可能となる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、二次構造情報や疎水性情報なども考慮した探索・アライメントを行なうため、従来のアミノ酸の文字のみを考慮したアライメントと比較すると、立体構造も考慮したアライメントが可能となり、より高精度なアライメント作成が可能となるできる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明のように二次構造情報や疎水性情報などを考慮して作成したアライメントを用いると、疎水エネルギーが安定化し、また二次構造が分断していないモデルを作成でき、モデルとしても真実構造に近いものができると考えられる。

また、本発明によれば、二次構造および類似性に基づいて同一の二次構造および同一の類似性を有するアミノ酸情報を同一の識別情報に分類するための二次構造・類似性分類情報を格納し、変換手段は、二次構造・類似性分類情報に基づいて、アミノ酸配列情報を識別配列情報に変換するので、二次構造・類似性分類情報を予め作成して格納しておき、変換時に当該情報を参照することにより、アミノ酸配列情報の二次構造およびアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したアライメントを効率よく得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、変換された識別配列情報を格納し、変換手段は、格納された識別配列情報から、アミノ酸配列情報に対応する識別配列情報を検索するので、変換された識別配列情報をデータベースなどに予め格納しておき、変換時に当該データベースなどを参照することにより、変換処理を効率化、高速化することができるようになる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、これにより、例えば公共のデータベース（例えばＰＤＢなど）に登録されたアミノ酸配列情報の二次構造を既存の二次構造判定プログラムなどを用いて判定した後、当該アミノ酸配列情報を本発明により二次構造・類似性分類情報に基づいて変換して作成した識別配列情報を予めデータベースに格納して利用することができるようになり、取得したアミノ酸配列情報に対応する識別配列情報を効率よく検索することができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、二次構造の組み合わせに応じた構造スコア値が割り当てられた二次構造置換マトリックス、および／または、識別情報の組み合わせに応じた類似性スコア値が割り当てられた類似性置換マトリックスを格納し、アライメント実行手段は、格納された二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、置換マトリックスを予め作成して格納しておき、アライメント実行時に当該置換マトリックスを参照して各スコア値に置換することにより、アミノ酸配列情報の二次構造および／またはアミノ酸配列情報を構成する各アミノ酸情報の類似性を加味したスコア値に置換することにより最適なアライメントを

効率よく得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、置換マトリックス基準アライメント実行手段は、予め定めた係数により重み付けされた二次構造置換マトリックスおよび／または類似性置換マトリックスに基づいて、識別配列情報を置換してアライメントを実行するので、例えば生物学的な知見などに基づいて二次構造置換マトリックスおよび／または類似性置換マトリックスに対して適切な係数を設定して重み付けをすることによって、生物学的な知見などを反映した最適なアライメントを効率よく得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、アライメントが実行された識別配列情報を構成する識別情報を、対応するアミノ酸情報に再変換するので、二次構造および類似性を加味してアライメントが実行されたアミノ酸配列情報を効率よく得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明によれば、二次構造・類似性分類情報は、二次構造を、αヘリックス構造、βシート構造、または、その他の構造の３つに分類し、類似性を、アラニン（Ａｌａ）とグリシン（Ｇｌｙ）からなる第１の群、アスパラギン酸（Ａｓｐ）とグルタミン酸（Ｇｌｕ）とアスパラギン（Ａｓｎ）とグルタミン（Ｇｌｎ）からなる第２の群、システイン（Ｃｙｓ）からなる第３の群、フェニルアラニン（Ｐｈｅ）とヒスチジン（Ｈｉｓ）とトリプトファン（Ｔｒｐ）とチロシン（Ｔｙｒ）からなる第４の群、イソロイシン（Ｉｌｅ）とロイシン（Ｌｅｕ）とメチオニン（Ｍｅｔ）とバリン（Ｖａｌ）からなる第５の群、リシン（Ｌｙｓ）とアルギニン（Ａｒｇ）からなる第６の群、プロリン（Ｐｒｏ）とセリン（Ｓｅｒ）とトレオニン（Ｔｈｒ）からなる第７の群の７つに分類するので、既知の置換マトリックスであるＢＬＯＳＵＭ６２に基づいてアミノ酸を７つの群に分類することによりアミノ酸情報の類似性を加味し、かつ、二次構造を３つに分類することによりアミノ酸配列情報の二次構造を加味して、アライメントが実行された識別配列情報を効率よく得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

さらに、本発明によれば、アミノ酸の類似性は、アミノ酸の疎水性、親水性、酸性、塩基性、荷電状態のうち少なくとも一つに基づく情報に関する類似性であるので、アミノ酸配列情報の立体構造などに影響するアミノ酸の性質（疎水性、親水性、酸性、塩基性、荷電状態など）の類似性を考慮してアライメントが実行された識別配列情報を効果的に得ることができる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体を提供することができる。

また、本発明をより効果的に用いるためにホモロジー検索において、ｅ値だけでなく、二次構造や疎水性相互作用によりモデルの立体構造をも考慮したモデルの順位付けが重要と考えられる。この全体の流れの中でｐｒｅｄ＿ｆａｓｔａを用いることにより、アライメント（ｅ値）、立体構造（二次構造・疎水性相互作用）を考慮したスコア付けを行うことが出来、このスコアが大きいほどアライメント的にもモデル構造的にも精度の高いモデルの構築が期待できるといえる。

