特集

蛋白モチーフ自動抽出システム：SODHO

SODHO: Automatic Protein-Motif Extraction System

あらまし

共通祖先を持つ相同蛋白のアミノ酸に共通して現れるモチーフと呼ばれる部分配列は，相同蛋白の分子進化の過程で保存されたものであり，蛋白の機能をアミノ酸配列から推定する上で重要な手がかりとなる。富士通は1990年より国立遺伝学研究所と共同研究を実施し，相同蛋白のアミノ酸配列間の保存プロファイル（配列パターン）を自動抽出しモチーフを自動推定する分子進化解析ソフト“SODHO”を開発してきた。当初のSODHOでは推定されたモチーフを正規表現系で表現していたため実利用上の制約があった。最近，正規表現系の隠れマルコフモデル化とモチーフ比較のための新規アルゴリズム開発を実現しSODHOをより実用的なシステムとした。現在，完全長cDNAプロジェクトなどから得られる未知・新規の蛋白からの目的蛋白探索に適用し成果を上げている。

本稿では，SODHOの実現技術および最近の成果について述べる。

Abstract

The partial sequences of amino acids, which appear in common with homologous proteins, are called motifs. Motifs are extremely important in predicting the functions of proteins. Since 1990, Fujitsu has been collaborating with NIG in developing a molecular evolutionary analysis software called SODHO for automatically extracting the preservative profiles (sequence patterns) of the amino acids of homologous proteins and generating motifs represented by regular expressions. Recently, we have been developing SODHO to use the Hidden-Markov-Model to represent motifs, which enables SODHO to be used as a more practical system. SODHO has been applied to the strange, new proteins obtained from full-length cDNA projects and it has provided excellent results. In this paper, we discuss the structure of SODHO and the latest results obtained by using it.

内藤公敏（ないとう　きみとし）

ライフサイエンス推進室 所属
現在，創薬研究ソフトウェアシステムの研究開発に従事。

ま　え　が　き

蛋白には機能や構造を担うモチーフと呼ばれる部位があり，相同蛋白（共通祖先を持つ蛋白群）間に共通なアミノ酸配列パターンとして保存されていることが知られている（**図**-1）。このことから筆者らは「相同蛋白間で保存されている共通アミノ酸配列パターン（以下，保存プロファイル）」をコンピュータ解析で抽出することで蛋白のモチーフを自動推定できると考え，そのコンピュータ解析手法の実現について国立遺伝学研究所（以下，遺伝研）と共同研究を実施してきた。共同研究の成果として，相同蛋白のアミノ酸配列から保存プロファイルを自動抽出しモチーフを自動推定する分子進化解析ソフトウェアが開発された。このソフトウェアは相同蛋白の「相同」から名前を取りSODHO（Sequence of DNA homologue）と名付けられた。共同研究ではSODHOを用いてアミノ酸配列既知の全蛋白を対象にモチーフを自動抽出し，その分子進化学的解析に成果をあげた[1]。しかし，当初のSODHOでは推定されたモチーフの表現に正規表現系を用いていたためモチーフをキーにしたアミノ酸配列の検索やモチーフ間相互の類似性の定量的把握などが困難であり実用ツールとしては不十分であった。最近，SODHOを実用ツールとする目的でモチーフ表現を正規表現系より情報量が多く利便性の高い「隠れマルコフモデル」へ拡張した。この結果，モチーフをキーにしたアミノ酸配列の検索やモチーフ間相互の類似性に関する定量的把握などの実用的機能が実現された。