本発明によれば、ＰＲＥＤ＿ＦＡＳＴＡは従来のホモロジー検索法と比べ、より遠縁の配列の探索が可能である。また、二次構造情報や疎水性情報なども考慮した探索・アライメントを行なうため、従来のアミノ酸の文字のみを考慮したアライメントと比較すると、より高精度なアライメント作成が可能であり、また立体構造も考慮したアライメントが可能になったといえる。

このように二次構造情報や疎水性情報を考慮して作成したアライメントを用いると、疎水エネルギーが安定化し、また二次構造が分断していないモデルを作成でき、モデルとしても真実構造に近いものができると考えられる。

ＰＲＥＤ＿ＦＡＳＴＡは、従来のホモロジー検索プログラムの短所を克服したホモロジー検索、アライメントプログラムであり、画期的なホモロジー検索の最終ツールとして生命科学分野の発展に大きく貢献するものと期待される。

また、本発明の適用後に、構造がよいと思われるが領域の短いモデル（局所構造）を領域の長いモデルの構造を用いて、その局所構造を生かしつつ伸長させることができ、特に上述した手順（４）においては、遺伝子の完全長モデルを構築することはより効果的な適用法であると考えられる。

【産業上の利用可能性】

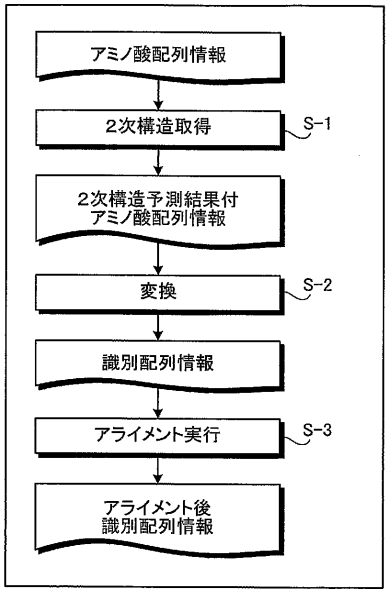
以上のように、本発明にかかる配列情報処理装置、配列情報処理方法、プログラムおよび記録媒体は、アミノ酸配列の二次構造および／またはアミノ酸の類似性を加味したアライメントを得ることができる。

これにより、本発明は、配列情報に対するホモロジー検索やアミノ酸配列のアライメントやタンパク質の立体構造などの解析を行うバイオインフォマティクス分野や生命科学分野において好適に利用することができる。

本発明は、産業上多くの分野、特に、医療、化学、製薬、食品等の分野で広く実施することができ、極めて有用である。

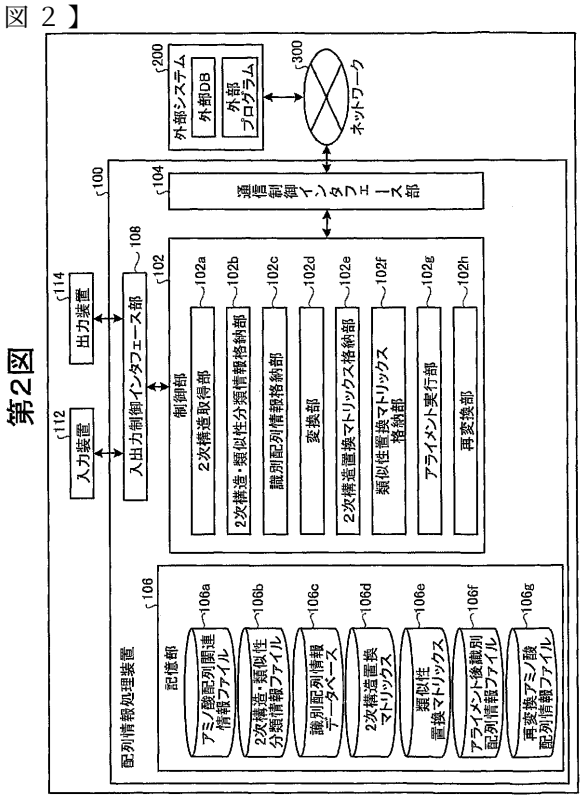
【図１】

第1図



【図２】

第2図



【図３】

第3図

アミノ酸配列関連情報ファイル
106a

アミノ酸配列識別情報		情報				
A001	アミノ酸情報	Ala	Phe	Trp	…	Lys
	2次構造情報	α	α	その他	…	β
	類似性情報	1	4	4	…	6
	識別情報	A	K	M	…	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮

【図４】

第4図

2次構造・類似性分類情報ファイル
106b

類似性情報	アミノ酸情報	2次構造情報		
		α	β	その他
1	Ala Gly	A	B	C
2	Asp Glu Asn Gln	D	E	F
3	Cys	G	H	I
4	Phe His Trp Tyr	K	L	M
5	Ile Leu Met Val	N	P	Q
6	Lys Arg	R	S	T
7	Pro Ser Thr	V	W	X

【図５】

第5図

識別配列情報データベース
106c

アミノ酸配列識別情報		情報				
K_001	アミノ酸情報	Ala	Phe	Trp	…	Lys
	識別情報	A	K	M	…	S
⋮	⋮	⋮	⋮	⋮	⋮	⋮

【図６】

第6図

2次構造置換マトリックス
106d

	α	β	その他
α	6	−6	−1
β	−6	6	−1
その他	−1	−1	6

【図７】

第7図

類似性置換マトリックス
106e

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	1	0	−1	0	1	−1	0	0	0	0	−1	−1	−1	0	−1	0	0	0	0	0	1	0	0
R	0	1	−1	0	0	−1	0	−1	−1	−1	−1	−1	−1	0	−1	1	1	0	0	0	0	0	0
N	−1	−1	1	−1	−1	1	−1	0	0	0	0	0	0	−1	1	−1	−1	0	0	0	−1	0	0
D	0	0	−1	1	0	−1	1	−1	−1	−1	−1	−1	−1	1	−1	0	0	0	0	0	0	0	0
C	1	0	−1	0	1	−1	0	0	0	0	−1	−1	−1	0	−1	0	0	0	0	0	1	0	0
Q	−1	−1	1	−1	−1	1	−1	0	0	0	0	0	0	−1	1	−1	−1	0	0	0	−1	0	0
E	0	0	−1	1	0	−1	1	−1	−1	−1	−1	−1	−1	1	−1	0	0	0	0	0	0	0	0
G	0	−1	0	−1	0	0	−1	4	4	4	−1	−1	−1	−1	0	−1	−1	0	0	0	0	0	0
H	0	−1	0	−1	0	0	−1	4	4	4	−1	−1	−1	−1	0	−1	−1	0	0	0	0	0	0
I	0	−1	0	−1	0	0	−1	4	4	4	−1	−1	−1	−1	0	−1	−1	0	0	0	0	0	0
L	−1	−1	0	−1	−1	0	−1	−1	−1	−1	1	1	1	−1	0	−1	−1	−1	0	−1	−1	0	−1
K	−1	−1	0	−1	−1	0	−1	−1	−1	−1	1	1	1	−1	0	−1	−1	−1	0	−1	−1	0	−1
M	−1	−1	0	−1	−1	0	−1	−1	−1	−1	1	1	1	−1	0	−1	−1	−1	0	−1	−1	0	−1
F	0	0	−1	1	0	−1	1	−1	−1	−1	−1	−1	−1	1	−1	0	0	0	0	0	0	0	0
P	−1	−1	1	−1	−1	1	−1	0	0	0	0	0	0	−1	1	−1	−1	0	0	0	−1	0	0
S	0	1	−1	0	0	−1	0	−1	−1	−1	−1	−1	−1	0	−1	1	1	0	0	0	0	0	0
T	0	1	−1	0	0	−1	0	−1	−1	−1	−1	−1	−1	0	−1	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	−1	−1	−1	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	−1	−1	−1	0	0	0	0	0	0	0	0	0	0
B	1	0	−1	0	1	−1	0	0	0	0	−1	−1	−1	0	−1	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	−1	−1	−1	0	0	0	0	0	0	0	0	0	0

【図８】

第8図

アライメント後識別配列情報ファイル
106f

アミノ酸配列識別情報	識別配列情報				
A001	A	B	C	…	P
⋮	⋮	⋮	⋮	⋮	⋮

【図９】

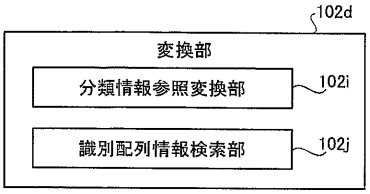
第9図

再変換アミノ酸配列情報ファイル
106g

アミノ酸配列識別情報	アミノ酸配列情報				
A001	Ala	Phe	Trp	…	Lys
⋮	⋮	⋮	⋮	⋮	⋮

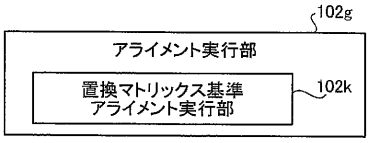
【図１０】

第10図



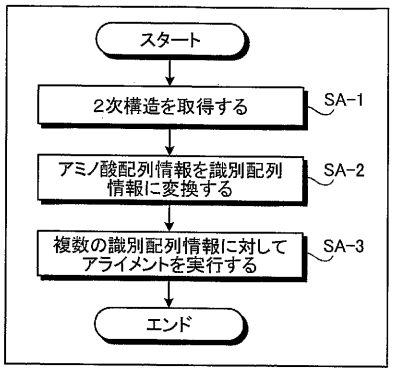
【図１１】

第11図



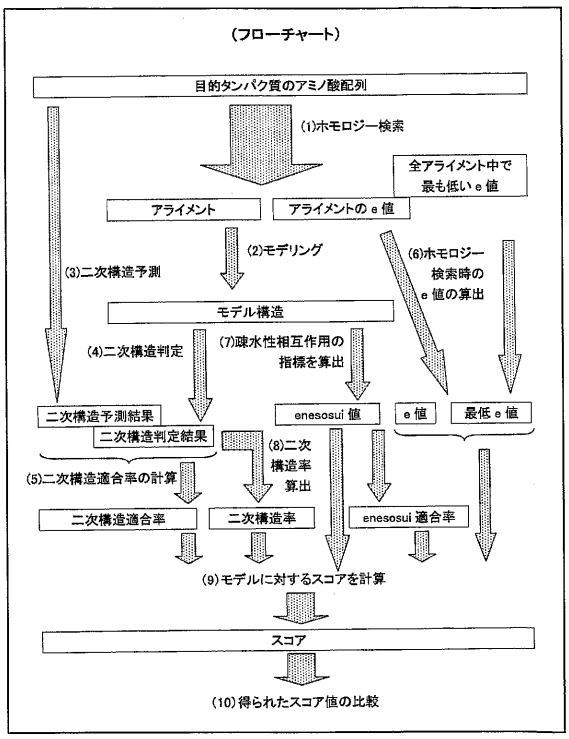
【図１２】

第12図



【図１３】

第13図



【図１４】

第14図

MA-1	MA-2	MA-3	MA-4	MA-5	MA-6
1	M	C	0.986	0.009	0.005
2	N	C	0.921	0.043	0.021
3	E	C	0.721	0.282	0.014
4	S	H	0.470	0.545	0.011
5	E	H	0.065	0.928	0.008