本稿ではSODHOの実現技術と開発成果について紹介する。

モチーフ自動推定

モチーフの自動推定は**図**-2に示すように，

（1）　既知の全蛋白アミノ酸配列収集

（2）　収集された蛋白の相同性分類

（3）　相同蛋白アミノ酸配列のマルチプルアラインメント

（4）　マルチプルアラインメント配列からの保存領域推定と保存プロファイル抽出

（5）　保存プロファイルの正規表現によるモチーフ生成

の手順で行われる。それぞれの処理，およびアルゴリズムの詳細について，以下に述べる。

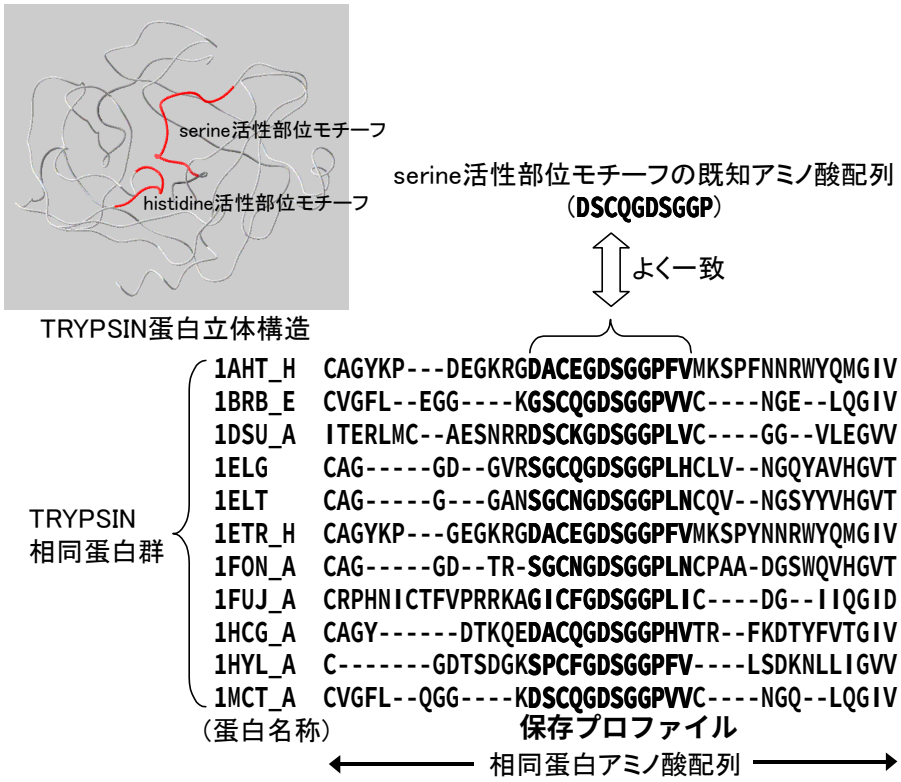

図-1　相同蛋白間で保存されている機能モチーフの例
Fig.1-Example of functional motif conserved among homologous protein.

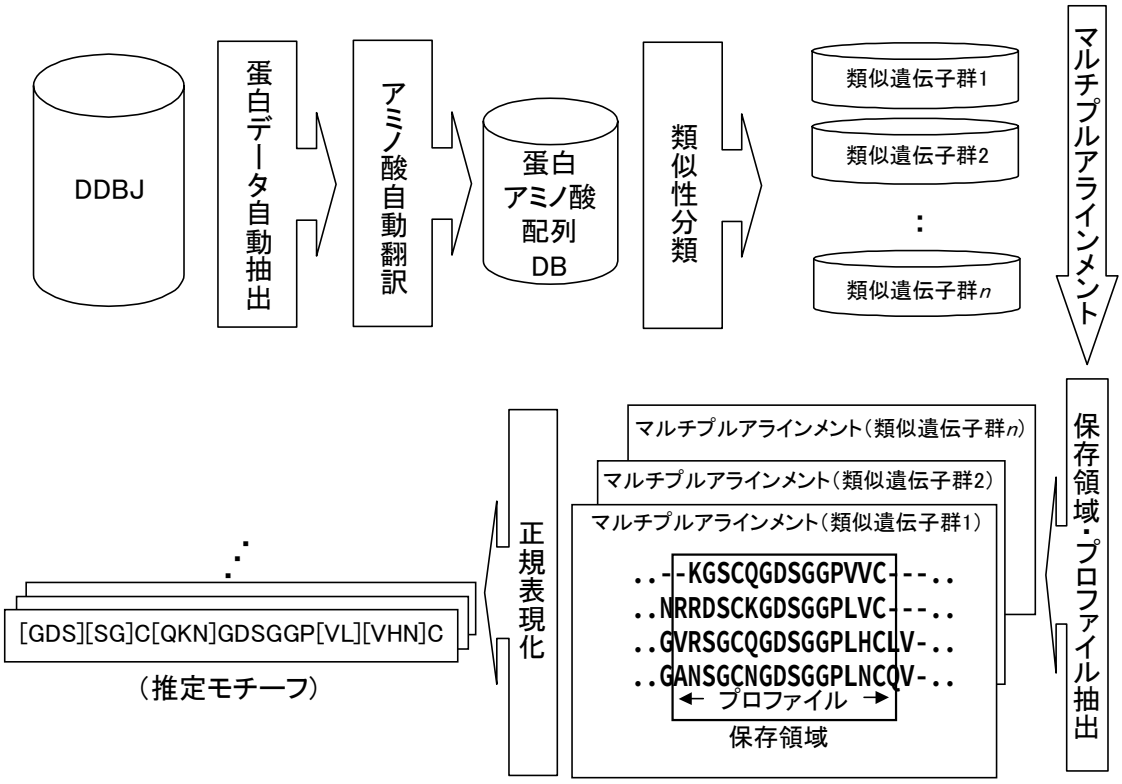

図-2　モチーフ推定手順
Fig.2-Procedure of predicting motifs.