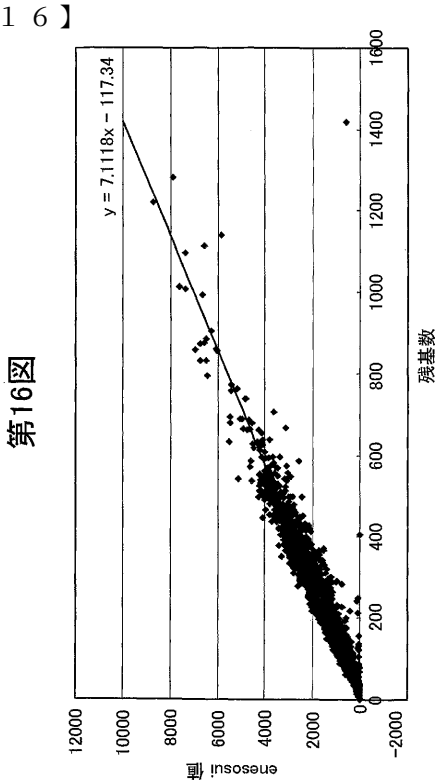
【図１５】

第15図

残基名	原子			
ALA	CB			
VAL	CG1	CG2		
PHE	CB	CD1	CD2	CZ
PRO	CG			
MET	CG	CE		
ILE	CG1	CG2	CD1	
LEU	CB	CD1	CD2	
ASP				
GLU	CB			
LYS	CB	CD		
ARG	CG			
SER				
THR	CG2			
TYR	CB	CE1	CE2	
HIS	CB			
CYS	CB			
ASN	CB			
GLN	CB	CG		
TRP	CB	CD2	CZ2	CZ3
GLY				

【図１６】

第16図



【図１７】

第17図

```
>test_sequence
MDIRTITSSDYEMVTSVLNEWWGGRQLKEKLPRLFFEHFQDTSFITSEHNSMTGFL
IGFQSQSDPETAYIHFSGVHPDFRKMQIGKQLYDVFIETVKQRGCTRVKCVTSPVNK
VSIAYHTKLGFDIEKGTKTVNGISVFANYDGPGQDRVLFVKNI
```

【図１８】

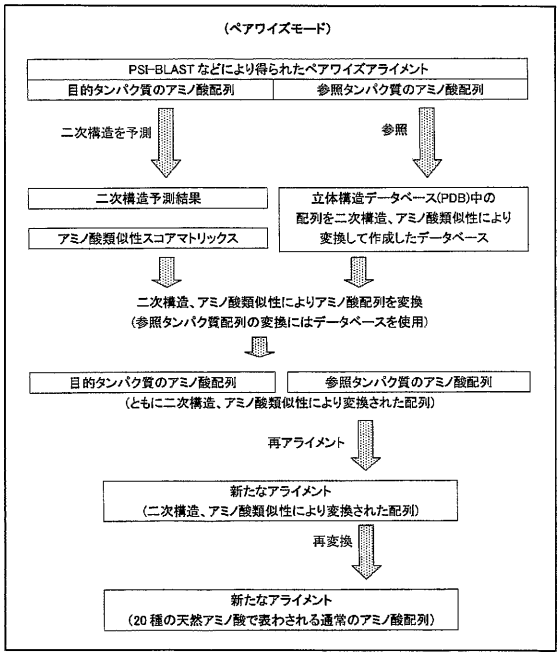
第18図

```
  MB-1    MB-2          MB-3        MB-4          MB-5          MB-6
5GCN_A length=166     score=721.23  Value=4e-41   Hom=15 round=2 rank=1   region=1
        MB-7   MB-8        MB-9     MB-10  MB-11 MB-12  MB-13   MB-14   MB-15
       13-156  24-162       PSI     PDB95 -961.000 0.88   -6.67   108.131  0.75

5GCN_A length=166     score=719.11  Value=4e-41   Hom=15 round=1   rank=1   region=1
        1-156   5-158       RPS     SCOP95 -997.000 0.77  -6.39   112.536  0.72
```

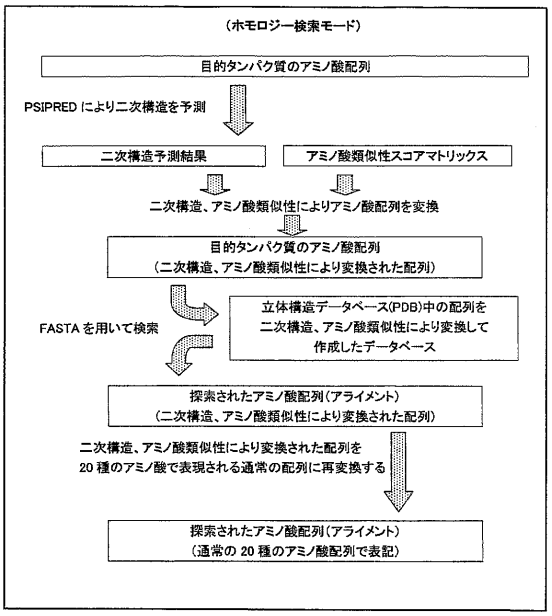
【図１９】

第19図



【図２０】

第20図



【図２１】

第21図

	グループに所属するアミノ酸（アミノ酸類似性による分類）	便宜上付けられるアミノ酸一文字表記		
		αヘリックス	βシート	その他
1	Ala Gly	A	B	C
2	Asp Glu Asn Gln	D	E	F
3	Cys	G	H	I
4	Phe His Trp Tyr	K	L	M
5	Ile Leu Met Val	N	P	Q
6	Lys Arg	R	S	T
7	Pro Ser Thr	V	W	X

【図２２】

第22図

	αヘリックス	βシート	その他
αヘリックス	6	−6	−1
βシート	−6	6	−1
その他	−1	−1	6

【図２３】

第23図

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	1	0	-1	0	1	-1	0	0	0	0	-1	-1	-1	0	-1	0	0	0	0	0	1	0	0
R	0	1	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	1	1	0	0	0	0	0	0
N	-1	-1	1	-1	-1	1	-1	0	0	0	0	0	0	-1	1	-1	-1	0	0	0	-1	0	0
D	0	0	-1	1	0	-1	1	-1	-1	-1	-1	-1	-1	1	-1	0	0	0	0	0	0	0	0
C	1	0	-1	0	1	-1	0	0	0	0	-1	-1	-1	0	-1	0	0	0	0	0	1	0	0
Q	-1	-1	1	-1	-1	1	-1	0	0	0	0	0	0	-1	1	-1	-1	0	0	0	-1	0	0
E	0	0	-1	1	0	-1	1	-1	-1	-1	-1	-1	-1	1	-1	0	0	0	0	0	0	0	0
G	0	-1	0	-1	0	0	-1	4	4	4	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0
H	0	-1	0	-1	0	0	-1	4	4	4	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0
I	0	-1	0	-1	0	0	-1	4	4	4	-1	-1	-1	-1	0	-1	-1	0	0	0	0	0	0
L	-1	-1	0	-1	-1	0	-1	-1	-1	-1	1	1	1	-1	0	-1	-1	-1	0	-1	-1	0	-1
K	-1	-1	0	-1	-1	0	-1	-1	-1	-1	1	1	1	-1	0	-1	-1	-1	0	-1	-1	0	-1
M	-1	-1	0	-1	-1	0	-1	-1	-1	-1	1	1	1	-1	0	-1	-1	-1	0	-1	-1	0	-1
F	0	0	-1	1	0	-1	1	-1	-1	-1	-1	-1	-1	1	-1	0	0	0	0	0	0	0	0
P	-1	-1	1	-1	-1	1	-1	0	0	0	0	0	0	-1	1	-1	-1	0	0	0	-1	0	0
S	0	1	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	1	1	0	0	0	0	0	0
T	0	1	-1	0	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	1	1	0	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	0	0	0	0	0	0	0	0	0
B	1	0	-1	0	1	-1	0	0	0	0	-1	-1	-1	0	-1	0	0	0	0	0	1	0	0
Z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	0	0	0	0	0	0	0	0	0	0

【図２４】

第24図

```
>test1
VLSAADKTNVKGVFSKISGHAEEYGAETLERMFTAYPQTKTYFPHFDLQHGSAQIK
AHGKKVVAALVEAVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAI
HHPSALTAEVHASLDKFLCAVGTVLTAKYR
```

【図２５】

第25図

```
>>1A4F_A A                                                (141 aa)
 initn: 892 init1: 892 opt: 892  Z-score: 134.4  bits: 30.7 E(): 0.065
Smith-Waterman score: 892;  91.489% identity in 141 aa overlap (1-141:1-141)

               10        20        30        40        50        60
test1  VLSAADKTNVKGVFSKISGHAEEYGAETLERMFTAYPQTKTYFPHFDLQHGSAQIKAHGK
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
1A4F_A VLSAADKTNVKGVFSKISGHAEEYGAETLERMFTAYPQTKTYFPHFDLQHGSAQIKAHGK
               10        20        30        40        50        60

               70        80        90       100       110       120
test1  KVVAALVEAVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALTAE
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
1A4F_A KVVAALVEAVNHIDDIAGALSKLSDLHAQKLRVDPVNFKFLGHCFLVVVAIHHPSALTAE
               70        80        90       100       110       120

              130       140
test1  VHASLDKFLCAVGTVLTAKYR
       :::::::::::::::::::::
1A4F_A VHASLDKFLCAVGTVLTAKYR
              130       140
```

【図２６】

第26図

```
>CAFASP9058
MANAPKGVKFSTSQTTEEIVRAKVSELVEQYDLIIGIPIDKEISASYANWKLVEKDY
ASKFHVLDARAVEMIIDWLINDIKGWLNTNPYSREGLDRFVEQYRKKTAAVLFVTD
TKPLVAGGRLSNLKSFIIKSLKFHLLISFLGENGKLQFFDKARSAHDAHKLAVKFL
KKQLLKHSAKLKRGAFLTTVFDEQTNTNLVQEFDKLLDHSINIEQSLLSPVICTHT
GLNSYVIVLQGE
```