●　既知の全蛋白アミノ酸配列収集

現在，解読された公開可能なDNA配列情報は全世界的規模で収集され，日本では遺伝研にあるDDBJ（DNA Data Bank of Japan）からDNAデータベースとして構築，公開されている。DNAデータベースでは蛋白としてアミノ酸配列に翻訳されるDNA配列情報はアミノ酸への翻訳領域と翻訳規則がアノテーション情報として付記されており，このアノテーション情報をもとにDDBJのDNAデータベースからすべてのアミノ酸翻訳領域をアノテーション情報に基づいてコンピュータで自動抽出しアミノ酸配列に自動翻訳することで，既知の全蛋白アミノ酸配列情報の自動収集を実現している。

●　相同蛋白分類

厳密な相同蛋白分類は解析対象の蛋白の分子進化系統（分子進化の入門テキストとしては参考文献[2]など）に関する情報が必要であるが情報の不十分な新規蛋白を含む全蛋白の厳密な相同分類は困難である。このため相同分類の近似としてアミノ酸配列の類似度による分類を用いることとした。アミノ酸配列の類似度は次の手順で計算される。

(1)　比較する二つの配列が最大類似となるように配列要素の順序は変えずに配列要素間にギャップを入れて整列する。これをアラインメントという。

(2)　アラインメントされた配列の配列要素ごとに類似度を計算し配列全体の類似度を求める。

以下，アミノ酸配列1（FLTKEAH）とアミノ酸配列2（FTKHALH）を例に説明する。まず，配列1と配列2のアラインメントを求める。自明な結果として以下を得る。

配列1　FLTKEA＊H
配列2　F＊TKHALH

配列1と配列2の類似度は配列の各位置におけるアミノ酸は，上記のアラインメント配列の個々の位置におけるアミノ酸の類似度の総和であり以下となる。

類似度＝S（F，F）＋S（L，＊）＋S（L，T）＋・・

上式で（X，Y）はアミノ酸Xとアミノ酸Yの類似度であり，宮田ら[3]により定められた値を用いた。＊はギャップを表し進化の過程で生じたアミノ酸の挿入あるいは欠損を意味している。アラインメントと類似度計算の厳密解法は2次元の動的計画法[4]となるがその演算量は対象とする二つの蛋白アミノ酸配列長の積程度となり，全蛋白間の類似度算出では膨大な演算量となる。計算時間の短縮と分類精度の確保を実現するため，まず対象とする

全蛋白アミノ酸配列について簡便法であるFASTA法[5]を用いて類似度分類を行った後，類似蛋白グループごとにグループ内のアミノ酸配列相互の類似度を動的計画法により厳密に計算し各類似グループに含まれる類似アミノ酸配列の本数が最大20となるようにした。

● **マルチプルアラインメント**

相同蛋白にある保存プロファイルを自動的に抽出するためには，図-1に示すように相同蛋白全体でのアミノ酸配列のアラインメントが必要である。これは相同蛋白分類の節で述べた類似度計算の二本の配列アラインメントをn本（$n \geqq 3$）に拡張することでありマルチチプルアラインメントと呼ばれている。このマルチプルアラインメントの厳密な解法はn次元の動的計画法となりその演算量は対象とする相同蛋白アミノ酸配列長の総積程度と莫大なものになる。このため，以下に示すペアワイズアラインメントという近似手法を取り演算量を減らすこととした。

(1)　相同蛋白から，まず一対のアミノ酸配列を抽出しアラインメントを行う。

(2)　(1)のアラインメント結果を用いて，ほかのアミノ酸配列とのアライメントを行う。

(3)　(2)をアラインメント対象配列に順次繰り返し，全配列のマルチプルアラインメントを得る。

以上のペアワイズアラインメントは配列の計算順序で結果が変わる可能性を持つという，計算精度上の大きな問題がある。保存プロファイル抽出を考えるとアラインメントは高精度の必要があり，SODHOではペアワイズアラインメントの精度を高めるため計算順序を分子進化系統樹（参考文献[2]など）に沿って行うこととした。実際には正しい分子進化系統樹は計算の最初の段階ではないため，まず上記の（1）から（3）の計算手順を配列の入力順など任意の順序で実施しその結果得たアラインメントをもとに分子進化系統樹を作成する，つぎに得られた分子進化系統樹に沿ってペアワイズアラインメントを再計算する，以上を進化系統樹の樹形が一定になるまで反復実行する，という分子進化系統樹によるアラインメント改良反復計算で行っている。