【図２７】

第27図

```
>>1E0J_F F                                                   (288 aa)
 initn: 344 init1: 205 opt: 727  Z-score: 128.0  bits: 31.3 E(): 0.14
Smith-Waterman score: 737;  18.797% identity in 266 aa overlap (1-236:19-274)

                                  10        20        30        40
CAFASP                  MANAPKGVKFSTSQTTEEIVRAKVSELVEQYDLIIGIPIDKE

1E0J_F PDGVVSALSLRERIREHLSSEESVGLLFSGCTGINDKTLGA-----RGGEVIMVTSGSGM
               10        20        30        40               50

              50          60        70        80        90
CAFASP ISASYANWKLVEKD--YASKFHVLDARAVEMIIDWLINDIKGWLNTNPYSREGLD-----

1E0J_F GKSTFVRQQALQWGTAMGKKVGLAMLEESVEETAEDLIGLHNRVRLRQSDSLKREIIENG
           60        70        80        90       100       110

          100       110       120       130       140       150
CAFASP RFVEQYRKKTAAVLFVTDTKPLVAGGRLSNLKSFIIKSLKFHLLISFLGENGKLQFFDKA

1E0J_F KFDQWFDELFGNDTFHLYDSFAEAETDRLLAKLAYMRSGLGCDVIILDHISIVVSASGES
          120       130       140       150       160       170

          160       170       180       190       200
CAFASP RSAHDAHKLAVKFLKKQLLKHSAKLKRGAFLTTVFDEQTNTNLVQE------------F

1E0J_F DERKMIDNLMTKLKGFAKSTGV-----VLVVICHLKNPDKGKAHEEGRPVSITDLRGSGA
          180       190            200       210       220       230

             210           220              230
CAFASP DKLLDHSINIEQSL----LSPVICTH------TGLNSYVIVLQGE

1E0J_F LRQLSDTIIALERNQQGDMPNLVLVRILKCRFTGDTGIAGYMEYNKETGWLEPSSYSG
             240       250       260       270       280
```

【図２８】

第28図

PSI-BALSTによるアライメント

>CAFASP8880
MSAVTVNDDGLVLRLYIQPKASRDSIVGLHG-DEVKVAITAPPVDGQANSHLVKFLGKQFRVAKSQVVIEKGELGRHKQIKI
>1JRM_A
MDCLREVGDDLLVNIEVSPASGKFGIPSYNEWRKRIEVKIHSPPQKGKANREIIKEFSETFG---RDVEIVSGQKSRQKTIRI

PRED_FASTAによるアライメント　（長くアライメントが得られた部分）

> CAFASP8880
TVNDDGLVLRLYIQPKASRDSIVGLHGDEVKVAITAPPVDGQANSHLVKFLGKQFRVAKSQVVIEKGELGRHKQIKI--INPQQIPPEVAALIN
>1JRM_A
LREVGDDLLVNIEVSPASGKFGIPSYNEWRKRIEVKIHSPPQKGKANREIIKEFSETFGRDVEIVSGQKSRQKTIRIQGMGRDLFLKLVSEKFG

【図２９】

第29図

>CAFASP8880
MSAVTVNDDGLVLRLYIQPKASRDSIVGLHG--DEVKVAITAPPVDGQANSHLVKFLGKQFRVAKSQVVIEKGELGRHKQIKI-INPQQIPPEVAALIN
>1JRM_A
MDCLREVGDDLLVNIEVSPASGKFGIPSYNEWRKRIEVKIHSPPQKGKANREIIKEFSETFG---RDVEIVSGQKSRQKTIRIQGMGRDLFLKLVSEKFG

【図３０】

第30図

>test3
VLSAADK---TNVKGVFSKISGHAEEYGAET
>1A4F_A
VLSAADKTNVKG---VFSKISGHAEEYGAET

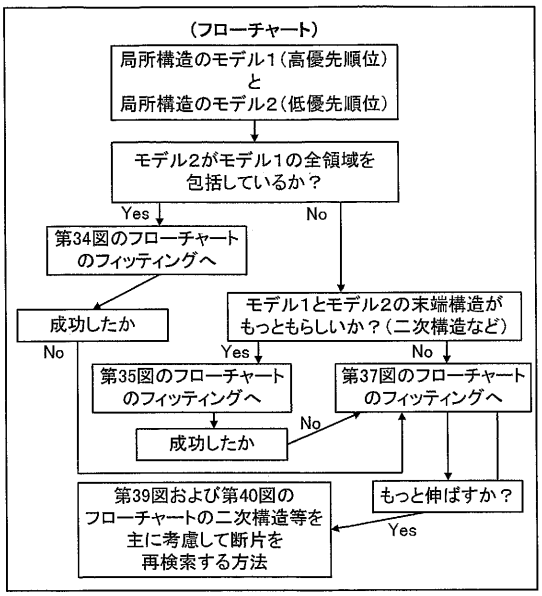
【図３１】

第31図

>test3
VLSAADKTNVKGVFSKISGHAEEYGAET
>1A4F_A
VLSAADKTNVKGVFSKISGHAEEYGAET

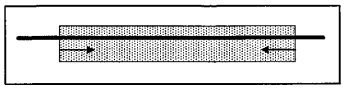
【図３２】

第32図



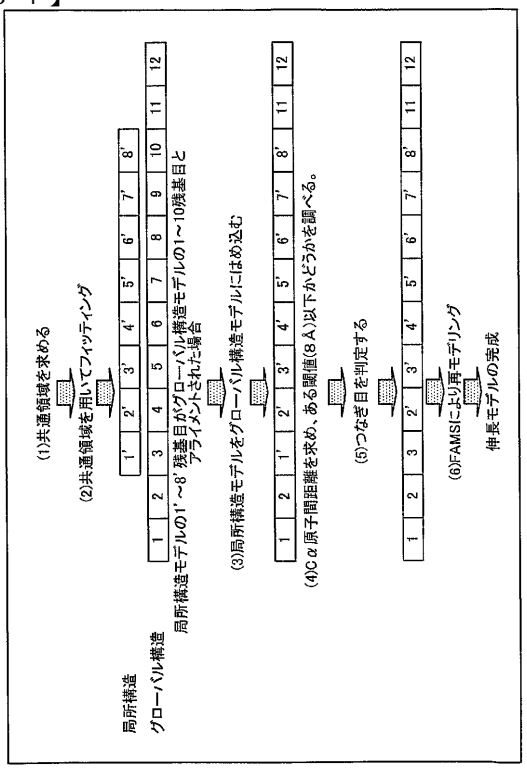
【図３３】

第33図



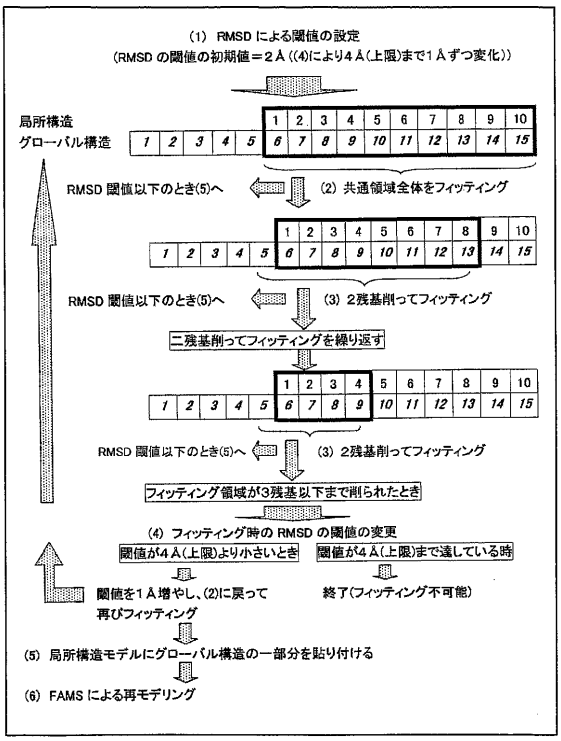
【図３４】

第34図



【図３５】

第35図



【図３６】

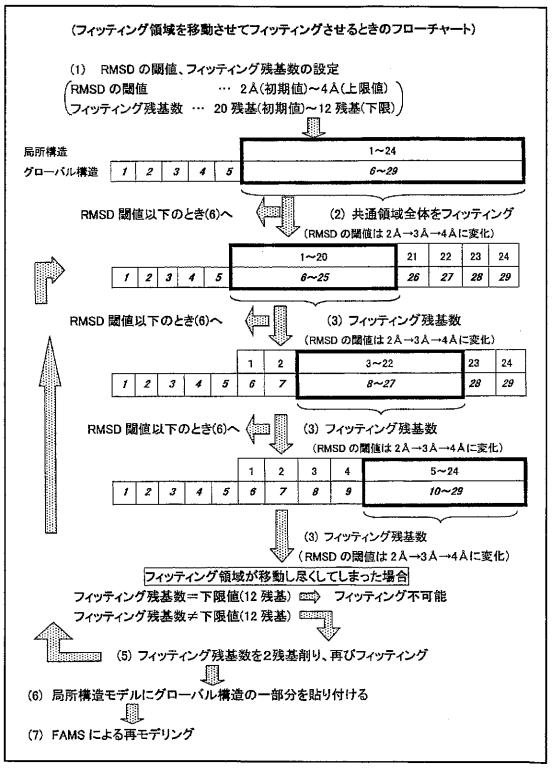
第36図

(RMSDの閾値とフィッティング残基数の変化)

RMSDの閾値	フィッティング残基数(伸長させない末端側から2残基ずつ削る)
2Å (初期値)	共通領域全体 → 伸長させる末端から20残基 → 18 → … → 4残基
3Å	共通領域全体 → 伸長させる末端から20残基 → 18 → … → 4残基
4Å (上限値)	共通領域全体 → 伸長させる末端から20残基 → 18 → … → 4残基