● **保存領域推定と保存プロファイル抽出**

保存領域推定とは，相同蛋白のマルチプルアラインメント配列についてアミノ酸の保存性の高い部分領域を推定することであり，配列全体にわたる大域的な保存度分布から局所的な保存領域を分離抽出できなければならない。また，進化距離の遠い蛋白間で保存されているアミノ酸は進化距離の近い蛋白間で保存されているアミノ酸より進化的に重要であると考えられるためアミノ酸の保存度について相同蛋白間の進化的距離による重み付けが必要である。以上を踏まえ，以下の計算手順で保存領域推定と保存プロファイル抽出を行っている。

(1)　各アミノ酸サイトでの保存度算出

保存度計算上，基本となるのは各アミノ酸配列サイトでの保存度である。これは，マルチプルアラインメントをもとに各配列サイトに出現する蛋白の類似性に各配列間の進化距離に基づく重み付けを加えて算出している。[1]

(2)　保存領域の推定

まず配列全体での大域的な保存度を求める。このためマルチプルアラインメント配列上に幅の大きなウィンドを設けウィンド内の平均的な保存度を（1）で求めた配列サイトごとのアミノ酸保存度を用いて算出し，その値をウィンドの中心位置の保存度とする。ウィンドを配列上で1アミノ酸サイトずつスキャンして計算することで配列全長上の大域的保存度分布を得る。つぎに，マルチプルアラインメント配列上に幅の狭いウィンドを設け大域的保存度分布同様の計算を行うことで局所的保存度分布を得る。最後に大域的保存度分布と局所的保存度分布を比較し局所的保存度分布が大域的保存度分布を上回っている領域を抽出し，保存領域と推定する。ウィンド幅については，試行錯誤の結果，大域的保存度分布計算に用いるウィンド幅を101アミノ酸，局所的保存度分布に用いるウィンド幅を11アミノ酸とした。以上の手順を模式的に**図-3**に示す。

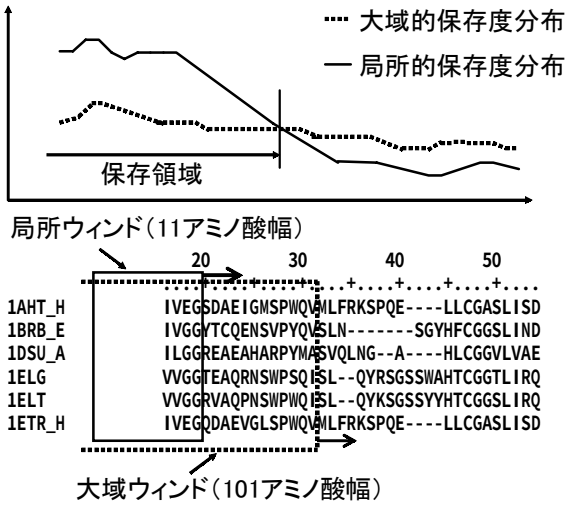

図-3　保存領域抽出
Fig.3-Extract the conservative region.

● 保存プロファイル抽出およびモチーフ推定

保存領域の推定で得られた保存領域内のマルチプルアラインメント配列を保存プロファイルとし，この保存プロファイルを正規表現したものをモチーフとした(図-2)。

モチーフの隠れマルコフモデル化

遺伝研との共同研究では，正規表現系モチーフにより蛋白の分子進化解析に成果を上げた[1]が，SODHOを実用ツールとして利用していく上で正規表現系モチーフの問題点が顕在化した。この解決策としてモチーフの正規表現を隠れマルコフモデル表現へ拡張した結果，SODHOの実問題への適用が可能となった。以下に，正規表現系モチーフの問題点と，その解決方法および成果について述べる。

● 正規表現系の問題点

正規表現系の最大の問題は保存プロファイル上に見出されるアミノ酸の出現頻度情報が欠落してしまう，ということである。例えば，図-2に示す正規表現系で表現されたモチーフの[GDS][SG]C[QKN]GDSGGP[VL][VHN]からは，[GSCKGDSGGPVH]という保存プロファイルのアミノ酸出現頻度から推定される出現確率が0.001953125と非常に低いモチーフと，[SGCQGDSGGPLH]という推定出現確率が0.0234375と高いモチーフが何の区別もなく同等に定義されてしまう。このことは正規表現系で表現されたモチーフではモチーフ検索結果の確からしさやモチーフの類似性解析などに定量的尺度を与えられず，モチーフの実用的な利用が困難であることを示している。

● 隠れマルコフモデルによる解決

正規表現系の問題点への解決策として，SODHOのモチーフ表現を正規表現系から隠れマルコフモデルへ拡張した。モチーフの隠れマルコフモデル（以下，HMMモチーフ）化およびその利用についてはフリーウェアの

表-1　SODHO推定モチーフ一覧

No.	クラスタ名	モチーフ長	GPCRDBでのアノテーション
1	AG22_HUMAN.6	12	TM（Transmembrane）4
2	5H1A_FUGRU.2	8	TM2とTM3の間の膜外
3	5H1B_MOUSE.2	15	TM2
4	AG22_HUMAN.3	20	TM2
5	AG22_HUMAN.2	20	TM1
6	O01611.9	15	（SOSUIによる解析結果を図-4に示す）
7	O16323.1	44	—

GPCRのアノテーションはPOTENTIAL

：PRINTS にエントリがあるモチーフ（GPCRRHODOPSN）

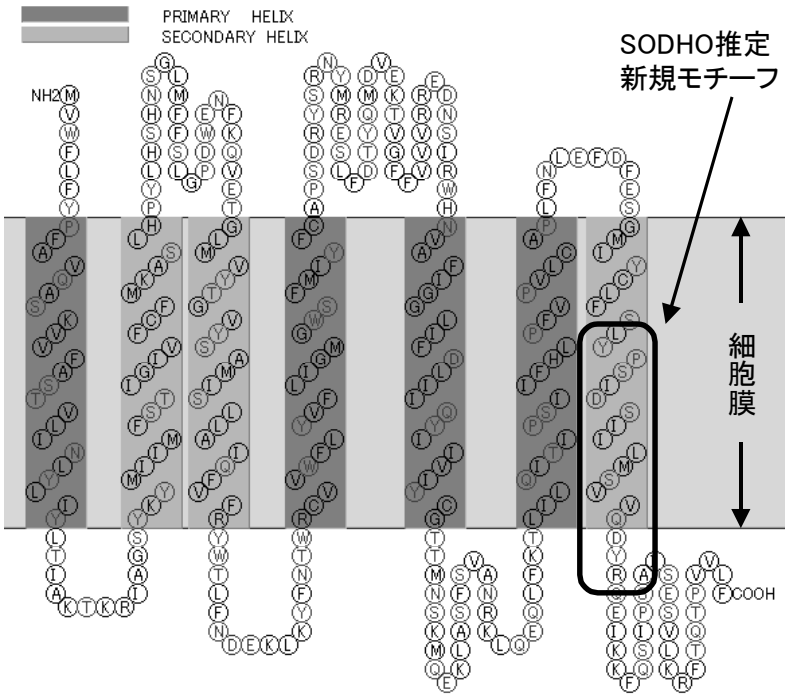

図-4　SOSUIによる新規モチーフ（表-1 No.6）解析結果
Fig.4-Result of analysis of new motif by SOSUI.

HMMER[6]を用いた。HMMにより，保存プロファイルに基づくHMMモチーフの生成と生成されたHMMモチーフをキーとする検索が実行可能である。さらにHMMモチーフ間の類似度求める新規アルゴリズムを開発した。[7] これによりモチーフの類似度に基づく分類が可能となった。

SODHO適用実験

モチーフ表現を隠れマルコフモデルへ拡張したSODHOを創薬ターゲットとして重要なGPCR（G-Protein Coupled Receptor）蛋白に適用しモチーフ解析に関する実証実験を行った。[8] 以下に実験結果とその考察を述べる。

● モチーフ検証

GPCR蛋白を収集しているGPCRDB[9]にSODHOを適用した結果，2948のHMMモチーフが自動推定された。このモチーフについてGPCRDB内でアノテートされている既知モチーフとの比較を行った。結果の一部を**表-1**に示す。表-1からSODHOが既知のモチーフを推定していると同時に独自新規のモチーフを推定していることが分かる。さらに，この独自新規モチーフが意味のあるものかどうかを検証するため，対象のGPCR遺伝子をSOSUI[10]で解析した。その結果，新規モチーフが膜貫通部位であることが予測された（**図-4**）。以上の結果，SODHOが既存モチーフを推定するだけでなく新規機能モチーフの予測にも有用であることが実証された。