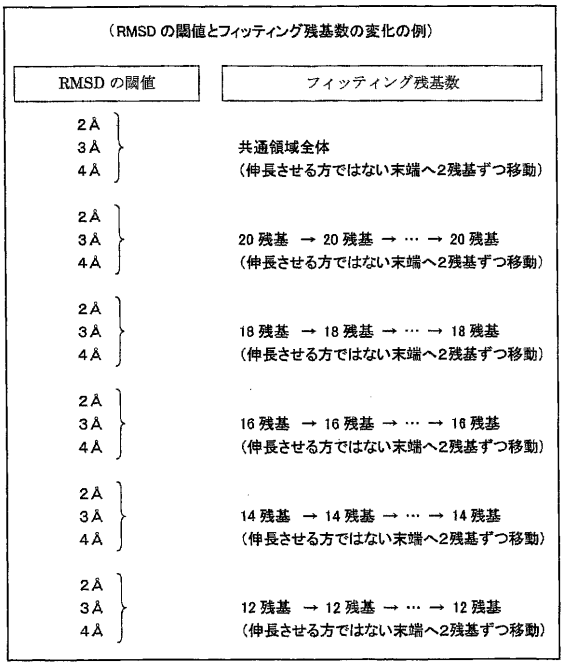
【図３７】

第37図



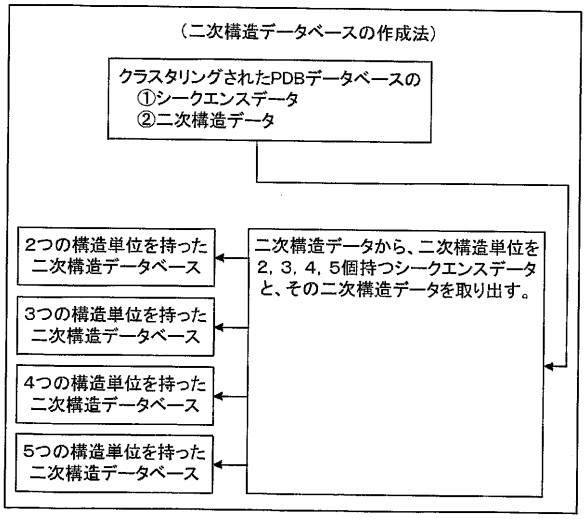
【図３８】

第38図



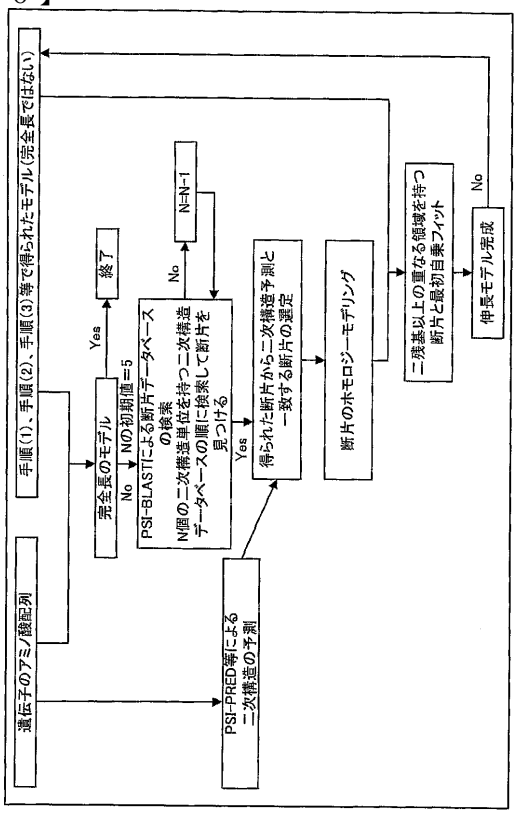
【図３９】

第39図



【図４０】

第40図



【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP03/15245

A. CLASSIFICATION OF SUBJECT MATTER
Int.Cl[7] G06F19/00, G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
Int.Cl[7] G06F19/00, G06F17/30

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Jitsuyo Shinan Koho 1922-1996 Toroku Jitsuyo Shinan Koho 1994-2004
Kokai Jitsuyo Shinan Koho 1971-2004 Jitsuyo Shinan Toroku Koho 1996-2004

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
JICST FILE(JOIS), WPI, INSPEC(DIALOG)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	FISCHEL-GHODSIAN, F. et al., 'Alignment of prote in sequences using secondary structure: a modifi ed dynamic programming method', PROTEIN ENGINEER ING, July 1990, Vol.3, No.7, pages 577 to 581, especially, pages 577 to 578, Methods	1-25

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search
03 February, 2004 (03.02.04)

Date of mailing of the international search report
17 February, 2004 (17.02.04)

Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

国際調査報告	国際出願番号 PCT／JP03／15245

A. 発明の属する分野の分類（国際特許分類（IPC））

Int. Cl⁷ G06F19／00, G06F17／30

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int. Cl⁷ G06F19／00, G06F17／30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922－1996年
日本国公開実用新案公報	1971－2004年
日本国登録実用新案公報	1994－2004年
日本国実用新案登録公報	1996－2004年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

JICSTファイル（JOIS），WPI，INSPEC（DIALOG）

C. 関連すると認められる文献

引用文献のカテゴリー＊	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求の範囲の番号
X	FISCHEL-GHODSIAN,F. et al.'Alignment of protein sequences using secondary structure:a modified dynamic programming method' PROTEIN ENGINEERING,July 1990,Vol.3,No7,p.577-581 especially p.577-578,Methods	1－25

☐ C欄の続きにも文献が列挙されている。 ☐ パテントファミリーに関する別紙を参照。

＊ 引用文献のカテゴリー
「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日	国際調査報告の発送日
03. 02. 2004	17. 2. 2004

国際調査機関の名称及びあて先	特許庁審査官（権限のある職員）	
日本国特許庁（ISA／JP） 郵便番号100－8915 東京都千代田区霞が関三丁目4番3号	高瀬 勤	5M 9069
	電話番号 03－3581－1101 内線 3597	

様式PCT／ISA／210（第2ページ）（1998年7月）

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HU,IE,IT,LU,MC,NL,PT,RO,SE,SI,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NI,NO,NZ,OM,PG,PH,PL,PT,RO,RU,SC,SD,SE,SG,SK,SL,SY,TJ,TM,TN,TR,TT,TZ,UA,UG,US,UZ,VC,VN,YU,ZA,ZM,ZW

（特許庁注：以下のものは登録商標）
Ｂｌｕｅｔｏｏｔｈ

（注）この公表は、国際事務局（ＷＩＰＯ）により国際公開された公報を基に作成したものである。なおこの公表に係る日本語特許出願（日本語実用新案登録出願）の国際公開の効果は、特許法第１８４条の１０第１項(実用新案法第４８条の１３第２項）により生ずるものであり、本掲載とは関係ありません。