● 隠れマルコフモデルモチーフ類似性分類

新規に開発した計算アルゴリズム[7]を用いて，類似度の閾値を75%として，GPCRで推定された全モチーフの類似性分類を行った。その結果，網羅的に推定された2948のHMMモチーフを1454の類似クラスタに分類することができた。以上の結果を類似クラスタ内のHMMモチーフの頻度分布として**図-5**に示す。

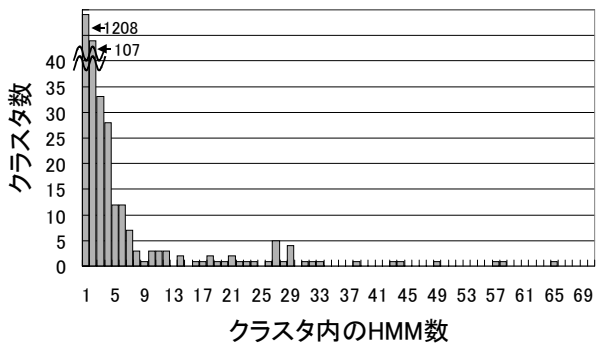

図-5　クラスタリング結果（頻度分布）
Fig.5-Result of clustering of HMMs.

む　す　び

SODHOによりコンピュータ解析のみでモチーフ（正確にはモチーフ候補）を自動推定することが実現できた。推定されたモチーフの表現をHMM化することで，SODHOを実用的なツールとすることができた。GPCR蛋白への適用実験の結果，既存のモチーフを推定できるだけでなく，新規の意味のあるモチーフを予測可能であることが検証された。さらに新規に開発された計算アルゴリズム[7]によりモチーフ間の類似性に基づく定量的解析が可能となった。

従来の，Pfam,[11]SMART,[12]PRODOM[13]などのモチーフ辞書の基本的な部分は，生物学的知識に基づき人手により作成されている。このためモチーフの網羅性や生成効率性に課題がある。SODHOではコンピュータ解析だけでモチーフを自動推定するため，従来のモチーフ辞書作成に比べて網羅性や効率性の観点で有利であると考えられる。一方，コンピュータ解析だけのため，推定されたモチーフの精度や偽陽性の存在などの課題がある。今後はウエット実験との比較検証を重ねモチーフ精度の向上や偽陽性の除去を行いながら対象蛋白を更に拡張し，創薬研究上流工程への適用を重ね有用性を実証し，実利用ツールとして更にブラッシュアップしていく予定である。

参 考 文 献

(1)　Y. Tteno et al.：Evolutionary Motif and Its Biological and Structural Significance. *Journal of Molecular Evolution*, Vol.44, p.38-43 (1997).

(2)　木村資生ほか：分子進化学入門，培風館.

(3)　Miyata T. et al.：Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, Vol.12, p.219-236 (1994).

(4)　Needleman S. B. et al.：A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins. *Journal of Molecular Biology*, Vol.48, p.444-453 (1970).

(5)　Lipman D. J. et al.：Rapid and Sensitive Protein Similarity Searches. *Science*, Vol.227, p.1435-1441 (1985).

(6)　Sean R. Eddy：Profile Hidden Markov Models for

biological Sequence Analysis.
http://hmmer.wustle.edu/

(7) 佐藤：類似性評価方法及び類似性評価プログラム．出願番号（特許2001-299218）．

(8) 松林：蛋白ファミリーからのモチーフ自動抽出．SIGMBI第20回研究会発表資料（2002.05.24）．

(9) G. Vriend et al.：GPCRDB: Information system for G protein-coupled receptors (GPCRs).
http://www.gpcr.org/7tm/

(10) Mitaku et al.：SOSUI about SOSUI.
http://sosui.proteome.bio.tuat.ac.jp/about-sosui.html

(11) Alex Bateman et al.：Pfam 3.1:1313 multiple alignments and profile HMMs match the majority of proteins．*Nucleic Acids Research*，Vol.27，p.260-262（1999）．

(12) Jorg Schultz et al.：SMART, a simple modular architecture research tool: Identification of signaling domains．Proc. Natl. Acad. Sci. USA，Vol.95，p.5857-5864（1998）．

(13) Florence Corpet et al.：ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons．*Nucleic Acids Research*，Vol.28，p.267-269（2000）．